The Phase-2 Upgrade of the CMS Level-1 Trigger Technical Design Report
The Phase-2 Upgrade of the CMS Level-1 Trigger

Technical Design Report

CMS Collaboration
Editors
A. Zabi, J. Berryhill, E. Perez and A. Tapper

Chapter Editors

Cover Design
Sergio Cittolin

Acknowledgements
This document would not exist without the devoted efforts of many colleagues, too many to be named explicitly, who contributed text and figures to this TDR. The very useful input from the CMS Trigger and Data Acquisition project, the CMS internal reviewers (D. Acosta, J. Alcaraz, T. Boccali, G. Boudoul, T. Camporesi, M. Capeans Garrido, S. Donato, S. Gennai, F. Hartmann, J. Kieseler, I. Mikulec, M. Pieri, C. Rovelli, S. Savin, P. Wittich), and the chair of the Phase-2 TDRs editorial board (C. Lourenco) helped to improve the quality of this document and is greatly appreciated. We would like to acknowledge the support of the CMS Management and the Upgrade Project Coordination team. We thank the Offline, Computing, Physics Performance and Datasets, Upgrade Physics Strategy Group, and all the Physics groups for their help in developing and executing the physics and performance studies. We would also like to thank O. Davignon and B. Radburn-Smith for their careful review of all the performance results included in this document, as well as K. Black, M. Konecki, P. Padley, for their help improving the document during the internal review by the project. We would like to thank Sergio Cittolin for providing the cover artwork inspired by this project.
Executive Summary

This Technical Design Report (TDR) describes the ongoing developments and plans towards the upgrade of the CMS Level-1 (L1) trigger for the High-Luminosity Large Hadron Collider (HL-LHC). The HL-LHC era constitutes the Phase-2 of the LHC operation succeeding to the Phase-1 exploitation period currently ongoing and ending in 2024. In its ultimate configuration, the HL-LHC will reach unprecedented performance in terms of instantaneous luminosity \((7.5 \times 10^{34} \text{ cm}^{-2}\text{s}^{-1})\) potentially leading to a total integrated luminosity of up to 4000 fb\(^{-1}\) after ten years of operations, scheduled to start in 2027. This previously unmatched amount of data opens the door to a rich and ambitious physics program including both high-precision measurements and searches for physics beyond the Standard Model.

To achieve the goals of the HL-LHC program, the Phase-2 upgrade of the L1 system utilizes technological advances to enhance the physics selectivity already at the hardware level of the data acquisition system. The intense hadronic environment corresponding to 200 simultaneous collisions per beam crossing, imposes serious challenges to the system requirements in order to maintain performance. To profit from the extended coverage and increased granularity of the upgraded CMS detector, the latency of the system is extended to 12.5 µs. The use of tracking and high-granularity calorimeter information is possible for the first time at L1. The maximum output bandwidth is 750 kHz to maintain the required performance. Modern processors are used to implement sophisticated algorithms including machine learning-based approaches to target the selection of specific final states.

As a starting point, the trigger algorithms have been studied with the minimum requirement of reaching the performance that CMS needs to perform the studies discussed in Ref. [1], a review that offers a detailed report on the physics program of the HL-LHC. Additionally, the algorithms were further developed to allow studies of several supplementary physics topics. Algorithms were simulated in software, and also prototyped in firmware, to estimate their hardware resource usage and corresponding latency.

Informed by the algorithmic requirements, a program of hardware R&D is ongoing with the aim of identifying hardware platforms and system architectures suitable for implementing the required algorithms. Developments are based on the advanced telecommunications computing architecture (ATCA) standard for electronics, using using state-of-the-art field-programmable gate arrays (FPGAs) and serial optical links running at speeds up to 25 Gb/s. Prototypes produced by several groups meet the requirements of the project and in many cases algorithms have been tested directly in prototype hardware. Plans for future development and testing are also summarized.

The physics and environmental considerations that lead to the choice of the conceptual design of the upgraded trigger system along with its technological implementation and its expected performance are provided. The organization of the project is presented, including the institutes participating, the estimated cost, schedule and milestones. The schedule includes the plan for installation, integration and commissioning of the system during the LHC long-shutdown 3 (LS3 starting 2025) in view of delivering triggers for the start of the LHC Run-4 in 2027.

Overall the studies contained within the report show that the physics requirements can be met within the technical constraints of the project with appropriate contingency.
Contents

1 Introduction and overview ................................. 1
  1.1 Physics motivations and CMS Phase-2 trigger upgrade .......... 1
  1.2 HL-LHC upgrade and triggering challenges .................. 4
  1.3 Phase-2 Level-1 trigger requirements and design .............. 6
    1.3.1 Phase-2 L1 upgrade conceptual design .................. 7
    1.3.2 Trigger algorithms for the HL-LHC ...................... 11
  1.4 Physics reach of the Level-1 Phase-2 Trigger ................ 18
    1.4.1 Preserving the CMS discovery potential at the HL-LHC .... 18
    1.4.2 Extending the CMS physics reach with the Phase-2 Level-1 trigger .......................... 24
  1.5 Outline of the Technical Design Report ..................... 29

2 Inputs to the L1 trigger system: the trigger primitives generation 31
  2.1 Track finder primitives .................................. 31
  2.2 Barrel calorimeter trigger primitives ....................... 36
    2.2.1 Electromagnetic barrel calorimeter ..................... 36
    2.2.2 Hadron barrel and forward calorimeters ................ 40
  2.3 High granularity endcap calorimeter trigger primitives ........ 42
    2.3.1 System architecture .................................. 42
    2.3.2 Reconstruction algorithms .............................. 44
    2.3.3 Implementation, latency and data volumes ................ 46
  2.4 Muon barrel trigger primitives ............................ 47
    2.4.1 Drift tubes ........................................ 47
    2.4.2 Barrel resistive plate chambers ....................... 49
    2.4.3 DT+RPC super-primitive combination .................... 50
    2.4.4 Muon barrel primitives performance .................... 52
    2.4.5 Future improvements ................................ 53
  2.5 Muon endcap trigger primitives ........................... 55
    2.5.1 Endcap resistive plate chambers ....................... 55
    2.5.2 Improved resistive plate chambers ..................... 55
    2.5.3 Endcap cathode strip chambers ......................... 56
    2.5.4 Gas electron multipliers .............................. 57
    2.5.5 Endcap muon trigger primitives performance ............. 59
    2.5.6 Future improvements ................................ 59
  2.6 External Triggers ........................................ 61
  2.7 Summary of trigger input bandwidth and latency ............... 62

3 Trigger algorithms ........................................ 65
  3.1 Introduction: triggering on physics objects at the HL-LHC .... 65
  3.2 Triggering on electrons and photons ......................... 66
    3.2.1 Introduction ........................................ 66
    3.2.2 Electron and photon reconstruction in the calorimeter barrel ........ 66
    3.2.3 Electron and photon reconstruction in the endcap calorimeter ........ 69
3.2.4 Exploiting L1 track-finder primitives for e/γ triggers .................................. 71
3.2.5 Firmware implementation and resources ............................................. 78
3.3 Triggering on muons .................................................................................. 80
  3.3.1 Introduction ......................................................................................... 80
  3.3.2 Standalone muon reconstruction in the barrel region ......................... 83
  3.3.3 Standalone muon reconstruction in the overlap region ....................... 90
  3.3.4 Standalone muon reconstruction in the endcap region ....................... 94
  3.3.5 Global muon trigger algorithms ......................................................... 99
3.4 Vertex reconstruction ................................................................................ 121
  3.4.1 Histogram-based vertexing algorithm .............................................. 121
  3.4.2 Track to vertex association ............................................................... 122
  3.4.3 Vertex finding performance .............................................................. 122
  3.4.4 Firmware implementation of vertexing algorithms ............................ 125
  3.4.5 End-to-end neural network vertex reconstruction model .................... 129
3.5 Particle-flow reconstruction ................................................................... 131
  3.5.1 Calorimeter inputs ............................................................................. 131
  3.5.2 Tracking and muon inputs ................................................................. 138
  3.5.3 Particle-flow and PUPPI algorithms ................................................ 139
  3.5.4 Particle-flow and PUPPI algorithm performance ............................... 143
  3.5.5 Firmware implementation ................................................................. 146
3.6 Triggering on jets, hadronic tau decays, and energy sums ....................... 156
  3.6.1 Jet finding algorithms ....................................................................... 156
  3.6.2 Hadronic τ algorithms ...................................................................... 166
  3.6.3 Missing transverse energy algorithms ............................................. 182
3.7 Global trigger algorithms ......................................................................... 189
  3.7.1 Cut-based trigger algorithms ............................................................. 189
  3.7.2 Machine learning based trigger algorithms ...................................... 192
  3.7.3 Overall Global Trigger resource usage ........................................... 195
3.8 Trigger algorithms for heavy ion collisions ........................................... 196
  3.8.1 Introduction ....................................................................................... 196
  3.8.2 Physics program ................................................................................ 196
  3.8.3 Trigger considerations ....................................................................... 197

4 Level-1 trigger menu ................................................................................. 201
  4.1 Inputs and design strategy of the Phase-2 Level-1 trigger menu ............ 202
    4.1.1 Muon trigger algorithm inputs ....................................................... 203
    4.1.2 Electron and photon trigger algorithm inputs .................................. 204
    4.1.3 Jets and jet-sum trigger algorithm inputs ....................................... 205
    4.1.4 Missing transverse energy trigger algorithm inputs ....................... 206
    4.1.5 Tau trigger algorithm inputs .......................................................... 206
  4.2 Simplified menu with Phase-1 thresholds ........................................... 207
  4.3 New L1 trigger algorithms to extend Phase-1 physics acceptance ........ 212
4.3.1 Physics with the extended pseudorapidity coverage of the L1 lepton triggers .............................................. 212
4.3.2 Physics with L1 soft and correlated muons .......................................................... 213
4.3.3 Physics of light mesons with L1 tracking ............................................................ 219
4.3.4 Physics with L1 displaced muons ................................................................. 221
4.3.5 Physics with L1 displaced jets ................................................................. 223
4.3.6 Dedicated VBF Higgs boson production L1 trigger algorithms based on machine learning techniques .............................................. 228
4.3.7 Summary of triggers for extended physics reach .............................................. 230
4.4 Study of the menu rate evolution with pileup .................................................... 230
4.4.1 Individual trigger rate evolution with pileup .............................................. 231
4.4.2 Alternative $E_T^{miss}$ and VBF triggers for extreme pileup ............................ 233
4.4.3 Menu rate and bunch-to-bunch pileup variations .............................................. 233

5 Conceptual design of the Phase-2 L1 Trigger .................................................. 237
5.1 Introduction ........................................................................................................... 237
5.2 Calorimeter Trigger system .............................................................................. 237
5.3 Muon Trigger system ......................................................................................... 242
  5.3.1 Endcap Muon Track Finder .............................................................................. 243
  5.3.2 Overlap Muon Track Finder ............................................................................ 244
  5.3.3 Barrel Muon Track Finder ............................................................................. 245
  5.3.4 Global Muon Trigger ..................................................................................... 245
5.4 Global Track Trigger system .............................................................................. 247
5.5 Correlator Trigger system ................................................................................... 248
  5.5.1 Correlator Trigger Layer-1: particle-flow candidates ........................................ 248
  5.5.2 Correlator Trigger Layer-2: object reconstruction ............................................. 252
5.6 Global Trigger system ......................................................................................... 253
  5.6.1 Inputs to the Global Trigger .............................................................................. 254
  5.6.2 Global Trigger architecture ........................................................................... 255
  5.6.3 Interfaces to TCDS, DAQ and HLT ................................................................ 257
  5.6.4 Using the 40 MHz Scouting System for trigger monitoring .............................. 258
5.7 Scouting system ................................................................................................... 259
  5.7.1 General architecture ....................................................................................... 259
  5.7.2 Inputs, memory, and connectivity requirements .............................................. 261
5.8 Overall trigger system architecture .................................................................... 263

6 Instrumentation of the Phase-2 trigger system .................................................. 269
6.1 Introduction ........................................................................................................... 269
6.2 System infrastructure and services .................................................................... 269
6.3 Hardware research and development ................................................................ 273
  6.3.1 ATCA processor: APx consortium .................................................................. 273
  6.3.2 ATCA processor: Serenity consortium ............................................................ 282
  6.3.3 ATCA processor: Ocean ................................................................................ 290
  6.3.4 Specific processing board: Barrel Muon Trigger (BMT) demonstrator .......... 294
6.4 Trigger demonstrators and slice tests .................................................. 299
  6.4.1 Calorimeter trigger demonstrators ................................................. 301
  6.4.2 Muon trigger demonstrators: muon track finder and global muon trigger  301
  6.4.3 Global track trigger demonstrator .................................................. 313
  6.4.4 Correlator trigger demonstrators .................................................... 320
  6.4.5 Demonstration of the 40 MHz scouting system ............................... 326
6.5 System-wide technical issues ............................................................ 328
  6.5.1 Hardware issues ............................................................................. 329
  6.5.2 Firmware and software deployment ................................................ 330
  6.5.3 Trigger links and system synchronization ....................................... 332
  6.5.4 Software-firmware interface ......................................................... 332
  6.5.5 Commissioning and parallel running .............................................. 334
  6.5.6 Monitoring and data validation ...................................................... 334
  6.5.7 Latency measurements and checks ................................................. 335
  6.5.8 System health check ..................................................................... 335
  6.5.9 Demonstrator systems in Run-3 ...................................................... 336

7 Organization, schedule, and costs ......................................................... 337
  7.1 Project organization and work breakdown structure ........................... 337
  7.2 Project schedule and milestones ....................................................... 338
    7.2.1 Schedule ...................................................................................... 338
    7.2.2 Milestone list ............................................................................. 340
  7.3 Institution interests and construction responsibilities .......................... 340
  7.4 Project costs ..................................................................................... 340

References ................................................................................................. 353
Chapter 1

Introduction and overview

The HL-LHC [2] presents the opportunity for a very rich and ambitious physics program, exploiting an integrated luminosity of 3000 fb$^{-1}$. The LHC will undergo major upgrades of its components leading to an increase of the instantaneous luminosity to $5 \times 10^{34}$ cm$^{-2}$s$^{-1}$, five times the machine’s original design value. In its ultimate configuration, the HL-LHC will reach a peak instantaneous luminosity of $7.5 \times 10^{34}$ cm$^{-2}$s$^{-1}$, increasing the average number of proton-proton collisions per bunch crossing (pileup) to around 200. The ultimate performance of the HL-LHC would enable the collection of 400 to 450 fb$^{-1}$ of integrated luminosity per year, potentially providing a total of 4000 fb$^{-1}$ to each of the CMS and ATLAS experiments.

The CMS detector requires a trigger and data acquisition system with exceptional performance to collect the required information-rich datasets in these challenging running conditions [3]. The upgrade of the trigger system will enhance the physics selectivity and maintain the performance necessary throughout the 10 year long HL-LHC program, which includes heavy ion operations.

1.1 Physics motivations and CMS Phase-2 trigger upgrade

Physics motivations: The CMS Phase-2 physics program plans to fully exploit the HL-LHC to perform searches for new physics beyond the Standard Model (BSM), achieve unprecedented high-precision measurements of the SM, including a significantly improved characterization of the Higgs Boson sector. A broad spectrum of physics analyses will become possible with the unprecedentedly large HL-LHC data samples and the new capabilities offered by the detector upgrade, such as its extended coverage. More advanced selection algorithms are required to maintain the effective selection of electroweak-scale processes with 200 pileup events. Sophisticated triggers will be required to select specific topologies such as VBS/VBF production, rare $B$-meson decays (based on usage of tracks in the L1 trigger for the first time), forward muon trigger for $\tau \rightarrow \mu \mu \mu$ (profiting from extended coverage), etc. Low mass resonances could be identified with a dedicated scouting system. In order to illustrate the discovery potential of the upgraded CMS detector, achieved through the efficient selection of the L1 system, benchmark signals have been selected and are presented in this document.

The CMS trigger system: The CMS experiment currently implements a sophisticated two-level triggering system composed of Level-1 (L1) [4], instrumented by custom hardware processor boards, and software High Level Trigger (HLT). The L1 receives information from calorimeter and muon systems generating an initial selection within a fixed latency of 4 $\mu$s, with a maximum output rate of 100 kHz. Upon reception of a L1 Accept (L1A) signal, the detector is fully read out and the selected event is reconstructed in the HLT. The HLT selection is based on this finer information, reducing the output rate to about 1 kHz on average. A first ma-
Chapter 1. Introduction and overview

A major upgrade of the L1 system was conducted during the long-shutdown 1 (LS1 2013–2015). The Phase-1 upgrade consisted in the complete replacement of the system that was deployed and successfully operated during Run-1 of the LHC (2010–2013). A new architecture with improved performance was installed to maintain a high efficiency for collecting the data, under the more challenging conditions experienced during Run-2 (2015–2018) and expected during Run-3 (2021–2024). All the hardware, interconnections, electronics boards, firmware and software were redesigned as described in the technical design proposal (TDR) [5] of the L1 trigger Phase-1 upgrade. Benefiting from the higher input granularity, more sophisticated and innovative object reconstruction algorithms were implemented. High-speed optical links were installed to rapidly collect all the information from sub-detectors contributing to the trigger, hence providing a full field view of the detector, adapted to the precise evaluation of global event quantities, such as pileup or energy sums. In addition, the new global trigger is capable of evaluating complex selection algorithms such as those involving the invariant mass of trigger objects. Using correlations in multi-object triggers has allowed CMS to further enhance the optimization of physics sensitivity while adapting to changing conditions and priorities. The performance of the L1 system during the Run-2 data-taking period is described in Ref. [6]. In spite of the increased luminosity ($2 \times 10^{34}$ cm$^{-2}$s$^{-1}$) and pileup, changes of proton filling schemes, and the effects of aging of the sub-detectors, the flexibility of the system has allowed the acceptance for physics to be maintained and a rich physics program to be pursued throughout the whole period. Many aspects of the Phase-2 upgrade of the L1 trigger system are inspired by the key technological choices made during the Phase-1 upgrade.

The Phase-2 upgraded CMS detector: In order to fully exploit the HL-LHC running period, major consolidations and upgrades of the CMS detector are planned [7]. The collision rate and level of expected pileup imply very high particle multiplicity and an intense radiation environment. The performance required on event object reconstruction to achieve the extraction of physics signatures relies on the implementation of higher granularity detectors along with robust readout electronics. The CMS collaboration plans to replace both the Strip and Pixel tracking detectors, with an Inner Tracker featuring small size pixel sensors and an Outer Tracker equipped with strip and macro pixel sensors, extending their coverage to $|\eta| = 3.8$ [8]. A narrower pitch will provide better transverse and longitudinal impact parameter resolution. The Outer Tracker will implement stacked strip modules, reducing the hit multiplicity and allowing track candidates for the trigger (L1 tracks) to be reconstructed up to $|\eta| = 2.4$, which opens major new possibilities for the L1 trigger. The readout electronics for the barrel calorimeters will be replaced to achieve finer granularity and provide timing information [9]. The endcap calorimeters will be replaced by the high-granularity calorimeter (HGCAL) [10], implementing over 6 million readout channels. This sampling calorimeter will provide shower separation and identification adapted to harsher conditions in the forward region of the detector. The muon detection system redundancy achieved through the combination of drift tubes (DTs), resistive plate chambers (RPCs), and cathode strip chambers (CSCs) will remain with consolidated electronics [11]. Additional improved RPC (iRPC) chambers and gas electron multiplier (GEM) chambers will be installed to extend the coverage up to $|\eta| = 2.4$ and 2.8, respectively [11]. A minimum ionizing particle (MIP) timing detector (MTD) [12] placed in front of the barrel and endcap calorimeters will provide precise timing measurement of charged tracks. Along with the sub-detector upgrades, a complete replacement of the trigger (L1 and HLT) and data acquisition (DAQ) system, with increased throughput, is planned. Similarly to Phase-1, the Phase-2 HLT will have access to the full granularity of the detector, with an input rate of 750 kHz in the ultimate HL-LHC scenario; the target for timing is 500 ms, measured on a current (2018) node. The selection algorithms will perform a rate reduction leading to an output bandwidth of 7.5 kHz. The most promising avenue of development is the use of high performance hardware;
for example, porting part of the reconstruction to run on GPUs [7].

The L1 Phase-2 upgrade: The Phase-2 upgrade of the L1 trigger system is designed not only to maintain the efficiency of the signal selection to the level of the Phase-1 performance, but also to significantly enhance, or enable, the selection of any possible new physics manifestations that could lead to unconventional signatures. High-precision measurements of physics processes will benefit from the extension of the available phase space such as enhanced trigger coverage in the forward region of the detector or the ability to exploit fully hadronic final states. Most importantly, state-of-the-art techniques used in offline reconstruction and analyses, such as the global event reconstruction based on particle-flow techniques [13], become possible at the L1 trigger, with the availability of L1 tracks delivered by the upgraded Outer Tracker. Moreover, these algorithms benefit from the increased granularity of the calorimeter information. The trigger processing capabilities allows their hardware implementation to be realized. Based on the Phase-1 experience, this functionality is obtained through the use of Field Programmable Gate Arrays (FPGA) coupled with high-speed optical links to retrieve the detector data and provide the systems interconnections. A flexible and modular architecture, relying on modern technologies, allows the data handling to be optimized to adapt to LHC running conditions and physics requirements. The Phase-2 upgrade of the trigger and DAQ system will keep a two-level strategy while increasing the L1 maximum rate to 750 kHz. The total latency will be increased to 12.5 $\mu$s to allow, for the first time, the inclusion of the tracker and high-granularity calorimeter information. Moreover, a longer latency will enable higher-level object reconstruction and identification, as well as the evaluation of complex global event quantities and correlation variables to optimize physics selectivity. The implementation of sophisticated algorithms using particle-flow reconstruction techniques or machine-learning based approaches can now be contemplated. In addition to these features, a 40 MHz scouting system harvesting the trigger primitives produced by sub-detectors and the trigger objects produced at various levels of the trigger system is proposed. The concept of trigger scouting has been introduced in CMS at the HLT [14]. It is based on the use of physics objects reconstructed as a by-product of the triggering process to perform data reduction and analysis, only storing high-level information for selected events, thus overcoming the rate to storage limitations of the DAQ. The Level-1 scouting system will use Level-1 trigger reconstructed objects and quantities in a similar way, selecting and analyzing them on the fly at the collision rate. This system has the additional advantage of allowing systematic search of correlations among multiple contiguous bunch crossing, and can be used to scrutinize the collision events and identify potential signatures unreachable through standard trigger selection processes. In order to successfully integrate and commission this complex upgraded L1 trigger, an approach similar to that adopted in the Phase-1 upgrade is chosen, where part of the system will run in parallel with the current system during Run-3 operations. The muon system in place now will remain in Phase-2 and is already used to test new algorithms and gain confidence in their development.

This TDR follows the interim TDR (iTDR) [3], which described the roadmap of the project, as well as the early research and development studies that led to the identification of the key features of this upgrade design. The iTDR reported the evaluation of the technical capabilities in terms of algorithms, firmware, and hardware required to achieve the L1 trigger features. Sophisticated algorithms were implemented in hardware through the extensive use of High-Level Synthesis (HLS) to produce the corresponding firmware. The decisive rate reduction obtained with the inclusion of tracking information was demonstrated along with the first physics menu. A full range of architecture designs were considered, directing the proposal towards the implementation of a central “Correlator Trigger” system where back-end information feeds advanced triggering algorithms. The iTDR also allowed the identification of potential tech-
nologies for FPGAs, optical links as well as the Advanced Telecommunications Computing Architecture (ATCA) as a platform. Advantages of generic processing engines were discussed and presented as an adequate choice for implementation, maintenance and flexibility of the design. Along with the L1 upgrade, major upgrades are planned for the DAQ and HLT.

1.2 HL-LHC upgrade and triggering challenges

In order to extend the LHC discovery potential, consolidations and upgrades of the machine and its injection chain have been planned during long shutdowns as displayed on the roadmap in Fig. 1.1. During LS1, interconnections between the LHC superconducting magnets were consolidated to permit the operation of the machine at 13 TeV of center-of-mass energy. The machine reached a record instantaneous luminosity of $2.1 \times 10^{34} \, \text{cm}^{-2}\text{s}^{-1}$, with average values of up to 55 proton proton collisions per crossing during Run-2 operations. During LS2, ongoing at the time of writing, the LHC is optimizing its parameters and luminosity production. The machine is being consolidated to potentially increase the center-of-mass energy to 14 TeV and sustain a maximum instantaneous luminosity of $2 \times 10^{34} \, \text{cm}^{-2}\text{s}^{-1}$ for longer periods of time during Run-3 operations. Phase-1 operations extend to 2024 with the plan to deliver between 350 fb$^{-1}$ and 500 fb$^{-1}$ of data to ATLAS and CMS. Major upgrades to the collider and its experiments will take place during LS3, after which Phase-2 operations will start.

![LHC / HL-LHC Plan](image)

Figure 1.1: LHC baseline plan for the next decade and beyond, showing the energy of the collisions (upper red line) and luminosity (lower red lines). The first long shutdown (LS1) in 2013–2014 allowed the design parameters of beam energy and luminosity to be reached. The second long shutdown (LS2), 2019–2020, will consolidate luminosity production and reliability as well as upgrade the LHC injectors. After LS3, 2025–2027, the machine will be in the High Luminosity configuration (HL-LHC).

The HL-LHC project is already half way through its developments targeting a start of operation in the second half of 2027. The baseline configuration of the upgraded collider should allow the collection of 3000 fb$^{-1}$ of integrated luminosity over ten years of operations. The peak instantaneous luminosity will steadily increase during Run-4 to reach a maximum of $5 \times 10^{34} \, \text{cm}^{-2}\text{s}^{-1}$.
as illustrated in Fig. 1.2, with an average of 140 simultaneous collisions per crossing. In this configuration, the instantaneous luminosity may be maintained at a fixed value throughout the duration of a physics fill. As all the equipment is being designed with a 50% margin with respect to instantaneous heat deposition and integrated radiation dose, the machine performance could possibly be pushed to achieve a peak instantaneous luminosity of \(7.5 \times 10^{34} \text{cm}^{-2}\text{s}^{-1}\). This would result in increasing the average number of collisions per bunch crossing to around 200. In this ultimate configuration, the integrated luminosity per year would exceed 400 fb\(^{-1}\). Assuming the experiments can handle this level of pileup, 4000 fb\(^{-1}\) would be collected at the end of the HL-LHC lifetime. This performance relies on key innovations pushing the limits of accelerator technology by implementing 11–12 T superconducting magnets, beam collimation, and rotation processes described in Ref. [2].

Figure 1.2: Forecast for peak luminosity (red dots) and integrated luminosity (blue line) in the HL-LHC era, according to the nominal HL-LHC parameters. The total integrated luminosity quoted assumes 160 days of physics operation of the LHC complex per year and a physics efficiency of 50%, estimated from the current Phase-1 performance and operation experience. Run-3 will end in 2024, while the LS3 will take place during the years 2025–2027.

The scenarios considered for the HL-LHC beam parameters have been derived based on the configurations used in Run-2. In 2016, although the number of bunches in the machine was smaller than the design value of 2808, the LHC has attained the nominal instantaneous luminosity of \(1 \times 10^{34} \text{cm}^{-2}\text{s}^{-1}\) through the reduced emittance achieved by the injectors and a reduced \(\beta^*\) of 40 cm at the interaction points IP1 (ATLAS) and IP5 (CMS). The nominal Run-2 filling scheme had 2556 bunches, grouped in trains of 48 bunches each, with 2544 of these colliding at the CMS interaction point. Due to frequent beam dumps caused by electron clouds in the LHC, a special filling scheme called “8b4e” had to be used during the 2017 running. In this specific scheme, short trains of 8 filled bunches followed by 4 empty bunches (without collisions) are injected in the machine and mitigate the formation of electron clouds. A maximum of 1916 bunches could be inserted in the LHC, forcing the bunch intensity to increase to achieve the same instantaneous luminosities. The instantaneous luminosity was levelled to a value
Table 1.1: The main HL-LHC parameters for proton collision operation are compared with the nominal LHC parameters [2]. The three schemes chosen are “standard 25 ns” (baseline configuration), “BCMS” and “8b4e”, already encountered during LHC Run-2 operations (see text). $N_b$ is the bunch population; $\beta^*$ is the beam beta function (focal length) at the collision point; $\epsilon_L$ is the longitudinal acceptance; $\epsilon_n$ is the transverse normalized emittance; $\mu$ is the average pileup (number of inelastic collisions in the same bunch crossing).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Nominal LHC Design Report</th>
<th>HL-LHC 25 ns standard</th>
<th>HL-LHC 25 ns BCMS</th>
<th>HL-LHC 8b4e</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beam energy in collision [TeV]</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>$N_b [10^{11}]$</td>
<td>1.15</td>
<td>2.2</td>
<td>2.2</td>
<td>2.3</td>
</tr>
<tr>
<td>Number of bunches per beam</td>
<td>2808</td>
<td>2748</td>
<td>2604</td>
<td>1968</td>
</tr>
<tr>
<td>Beam current [A]</td>
<td>0.58</td>
<td>1.09</td>
<td>1.03</td>
<td>0.82</td>
</tr>
<tr>
<td>Minimum $\beta^*$ [m]</td>
<td>0.55</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>$\epsilon_n$ [µm]</td>
<td>3.75</td>
<td>2.50</td>
<td>2.50</td>
<td>2.20</td>
</tr>
<tr>
<td>$\epsilon_L$ [eVs]</td>
<td>2.50</td>
<td>2.50</td>
<td>2.50</td>
<td>2.50</td>
</tr>
<tr>
<td>Peak luminosity with crab cavities</td>
<td>(1.18)</td>
<td>12.6</td>
<td>11.9</td>
<td>11.6</td>
</tr>
<tr>
<td>[10^{34}\text{cm}^{-2}\text{s}^{-1}]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levelled luminosity for $\mu = 140$[10^{34}\text{cm}^{-2}\text{s}^{-1}]</td>
<td>-</td>
<td>5.32</td>
<td>5.02</td>
<td>5.03</td>
</tr>
<tr>
<td>(inelastic) collisions/crossing $\mu$</td>
<td>27</td>
<td>140</td>
<td>140</td>
<td>140</td>
</tr>
<tr>
<td>(with levelling and crab cavities)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum line density of pileup events during fill [events/mm]</td>
<td>0.21</td>
<td>1.3</td>
<td>1.3</td>
<td>1.3</td>
</tr>
</tbody>
</table>

of $1.5 \times 10^{34}$ cm$^{-2}$s$^{-1}$ to avoid more than 60 simultaneous collisions per bunch crossing. The mitigation of beam dumps induced by electron clouds in 2018 allowed the return to the nominal filling scheme. Other configurations, such as “Batch Compression Merging and Splitting (BCMS)”, have been used to reduce the transverse emittance, in order to avoid a larger than expected emittance growth in the LHC during injection, ramp and squeeze. Table 1.1 compares the nominal LHC parameters with those of the three potential configurations to be considered for the operation of the HL-LHC. The machine parameters have direct implications on the operation and physics performance of the L1 trigger. Challenges, related to the machine filling scheme and pileup, overcome during Phase-1 are described in Ref. [6]. Studies performed in this TDR address the implications of large pileup conditions on triggering performance and exercise the robustness of the design proposed. Unless specified otherwise, all studies presented in this TDR consider pileup conditions of 200, corresponding to the HL-LHC ultimate configuration (not shown in Table 1.1).

1.3 Phase-2 Level-1 trigger requirements and design

With the intense HL-LHC running conditions in terms of instantaneous luminosity and pileup events, the detectors and their associated readout electronics will grow in complexity. As a result, trigger architectures will also become increasingly complex in order to provide sophisticated selection algorithms to ensure the highest possible acceptance for physics already at the hardware level of the data acquisition system. The key technological features of the CMS L1 upgraded trigger system will be the following:

- The extensive use of state-of-the-art FPGAs and processors to achieve optimized
reconstruction, identification, isolation and energy calibration of trigger candidates using high granularity information from the detector.

- The use of high-speed optical links to facilitate the aggregation of data from across the entire detector to allow complex global processing. The availability of the data on the same processing board provides a complete view of the detector for the purpose of the precise evaluation of global quantities such as missing transverse energy or the implementation of pileup mitigation techniques. Moreover, the use of more sophisticated correlations to select specific topologies such as Vector Boson Fusion, where jets are spread over a large polar angle, are facilitated.

- The implementation of a flexible and modular architecture, which can be reconfigured to adapt to different HL-LHC running conditions and physics needs. Extra resources allow more sophisticated quantities to be computed, to give a richer physics menu and increase selectivity.

This section provides a summary of the features of the baseline architecture for the L1 upgraded system and the main characteristics of the algorithms under development. These topics are further developed in Chapter 3 for the algorithms and Chapter 5 for the system design.

### 1.3.1 Phase-2 L1 upgrade conceptual design

The CMS L1 Phase-2 trigger system is designed to benefit from the new features provided by the upgraded sub-detectors to sustain a high efficiency of physics event selection in the very high luminosity regime. The functional diagram of the architecture and data flow of the Phase-2 trigger system is presented in Fig. 1.3. With the 12.5 μs latency, not only is information from the calorimeters and muon detectors used (as in the Phase-1 system), but the information from the new tracker and high-granularity endcap calorimeter can also be included. The total output bandwidth considered is 750 kHz. Given the complexity and large data volume produced by the detector, a significant fraction of the computing of trigger quantities, such as trigger primitives completed by particle identification variables, takes place in the detector backend electronics.

The key feature of the proposed system is the introduction of a correlator layer, which implements sophisticated algorithms producing higher-level trigger objects resulting from the combination of the information of multiple sub-detectors to achieve enhanced selectivity, approaching that of the HLT. To achieve optimum flexibility of the design with the required robustness, four independent data processing paths are implemented: tracking, calorimetry, muon systems, and particle-flow techniques. This division reflects the need to generate complementary types of trigger objects to achieve the best physics selectivity. Each group of objects targets performance for specific physics requirements that can easily be optimized while providing independent trigger criteria essential to the early commissioning of the CMS detector.

**Calorimeter Trigger path:** A barrel calorimeter trigger (BCT) and the HGCAL backend are used to process high-granularity information from the calorimeters to produce high-resolution clusters and identification variables to be used for later processing. Outputs from the BCT, HGCAL and the hadron forward calorimeter (HF) are sent to a global calorimeter trigger (GCT), where calorimeter-only objects such as $e/\gamma$\textsuperscript{1} candidates, hadronically decaying tau leptons ($\tau_h$) candidates, jets and energy sums are built.

**Track Trigger path:** Tracks from the Outer Tracker are reconstructed in the track finder (TF)

\textsuperscript{1}Referring here to photons and electrons not distinguishable without tracking information as it is the case in the current L1.
Chapter 1. Introduction and overview

Figure 1.3: Functional diagram of the CMS L1 Phase-2 upgraded trigger design. The Phase-2 L1 trigger receives inputs from the calorimeters, the muon spectrometers and the track finder. The calorimeter trigger inputs include inputs from the barrel calorimeter (BC), the high-granularity calorimeter (HGCAL) and the hadron forward calorimeter (HF). It is composed of a barrel calorimeter trigger (BCT) and a global calorimeter trigger (GCT). The muon trigger receives input from various detectors, including drift tubes (DT), resistive plate chambers (RPC), cathode strip chambers (CSC), and gas electron multipliers (GEM). It is composed of a barrel layer-1 processor and muon track finders processing data from three separate pseudorapidity regions and referred to as BMTF, OMTF and EMTF for barrel, overlap and endcap, respectively. The muon track finders transmit their muon candidates to the global muon trigger (GMT), where combination with tracking information is possible. The track finder (TF) provides tracks to various parts of the design including the global track trigger (GTT). The correlator trigger (CT) in the center (yellow area) is composed of two layers dedicated to particle-flow reconstruction. All objects are sent to the global trigger (GT) issuing the final L1 trigger decision. External triggers feeding into the GT are also shown (more in Section 2.6) including potential upscope (mentioned as “others”) such as inputs from the MTD. The dashed lines represent links that could be potentially exploited (more details are provided in the text). The components under development within the Phase-2 L1 trigger project are grouped in the same area (blue area). The various levels of processing are indicated on the right: trigger primitives (TP), local and global trigger reconstruction, particle-flow trigger reconstruction (PF) and global decision.

processors as part of the detector backend. The reconstructed track parameters and track reconstruction quality flags are provided to the trigger system to achieve precise vertex reconstruction and matching with calorimeter and muon objects. This key feature maximizes the trigger efficiency while keeping the trigger rate within the allowed budget. A global track trigger (GTT) will be included, to reconstruct the primary vertices of the event along with tracker-only based objects, such as jets and missing transverse momentum. The GTT can also be used
1.3. Phase-2 Level-1 trigger requirements and design

to propagate extra copies of tracks to any other part of the trigger system if need be. Transiting track information through GTT could allow the number of interconnections to be reduced and therefore simplify the logic on the receiving end. Dedicated studies using hardware demonstrators will help evaluate the implication of such an option in terms of latency.

**Muon Trigger path:** While the muon system will maintain its redundant infrastructure of detection chambers, additional stations will be installed to extend the coverage up to \(|\eta| = 2.8\).

The processing of trigger primitives (TP) by muon track finder algorithms is organized as in the Phase-1 system covering three separate regions: barrel (barrel muon track finder, BMTF), overlap (overlap muon track finder, OMTF) and an endcap (endcap muon track finder, EMTF). Standalone muons and stubs as well as L1 tracks are sent to a global muon trigger (GMT). A muon stub contains reconstructed local information extracted from the detector hits in each of the muon stations. It includes position, bend angle, and timing, depending on the station. Beyond the removal of muon duplicates and misreconstructed muons, the main feature of the GMT is the generation of track-matched muons and L1 tracks matched to muon stubs, the so-called tracks plus muon stubs. This is achieved through the propagation of tracker tracks through the layers of the muon system. Tracks can either be received directly from the TF or through the GTT. Interconnections established between GTT and GMT offer the possibility to provide the vertex information to the GMT algorithms if required. Motivated by robustness and latency considerations, the baseline design foresees the tracks to be sent to the GMT directly from the TF.

**Particle-Flow Trigger path:** The correlator trigger (CT) occupies a central role in the design. The CT implements sophisticated algorithms to produce higher-level trigger objects, applies particle identification, and provides a sorted list of objects to the global trigger. Among the algorithms considered to achieve global event reconstruction, the particle-flow reconstruction algorithm, widely used by CMS in other areas, has been chosen and its performance and implementation studied in depth. The structure of the CT is organized in two layers with a first layer, referred to as “Layer-1” producing the particle-flow candidates, which are constructed from the matching of calorimeter clusters and tracks, and a second layer, called “Layer-2”, building and sorting final trigger objects and applying additional identification and isolation criteria. Layer-1 could also host a simplified version of the Pileup Per Particle Identification [15] (PUPPI) algorithm, which can be used to mitigate the degradation of the energy resolution caused by pileup. PUPPI candidates are also transmitted to Layer-2 to be used in trigger object reconstruction (see Section 1.3.2). The GTT transmits the event primary vertex candidates to the CT Layer-1. In addition, the Layer-2 can receive tracker-only objects from the GTT, track-matched muons (and/or muon track plus stubs) from the GMT, as well as calorimeter-only objects from the GCT, and apply further isolation criteria, for example. Isolation of muons may also be calculated in the GMT, depending on available resources.

**Global Trigger:** Outputs from the GCT, GMT, GTT, and CT are combined in the global trigger (GT), which calculates a trigger decision based on a menu of algorithms. The GT has resources to evaluate sophisticated correlation variables among various types of objects to increase the selectivity. The Level-1 Accept signal is transmitted to the Trigger Control and Distribution System (TCDS), which distributes it to the detector backend systems, initiating the readout to the DAQ. The GT provides the interface to external triggers (more in Section 2.6) such as triggers for the precision proton spectrometer (PPS), beam position and timing monitors (BPTX), and luminosity and beam monitoring (BRIL) detectors. A potential upscope of the trigger would include other external inputs, discussed later in this paragraph.

**Scouting System:** The 40 MHz Level-1 Scouting System captures part or all of the intermediate
trigger data streams from the different trigger layers using spare optical outputs of the different processing boards, and enables trigger-less analysis of Level-1 objects at the bunch-crossing rate. The scouting input boards, hosted on standard computers, perform zero suppression and local preprocessing of trigger data, and feed it to host memory. Subsequent processing is carried out asynchronously in a distributed fashion across a high-performance interconnect, and produces feature streams that can be stored and used for analysis or monitoring. These feature streams only contain high-level information derived from the trigger objects, and thus retain the limitations of the Level-1 reconstruction in terms of resolution and purity. Thanks to the limited size of the information stored per bunch crossing, they can however be stored at high-rate and also allow analysis of multi-bunch-crossing correlations. In particular, capturing the inputs and outputs of the Global Trigger at 40 MHz will enable detailed diagnostics of the trigger system at large. It will also enable the detection of anomalies in quasi-real-time in most of the lower level systems, ease the testing of novel GT algorithms, and provide alternative luminosity measurements based on physics processes. Besides its diagnostic and monitoring capabilities, the Scouting System will be used to study physics channels lacking a well defined signature for effective Level-1 rate reduction, in cases where the full detector acceptance and/or full detector resolution is either already approached with L1 primitives and L1 algorithm outputs or not strictly necessary for a competitive measurement.

Architecture overview: This architecture proposal is the result of studies conducted to optimize the number of processing boards and board interconnections, as well as latency, while complying with the need for flexibility and robustness. A division of labor achieved through the implementation of global triggers (GCT, GMT and GTT), allows the reduction of the FPGA resources required to implement the particle-flow algorithm in the CT. Enough headroom is available to further optimize the algorithms, as was the case during the Phase-1 L1 trigger operation. The physical implementation of the architecture is described in Chapter 5. Regional and time-multiplexed\(^2\) architecture options are both evaluated. The design exploits the availability of new hardware technologies to ensure the computing power along with high-speed data transfer (up to 28 Gb/s) to provide a global detector view. The system will handle a huge amount of input data bandwidth (63 Tb/s compared to \(\sim 2\) Tb/s today). Inspired by the technological choices made during the Phase-1 upgrade of the L1 trigger, the data processing units are mainly designed as generic stream-processing engines instead of custom-designed processors. Generic processors can be used to perform any task within the system, allowing to consider original trigger architectural options adapted to the requirements inherent of a flexible Correlator design. The electronics boards under development feature large FPGAs such as Xilinx VU9P equipped with 28 Gb/s transceivers. A small level of customisation remains to address specific requirements such as lighter processing tasks or fast memory access, etc. More on this topic is described in Chapter 6. Results of hardware demonstrators are included as well. The paragraph 1.3.2.1 describes the inputs to the L1 Phase-2 trigger system and their use by the sophisticated trigger reconstruction algorithms under development.

Potential future upscopes: The design of the Phase-2 L1 trigger system provides provision for more input bandwidth and latency for later potential upscopes, for example from the MTD [12]. Timing information from MTD could help selecting displaced signals and complement the timing information provided by the calorimeters and muon systems.

\(\text{\footnote{In a time-multiplexing (TM) approach, } N_{TM} \text{ processors run identical algorithms, on different events. The same data may also be sent to multiple boards, which run different algorithms as described in Ref. [16].}}\)
1.3. Phase-2 Level-1 trigger requirements and design

1.3.2 Trigger algorithms for the HL-LHC

The envisaged L1 system will more closely replicate the full offline object reconstruction, instead of making use of simple subsystem variables, to make a better optimized selection. The Phase-2 trigger algorithms foreseen can be used to reconstruct a large variety of objects: standalone, which are reconstructed from single detector information (including tracker-only objects), standalone matched to L1 tracks, and particle-flow [13]. The trigger decision can rely on the complementarity of these objects to achieve the best possible efficiency while keeping the trigger rate under control. The trigger system described in this document aims to achieve the most optimized selection of collision events through a global event reconstruction that has proven to be efficient to reach the highest sensitivity in offline data analyses. Given the physics drivers listed in Section 1.1, the trigger object requirements are not only driven by the need to maintain physics selection thresholds to match those of Phase-1, but also by having to provide the selection of exotic signatures, including displaced objects. The algorithms described here are developed keeping in mind potential features that could help expand the physics reach while providing robustness in view of effects introduced through detector ageing, for example. The algorithm implementation in firmware (see Chapter 3) greatly benefits from the introduction of High-Level-Synthesis software that could be used to design advanced machine learning trained variables or even iterative processes in the core of the trigger system. This section provides an overview of the baseline algorithms that have been developed with the minimum requirement of meeting the challenges of the HL-LHC. New developments and further improvements are certainly foreseen beyond the scope covered by this TDR. A short description of the trigger primitives inputs is reported as well. In view of the potential ultimate configuration of the HL-LHC machine, all studies were carried out with simulated data samples containing 200 pileup events, unless specified otherwise.

1.3.2.1 Inputs to the L1 trigger

The Phase-2 upgraded calorimeter trigger will benefit from the enhanced granularity provided by the upgraded barrel [9] and endcap calorimeters. The ECAL barrel (EB), Very Front-End (VFE), and Front-End (FE) electronics will be fully replaced with boards equipped with high-speed optical links (5 Gb/s, compared to 800 Mb/s in the Phase-1 system) in order to transmit the single crystal energy information for triggering purposes, achieving a granularity increased 25 times with respect to the Phase-1 system. The HCAL barrel (HB) on-board electronics readout will be replaced to retrieve the depth information, providing 4 times more granular information with respect to the current system. Trigger primitive information from both ECAL and HCAL will carry timing information with an expected resolution of 30 ps and 500 ps for objects with $p_T > 50$ GeV, respectively. Similar trigger primitives information as in Phase-1 is expected from the hadron forward (HF) calorimeter. The endcap calorimeters will be replaced by the HGCAL instrumented with over 6 million readout channels [10]. It is a 3D sampling calorimeter with alternating layers of silicon sensors/scintillator and copper/lead absorbers. The goal is to achieve unprecedented spatial resolution and shower separation to optimise the matching with tracks. About half of the 50 layers (28 electromagnetic and 22 hadronic), representing a total of roughly 990 000 trigger channels, can be exploited for triggering purposes, hence increasing the granularity by a factor of more than 500 compared to the Phase-1 system. The HGCAL trigger primitives consist of the summed energy of 4 adjacent channels. Channels correspond to a readout cells of roughly $1 \times 1$ cm$^2$ paving each calorimeter layer. The trigger cell information are grouped and sent to the backend electronics responsible for producing the trigger primitives. This system is organised in two layers where the first layer is used to calibrate and reorganise the data in time-multiplexed fashion and the second layer reconstructs
3D clusters. The 3D clustering algorithm is performed on a full depth view of the detector, hence exploiting the longitudinal and transverse shower profiles specific to each physics object. Clusters are transmitted to the Phase-2 L1 trigger along with variables to discriminate electromagnetic, hadronic and pileup-induced energy deposits.

One of the key features of the Phase-2 trigger upgrade system is the addition of charged particle track reconstruction. Standalone reconstructed trigger objects will suffer from high trigger rates and, therefore, reconstruction algorithms will greatly benefit from matching with tracks. The higher $p_T$ resolution of the tracks reduces contributions to the rate coming from $p_T$ below the trigger threshold. Tracking information can also be used to compute photon and lepton isolation, improve jet and energy sum trigger reconstruction, provide vertexing, estimate the level of pileup (not only calorimeter-based estimators as in Phase-1), etc. CMS plans to replace both the silicon Pixel and Strip tracking detectors. The Outer Tracker on-detector electronics will generate “stubs” that the track finder system will associate into track candidates to be transmitted to the L1 trigger system. Stubs are computed from closely-spaced silicon-sensor modules composing each tracker layers. The bending of tracks between each side of the modules is used to discard hits from low $p_T$ tracks and reduce, on detector, the hit rate to a manageable level. The track finder coverage in pseudorapidity extends up to $|\eta| = 2.4$. The overall latency is estimated to 5 \(\mu\)s (including 1 \(\mu\)s for data transmission). A total of around 1000 tracks with $p_T > 2$ GeV can be reconstructed per event at 200 pileup in the Phase-2 LHC environment. The track finder primitives are reconstructed with a hybrid algorithm performing a Kalman filter fit on associated stubs to a given seed or “tracklet” (also formed from adjacent layers stubs). These L1 tracks are transmitted to the Phase-2 L1 trigger and contain all the fitted parameters ($p_T$, $\eta$, $\phi$, $d_0$, and $z_0$), along with the track quality. Another approach, referred to as "Extended L1 tracking" in the text, could be used to reconstruct displaced trajectories originating from potential BSM signatures. The principle lies in the combination of three layers and/or disks, called triplet seeds, that can be formed without beamspot constraint in addition to the tracklet seeds. While not in the current baseline, studies are being conducted to evaluate the possibility of such an algorithm even for standard objects.

Although most of the muon detectors will remain in place, their associated readout electronics will be replaced, providing finer information and improved timing from 25 ns to 1.5 ns. This information can then be used to develop more sophisticated trigger primitives algorithms with enhanced performance, close to present offline reconstruction. Improved DT trigger primitive generation is based on hit information in each layer and correlations among super-layers allowing enhanced bunch crossing assignment. Each RPC hit will carry time information with a granularity of one sixteenth of a 25 ns BX period. Trigger primitives from DT and RPC are combined into super-primitives to exploit the redundancy and complementarity of these two sub-detectors, thus achieving better performance and being resilient to ageing-induced effects. The CSC on-chamber electronics, located in the inner rings will also be replaced to handle higher trigger rates and potentially improve robustness of the trigger primitive generation. A larger improvement in the bending direction resolution could be achieved by combining CSC and GEM trigger hits. With improved RPC chambers, the full redundancy of the muon spectrometers is performed in the forward region of the detector. Trigger primitives are computed from clustered strip hits providing excellent timing resolution. The GEM chambers are essential to achieve a precise measurement of the muon bending angle in the forward region $1.6 < |\eta| < 2.8$. GEM trigger primitives are obtained from the clustered information generated by dedicated trigger pads mounted on the detectors. These TPs can then be combined with CSC or directly transmitted to the Phase-2 endcap muon track finder (EMTF). Both the GE1/1 chambers installation and the CSC on-chamber electronics replacement are being performed
1.3. Phase-2 Level-1 trigger requirements and design

during LS2.

1.3.2.2 Triggering on electrons and photons

Many standalone electron and photon trigger reconstruction techniques are being investigated to optimise both the response and the position resolution for the purpose of achieving the highest possible track matching efficiency. The proposed algorithms utilize both identification criteria and isolation variables based on calorimeter, as well as tracking information to reduce the background level. Given the intense running conditions foreseen, the algorithms are designed to be pileup resilient.

The electron finder in the barrel region uses the crystal information from the ECAL. A $5 \times 3$ crystal matrix ($\Delta\phi \times \Delta\eta = 0.087 \times 0.052$) is used to define the maximum size of the electron footprint in the ECAL. As in the Phase-1 algorithm approach, the extension in $\phi$ is motivated by the necessity to recover energy lost through bremsstrahlung. An improved position resolution is achieved using a weighted-energy sum around the seed crystal (the seeding threshold used is $E_T > 1 \text{ GeV}$). Extra shower shape features are used as identification criteria and the matching of the clusters with tracks is performed using an extrapolation to the ECAL surface.

The starting point of the electron reconstruction algorithm in the endcap region is the cluster reconstructed in the backend electronics of the HGCAL. Further identification of the electromagnetic object is performed through a multivariate approach optimized to exploit the input variables transmitted from the HGCAL. Dedicated boosted decision trees (BDTs) are trained on signal and background to achieve an optimal signal efficiency while rejecting pileup-induced clusters. Bremsstrahlung recovery is performed as well as an energy calibration of the final e/\(\gamma\) candidate. The availability of the tracking information facilitates the reconstruction of isolated photon candidates. The reconstruction of electron tracks with the TF is inefficient because of the radiation induced by bremsstrahlung throughout the tracker material. The extended tracking, originally designed to reconstruct displaced trajectories, could help recover track reconstruction efficiency.

1.3.2.3 Triggering on jets, taus and energy sums

Triggering on hadronic signals has always represented a challenge for detectors operating in an intense hadronic environment. Algorithms developed for the Phase-1 Level-1 trigger system are optimized to provide thresholds adequate for physics using calorimeter-only information [6]. The rate of such triggers measured on Run-2 collision data and, in particular, trigger conditions based on missing transverse energy, display a strong dependence on the level of pileup events and, to a certain extent, the filling scheme used by the machine. Dedicated pileup mitigation techniques have been developed to maintain the trigger rate to acceptable levels, preventing a dramatic rise of thresholds. The level of pileup expected at the HL-LHC will seriously impact the performance of hadronic triggers. This motivates the detailed studies conducted here to provide pileup mitigation algorithms exploiting the full capabilities of the Phase-2 detectors. As calorimeter-only algorithms are expected to have high thresholds, complementary approaches are proposed with tracker-only information, track-matched calorimeter objects and higher-level objects reconstructed with particle-flow and PUPPI inputs. Calorimeter-only jet finding algorithms use barrel ECAL and HCAL information, endcap HGCAL and forward HF information. Although various configurations were considered, a simple square geometry of $7 \times 7$ trigger towers (see Section 3.6 for the trigger tower definition) around a local maximum gives acceptable performance while keeping the pileup contribution to a minimum. More details are provided in Section 3.6. The jet window definition corresponds approximately to the cone size of 0.4 used by the offline anti-$k_T$ algorithm [17]. Similarly to the
Phase-1 algorithm, a tower-by-tower pileup correction depending on the level of pileup and $\eta$ is applied prior to jet clustering.

Tracker-only jet finding is performed on a set of tracks from the track finder [8] passing purity requirements to keep the trigger object resilient to pileup. Track clustering makes no use of the primary vertex information. In order to optimize the latency, primary vertex computation and jet clustering are performed in parallel. Note that more robustness against pileup is obtained by considering a smaller $z$-range of tracks from the main interaction point. The clustering of tracks in the $\eta - \phi$ plane is performed using a nearest-neighbor approach in two one-dimensional steps. The maximal jet size corresponds to $\Delta R = \sqrt{\Delta \eta^2 + \Delta \phi^2} = 0.3$ while the jet $p_T$ is computed as the sum of each track $p_T$ associated to it. The extended tracking mentioned previously could provide displaced tracks that could be clustered to reconstruct displaced jet objects. The performance of such objects is studied in this document as well. The particle-flow based jet finding consists of building jets from particle-flow candidates grouped into pseudo trigger towers (equivalent to $0.083 \times 0.087$ in $\eta \times \phi$), which are then clustered into a $7 \times 7$ tower window around a local maximum. The jet momentum is computed as the sum of the objects’ momentum within $7 \times 7$ window and the jet position is that of its seeds ($\eta, \phi$) coordinates. Various pileup mitigation techniques were investigated based on jet energy correction scale factors or pileup energy subtraction. Each jet finding algorithm has a dedicated energy calibration, detailed in Chapter 3. The jet $p_T$ and direction can then be used to compute $H_T$ (scalar sum of jet $p_T$) and $H_T^{\text{miss}}$ (vector sum of jet $p_T$).

The benefit of developing a dedicated hadronically decaying tau ($\tau_h$) reconstruction and identification algorithm has been fully demonstrated during Run-2 of the LHC [18]. As for the jet finding algorithm, calorimeter-based $\tau_h$ are built from trigger towers. Given that $\tau_h$ are narrow jets and that several decay products may be producing more than one cluster spatially separated along the $\phi$ direction due to the magnetic field, a $3 \times 5$ tower window ($\Delta \eta \times \Delta \phi = 0.261 \times 0.435$) is chosen to optimise $p_T$ resolution. Similarly to the Phase-1 algorithm, a window of $7 \times 7$ trigger towers around the core objects defines the isolation region that is used to identify further the $\tau_h$ candidate while maintaining the rate under control. Although this approach performs well within the entire calorimeter acceptance, the enhanced granularity of the HG-CAL detector allows the implementation of advanced identification techniques exploiting the $\tau$ shower characteristics. The shower profile being different to that of pileup-induced particles, stringent discrimination can be achieved. A dynamic clustering of 3D-clusters gives optimum response, while trained BDTs can provide dedicated energy calibrations for each of the $\tau$ decay modes. Algorithms matching the current offline reconstruction of $\tau_h$, called ”hadron plus strips” [19] have also been developed. These algorithms have been studied considering both PUPPI or particle-flow particle candidates. In the case of tau leptons, the isolation variable is essential to further reduce the background. The track isolation is calculated from tracks that are within $|\Delta z| < 1.0$ cm of the $\tau_h$, where $z$ is the distance of the tracks along the beam line from the interaction point. When PUPPI is utilized, this can further be enhanced by including the neutral candidates as well. Identification can also rely on neural network approaches that, despite the complication of firmware implementation, give excellent performance when combined with the PUPPI-based approach.

Triggering on missing transverse energy ($E_T^{\text{miss}}$) is a particularly challenging task for detectors operating in hadronic environment, especially when the average expected pileup is 200. This quantity is a key input for many signatures, including beyond Standard Model processes, in the L1 trigger. The use of L1 tracks is essential to achieve manageable rates for moderate

---

3To a certain extent, the HCAL longitudinal segmentation could be used to perform similar discrimination of pileup-induced jets (more in this discussed in Section 2.2.2)
1.3. Phase-2 Level-1 trigger requirements and design

thresholds. The algorithms pursued are either tracker-based or PF-based. The tracker-based approach considers tracks originating from the primary vertex and applying dedicated selection to reject misreconstructed tracks. The rate is considerably reduced using this approach. With particle-flow, the information of all sub-systems is used and further mitigation of pileup contributions is obtained with PUPPI. Thresholds applied to particle-flow and PUPPI inputs in various $\eta$ regions can be adjusted to optimize performance.

1.3.2.4 Triggering on muons

The overall structure of the muon system for Phase-2 remains similar to the current one. It is composed of three partially-overlapping sub-detectors (CSC, DT, and RPC), whose signals are combined to reconstruct muons and measure their transverse momenta. Additional muon stations, such as iRPC, GEM and ME0, are installed in the forward region to extend the acceptance to $|\eta| = 2.4$ and $|\eta| = 2.8$, respectively. Following the approach of the Phase-1 trigger upgrade, the reconstruction of standalone muons uses information from all available sub-detectors simultaneously to build tracks in three distinct pseudorapidity regions, improving the muon reconstruction and increasing signal efficiency while reducing background rates. Given the improved sub-detector electronics readout for Phase-2, the muon chambers will provide finer information along with precise timing ($\sim 1.5$ ns) that can be exploited by the muon track finding algorithms. Each track finder uses an optimised track reconstruction algorithm and $p_T$ assignment logic, and assigns a track quality corresponding to the estimated $p_T$ resolution. Similarly to the existing system, the BMTF uses DT and RPC trigger primitives to reconstruct segments merged to obtain a muon candidate. The RPC fired strips are clustered before being used by the muon track finders. A track finding approach based on a Kalman Filtering technique called Kalman barrel muon track finder (KBMTF) has been developed and tested on Run-2 collision data. As tracks can be reconstructed by KBMTFT with and without a constraint forcing them to originate from the primary vertex, displaced muons can be reconstructed with acceptable $p_T$ resolution resulting in higher acceptance. In the overlap region, the OMTF receives data from DT, RPC and CSC stations and reconstructs tracks by associating hits, using generated patterns from simulated events. This naive Bayes-classifier approach identifies the most likely muon $p_T$. The muon endcap track finding algorithms exploit the information from up to 12 muon stations. In addition to CSC and RPC trigger primitives, the Phase-2 EMTF++ system proposed for this upgrade receives information for GEM (including ME0) and iRPC detectors. The standalone reconstruction algorithm looks for correlated CSC trigger primitives through multiple stations compatible with a muon track corresponding to predefined patterns. Consistent RPC primitives are associated to this track candidate and a trained deep neural network (DNN) for $p_T$ assignment, with and without beam constraint, is implemented. GEM detectors cover $|\eta| > 1.6$ (GE1/1) and $|\eta| > 1.8$ (GE2/1). For $|\eta| > 1.6$, EMTF++ can use combined GEM-CSC bend angle. The muon reconstruction in the forward region beyond 2.1 in pseudorapidity uses the position and bend angle from ME0 (up to $|\eta| < 2.8$) and CSC (up to $|\eta| < 2.4$).

The availability of tracks from the Outer Tracker allows another category of muons, with increased acceptance at low $p_T$ or originating from regions with limited detector coverage, to be considered. The matching of standalone muons and tracks is performed optimally in each pseudorapidity region, so that, as in the offline or Phase-1 HLT cases, misreconstructed muons are reduced, and the $p_T$ measurement accuracy is improved. Another complementary approach consists of propagating the tracks from the Outer Tracker into the muon detectors and associate stubs from at least two layers of the muon stations. This algorithm shows optimum performance for a large variety of physics signals while maintaining efficiency and providing robustness against detector aging. The possibility to correlate tracking and muon stubs information is used to produce trigger objects adequate to identify heavy stable charged particles.
Chapter 1. Introduction and overview

(HSCPs). Given the particularity of this signal, the candidate L1 track is matched to muon stubs from the same event or subsequent ones. Other combinations of L1 tracks and muon stubs can form topological objects appropriate for the selection of \( \tau \rightarrow \mu\mu\mu \) events, for example.

1.3.2.5 Reconstruction of the primary vertex and particle-flow objects

The reconstruction of the primary collision vertex is an essential ingredient of the physics object reconstruction. The provision of tracks from the track finder allows the main interaction point and its associated tracks to be identified. Various algorithms were developed and tested on different types of final states. The studies considered busier events such as \( \bar{t}t \) producing jets with high track multiplicity and processes with less tracks such as \( Z \rightarrow \mu\mu \). The algorithms range from a simple histogram-based approach to more sophisticated ones based on machine learning techniques or even iterative implementations. Firmware considerations help select the optimum one. While performance can reach 85\% of vertex identification efficiency for \( \bar{t}t \) events, its lower score for \( Z \rightarrow \mu\mu \) can be recovered if multiple vertex candidates (at least 3) are considered. The consequence in terms of firmware resources needs careful studies.

The availability of tracking information and the enhanced calorimeter granularity at the Level-1 trigger allow us to contemplate the exciting possibility to propose algorithms matching the performance of those implemented at the HLT. The benefit of introducing a full event description (such as the one obtained through the particle-flow reconstruction algorithm [13]) have been demonstrated with the Phase-1 HLT system. The Phase-2 trigger Correlator processor (see Fig. 1.3) will be used to combine all detector information to produce a list of candidates from which higher-level trigger objects are constructed, such as identified prompt leptons, photons and jets, as well as global quantities such as missing transverse energy or hadronic transverse energy sum \( (H_T) \) at L1. Dedicated energy calibration factors are derived for each sub-detector input leading to an improved response for single particles. In addition to the particle-flow algorithm, a simplified version of the offline PUPPI algorithm can be implemented. PUPPI relies on the knowledge of the vertex position but can use other mechanisms to discriminate against pileup outside the tracker volume. In addition to the particle-flow candidates, PUPPI particle candidates can also be used to reconstruct complex objects or isolation variables. The firmware design is simple and able to process all input objects in parallel for pipelined execution on FPGA, in contrast to the sequential approach used offline, adapted to CPU. The firmware structure chosen has direct consequences on the correlator architecture organization and, in particular, the regionalization of the detector to perform optimal processing with a short and fixed latency.

1.3.2.6 Global Trigger Algorithms

Global trigger algorithms refer here to algorithms based on correlations among physics objects or using more sophisticated variables such as invariant masses. The ability to trigger on such variables has significantly enhanced the selectivity of the Phase-1 trigger system and therefore this feature is foreseen for its Phase-2 upgrade. Naturally, these algorithms are implemented in the GT, where tailored triggers for specific physics analyses are provided. In addition, the GMT, GCT and GTT systems are also used to generate quantities based on information only available upstream of the correlator trigger and the GT. Specific objects or variables can be propagated through direct connection and combined with other physics objects. For example, the GTT can compute invariant masses of combinations of tracks passing certain quality cuts to trigger on particular light resonances. Provision is made to eventually receive timing information from MTD to flag out-of-time physics objects predicted in beyond Standard Model processes. Other global quantities such as centrality, commonly used for heavy ion triggering purposes,
can be implemented. Machine learning approaches are also pursued as alternatives to simple cut-based trigger conditions on object combinations. Specific software such as hls4ml [20] has rendered this type of techniques implementable in FPGAs. Preliminary results on selecting signatures such as the production of the Higgs boson through VBF, with the H → bb and invisible decay (H → inv) channels, show significant improvements with respect to standard triggers.

1.3.2.7 Examples of the Phase-2 L1 trigger algorithms performance

In this Section, examples of the Phase-2 Level-1 trigger algorithm performance are presented. The performance of all algorithms proposed are presented in Chapter 3. Figure 1.4 (left) displays the $H_T$ trigger efficiency for simulated $t\bar{t}$ events with 200 average pileup events as obtained for a fixed rate of 10.5 kHz. The performance of $H_T$ computed with jets reconstructed with the PUPPI algorithm is compared to calorimeter-only, tracker-only quantities. The PUPPI-based $H_T$ approach outperforms the other algorithms by optimally exploiting both the tracking information and enhanced detector granularity. The muon trigger performance presented in Fig. 1.4 (right) demonstrates the performance of the barrel standalone muon trigger reconstruction algorithm. This algorithm can provide a sustainable rate for muons originating from long-lived particles that appear as displaced trajectories in the detector. The impact parameter in the transverse plan is referred to as L1 $D_{xy}$.

---

*Figure 1.4: Left: The signal efficiency is shown for $t\bar{t}$ events selected by $H_T$ triggers utilizing different sets of inputs at a fixed selection rate of 10.5 kHz. The performance of the $H_T$ trigger quantity computed with PUPPI jets, tracker-only jets and calorimeter jets are compared. Right: Single muon rate as a function of the L1 muon $p_T$ for the Phase-1 BMTF (red), the Phase-2 prompt (black) and the displaced KBMTF (blue) reconstruction algorithms. The displaced KBMTF muon algorithm takes the maximum of the vertex constrained and unconstrained $p_T$. Displaced muon rate for L1 $D_{xy} \geq 75$ cm is also shown (magenta) as a function of the L1 muon $p_T$ threshold. $D_{xy}$ refers here to L1 trigger reconstructed quantity.*

---

4The method used to compute the rates throughout this document is explained in Section 3.1.
Chapter 1. Introduction and overview

1.4 Physics reach of the Level-1 Phase-2 Trigger

The HL-LHC will provide the experiments with the largest dataset ever collected. Prospect studies of the direct and indirect discovery potential offered by HL-LHC have been documented in a “Yellow Report” [1]. For some of the studies reported in this document, it is assumed that the selection criteria used in Run-2 analyses can be maintained, while for others, larger physics acceptance compared to the current one was considered. Consequently, the analyses reported in Ref. [1] manifest how crucially the HL-LHC physics program relies on the capabilities of the trigger system to maintain comparable physics acceptance to that of Phase-1 and to collect further enriched datasets covering an extended phase-space, not reachable previously. These requirements are discussed in Section 1.4.1. The extended physics reach achieved through the new features of the Phase-2 L1 trigger system is presented in Section 1.4.2. The gain in acceptance with respect to reference Phase-1 physics channels is described, followed by a presentation of the improvement expected on several SM measurements and BSM searches. Throughout this section, benchmark physics channels were chosen for these studies. Targeted trigger thresholds delivered by the upgraded Phase-2 Level-1 trigger system are specified.

1.4.1 Preserving the CMS discovery potential at the HL-LHC

In order to substantiate the capabilities of the Phase-2 L1 Trigger to maintain the current physics acceptance in the extreme scenario of 200 average pileup events per crossing, a simplified L1 trigger menu, covering most of the CMS Run-2 physics selection strategy, was developed (see Chapter 4). The development of the trigger menu represents an important step in demonstrating the successful implementation of the Phase-2 trigger strategy. A typical L1 trigger menu is composed of a set of criteria applied to single-, multi-, and cross-objects targeting a variety of final states. Typical criteria include lower thresholds on the object’s transverse momentum and conditions on its pseudorapidity. Trigger algorithms include: single lepton triggers, double lepton and photon triggers, hadronic triggers (based on jet $p_T$, scalar sum of jet $p_T (H_T)$ or missing transverse energy ($E_{T \text{miss}}$)), cross lepton triggers, cross hadronic-lepton triggers, specific soft leptons triggers (targeting B-physics signals) and specific hadronic triggers (targeting VBF topologies). The menu presented in Chapter 4 contains an ensemble of such algorithms, with $p_T$ thresholds set to the ones that were used at the end of Run-2. A similar exercise was performed previously in Ref. [21], with equivalent Phase-2 LHC running conditions but using the Phase-1 trigger design and algorithms, hence not including tracking capabilities. The total Level-1 trigger rate attained was approximately 1500 kHz for 140 pileup events and almost 4000 kHz for 200 pileup events, well beyond the 750 kHz originally foreseen. This study alone motivated the inclusion of tracking information into the Phase-2 Level-1 trigger. What follows is not intended to summarise the menu developments but to illustrate the capability of the upgraded trigger to provide low enough thresholds for a few selected CMS Phase-2 physics benchmarks. To emphasise further the need for this upgrade, the L1 trigger thresholds used in Run-2, and allowed by the proposed system, are compared to the reference thresholds the current trigger would need to apply to reach the same target rate under HL-LHC running conditions.

1.4.1.1 Final states requiring lepton trigger algorithms

- **Higgs boson associated production with a W/Z boson (HW/Z).** Muons and electrons from the decay of the W/Z bosons are particularly important to trigger the selection of events in which the Higgs boson is produced in association with a W or a Z boson and decays into hadronic objects with soft $p_T$. Triggers targeting this final state generally do not include conditions on the hadronic objects produced through the Higgs decay, given the large associated background rate. As a result of their
clear signature, lepton triggers are of utmost importance in a hadronic environment. Hence, triggers based solely on leptons allow for an inclusive selection that can target a broad range of final states; they have been successfully used for several Phase-1 analyses. This is particularly relevant for searches for physics beyond the SM, such as searches for Higgs boson decays into light pseudoscalars that subsequently decay into hadrons. The analyses targeting such rare processes see their acceptance significantly limited by any increase of the lepton $p_T$ thresholds of the single- or double-lepton triggers used to select these events.

- **Single top-quark or top-quark pair production ($t\bar{t}$).** Precision measurements of the top-quark properties are an important test of the Standard Model. In particular, flavor changing neutral current (FCNC) couplings of the top-quark, which are highly suppressed in the SM and for which the current experimental constraints are far from approaching the SM predictions, offer interesting opportunities to probe BSM models, for example via studies of single top-quark production. Other processes involving the top-quark such as the associated production of the Higgs boson with a pair of top-quarks ($t\bar{t}H$) or the production of four top-quarks, would equally rely on lepton triggers.

- **Higgs boson pair production** $HH \rightarrow \tau\tau bb \rightarrow \ell\nu\nu_T\tau\nu_T\nu\nu_T bb$. The study of Higgs boson pair (HH) production is a crucial test of the Standard Model as it grants direct experimental access to the Higgs boson self-coupling and thus the characteristics of the Higgs potential itself. The most recent prospect for the study of HH production at HL-LHC with the CMS detector results in an expected significance for the Standard Model signal of $2.6\sigma$ [1], assuming the capability of the Phase-2 L1 trigger to achieve comparable thresholds as during LHC Run-2 operations. The measurement of the Higgs boson self-coupling requires the detection of Higgs bosons pairs with an invariant mass close to the kinematic production threshold of $2 \times m_H$, where a deviation of the self-coupling from its SM value gives the largest variations. This results in soft objects in the final state where leptons can provide a clean signature for triggering. In the $HH \rightarrow \tau\tau bb$ state, one of the leading channels for HH production observation at the HL-LHC, final states where the $\tau\tau$ system decays to $\ell\nu_T\tau_h$ and neutrinos can give access to the low HH invariant mass region for the self-coupling determination and can be recorded with single-lepton triggers.

Figure 1.5 (left) shows the inclusive $p_T$ distributions of electrons and muons coming from the $HH \rightarrow b\bar{b}\tau\tau \rightarrow b\bar{b}\tau_T\tau_h (\ell = e, \mu)$ process, single-lepton $t\bar{t}$ and single-top production at a center-of-mass energy of 14 TeV; $\tau_T$ refers to leptonically decaying $\tau$ leptons. The single electron/muon trigger thresholds allowed by the upgraded Phase-2 L1 trigger are compared to the ones that would be expected without the use of tracking information when reconstructing those objects. By definition, the meaning of these trigger thresholds is that the efficiency of the single-lepton trigger, as a function of the $p_T$ of offline reconstructed leptons, reaches 95% of its plateau efficiency at the quoted value. The target thresholds for single leptons of 15 GeV (muon) and 28 GeV (electron) can be sustained with the Phase-2 algorithms, which associate L1 tracks with standalone calorimeter or muon objects, while with standalone objects only, these thresholds would have to be increased to 24 GeV and 62 GeV respectively, for an equivalent rate of 42 kHz (muon) and 24 kHz (electron). The significant gain in acceptance for the aforementioned analyses is clearly visible in Fig. 1.5 (left). This gain is particularly essential to the search for Higgs boson pair production $HH \rightarrow b\bar{b}\tau\tau \rightarrow b\bar{b}\tau_T\tau_h$, where the lepton from the $\tau$ decay displays a soft $p_T$ spectrum. Additionally, Fig. 1.5 (right) illustrates how the expected exclusion limit at 95% confidence level (CL) on the FCNC $t \rightarrow u\bar{g}$ branching fraction would degrade with an
increased lepton $p_T$ threshold used in the analysis [1]. See Section 4.2 for more details.

Figure 1.5: Left: Simulated distributions of transverse momentum of the electron and muon produced in HH, single top-quark, and the semileptonic decays of $t$ quark. The vertical lines correspond to the offline $p_T$ thresholds at which the single object trigger efficiency reaches 95% of the efficiency plateau. The solid vertical lines correspond to the trigger thresholds provided by the Phase-2 L1 trigger system (at 200 pileup) matching the thresholds currently deployed by the L1 menu for Run-2. The dashed vertical lines correspond to the trigger thresholds required to achieve the same rate using trigger algorithms that do not make use of L1 tracks. Right: The expected exclusion limits at 95% CL on the FCNC $t \rightarrow ug$ branching fraction as a function of the offline leptons $p_T$ threshold.

### 1.4.1.2 Final states requiring double-photon trigger algorithms

- **Higgs boson pair production** ($HH \rightarrow \gamma\gamma bb$). As shown in Ref. [1], one of the most sensitive decay channels to access di-Higgs production is $HH \rightarrow \gamma\gamma bb$, where the event selection relies on a double-photon trigger with thresholds as low as those used in Phase-1. Harvesting these rare events would contribute to obtaining evidence of the HH process, which constitutes one of the main goals of the CMS Phase-2 physics program.

- **Higgs boson decay into photons** ($H \rightarrow \gamma\gamma$). This final state benefits from the complete reconstruction of the Higgs boson kinematics and from the clean signature of the diphoton invariant mass. Hence, this channel is particularly suited to perform the measurement of the Higgs boson differential cross sections and in particular of the Higgs boson $p_T$ spectrum, including the low $p_T$ regime. During Phase-1, this major discovery channel relied on the double-photon trigger. The baseline strategy to pursue this analysis remains similar to Phase-1 and therefore the trigger thresholds applied should sustain a selection as inclusive as possible.

Figure 1.6 displays the inclusive $p_T$ spectrum of the sub-leading photon in single and double Higgs boson final states. In the case of Higgs boson pair production, one of the Higgs bosons decays into photons. The trigger threshold on the sub-leading leg of the double-photon trigger allowed by the upgraded Phase-2 L1 trigger is compared to the one expected without any tracking information used. The Phase-2 photon objects reconstruction exploits both
the full granularity of the Phase-2 calorimeters and the L1 tracking information. The latter al-

1.4. Physics reach of the Level-1 Phase-2 Trigger

1.4. Physics reach of the Level-1 Phase-2 Trigger

ows track-isolation requirements to be used to select isolated photons, which gives improved

21

performance compared to isolation criteria based on calorimeter information alone as imple-

21

mented by the Phase-1 L1 trigger. Exploiting the new features of the Phase-2 L1 trigger, a

21

threshold of 12 GeV can be implemented while a threshold of 20 GeV would be needed with-

21

out the tracker information, for an equivalent rate of 50 kHz (see Section 4.2 for more details).

40

80

120

$[\text{GeV}]$

20

40

60

80

100

120

$\text{p}_{T}$

3.5

3.0

2.5

2.0

1.5

1.0

0.5

0.0

CMS Phase-2 Simulation

14 TeV

Thresholds for a rate of 50 kHz

Figure 1.6: Simulated distribution of transverse momentum of the lowest $p_T$ photon from

HH $\rightarrow$ $\gamma\gamma b\bar{b}$ and $H \rightarrow \gamma\gamma$ processes. The vertical lines correspond to the offline $p_T$ thresh-

olds at which the single object trigger efficiency reaches 95% of the efficiency plateau. The solid

vertical lines correspond to the trigger thresholds provided by the Phase-2 L1 trigger system

(at 200 pileup) matching the thresholds currently deployed by the L1 menu for Run-2. The

dashed vertical lines correspond to the trigger thresholds required to achieve the same rate

using trigger algorithms that do not make use of L1 tracks.

1.4.1.3 Final states requiring double-$\tau$ lepton trigger algorithms

- Higgs boson pair production ($HH \rightarrow \tau\tau b\bar{b}$). As mentioned, one of the most sensi-

1.4.1.3 Final states requiring double-$\tau$ lepton trigger algorithms

tive channels for the search of double Higgs boson production at HL-LHC is $\tau\tau b\bar{b}$. Most of the analysis sensitivity is provided by the final states where both $\tau$ leptons
decay into hadrons and a neutrino. The HH signal acceptance in this final state is

limited by the L1 trigger thresholds. During the Phase-1 running, this channel relied

on a double-$\tau$ trigger with low enough thresholds to provide an event selection as inclusive as possible. While more features of this specific final state could be ex-

ploited at Level-1, using in particular the additional jets from the $H \rightarrow b\bar{b}$ decay,
a similar strategy is used here for the sake of evaluating the impact of Phase-2 $\tau$

reconstruction on the signal acceptance.

Figure 1.7 (left) shows the inclusive $p_T$ distribution of the sub-leading $\tau$ in events where two

Higgs bosons are produced, one decaying into a pair of $\tau$ leptons and the other into a pair of

$b$-quarks ($HH \rightarrow \tau\tau b\bar{b}$). The threshold on each leg of the double-$\tau$ trigger, allowed by the

Phase-2 L1 trigger, is compared to the expected threshold without any tracking or particle-

flow information used by the reconstruction algorithm. These thresholds are defined as the $p_T$
of the offline reconstructed tau at which 50% of the single object trigger efficiency plateau is

reached. The $\tau$ trigger algorithm deployed by the Phase-2 L1 trigger is based on a particle-flow
reconstruction approach combined with a dedicated neural network discriminator to select isolated $\tau$ candidates. The mitigation of pileup performed by PUPPI (see Section 3.5) allows the (36, 36) GeV threshold used at Run-2 to be maintained despite the intense running conditions. For an equivalent rate of 6.6 kHz, the threshold achieved by the $\tau$ reconstruction algorithm using calorimeter-only information (similarly to the Phase-1 system) would correspond to (90, 90) GeV. This figure establishes the dramatic impact of the thresholds applied to $\tau_h$ trigger objects on the signal acceptance. See Section 4.2 for more details.

1.4.1.4 Final states requiring multi-jet and $H_T$ trigger algorithms

- **Higgs boson pair production ($HH \rightarrow b\bar{b}b\bar{b}$)**. Another essential channel for the search of double Higgs boson production at the HL-LHC is the fully hadronic final state where each Higgs boson decays into a pair of $b$-quarks. During LHC Run-2, a dedicated L1 algorithm was implemented for this particular channel, requiring at least four jets above certain thresholds and a minimum total hadronic activity ($H_T$). Such a multi-jet trigger requirement allowed low thresholds on the individual jets to be sustained. This is particularly relevant in this context as the sensitivity to the Higgs boson self-coupling depends on the measurement of Higgs boson pairs with invariant masses close to the production threshold, that result in soft final state objects. The upgrade of the Level-1 trigger to provide low jet trigger thresholds will be crucial to achieve a measurement of the Higgs boson self-coupling and maintain the experimental sensitivity in the four $b$-jet channel.

A study of the sensitivity to HH production in the four b-jet channel has been performed by extending the results originally presented in Ref. [1]. The projected sensitivity to the Standard Model Higgs boson pair production is studied as a function of the minimal $p_T$ threshold applied in the selection of the four jets, that is driven by the requirement imposed by the trigger. An integrated luminosity of 3000 fb$^{-1}$ and 200 pileup interactions in the event are assumed in the simulation of the signal and background processes and of the CMS detector response. The study concludes that, despite the large multi-jet background for low $p_T$ jets, the experimental sensitivity largely benefits from low thresholds. The result is summarised in Fig. 1.7 (right) that reports the expected loss in signal significance as a function of the minimum jet $p_T$ threshold applied. The vertical lines at 40 GeV and 65 GeV denote, respectively, the threshold that can be achieved with the Phase-2 upgrade, and an example threshold that could be sustained with a calorimeter-only jet trigger algorithm at the price of a much larger bandwidth allocated to such a multi-jet trigger.

1.4.1.5 Final states requiring $E_T^{\text{miss}}$ trigger algorithms

- **BSM models with compressed mass spectra.** In natural Supersymmetry scenarios, the higgsinos $\tilde{\chi}_1^\pm$, $\tilde{\chi}_2^0$ and $\tilde{\chi}_1^0$ can be the lightest SUSY states. Given that their mass splitting is driven by radiative corrections and corresponds to values of only up to a few GeV, their decay into the lightest undetectable neutralino $\tilde{\chi}_1^0$ happens through the emission of very soft SM particles. In order to render this final state observable at the HL-LHC, the analysis strategy requires a jet from initial state radiation (ISR) to boost the event topology and induce enough $E_T^{\text{miss}}$ from the massive $\tilde{\chi}_1^0$ resulting in reduced signal cross section. Compressed mass spectra are also predicted in models describing vector like fermions, as proposed in Ref. [22]. Searches for such new heavy fermions, via their decay into a muon and a scalar dark matter candidate, are of particular interest when the new fermion and the dark matter candidate are very close in mass, with a difference of less than 20 GeV [22]. The main trigger
Figure 1.7: Left: Simulated distribution of transverse momentum of the lowest \( p_T \) \( \tau_h \) candidate in \( HH \to \tau \tau b \bar{b} b \bar{b} \) decays. The vertical lines correspond to the offline \( p_T \) thresholds at which the single object trigger efficiency reaches 50\% (for \( \tau \)) of the efficiency plateau. The solid vertical line corresponds to the trigger threshold provided by the Phase-2 L1 trigger system (at 200 pileup) matching the thresholds currently deployed by the L1 menu for Run-2. The dashed vertical line corresponds to the trigger threshold required to achieve the same rate using trigger algorithms that do not make use of L1 tracks or particle-flow candidates. Right: Expected loss in signal significance for the CMS Phase-2 \( HH \to bbbb \) analysis as a function of the minimum jet \( p_T \) threshold implemented by the multi-jet trigger algorithm used to select these events. The green and red lines indicate the thresholds that can be achieved by a jet trigger algorithm with and without using L1 tracking and particle flow inputs, respectively.

algorithms used to target these final states are based either on a minimum threshold on the event’s \( E_T^{\text{miss}} \), or on a cross-object trigger algorithm that requires \( E_T^{\text{miss}} \) and low \( p_T \) muons. The typical offline requirements on \( E_T^{\text{miss}} \) are driven by the trigger selection and, in Phase-1, were of 200 GeV for the pure \( E_T^{\text{miss}} \) trigger and of 125 GeV for the cross-object one. The signal acceptance for these exotic signatures is significantly reduced, would the L1 \( E_T^{\text{miss}} \) thresholds not be maintained to these values. For example, the relevant parameter space of the model proposed in Ref. [22] can be explored with the Phase-2 dataset only if the thresholds of the cross-object trigger are kept to their Phase-1 values.

- **Higgs boson associated production** \( (ZH \to \nu \nu b b) \). The SM \( H \to b \bar{b} \) events are mostly accessible through the associated production of the Higgs boson with a \( Z/W \) boson. The leptonic decays of the \( Z \) and \( W \) bosons are exploited at trigger level to achieve manageable rates given the large QCD background expected in this hadronic decay mode of the Higgs. This channel significantly contributed to achieve the observation of the Higgs boson decay into a pair of b-quarks during LHC Run-2 [23]. The associated production with a \( Z \) boson decaying into a pair of neutrinos is a major channel targeted at HL-LHC. The neutrinos produce significant \( E_T^{\text{miss}} \) that can be used to select events at trigger level and drastically reduce the background contribution.

Figure 1.8 shows the \( E_T^{\text{miss}} \) distribution for the final states mentioned above. The Phase-2 L1 \( E_T^{\text{miss}} \) reconstruction algorithm makes use of the tracking information, of the particle-flow re-
Figure 1.8: Simulated distribution of the missing transverse momentum $E_{\mathrm{miss}}^T$ from SUSY production of $\chi_1^+ \chi_2^0 \rightarrow W^+ Z \chi_1^0 \chi_1^0$ with $m(\chi_1^+ , \chi_2^0 ) = 300 \text{ GeV}$ and $m(\chi_1^0 ) = 292.5 \text{ GeV}$, and for $ZH \rightarrow v\bar{v}bb$ events. The vertical lines correspond to the offline $p_T$ thresholds at which the single object trigger efficiency reaches 90\% (for $E_{\mathrm{miss}}^T$) of the efficiency plateau. The solid vertical line corresponds to the trigger threshold provided by the Phase-2 L1 trigger system (at 200 pileup) matching the threshold currently deployed by the L1 menu for Run-2. The dashed vertical line corresponds to the trigger threshold required to achieve the same rate using a trigger algorithm that does not make use of particle-flow candidates.

construction and of PUPPI particle candidate inputs. This algorithm allows the trigger threshold of 200 GeV, used at Run-2, to be maintained at Phase-2, this threshold being defined as the value of the offline $E_{\mathrm{miss}}^T$ at the trigger efficiency reaches 90\% of its plateau value. More details are given in Section 4.2.

1.4.2 Extending the CMS physics reach with the Phase-2 Level-1 trigger

Besides maintaining low or medium thresholds for baseline triggers, which, as shown in the previous paragraphs, is crucial to ensure a high acceptance for many benchmark physics processes, the upgraded L1 trigger system further extends the physics reach of the experiment by allowing innovative trigger strategies to be deployed. Searches for physics beyond the Standard Model and precision measurement analyses will both benefit from such new techniques, that rely on sophisticated algorithms, to cover an extended phase space. The key features of the Phase-2 L1 trigger that make these new strategies possible include: the availability of tracking information and the determination of the event primary vertex; the extension of muon track reconstruction in the extended fiducial region covered by the upgraded muon detectors; the reconstruction of higher-level trigger objects with particle-flow techniques; the reconstruction of displaced objects; the inclusion of timing information; and the possibility of implementing machine learning approaches in modern programmable logic units. Examples that illustrate each case are detailed in Chapter 4. A summary is given in this section.

1.4.2.1 Physics of light mesons with tracking at the Level-1 Trigger

Rare B meson decays are a powerful probe for BSM physics since the higher order contribution from new particles may dramatically increase the rate of these decays. The same holds for Lepton Flavor Violating processes. Low $p_T$ thresholds (of a few GeV) are mandatory for the
multi-lepton triggers that target such processes. However, the online selection of these signal events, usually performed via double-muon triggers, has become a challenge already during the Run-2 of the LHC. The Phase-2 L1 trigger upgrade overcomes the limitations encountered during Phase-1 through the extensive usage of L1 Tracks, which have an excellent resolution in $p_T$ (allowing the thresholds to be lowered) and in position (allowing variables such as invariant masses to be used). In order to illustrate the gain brought in by the Phase-2 trigger system for such final states, the FCNC process $B_s \rightarrow \phi \phi \rightarrow 4K$ and the LFV process $\tau \rightarrow \mu \mu \mu$ have been considered. The techniques discussed here can easily be used to address several other analyses with light mesons, such as rare Higgs boson decays ($H \rightarrow \rho \gamma, H \rightarrow \phi \gamma$).

- The $B_s \rightarrow \phi \phi \rightarrow 4K$ process, which is forbidden at tree level and highly suppressed at higher orders in the Standard Model, is a good indirect probe for BSM physics, especially via measurements of CP violating asymmetries in this decay. In Phase-1, a dedicated trigger for this process could not be built. A Phase-2 trigger algorithm, that demands four low-$p_T$ L1 tracks (2 GeV) and reconstructs the invariant masses of the $\phi$ and $B_s$ mesons, provides 30% signal efficiency. As an illustration, Fig. 1.9 (left) shows that the resolution on the $B_s$ mass achieved with the L1 tracks is nearly as good as the one that is obtained from offline tracks. This trigger is described in detail in Section 4.3.3. This channel serves here as an example for similar processes.

- The LFV process $\tau \rightarrow \mu \mu \mu$ is a flagship analysis of the HL-LHC physics program. At the LHC, most of the $\tau$ leptons are produced via the decays of $D$ and $B$ mesons, and the decaying muons appear collimated, predominantly in the forward region of the detector. In Phase-1, the triggering strategy at Level-1 was based on the requirement of two standalone muon tracks with $|\eta| < 1.6$. In Phase-2, the trigger will require muon-jets, comprised of three objects, a combination of standalone unmatched stubs and track-matched muons, and impose restrictions on their maximum spatial separation and on their invariant mass. The extended acceptance ($|\eta| < 2.4$) together with the exclusive requirements that allow the $p_T$ thresholds to be lowered, result in...
1.4.2.2 Physics with the extended pseudorapidity coverage of the Level-1 lepton triggers

While the fiducial volume of the track finder is limited to $|\eta| < 2.4$, standalone lepton trigger objects such as $e/\gamma$ and muons can be reconstructed at higher pseudorapidity, exploiting the coverage of the muon trigger chambers (up to $|\eta| = 2.8$, larger than that of Phase-1) and of the HGCAL (up to $|\eta| = 3$). Several physics processes will benefit from triggers on leptons at high $|\eta|$, such as:

- The measurement of the $t\bar{t}$ differential cross section at high rapidity of the $t\bar{t}$ system, which brings interesting constraints on parton distribution functions [1] with strong implications on the precision of predictions for production of heavy BSM particles through gluon fusion. Semi-leptonic events are traditionally used for this measurement, triggered by single lepton triggers, and extending the trigger $|\eta|$ coverage from 2.4 to 2.8 increases significantly the event yields in the most interesting forward region.

- Measurements of Double Parton Scattering (DPS), performed in events where two same-charge W bosons decay into muons or electrons, will provide useful information on transverse and longitudinal parton correlations within the proton. Increasing the $|\eta|$ acceptance of the single- or double-lepton triggers allows the DPS differential cross section to be measured in an extended range in pseudorapidity that translates into stronger constraints on parton correlations. Similarly, the final state associated with the WW Vector Boson Scattering (VBS) could profit from the increased coverage of these lepton triggers.

- As mentioned earlier, in the $\tau \rightarrow \mu\mu\mu$ process, the decaying muons are mostly produced at high $\eta$ with low transverse momentum. Variations of the trigger described in the previous paragraph can be built, whereby one or two standalone muons are required (instead of track-matched muons). The increased signal acceptance brought in by the extended pseudorapidity coverage improves the overall sensitivity on the $\tau \rightarrow \mu\mu\mu$ branching fraction by a factor of $\sim 1.2$, as detailed in Ref. [11].

1.4.2.3 Physics with displaced muons reconstructed by the Level-1 trigger

Many extensions of the Standard Model predict new particles with a long lifetime. The resulting displaced signatures are particularly challenging for the trigger. The L1 muon reconstruction used during Run-2 assumed that the muon originates from the nominal interaction point, leading to a very small efficiency for displaced muons. In Phase-2, the algorithms targeting prompt muons will be complemented by dedicated algorithms that relax this constraint, as described in Section 3.3. The gain in acceptance has been quantified using the example of the production of dark photons, whose decay leads to displaced muons [24]. While the analysis of the Run-2 data has no acceptance when the decay length of the dark photon is larger than about 10 cm [25], due to trigger requirements, the Phase-2 displaced standalone muon algorithm maintains acceptance up to at least 100 cm. Details are given in Section 4.3.4.

1.4.2.4 Physics with displaced jets reconstructed by the Level-1 trigger

Similarly, displaced jets may also be a characteristic feature of BSM processes. For example, the H(125) Higgs boson may decay into a pair of light long-lived pseudoscalars (a), which, in turn, decay dominantly into $b$ quarks, $H \rightarrow aa \rightarrow bbb$. Due to the relatively small Higgs mass,
the resulting hadronic final state is in general too soft for a standard multi-jet trigger strategy to be effective. During Phase-1, such decays can be probed only in the mode where the Higgs is produced in association with a W or a Z boson, at the price of a reduced cross section. In Phase-2, these decays will be accessible irrespective of the Higgs production mode through the availability of the new triggers, which exploit the long lifetime of the hidden sector state $a$ to select soft, displaced jets:

- Using tracking information: a dedicated displaced tracking algorithm reconstructs L1 tracks without any beamspot constraint. Such tracks can be subsequently clustered to reconstruct displaced jets. As detailed in Section 4.3.5.1, such a trigger provides acceptance for the aforementioned process for a $a$ decay length up to 5–10 cm. As an illustration, Fig. 1.9 (right) shows the number of signal events selected by such a trigger in 3000 fb$^{-1}$ of integrated luminosity.
- Using timing information: for longer lifetimes of the $f$, timing information provided by the barrel calorimeters improves significantly the signal acceptance, as discussed in Section 4.3.5.2. In addition, the pointing capabilities of the HGCAL could be exploited, but this option has not been explored at the time of this writing.

### 1.4.2.5 VBF Higgs boson production trigger algorithm based on machine learning techniques

The Higgs boson production through Vector Boson Fusion is essential for many measurements in the Higgs sector due to its distinct topology. The improved energy resolution of hadronic objects provided by particle-flow reconstruction at L1 is certainly beneficial to the development of triggers targeting specific signatures such as VBF. As demonstrated in Chapter 4, the current thresholds used by the Phase-1 L1 trigger for the dedicated VBF trigger algorithm can remain similar for HL-LHC as a consequence of the use of particle-flow reconstructed jets. Moreover, the signal acceptance, for example on the VBF $H \rightarrow$ invisible and VBF $H \rightarrow b\bar{b}$ processes, can be significantly improved by including more topological information. Since the Phase-1 trigger upgrade, Level-1 menus have been expanded to include selection criteria on more sophisticated variables. A natural continuation of this evolution towards more complexity consists in using modern machine learning tools to build more powerful multivariate discriminators. The software tools to synthesize such algorithms into FPGA firmware now exist. This topic is discussed in Section 3.7 and a multivariate trigger algorithm that targets the VBF $H \rightarrow$ invisible and VBF $H \rightarrow b\bar{b}$ signals is detailed in Section 4.3.6.

### 1.4.2.6 Physics with the 40 MHz Scouting

The following physics processes can potentially benefit from a 40 MHz scouting-based analysis using the architecture outlined in Section 5.7.1, and the plan is to investigate them further, in some cases also profiting from work that can be carried out already in Run-3, as discussed for example in Section 6.4.2.2.

- Higgs rare decays: the exclusive channels $H \rightarrow J/\psi \gamma$, $H \rightarrow \phi \gamma$, $H \rightarrow \rho \gamma$ can in principle all be studied with a single-photon trigger, provided that the threshold is low enough. However, given the prospects in terms of statistics from the SM BR prediction, lowering the photon threshold by combining it with a track pair may help. The limits of track-pair resonance candidate selection in the Global Track Trigger need to be studied. Scouting, albeit with limited mass resolution, could provide an alternative in some cases, selecting and storing all Level-1 photons with relaxed or no threshold and all Level-1 track pairs within a broad mass window.
• Displaced Muons, LLP: as several BSM models predict the creation of new particles with a substantial lifetime, displaced muons would represent a clear signature for such new physics. At the same time, the Level-1 must provide optimal efficiency for prompt signatures. Studies could concentrate on the possibility of relaxing quality cuts for example on the muon-track matching, and/or using standalone muons combined with calorimeter information. A stream containing only selected level-1 displaced muon candidates without rate limitation can be used to detect anomalous production of displaced muons at low-$p_T$. The possible benefits of cross-BX correlations, only marginally accessible to the standard Level-1, could also be investigated in the context of LLP studies.

• Flavor anomalies, $\tau$ physics: processes that violate Lepton Flavor Unitarity are usually challenging because of the high background rates involved, but may not require extreme resolutions. Likewise, searches for e.g. $W' \to \tau\nu$ could benefit from scouting. In general, searches with single-$\tau$ final states are notoriously hard to perform because of the difficulties of controlling a single-$\tau$ trigger rate, and the typical signal acceptance is quite low. Studies would focus on efficient $\tau$ identification and improving the quality of the reconstruction performed with Level-1 data, possibly by using less conventional techniques such as Machine Learning.

• $B_s \to \tau\tau$: similarly to single-$\tau$ final states, requires high efficiency $\tau$ selection at low-$p_T$.

• Soft hadron final states: some hidden sector models predict light mediators with small couplings, producing high multiplicity final states consisting of spherically-symmetric distributions of prompt or non-prompt soft particles spread over a large region with $p_T \lesssim 100$ MeV (soft bombs). A classic trigger approach is clearly not viable or severely limited in acceptance [26], while the scouting system could be used to analyze the spatial correlations of Level-1 tracks down to the lowest $p_T$, only storing high-level parameters of the track distribution instead of individual tracks.

• QCD measurements with high statistics: these are the classic application of scouting methods and analyses can be clearly adapted to use Level-1 jets, $E_T^{\text{miss}}$ and $E_T$ sums, in particular from particle-flow, to produce results in regions of phase space inaccessible to the standard Level-1.
1.5 Outline of the Technical Design Report

This TDR describes the evaluation of the options discussed prior to this document in order to propose the first conceptual design of the Phase-2 L1 trigger architecture and its instrumentation. The introduction and overview (Chapter 1) summarizes the challenges imposed by the HL-LHC machine and physics program. The requirements of the L1 trigger in terms of algorithms and architecture are described along with the physics reach of the system proposed through the description of the foreseen physics menu. Chapter 2 covers the interfaces with the sub-detectors and their update since the iTDR in view of the requirements of the trigger algorithms performance. A complete review of the trigger algorithms and their firmware implementation is provided in Chapter 3. Algorithms to identify standard physics objects display enhanced efficiency through the matching with tracks. More complex objects such as jets or missing transverse energy ($E_T^{\text{miss}}$) are reconstructed with a Particle-Flow approach. Specific algorithms to address the need for triggering on displaced objects are also presented. The complementarity of the various flavors of algorithm proposed is addressed in Chapter 4. The menu exercise consists of compiling a selection strategy that covers the physics needs of the CMS Phase-2 physics program while remaining within the 750 kHz rate budget. The robustness of this strategy is tested against harsher conditions. Features exploiting the selection of exotic signals are also part of this exercise. A complete description of the conceptual design of the L1 architecture is detailed in Chapter 5. This Chapter exposes the motivations for the proposed structure considering integration, operation, maintenance and commissioning aspects. Each sub-component organization has been developed in order to optimize latency, resources and number of links. A full architecture exercise displaying the flexibility and potential evolution of the design conclude this chapter. Chapter 6 focuses on the hardware research and development (R&D) considered to instrument the upgraded L1 trigger system. The current status of R&D lines is reviewed along with results from on-board validation tests conducted on dedicated benches. This Chapter addresses the feasibility of this design by including results from hardware demonstrators and slice tests realized with target electronics boards. And finally, Chapter 7 contains the details of the schedule, cost and funding of the project and other management aspects.
Chapter 2

Inputs to the L1 trigger system: the trigger primitives generation

As previously described, the Phase-2 L1 trigger system will interface with a large range of sub-detectors including calorimeters, muon detectors and the tracking system as illustrated in Fig. 1.3. The trigger primitives are generated in the backend electronics systems of each of these sub-detectors. This Chapter gives a short description of the system functions and organization. The baseline trigger primitive algorithms and their performance in terms of efficiency and resolution in the foreseen HL-LHC running conditions are also presented. As these input primitives are used by the trigger object reconstruction algorithms running downstream, basic features such as minimum energy thresholds, selective readout schemes, multiplicities, pileup dependence etc., are also mentioned. In some cases, ageing studies and the evaluation of their potential impact on trigger performance have been included. The Chapter concludes with a summary of the trigger input bandwidth and latency (including the expected number of output links and input format) — an essential input to the development of the Phase-2 trigger architecture. The content of this Chapter relates the current status of these interfaces and their expected characteristics at the time of writing. Potential future improvements in the primitive generation algorithms or their associated systems are mentioned.

2.1 Track finder primitives

One of the key features of the CMS upgrade for the HL-LHC is the addition of track information in the L1 trigger. The identification of charged particle trajectories at 40 MHz is enabled through a unique design of the Phase-2 Outer Tracker [8] through “$p_T$ modules” that perform hit correlations between two closely spaced silicon sensors to filter signals of particles with a $p_T$ above approximately 2 GeV, allowing a rapid data-reduction of about one order of magnitude\(^1\). The pairs of correlated hits (“stubs”) are sent to the off-detector, backend L1 track finding system at 40 MHz.

In Ref. [8], different approaches and technologies for performing the L1 track finding are discussed. The L1 track finding algorithm described here (“hybrid” algorithm) combines the different approaches to provide optimal performance and is being implemented in an all-FPGA system based on an ATCA platform [27], extensively utilizing parallel processing both geometrically and in time. The foreseen system divides the Outer Tracker in nine $\phi$ sectors, where each $\phi$ sector is processed by a single FPGA. In addition, a time-multiplexing factor of 18 is used, so that each processing board will receive a new event every 450 ns, and the total system corresponds to 162 track finder processing boards. Stubs are duplicated near the $\phi$ sector boundaries to avoid sideways communication across track finding processing boards. The geometry of the

\(^1\)Only the Outer Tracker is used for the L1 track finding, consequently, coverage is provided for $|\eta| < 2.4$. 

31
Figure 2.1: Top: Sketch in the \( r - z \) plane of one quarter of the upgraded tracker for HL-LHC, showing the Outer Tracker (in blue and red) that is used for the L1 tracking, as well as the Inner Pixel detector (in green and yellow), not used for L1 [8]. Bottom: Tracker geometry in the \( x - y \) plane, showing the data flow through the data trigger and control boards (DTCs) to the track finding processor boards (TFPs). The L1 track finding system is divided into nine \( \phi \) sectors, each processed by one track finding board for each of the 18 time-multiplexing (TM) slices.

Outer Tracker and a schematic overview of the L1 track finding processing system are shown in Fig. 2.1.

The hybrid algorithm first forms seeds (“tracklets”) from stubs in two adjacent layers or disks, combined with a constraint to the beamspot, and calculates the tracklet parameters [28]. Only tracklets that are consistent with a particle with \( p_T > 2 \text{ GeV} \), \(|\eta| < 2.4\), and \(|z_0| < 15 \text{ cm}\) are considered. The Outer Tracker consists of six barrel layers (L1-L6) and five disks (D1-D5). The seeding combinations used here are L1+L2, L3+L4, L5+L6, L1+D1, L2+D1, D1+D2, D3+D4, and in the transition region between the barrel and the endcap, also L2+L3 seeds are included. The usage of multiple parallel seeding combinations ensures high track finding efficiency and redundancy. Next, the tracklets are projected to other layers and disks, both inward and outward, to search for additional matching stubs in narrow windows around the projection. A minimum of four stubs are required to form an L1 track. Track candidates that share stubs are merged prior to fitting. In the final step, track candidates are fitted using a Kalman filter algorithm to identify the best stub candidates and calculate the track parameters \( (p_T, \eta, \phi, z_0) \) [29]. The algorithm performance is studied using simulated \( t\bar{t} \) events with an average pileup of 200. The tracking efficiency as a function of \( \eta \) for different \( p_T \) ranges and the track \( z_0 \) resolution versus \(|\eta|\) are shown in Fig. 2.2. The tracking efficiency is defined with respect to truth-level particles from the primary interaction that produced stubs in at least four layers/disks.

Each L1 track is described by a 96 bit word containing information about the fitted track parameters and track quality. The track word content, shown in Table 2.1, contains enough precision to capture the full possible resolution from the L1 tracking system, as well as various measures.
2.1. Track finder primitives

Figure 2.2: Left: L1 tracking efficiency as a function of $\eta$ for tracks above 2 GeV (filled black markers) or 8 GeV (open red markers) for tracks in $t\bar{t}$ events overlaid with an average pileup of 200. The efficiency is defined with respect to truth-level particles that produced stubs in at least four layers/disk. Right: Track $z_0$ resolution as a function of $|\eta|$ for tracks in $t\bar{t}$ events with an average pileup of 200. The $z_0$ resolutions correspond to intervals that encompass 68% (filled black markers) or 90% (open red markers) of all tracks with $p_T > 2$ GeV.

The track finding system is capable of reconstructing charged particle tracks with $p_T > 2$ GeV. Sufficient bandwidth must be built into the system to transmit these tracks to the downstream system without significant loss. Figure 2.3 shows the maximum number of tracks reconstructed in any of the nine $\phi$ sectors of the track finding system. To keep truncation to the level of $< 10^{-4}$, we must maintain bandwidth to transmit a maximum of 110 tracks per sector. This estimate is based on the default tracking described above. As described below, an “extended tracking” option is under consideration, with the potential to reconstruct tracks displaced from the beamspot. Inclusion of the displaced tracks increases the number of tracks per sector by $\sim 40\%$. We further provision for an additional 20% more tracks as a safety margin for worse conditions than what is currently expected by the simulation. To maintain truncation below $< 10^{-4}$, an upper limit of 185 tracks per sector must instead be considered, or a total bandwidth per sector of 185 tracks times 96 bits, or 17 760 bits per $\phi$ sector. If transmitted over 18 BX, these tracks can be delivered on two fibers at 25 Gb/s.

With clear physics motivation for exploring signatures of long-lived particles, an extension to the L1 track algorithm is being explored to identify trajectories from such long-lived particles that would appear as “displaced” in the detector. This extended L1 tracking includes so-called “triplet” seeds, formed from a combination of three layers and/or disks but without any constraint to the beamspot, in addition to the regular tracklet seeds. In addition, a five-parameter Kalman Filter fit is used to determine the transverse impact parameter ($d_0$) of the track. The improvement in tracking performance for long-lived particles using the extended tracking can be seen in Fig. 2.4. It shows the track finding efficiency, again defined with respect to truth-level particles that produced stubs in at least four layers/disk, as a function of the track $d_0$, comparing the baseline and extended tracking options for single displaced muons with $|\eta| < 2.0$ and a flat $p_T$ spectrum within 2–20 GeV, in events with PU = 0. The extended tracking is shown for a working point corresponding to an optimization for $d_0$ of $\pm 5$ cm – the ultimate working point is under study. The extended tracking also has the potential to improve the tracking efficiency for electrons, where the trajectories are often distorted due to brehmsstrahlung. The preliminary
Chapter 2. Inputs to the L1 trigger system: the trigger primitives generation

Table 2.1: Content of the L1 track word, which includes the charge divided by the radius of curvature \( (q/R) \), azimuthal direction \( (\phi) \), tangent of the rapidity angle \( (\tan(\lambda)) \), longitudinal impact parameter \( (z_0) \), transverse impact parameter \( (d_0) \), \( \chi^2 \) per degree of freedom \( (\chi^2/dof) \), \( \chi^2 \) computed from track stub bend angles \( (\text{bend-}\chi^2) \), summary of tracker layers with hits \( (\text{hit mask}) \), multivariate analysis \( (\text{MVA}) \) discriminator to determine track quality \( (\text{track quality MVA}) \), and space for two additional multivariate discriminators for specialized track quality, such as for electrons \( (\text{other quality MVAs}) \).

<table>
<thead>
<tr>
<th>Track parameter</th>
<th>Number of bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q/R )</td>
<td>15</td>
</tr>
<tr>
<td>( \phi )</td>
<td>12</td>
</tr>
<tr>
<td>( \tan(\lambda) )</td>
<td>16</td>
</tr>
<tr>
<td>( z_0 )</td>
<td>12</td>
</tr>
<tr>
<td>( d_0 )</td>
<td>13</td>
</tr>
<tr>
<td>( \chi^2/dof )</td>
<td>4</td>
</tr>
<tr>
<td>( \text{bend-}\chi^2 )</td>
<td>3</td>
</tr>
<tr>
<td>hit mask</td>
<td>7</td>
</tr>
<tr>
<td>track quality MVA</td>
<td>3</td>
</tr>
<tr>
<td>other quality MVAs</td>
<td>6</td>
</tr>
<tr>
<td>track isValid</td>
<td>1</td>
</tr>
<tr>
<td>spare</td>
<td>4</td>
</tr>
<tr>
<td>total</td>
<td>96</td>
</tr>
</tbody>
</table>

![Fraction of events](image)

Figure 2.3: Maximum number of prompt tracks with \( p_T > 2 \) GeV found in any single \( \phi \) sector of the L1 track finder for simulated \( t\bar{t} \) events overlaid with an average pileup of 200.
2.2. Barrel calorimeter trigger primitives

Figure 2.4: Preliminary L1 tracking efficiency as a function of $d_0$ for tracks originating from displaced muons (PU = 0) with $|\eta| < 2.0$ and $2 < p_T < 20$ GeV. The efficiency is defined with respect to truth-level particles that produced stubs in at least four layers/disks. The black (filled) points show the baseline tracking (with a $d_0 = 0$ constraint), the green (triangle) points show the baseline tracking but using a 5-parameter track fit that allows for a nonzero $d_0$, and the red (open) points show the extended tracking using triplet seeds (for a $\pm 5$ cm optimization) and a 5-parameter track fit.

Figure 2.5: Left: Preliminary L1 tracking efficiency as a function of $\eta$ for tracks with a flat $p_T$ spectrum within 2–100 GeV and $|\eta| < 2.4$ in single electron events (PU = 0). Right: Preliminary L1 tracking efficiency as a function of $p_T$ for tracks with a flat $p_T$ spectrum within 2–100 GeV and $|\eta| < 2.4$ in single electron events (PU = 0). The efficiency is defined with respect to truth-level particles that produced stubs in at least four layers/disks. The black (filled) markers show the efficiencies for the baseline tracking, while the red (open) markers show those of the extended tracking.

improvement in the electron tracking efficiency is shown in Fig. 2.5 for events without pileup$^2$. The most significant improvement is in the barrel region as a consequence of the triplet seeds used. Work is presently ongoing to understand the additional FPGA processing resources cost associated with these potential improvements.
Figure 2.6: The upgraded ECAL Barrel system: the first amplification stage is a Trans-Impedance amplifier (TIA), required to handle signals of up to 2 TeV. Two gain stages follow the TIA to digitize the dynamic range with two 12-bit ADCs. The Preamp and ADC stages constitute the Very-Front End electronics (VFE). The Front-End (FE) card will serve five VFE cards with 5 channels each as in the Phase-1 ECAL Barrel. The FE receives and distributes the clock and control information to the front-end system and receives the channel data streams from the VFE. The FE finally transmits data to the off-detector barrel calorimeter processor (BCP) via 10.24 Gb/s optical links. The BCP will house two FPGAs, which will be programmed to provide the necessary data processing, among which the trigger primitive generation.

2.2 Barrel calorimeter trigger primitives

2.2.1 Electromagnetic barrel calorimeter

The increased trigger latency and rate of CMS at the HL-LHC require the ECAL barrel readout and backend electronics to be upgraded. For Phase-1, data are pipelined in the readout electronics and the ECAL barrel trigger primitive generator (EB TPG), which is located on-detector, produces trigger tower sums of $5 \times 5$ crystals. For Phase-2, the avalanche photo diodes (APDs), the geometry and the mechanics of the detector remain unchanged [30]. Crystals will still be grouped in $5 \times 5$ with the difference that single crystal data will be streamed from the on-detector to the backend electronics. Besides the highly improved granularity, the new Very Front-End electronics will allow for a finer sampling at 160MHz to be performed, instead of 40 MHz. Furthermore the new electronics will provide a much faster (w.r.t. Phase-1) signals shaping. The two new features will be a strong tool for rejection of the larger pileup level expected at the HL-HLC. Once data reach the backend processors, they are used for the EB trigger primitive generation and buffered before going to DAQ while waiting for the Level-1 trigger accept. A detailed description of what is discussed in this Section can be found in the ECAL barrel upgrade TDR [9].

Because of the mechanical constraints, each front-end card will collect data from a $5 \times 5$ array of crystals at a 160 MHz sampling frequency (Fig. 2.6). Twelve such cards send data to one of the two FPGAs sitting on the backend barrel calorimeter processor (BCP) via 48 upstream links and 12 downstream links. Each FPGA covers a region of 300 crystals corresponding to $\eta \times \phi = 0.26 \times 0.35$. A total of 108 BCPs, housed in 12 crates, cover the full ECAL barrel. As shown in Fig. 2.7, each crate covers one sixth in $\phi$ of each $\eta$ side of the detector (the number

\[\text{2Similar performance is obtained for single electrons with an average pileup of 200.}\]
of crates and the mapping have been optimized since the ECAL Barrel Upgrade TDR was published).

This architecture allows for sharing boundary data between neighboring BCPs, which is achieved via dedicated optical fiber router boxes, using technology developed for the Phase-1 calorimeter trigger upgrade [31]. Isolated and very narrow anomalous signals ("spikes") [32], produced by the interaction of hadrons with the core of the APDs and as large as 100 GeV, are regularly observed in the ECAL, with a frequency proportional to the instantaneous luminosity. If not suppressed, they would badly affect the electromagnetic Level-1 Trigger [9]. Data sharing between neighboring BCPs is required to measure the energy shared among close crystals, in order to reject the spikes, as well as to build the clusters that may optionally be sent to the Level-1 trigger. The most effective method to reduce the rate of spikes in Phase-2, however, will be provided by an online analysis of the sharper signal shape of the TIA (Fig. 2.8). The method will reduce the spike rate to a few Hz at HL-LHC up to an integrated luminosity of up to 4500 fb⁻¹.

Although the BCP design around the main units (the FPGAs) is optimized for the ECAL barrel needs and carried out by the ECAL barrel community, it shares common infrastructure and software like the IPMC, the ELM and the ESM developed within the APx consortium as shown in Section 6.3.1.

Two options for the EB trigger primitive words have been investigated: a baseline single crystal primitive word, and an optional cluster primitive word. In both cases, the EB trigger primitive must include the calibration of the input data, as well as the shape analysis of the input pulses to extract the transverse energy (\(E_T\)) and time information. The baseline EB trigger primitive is a 16-bit word for each of the 61 200 crystals that encodes \(E_T\), timing, and a spike flag (Table 2.2 (left)). The spike flag bit is determined by the signal shape analysis: a linear discriminant allows for the spike signal to be disentangled from genuine scintillation signals, as was demonstrated during a testbeam in 2018. The EB trigger primitive will contain also timing information; in a preliminary version, the approach is to measure the time with respect to the beam crossing of the triggered event in a window of 2 ns. With the new signal shape provided by the TIA, the expected offline time resolution is about 30 ps for energies above 50 GeV. The achievement of such a resolution is, however, dependent on an extremely precise and uniform clock distribution to the whole detector. Achieving the same level of resolution online for the Level-1 trigger is additionally limited by the accuracy of calibrations. Taking all this into account, five bits have been preliminarily allocated in the trigger primitive, which would encode 60 ps resolution. The trigger primitive data amount to a total of 979 200 bits per bunch crossing, which corresponds to a rate of 39 168 Gb/s, if no zero suppression is applied. Data will be sent across a total of 3060 optical fibers, corresponding to 90 backend cards with thirty 16 Gb/s links and 18 backend cards with twenty 16 Gb/s links. The latency of the EB system is foreseen to be within 1.5 \(\mu\)s, but the detailed study is ongoing and not finalized by the time of this TDR.

Cluster trigger primitive words, generated directly in the EB backend electronics, are considered as a possible future option and studies are ongoing. Such a capability could be useful later if the Level-1 trigger becomes constrained by the bandwidth, or it requires a redundant source of information. A possible choice of format for a cluster trigger primitive is a 40-bit word, which encodes the number of crystals belonging to the cluster, the cluster \(E_T\), time, spike identification flag, as well as \((\eta,\phi)\) coordinates for the cluster maximum (Table 2.2 (right)). Figure 2.9 shows the multiplicity distributions of cluster trigger primitives built as \(3 \times 5\) crystal matrices measured in Zero Bias simulated events with an average number of pileup events of 200, with ECAL barrel conditions corresponding to the detector after 1000 fb⁻¹. The plot
Chapter 2. Inputs to the L1 trigger system: the trigger primitives generation

Figure 2.7: a) BCP to Level-1 trigger interface. Each BCP will include 2 FPGAs. Each FPGA will process data from 12 towers (300 channels); b) Mapping between the whole ECAL barrel and the backend Advanced TCA (ATCA) crate. Nine BCPs in one crate cover three supermodules, with six crates per positive (negative) $\eta$. 

...
2.2. Barrel calorimeter trigger primitives

Figure 2.8: Average pulse shapes from a spike in the APD and scintillation signal. These pulse shapes are based on the measurements with TIA prototypes in test beams. The amplitude samples from the digitization at 160 MHz sampling frequency are shown. The TIA fast response will allow the two to be disentangled.

Table 2.2: ECAL Barrel trigger primitive word definition. The left table shows the preliminary baseline version where each crystal is encoded. The right table shows the preliminary definition for possible additional clusters.

<table>
<thead>
<tr>
<th>Quantity per crystal</th>
<th>N bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_T$</td>
<td>10</td>
</tr>
<tr>
<td>timing</td>
<td>5</td>
</tr>
<tr>
<td>spike flag</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quantity per cluster</th>
<th>N bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_T$</td>
<td>10</td>
</tr>
<tr>
<td>time</td>
<td>5</td>
</tr>
<tr>
<td>$\eta$</td>
<td>8</td>
</tr>
<tr>
<td>$\phi$</td>
<td>8</td>
</tr>
<tr>
<td>$N_{\text{crystal}}$</td>
<td>8</td>
</tr>
<tr>
<td>spike flag</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
</tr>
</tbody>
</table>
Chapter 2. Inputs to the L1 trigger system: the trigger primitives generation

Figure 2.9: Expected multiplicity of $3 \times 5$ cluster trigger primitives for different $E_T$ thresholds.

shows the multiplicity of $3 \times 5$ matrices built around a seed crystal with $E_T \geq 1$ GeV including neighboring crystals with $E_T \geq 500$ MeV for various cluster $E_T$ thresholds. The spike flagging in clusters will almost certainly be achieved with the “swiss-cross” algorithm [32]. The details of this algorithm implementation remain to be defined and are currently being studied.

2.2.2 Hadron barrel and forward calorimeters

The Phase-2 upgrade of the HCAL Barrel (HB) and Forward (HF) calorimeters consists of replacing the off-detector electronics. The number of readout channels, the transverse ($\eta$-$\phi$) segmentation, the number of longitudinal readout depths, and the data+TPG latency will remain as after the Phase-1 upgrade. The segmentation of HB in ($\eta$, depth) for a fixed $\phi$ is shown in Fig. 2.10. The segmentation in ($\eta$, $\phi$) at a fixed depth is shown in Fig. 2.7, except that each HCAL half-barrel has only 16 pseudorapidity segments. The TPG electronics will use the same hardware that is being developed for the EB.

In HB, a BCP will cover four $\phi$ segments across all $\eta$, receiving 66 front-end data links each at 5.0 Gb/s. In all 2304 trigger towers in HB, each covering $0.087 \times 0.087$ in $\Delta \eta \times \Delta \phi$, signals from four depth segments (or three for towers with the highest $\eta$) are sampled at 40 MHz and corrected for pedestal, gain, and response. The depth samples for each tower are then summed and processed by an algorithm that detects a peak and subtracts out-of-time pileup. In addition to the tower $E_T$, the HB TP contains several feature bits that will facilitate encoding of (a) longitudinal shower profile data, for use in calibration, lepton isolation, and identification of Minimum Ionizing Particles (MIPs), (b) shower time data constructed from the 2 bits of the Time-to-Digital-Converter (TDC) information available in each constituent channel of the trigger tower. Configurable look-up tables will determine which time windows within the BX of interest are represented by the available TDC codes. These time window boundaries will have a granularity of 0.5 ns.

In the case of HB, both the pulse shape and the increased longitudinal segmentation (4 depths)
will provide some discrimination against pileup backgrounds. Because the HCAL electronics signal integrates over more than 25 ns, out-of-time pileup interactions can affect the detector response. To obtain a proper estimate of the signal amplitude for a given bunch-crossing, a pulse shape filtering scheme will be applied with coefficients that are optimized in order to minimize the effects of out-of-time pileup. Phase-1 upgrades of the HB detector also provide the possibility to implement fine-grain algorithms that carry information of energy depositions from the various layers of the calorimeter that may also be of use to the trigger for PU mitigation purposes. PU deposits have a distinctive depth profile (e.g., more energy in the first layers), and the use of HCAL depths may help discriminate deposits induced by jets from hard interactions against (in-time and out-of-time) pileup deposits. This is illustrated in Fig. 2.11, showing characteristic longitudinal depth profiles of HB trigger primitives associated with jets from harder interactions versus jet events from pure pileup interaction events in comparison. The plots show that the separation is pronounced for low energy TPs, below 10 GeV (left plot), as expected.

Fine-grained algorithms can also be formed using the TDC information to mark hits with early or late times. These algorithms will allow high efficiency hadronic triggers that capture decays from long-lived particles, with lifetimes up to 1-2 m relevant to decays prior to or within the HCAL barrel volume, to be developed.

The HF TP definition will remain as in Phase-1 (see Table 2.3). Signals from long and short fibers in each tower, sampled at 40 MHz, are used to determine the tower energy, along with time measurements from each channel's TDC. The $E_T$ reconstruction algorithm includes suppression of the collision-induced anomalous signals that arise when charged particles interact directly with the photomultiplier tube windows. Two feature bits are available for each HF TP. One is used to indicate that the ratio of the energy measured in the long versus short fibers is consistent with the deposit of an electromagnetic shower, while the other is an ADC-over-threshold indicator with individual thresholds per channel to define minimum-bias triggers.

Figure 2.10: The grouping of scintillating layer segments into electronics channels in one azimuthal slice of an HCAL half-barrel. The scintillating tiles are numbered 0–16, the pseudorapidity segments 1–16, and the readout depths 1 (blue), 2 (yellow), 3 (green), and 4 (magenta). FEE indicates the location of the front-end electronics.
Chapter 2. Inputs to the L1 trigger system: the trigger primitives generation

Figure 2.11: HCAL barrel Trigger Primitives depth profiles for jets from hard (QCD) interactions versus jet events from pure PU interactions with PU200 conditions in Phase-2 simulation. Left: HB TP transverse energies are restricted between 0.5 and 10 GeV, and Right: HB TP transverse energies are restricted between 10 and 128 GeV.

Table 2.3: Baseline Barrel HCAL (HB) and Forward HCAL (HF) tower TP word definition.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>N bits (HB)</th>
<th>N bits (HF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_T$</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Feature bits</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>10</td>
</tr>
</tbody>
</table>

2.3 High granularity endcap calorimeter trigger primitives

The Phase-2 high-granularity calorimeter (HGCAL) (also referred to as CE, for calorimeter endcap) is a sampling calorimeter with a high transverse and longitudinal granularity and will cover the pseudorapidity region from 1.52 to 3. It uses silicon as the active material in the region receiving the largest lifetime dose of ionizing radiation and plastic scintillator deeper into the HGCAL and further from the beam axis, as indicated in Fig. 2.12. Each endcap has 50 sensitive layers, with 28 in the electromagnetic section (CE-E, for calorimeter endcap electromagnetic) and the remaining 22 in the hadronic section (CE-H, for calorimeter endcap hadronic). Among these, all the CE-H layers but only alternate layers in the CE-E are equipped to transmit trigger data. A complete description of the detector design as of 2017 can be found in the HGCAL TDR [10], although some details have changed as the design evolves towards the HGCAL engineering design report (EDR) due in 2021.

The implementation of the updated geometry in the simulation being under active development at the time of this writing, the geometry described in the HGCAL TDR is used here. The main differences in the HGCAL TDR geometry are a slightly larger number of layers in the CE-H (24 instead of 22) and a simplified envelope of the detector close to the beam axis.

2.3.1 System architecture

The raw input data to the calorimeter trigger primitive generator (HGCAL-TPG) in the silicon section are the energies in hexagonal cells summed into trigger cells with a granularity of approximately 4 cm$^2$. In the scintillator section, trigger cells are groups of scintillating tiles covering an azimuthal angle of 2.5 degrees, corresponding to dimensions from about 4 cm to 10 cm,
2.3. High granularity endcap calorimeter trigger primitives

with similar radial dimensions. The charge deposited in these trigger cells is compressed to 7 bits with a floating point format. No timing information is kept due to the excessively large bandwidth that would be required to send precise enough timing data. Furthermore, not all the trigger cells can be sent out of the detector and only those above an energy threshold are selected. A value of this threshold between 1 and 2 MIP/sin$\theta$ will be used, depending on the exact trigger cell occupancy. To compensate the resulting loss of energy, the unselected channels are summed in each HGCAL module (module sums, corresponding to an area of 48 trigger cells in the silicon section and a similar number in the scintillator section). Other data reduction strategies are under study, and could be used instead of the threshold selection. This trigger cell summation, energy compression and selection, and the module sums, are implemented in two custom on-detector ASICs, the HGCROC (High-Granularity Calorimeter ReadOut Chip) ASIC followed by the ECON-T (Endcap CONcentrator TPG) ASIC. The resulting data are sent out of the detector to the backend electronics on about ten thousand 10 Gb/s optical links using the lpGBT [33] (low power GigaBit Transceiver) link protocol.

The core of the HGCAL-TPG runs on off-detector FPGAs in a two-stage system. Each FPGA in the first stage receives up to 72 lpGBT links from on-detector and covers about 1.4% of one endcap. The first stage mainly implements repacking and calibration of the input data and sends the data to the second stage on 16 Gb/s links in a time multiplexing fashion. Each board of the second stage processes one out of 18 bunch crossings and covers one sector corresponding to one third of one endcap (120 degrees). Data close to the boundaries of these sectors are

Figure 2.12: Longitudinal cross section of the upper half of one endcap calorimeter. It consists of an electromagnetic (CE-E) compartment followed by a hadronic (CE-H) compartment. The CE-H is instrumented partly with silicon sensors and partly with scintillator tiles.
duplicated and sent to two Stage-2 boards. This architecture, illustrated in Fig. 2.13, allows trigger cells to be processed over large regions covering the full depth of the detector, enabling the implementation of 3D clustering algorithms.

![Figure 2.13: HGCAL-TPG system for one endcap, composed of two processing stages, the other endcap being covered by an exact copy. The Stage-1 receives trigger data from on-detector electronics and is connected to the Stage-2 in a time multiplexing fashion. The Stage-2 is connected to the Central L1T.](image)

### 2.3.2 Reconstruction algorithms

The Stage-2 implements a clustering algorithm divided in two steps: the seeding of the clusters and the actual clustering around the identified seeds. For each of these two steps, position quantities divided by the layer depth $z$ are used such that a particle coming from the center of the detector and following a straight line has the same coordinates in all the detector layers.

- **Seeding**: trigger cell energies are projected and summed into histogram bins in the $r/z - \phi$ plane, where $r$ is the radial distance from the beam axis. A smoothing is applied to the raw histogram in order to reduce fluctuations and seeds are defined as local maxima above a threshold of about $E_{\text{seed}} = 10 \text{ MIP}/\sin \theta$. The positions of the seeds are defined as the energy-weighted barycenters of the positions of all the trigger cells in the bin containing the seed. The seeding parameters are chosen to avoid the reconstruction of multiple seeds for single showers, though this can still happen, in particular at high pseudorapidity and for hadronic showers. Nevertheless, nearby clusters are merged together downstream when building trigger objects from HGCAL trigger primitives, which limits the impact of energy splitting on the trigger objects performance.

- **Clustering**: trigger cells are attached to seeds within a distance in the $x/z - y/z$ plane varying between 0.015 in the first layers to 0.050 in the last layers. If a trigger cell can be attached to more than one seed, the closest candidate is chosen. The choice of the cluster size is a compromise between shower containment and energy
contamination from soft pileup.

The position of the cluster is computed as the barycenter of the trigger cell positions, weighted with their energy. Two energy interpretations are associated to each cluster: one interpretation for electromagnetic (EM) clusters and one interpretation for hadronic (HAD) clusters. The EM-interpretation uses only the energy deposited in the CE-E and only in the core of the cluster, within a distance of 0.015 (in $x/z - y/z$) from the cluster center. In addition, different layer weights are used to calibrate the energy deposited in the active material. This EM-interpretation is used in later stages to build electron and photon trigger candidates, while the HAD-interpretation is used otherwise. In addition to the energies and position of the clusters, several quantities encoding the shape and substructure of the clusters are computed. These are typically quantities containing information on the longitudinal development of the shower and its transverse size (which have a discriminating power between electromagnetic, hadronic and PU showers), as well as information on possible overlapping showers reconstructed as single clusters. Only some of the energy and shape quantities are sent by default, while optional quantities are sent if the cluster passes some quality or energy criteria. These criteria are not yet defined and the studies shown here assume that all quantities are available for all the clusters.

An $E_T$ threshold of 1 GeV, using the HAD-interpretation, is applied. It is appropriate for the matching with L1 tracks and yields an average number of clusters of about 200 per bunch crossing, dominated by low-energy PU clusters. Figure 2.14 (left) illustrates the distribution of the number of selected clusters for different transverse energy thresholds, while Fig. 2.14 (right) shows the average number of reconstructed clusters for $E_T > 1$ GeV as a function of the number of PU interactions.

![Figure 2.14](image-url)

Figure 2.14: Left: Distribution of the number of reconstructed clusters above different thresholds in the whole HGCAL. Right: Average number of reconstructed clusters for $E_T > 1$ GeV as a function of the number of PU interactions. Simulated events containing an average of 200 minimum bias collisions are used in both cases.

In parallel with the cluster reconstruction, a fixed number of projective towers is formed. These towers are meant to contain all the energy not included in any cluster. Two types of objects are taken as inputs: the module sums and the trigger cells which have not been clustered. It ensures that there is no loss of energy, which is useful in particular to attenuate the impact of thresholds, for instance on jet reconstruction. These towers will have a typical size in the $\eta - \phi$ plane of $\pi / 36 = 0.0873$, which matches the geometry of the barrel calorimeter towers. However at
high pseudorapidity this tower area becomes comparable with single trigger cells, so coarser
towers are foreseen, in particular outside the tracker acceptance. Given that the geometry of
the module sums does not match the geometry of the towers, module sum energies may be
split according to their overlap with the tower areas.

The different HGCAL-TPG processing steps described above, and their relationships, are sum-
marized in Fig. 2.15. The steps performed in the frontend electronics and in the backend elec-
tronics are both depicted.

Figure 2.15: Processing steps of the HGCAL-TPG in the frontend and in the backend electronics.
The primitives produced are a variable number of 3D clusters and a fixed number of projective
towers. 3D clusters are built from trigger cells passing a selection in the frontend, while towers
are summing, in a coarser manner, all the remaining energies that have not been aggregated
into clusters.

2.3.3 Implementation, latency and data volumes

The main task of the Stage-1 is to prepare the trigger cell data in a suitable format for the seed-
ing and clustering in the Stage-2. It synchronizes the asynchronous input data, calibrates the
trigger cell energies and sorts them in the radial direction, suitable to fill the seeding histogram
bins in order. It also starts to create tower sums from the available module sums. The main
latency drivers are the synchronization of the inputs and the sorting of the output trigger cells.

The Stage-2 first distributes trigger cells to $\phi$-columns and sums their transverse energies in the
seeding histogram bins. A first 1D smoothing kernel is applied in the $\phi$ direction, followed by
a second one in the radial direction. Local maxima in a $3 \times 3$ bins window are then identified
as seeds. The radial ordering of the input data is important to be able to pipeline efficiently the
different algorithm blocks and start to process trigger cells as they arrive. Input trigger cells are
kept until the seeds are formed and the different cluster properties (position, energies, shapes)
are computed from the trigger cells associated to the clusters. The Stage-2 also finalizes the
towers by adding the partial sums computed in the Stage-1, as well as the remaining trigger
cells not associated to clusters.

Only parts of the Stage-1 and Stage-2 processing blocks have been implemented and the ex-
ant latency of the system is not known at the moment, but the total HGCAL-TPG latency is
foreseen to be within $5 \mu s$. This estimate includes about $1 \mu s$ in the frontend ASICs (mainly
2.4 Muon barrel trigger primitives

2.4.1 Drift tubes

A schematic view of a quadrant of the CMS muon system is shown in Fig. 2.16. The barrel muon system is equipped with 250 DT chambers covering up to $|\eta| < 1.2$. It is divided into 5 wheels with identical layout, arranged one parallel to the other along the CMS global $z$ axis, and called W-2, W-1, W0, W+1 and W+2. Within each wheel, chambers are organized in four concentric station rings, labeled from inside-out as MB1 to MB4, and segmented into 12 sectors along the CMS global $\phi$ coordinate. Chambers from the three innermost DT stations are equipped with three superlayers (SL), each consisting of four, half staggered, layers of parallel DT cells. Two SLs measure the muon trajectory along the bending plane ($r - \phi$) whereas the remaining SL measures the position along the longitudinal ($r - \theta$) plane. The MB4 chambers are instead equipped only with two $r - \phi$ SLs.

The Phase-2 upgrade of the DT system is described in detail in Ref. [11]. It foresees a replacement of the chamber on-board electronics, which is presently built with components that are: (i) not sufficiently radiation hard to cope with HL-LHC conditions, and (ii) not designed to cope
with the expected increase of L1T rate foreseen for Phase-2 operation. DT chambers themselves
will not be replaced, hence the existing detectors will operate throughout Phase-2. In the up-
graded DT architecture, time digitization (TDC) data will be streamed by the new on-board
DT electronics (OBDT) directly to a new backend electronics, hosted in the service cavern, and
called barrel muon trigger layer-1 (BMTL1). Event building and trigger primitive (TP) genera-
tion will be performed in the BMTL1 using the latest commercial FPGAs. This will allow us to
build L1 TPs exploiting the ultimate detector time resolution (few ns) improving BX identifica-
tion, spatial resolution and reducing the probability to produce multiple trigger segments per
chamber for a given crossing muon (ghosts), with respect to legacy TPs [34].

Two algorithms are presently being evaluated as candidates to perform DT TPs generation,
which is carried out independently in each chamber. Both are implemented in software (as
C++ emulators) and validated in real hardware demonstrators, as described in Section 6.4.2.1.
They assume that muons follow a straight path inside a chamber and rely on the mean-timer
property, holding for triplets of half staggered drift cells characterized by constant drift velocity,
that allows the bunch-crossing of origin of an incoming muon to be identified [35].

The first algorithm, called analytical method (AM), has been designed following a very direct
hardware-oriented approach. It operates in three steps, called grouping, fitting and correlation.
Starting from a group of 10 nearby cells distributed across the four layers of a DT SL, the
grouping step selects patterns of 3 or 4 fired DT cells compatible with a straight line.
From those patterns, the fitting step exploits the mean-timer property to compute unambigu-
ously the BX corresponding to every subset of 3 cells within each pattern. For cases with 4-cells
patterns, the BX of each triplet is combined with the others using an arithmetic mean. Track
parameters are then computed using exact formulas from $\chi^2$ minimization.
Finally, a correlation between the two $r - \phi$ SLs is attempted. A potential match between
the primitives of the SLs in the same chamber is looked for, within a window of $\pm 25$ ns. If a match
is found, the track parameters are recalculated, either as an arithmetic mean of the ones of each

Figure 2.16: Longitudinal view of a quadrant of the CMS Phase-2 muon system. Different
colors in the figure refer to different sub-detectors: DT - orange, RPC - light blue, CSC - green,
iRPC - purple, GE - red, ME0 - orange.
SL (position, time), or as ratio of the difference between the positions of each primitive and the distance between the two SLs (direction). If no match is found, all per-SL primitives are forwarded to the next stage to maintain high efficiency.

The second algorithm, called histogram-based (HB) method, operates BX identification using a histogram-based mean-timer technique (called majority mean-timer, MMT) [36] and performs track segment reconstruction using a compact Hough transform (CHT) method [37].

In MMT, all meaningful triplets in a set of preclustered channels (called a macrocell) are used to identify the muon bunch-crossing by exploiting the mean-timer property. The most voted result among all triplets is chosen as candidate BX within a given macrocell. Votes are filtered, before being counted, in case they are compatible with a meaningful quadruplet pattern.

Within the CHT, all permutations of TDC count pairs from different layers are processed in parallel to compute track slope hypotheses [38]. Three histograms, two with pairs from each \( r - \phi \) SL and one from pairs built using both \( r - \phi \) SLs, are filled with such hypotheses. The results are combined, after applying a threshold cut, and the most voted hypothesis, common to all histograms, is chosen as track slope for the TP. The TP’s track intercept position is finally calculated out of the already computed crossing time and track slope.

In both algorithms, the multiplicity and permutation of hits used to build a TP define the trigger segment quality. Highest quality TPs are the ones where patterns of 3 or 4 hits from both \( r - \phi \) SLs are exploited to assess the BX and get combined in a single TP (correlated triggers). Alternatively, the BX can be identified and a TP can be built out of 3 or 4 hits from only one SL (uncorrelated triggers). For uncorrelated triggers, one or two hits from the second \( r - \phi \) SL may be used to confirm and improve a trigger segment. These confirmed triggers get also flagged with dedicated qualities.

### 2.4.2 Barrel resistive plate chambers

The barrel RPC system consists of 480 chambers organized in four stations called RB1 to RB4. The two innermost barrel stations, RB1 and RB2, are instrumented with RPCs both in the inner and outer faces of the corresponding DT chambers, while only one RPC layer is present in the outermost stations, RB3 and RB4. The spatial resolution of the RPCs in the azimuth direction is of the order of the centimeter and the average efficiency for one layer is above 95%, measured from Run-2 collision data [39].

In order to better exploit the intrinsic time resolution of the RPC system (1.5 ns), and ensure the robustness of its readout throughout the HL-LHC era, the off-detector electronics (called Link System) will be replaced during LS3. Regarding Level-1 Trigger, the most relevant aspect of this upgrade is the increase of the readout frequency from 40 MHz to 640 MHz (reading out the detector sixteen times per bunch-crossing). As a consequence, each RPC hit provided to the muon track finders will have an additional time information featuring a granularity of one sixteenth of bunch-crossing (sub-BX timing). More details on the RPC Link System upgrade can be found in Ref. [11].

Each link board (LB) reads 96 strips and the Link System arranges them by group of three: two Slave LBs (SLB) and one Master LB (MLB) collecting also the information from the two associated SLBs and ensuring the communication with the BMTL1. The barrel MLBs will be connected to this backend by 300 unidirectional optical links (5 per 30° sector per wheel) transmitting data at a rate of 10.24 Gb/s. The expected latency before RPC data gets to the BMTL1 (described in Fig. 5.4) corresponds to 43 bunch-crossings, shared as follows: 9 between RPC and the LB system, 12 in the LB system and 22 to reach the backend via 110 meter long fibers.
Table 2.5: Data frame format for the communication between one RPC master link board and the BMTL1. The information relative to the hit (link board number, strip number and sub-BX timing) will be replicated for each hit that occurred during the bunch-crossing, with the possibility of sending information for up to fourteen hits per MLB.

<table>
<thead>
<tr>
<th>Data field</th>
<th>Header of hits</th>
<th>Num. LB</th>
<th>Num. Strip</th>
<th>Num. Sub-BX</th>
<th>Time crossing</th>
<th>Alignment bit</th>
<th>Partition delay</th>
<th>End of data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bits</td>
<td>24</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>4</td>
<td>12</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

The data frame format for RPC hits is described in Table 2.5, and allows up to fourteen hits per bunch-crossing per MLB to be sent. In rare occurrences, where the number of hits would exceed this threshold, the remaining hits will be sent over subsequent bunch-crossings (up to eight) assigning them a special partition delay flag. The maximum hit rate in RPC, extrapolating from current data to HL-LHC conditions and including a safety factor of three, has been estimated to be 600 Hz/cm² [11]. At this rate, the data format described in Table 2.5 will require 0.936 Gb/s per link for physics information to be buffered in the BMTL1.

A muon crossing an RPC chamber induces, in general, a signal in more than one strip. Therefore, before using RPC information in the Level-1 trigger, a clustering of the RPC single hits will be performed in the BMTL1 (assigning cluster position with half-strip resolution), together with the DT TP building, and the DT+RPC combination into the so-called super-primitives described in the next section.

### 2.4.3 DT+RPC super-primitive combination

After building the DT segments and clustering the RPC hits, the BMTL1 will merge information from both sub-detectors into combined super-primitives (SP) according to the data flow shown in Fig. 2.17. This exploits the redundancy and complementarity of the two sub-systems, combining the DT spatial resolution with the RPC time resolution, thus increasing the trigger primitive performance and making them more robust against detector aging, as demonstrated in Section 2.4.4.

The baseline SP combination algorithm works as follows. The RPC clusters are translated into DT coordinate conventions and the DT primitives are matched to RPC TPs based on the azimuth angle, within a three units bunch-crossing window centered around the DT primitive’s bunch-crossing. For DT TPs that are matched to an RPC cluster, the SP BX and sub-BX timing are set based on RPC information. In the first two stations, the algorithm can also build an RPC-only TP out of hits from the two RPC layers, in chambers where no DT primitive is found. Furthermore, in the whole barrel, RPC clusters not matched to any DT segments nor used to build an RPC-only segment are passed to the track finders. To keep track of the full SP content and provide maximal flexibility to track finders, a dedicated data field describing the RPC use within a SP is provided in the SP data format. This three-bits flag is described in Table 2.6.

The full format used by the emulators of Phase-2 barrel trigger primitive generators for track segment objects is summarized in Table 2.7, where the number of bits required for each data field is reported.

Each track segment will consist of 64 bits. The present format only stores information for segments in the transverse view, whereas the format for segments reconstructed in the longitudinal view is not defined yet. In order to perform rate bandwidth computations, we assume the same format will be used for segments reconstructed in both views. This is a conservative assumption, as primitives from the longitudinal view, built using a single SL, will need to store
2.4. Muon barrel trigger primitives

Figure 2.17: Architecture for the muon barrel trigger primitive generation. DT hits are sent to the barrel muon trigger layer-1 via the on-board DT electronics (OBDT [11]), while RPC barrel hits are sent via the Link System. The dashed line represents the separation between the underground experimental cavern (UXC) and the underground service cavern (USC). The barrel muon trigger layer-1 will cluster the RPC hits, build the DT trigger segments and combine both collections into super-primitives. These super-primitives will then be sent to the Barrel Muon Track Finder (BMTF) and the Overlap Muon Track Finder (OMTF) via high performance optical links.

Table 2.6: Description of the barrel super-primitive (SP) RPC flag meaning.

<table>
<thead>
<tr>
<th>RPC Flag</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>DT segment that could not be matched to any RPC cluster</td>
</tr>
<tr>
<td>1</td>
<td>SP whose time and bunch-crossing are defined by RPC</td>
</tr>
<tr>
<td>2</td>
<td>segment made out of RPC clusters only</td>
</tr>
<tr>
<td>3</td>
<td>RPC cluster not matched to DT nor used to build an RPC only segment</td>
</tr>
<tr>
<td></td>
<td>(no direction information)</td>
</tr>
<tr>
<td>4</td>
<td>DT segment matched to an RPC cluster but with all</td>
</tr>
<tr>
<td></td>
<td>quantities left untouched</td>
</tr>
<tr>
<td>5</td>
<td>RPC cluster matched to a DT segment or used for an RPC only segment</td>
</tr>
</tbody>
</table>

In order to maintain high efficiency for close-by muons, in particular in regions where the background rate is large, we assume that up to four track segments per chamber per bunch crossing, in each of the transverse/longitudinal views, can be reconstructed and delivered to the Layer-2 Muon backend.

Considering one muon barrel 30° sector as the region of interest for a BMTL1 module, the maximum output payload from each board will consist of 16 segments in the transverse view and 12 segments in the longitudinal view, totaling 1792 bits/BX, corresponding to about 72 Gb/s.

Such a payload (1792 bits/sector/BX) can be accommodated into two/three/five 40/25/16 Gb/s links per sector. Since a time-multiplexing of 18 will be used between the BMTL1 and the BMTF, this could be accommodated into one 16 or 25 Gb/s fiber per sector. This is a feasible solution given present transceiver technologies and the overall input link count that the BMTF will be able to receive. An intermediate layer acting as concentrator is also considered a possibility in case an increase of the bandwidth is needed.
Chapter 2. Inputs to the L1 trigger system: the trigger primitives generation

Table 2.7: Description of the barrel super-primitive data frame format. Including 5 spare bits, each super-primitive will consist of 64 bits.

<table>
<thead>
<tr>
<th>Data field</th>
<th>Position</th>
<th>Bending angle</th>
<th>BX in-orbit</th>
<th>Sub-BX time</th>
<th>Chi2</th>
<th>Quality</th>
<th>Superlayer</th>
<th>RPC info</th>
<th>Spare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bits</td>
<td>17</td>
<td>12</td>
<td>12</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

2.4.4 Muon barrel primitives performance

The performance of the barrel muon SPs was studied with Monte Carlo simulations in different pileup conditions, including a worst case scenario with an average number of overlapping collisions of 300 per BX. Performance metrics were confirmed to be stable as a function of pileup in the whole probed range, hence, in the following, results are reported only using samples with an average pileup of 200.

Effects of detector ageing and failures are also accounted for within this study. The assessment of overall chamber longevity and potential system failures is described in detail in Ref. [11]. In general, in the barrel region of the muon system, only DTs have measured loss of efficiency in accelerated radiation ageing, which has been used to estimate the performance over HL-LHC running. Given the multiplicity of layers comprising a DT chamber, the effect on DT trigger-segments has been estimated to be small almost everywhere, with the exception of the MB1s of the external wheels (10% of the chambers), where the effect on TPs is estimated to be relevant in the running period close to the end of Phase-2, if safety factors of 2 or larger are considered. Though the irradiation program of spare RPC chambers shows so far no sign of detector performance degradation, failures have been observed over the years of operation due to electronics deficiency or to chambers switched off in order to mitigate the ecological and financial impact of gas leaks (only a fraction of the leaky chambers being switched off). Most of these failures can be repaired during extended technical stops or long shutdowns, hence the worst detector conditions are encountered at the end of data-taking eras. Unless explicitly stated otherwise, all the performance curves consider the RPC system with realistic electronics failures expected before long shutdowns. Despite the difficulty of predicting such random events, we provide pessimistic leak scenarios switching off the 30 currently unrepaired barrel chambers and adding extra ones assuming three new leaks per year of operation and a repair success of 50% every long shutdown. This scenario is pessimistic for two reasons: first, several actions are being taken to mitigate this issue and second, it assumes all leaky chambers will be turned off.

In the following, a scenario that targets the ageing/failures expected at the end of the HL-LHC running is considered. It is computed assuming a baseline of 3000 fb⁻¹ of data collected at an instantaneous luminosity of $5 \times 10^{34}$ cm⁻²s⁻¹ and considering a safety factor of 2 both on instantaneous and integrated luminosities. Within this last scenario the average DT hit efficiency within a given chamber is ~ 71% in the MB1s of W+/−2, whereas it ranges from 84% to 97% in the rest of the detector, and the RPC system is considered to have 50 barrel chambers off.

The Phase-2 Level-1 Trigger will profit from an unprecedented time resolution achieved from the sub-BX granularity provided by the new electronics. Besides improving the BX identification efficiency of the TPs, this feature can be exploited in the development of a dedicated Long Lived Particles triggers, such as the one for Heavy Stable Charged Particles (HSCP) relying on time of flight, as described in Section 3.3.5. The muon system plays an important role in such trigger algorithms given its location far away from the interaction point, which brings a great lever arm to compute the time of flight. Furthermore, its redundancy provides many
measurements, thus increasing the efficiency and helping to keep the rates of time-based algorithms under control. The time resolution of the muon barrel SPs, shown in Fig. 2.18 (left), is estimated from a MC simulation based on the time coordinate of SPs associated to generated muons with $p_T > 20$ GeV, assuming that every chamber is calibrated for prompt muon time arrival and that the smearing due to the primary vertex position is negligible. The simulation includes a uniform smearing to account for the signal propagation along the strip/wire and a gaussian smearing to account for detector/electronics effects and assumes a perfect inter-chamber calibration.

Figure 2.18 (right) shows the efficiency to reconstruct a barrel SP and correctly assign its BX. It is computed with respect to offline reconstructed segments geometrically matched with incoming generated muons and no minimal quality threshold cut is applied to the SPs. Each bin of the plot represents a ring, which refers to chambers from a given muon barrel wheel and a given station. In the no-ageing case, the TP efficiency is $\sim 98\%$ for DT-only primitives and increases to above $\sim 99\%$ for combined DT+RPC SPs. In most of the detector, efficiency for DT-only TPs remains higher than 90%, even if ageing is considered. The main exception being the efficiency in MB1s of the external wheels, which degrades significantly for DT-only triggers because of the reduced hit efficiency expected for those chambers. Anyhow, through the use of DT and RPC combined SPs, it is possible to restore efficiency above 90% also in this critical region.

The spatial resolution for position and direction of the DT TPs, computed with respect to simulated hits, is presented in Fig. 2.19 (left) and Fig. 2.19 (right), respectively. The layout of the plots is identical to the one used for Fig. 2.18 (right). As and example, MB2 stations show a position resolution of $\sim 0.3$ mrad ($\sim 0.6$ mrad) for the AM (HB) method, which reflects an improvement of a factor of $\sim 6$ (3) with respect to the legacy TPs [34]. A similar improvement is observed for the direction resolution, which is $\sim 1$ mrad (2 mrad) for the Phase-2 AM (HB) TPs. The same conclusion holds for all stations. The different resolution between the AM and HB algorithms is due to the chosen granularity for the HB method, driven by the optimization criteria stated in [37], where it is shown that a higher resolution can be obtained with a finer granularity, though using a larger amount of computational resources.

### 2.4.5 Future improvements

The descriptions of the barrel muon trigger primitives provided up to now correspond to the baseline upgrade program. Improvements to this baseline are being investigated and discussed in what follows.

Firstly, few solutions to preserve barrel TPs performance in case of detector failures and ageing are being considered. For example, the time-matching strategy, used to correlate information from the two DT $r - \phi$ SLs of a given chamber, can also be exploited to match with other primitives from the $r - \theta$ SL of the same chamber, or with SLs in neighboring stations. Moreover, alternatives for the grouping step of the AM DT TP generation, that allow the information from both $r - \phi$ SLs to be combined at the first stage of the DT TP reconstruction, are being studied. Both approaches aim at maximizing the efficiency to build standalone DT TPs in case of low hit multiplicity within a single SL.

Strategies which exploit RPC timing directly into the TP building, alleviating the interdependence of time and spatial measurements inherent to the drift tube technology, are also considered as potential options to mitigate the impact of ageing and failures. Nevertheless this option has been left as a future effort to ensure, as a first priority, that muon triggers can work effectively without explicit sub-system interdependence.
detected signals must propagate along the length of a DT wire or RPC strip before being read.

Additionally, the position measurement of DT hits occurring at different times, increasing purity and resolution of the RPC TPs. The efficiency is computed with respect to reconstructed segments matched geometrically with generator level muons with $p_T > 20$ GeV. Results are computed with and without including detector ageing and failures.

Finally, general improvements of the baseline algorithms are also going to be studied. One refinement consists of using sub-BX timing precision in the RPC clustering to separate spatially adjacent hits occurring at different times, increasing purity and resolution of the RPC TPs. Additionally, the position measurement of DT $r - \theta$ TPs can be used to correct for a uniform time smearing of RPC or DT TPs built in the $r - \phi$ view, which originates from the fact that detected signals must propagate along the length of a DT wire or RPC strip before being read.
2.5 Muon endcap trigger primitives

2.5.1 Endcap resistive plate chambers

A diagram representing a quadrant of the CMS muon system is presented in Fig. 2.16. The current endcap RPC system consists of 576 chambers organized in four stations called RE1 to RE4 and covering a pseudorapidity region from 0.9 to 1.9. The readout system of the endcap chambers will undergo the same upgrade program as the barrel chambers described in Section 2.4.2. This readout system will send single hit data to the RPC endcap backend following the protocol defined in Table 2.5, and will achieve the same latency as the barrel link system. In total, 192 unidirectional 10.24 Gb/s optical links (2 per 30° sector per station) will ensure the data transfer between the RPC endcap MLB’s and the RPC endcap backend, which will perform the single hit clustering and act as a concentrator. With the maximum hit rate quoted in Section 2.4.2, 0.53 Gb/s of physics data should be buffered per link in the RPC endcap backend.

Table 2.8 shows a proposal for the RPC cluster data format. In order to maintain a high TP efficiency in regions with large background hit rate and to be able to reconstruct events with a particular topology (displaced jets, collimated muons, etc.), we expect to send four hits per chamber per bunch-crossing. The RPC system sends information from 36 chambers per EMTF 60 degree sector3 totalling 2304 bits/BX or 92.16 Gb/s. Assuming 25 Gb/s links, these payloads can be accommodated with four links to each of the track finder sector, including some margin. Sending data from RE1/2, RE1/3, RE2/3, RE3/3 and RE4/3 to OMTF leads to a payload of 3840 bits/BX (OMTF sector covers 120 degrees, i.e. 60 RPC chambers), which implies that 7 links per processor are required.

2.5.2 Improved resistive plate chambers

In order to increase the redundancy of the muon system in the challenging forward region, new improved RPCs (iRPC) chambers will be installed in stations 3 and 4 (RE3/1 and RE4/1), extending the RPC pseudorapidity coverage to $|\eta| < 2.4$. The reduced bakelite resistivity and gap thickness (1.4 mm compared to 2 mm in the present RPC system) allows the detector to withstand the high rates anticipated in RE3/1 and RE4/1. The latter also improves the time resolution as shown in Section 2.5.5 and the high sensitivity electronics ensure to keep the efficiency as high as before, even though less charge is collected per hit. This new electronics features a readout at both ends of the strips allowing us to derive the hit radial position based on difference in signal time arrival.

Both RE3/1 and RE4/1 will be instrumented with 18 chambers (for a total of 72 chambers considering both endcaps) equipped with two frontend boards (FEBs). Each of these FEBs reads out 48 strips with three PetiROC ASICs [40] (one PetiROC has 32 channels to read both ends of 16 strips). A proposal for the data frame format generated by the FEB is shown in

---

Table 2.8: Description of the endcap RPC clusters data frame format. The cluster position can be retrieved from the first fired strip and the cluster size. The maximal cluster size value will be assigned to any cluster having nine or more fired strips.

<table>
<thead>
<tr>
<th>Item</th>
<th>First strip</th>
<th>Cluster size</th>
<th>Sub-BX timing</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of bits</td>
<td>7</td>
<td>3</td>
<td>6</td>
<td>16</td>
</tr>
</tbody>
</table>

out. Such enhancement would improve the sub-BX TP time resolution.

---

3RE1/2, RE2/2, RE3/2, RE3/3 RE4/2 and RE4/3 are equipped with chambers covering 10 degrees.
Table 2.9: Physics data frame format for the communication between iRPC front-end and back-end. The timing information is obtained by summing coarse and fine time.

<table>
<thead>
<tr>
<th>Item</th>
<th>Edge</th>
<th>ASIC</th>
<th>Channel</th>
<th>Coarse time</th>
<th>Fine time</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of bits</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>16</td>
<td>8</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 2.9. In total, 144 optical links at 4.8 Gb/s will ensure the communication with the iRPC backend, where the single hits will be clustered. The physics data transmission rate between iRPC FEBs and backend is 2.5 Gb/s per link assuming an average cluster size of 3 and a hit rate of 2000 Hz/cm$^2$.

The iRPC TP building works as follows. Adjacent hits are first clustered separately from both ends of the strips, within a 1.5 ns time window to account for possible different signal rising time. Clusters from both ends are then combined, allowing us to have a missing signal from one edge and finally a filtering is applied to split clusters made of hits with incompatible radial distance. The data frame format for iRPC clusters will follow the one for RPC clusters (described in Table 2.8), adding 8 more bits to encode the cluster radial position. One EMTF sector receives data from six iRPC chambers$^4$, which leads to a payload of 23.04 Gb/s assuming that 4 clusters per chamber will be sent every BX. This can be accommodated with e.g. three 10 Gb/s or two 16 Gb/s links per EMTF sector.

2.5.3 Endcap cathode strip chambers

The endcap CSC system consists of 540 chambers organized in four stations called ME1 to ME4 and covering a pseudorapidity region from 0.9 up to 2.4. Each CSC chamber contains six layers, and each layer contains cathode strips that run radially, and wires that run approximately orthogonal to the cathode strips. The cathode strips vary in width from 4 to 16 mm, depending on the radius from the beam line and the specific location of the chamber within CMS. The anode wires are directly wired together in sets of 5 to 17 wires per readout channel, covering 1.6 to 5.4 cm radially. Cathodes and anodes are both instrumented with trigger, as well as readout electronics.

The cathode trigger electronics of the CSC require a certain minimum charge deposition to register a muon hit. Although the threshold is well below minimum ionizing, the random noise level of cathode hits is very low. For the trigger, the muon hits are further localized with an accuracy of one-half of a strip on each chamber layer by analog comparison of charge deposition on each strip with its neighbors, as well as comparison of the neighbor strips – the technique is built into Comparator ASIC chips. The CSC anode electronics frontend has constant-fraction discriminators that create muon hits with very little time walk; the CSC anode hits are registered each LHC bunch crossing.

The baseline CSC trigger primitives, known as local charged tracks (LCT), constitute the input from the CSC trigger motherboards (TMBs) to the L1 muon trigger track finders. The LCT are a coincidence between straight-line patterns found in anode and cathode electronics. Both anode and cathode trigger electronics require at least four layers to contain hits within patterns. As the magnetic field is solenoidal, muons do not bend significantly in the $R-z$ plane measured by CSC anode wires, and so there is essentially only one CSC trigger anode pattern that merely indicates that the muon appeared to have originated roughly from the CMS collision point. On the other hand, muons do bend in the $r-\phi$ plane, and a set of nine cathode trigger patterns is used to indicate the amount and direction of bending. The bending is inversely related to

$^4$RE3/1 and RE4/1 are equipped with chambers covering 20 degrees.
momentum, and is largest in the first CSC station, i.e. the one closest in $z$ to the interaction point. Occupancy of CSC chambers by tracks and neutron-induced hits is much higher in the inner ring closest to the beam line, and is higher in the first CSC station than in the other stations.

The performance of the CSC trigger primitive generation has been excellent in LHC running thus far. The CSC trigger primitives are generated with 98% efficiency in all stations except ME1/1, where it is around 94%, localize the muon positions in the $\phi$ (bend) direction within an RMS of 0.174 strips, and find the correct bunch crossing for well over 99% of the muons.

The major upgrade to the CSC will replace on-chamber cathode boards on the inner rings of chambers ($1.6 < \eta < 2.4$) in order to handle higher trigger and output data rates, and FPGA mezzanine boards on most of the on-chamber anode boards in order to cope with higher L1 trigger latency. Corresponding off-chamber boards that receive trigger and readout data will also be replaced to handle the higher data rates.

The CSC trigger primitives are sent to the L1 trigger track finders via optical links from muon port cards (MPC) located in crates on the periphery of the endcap muon system. There are 60 MPCs, and each sends the LCT information for up to two muons to the trigger over 8 links transmitting at 3.2 Gb/s with standard 8B/10B encoding.

The algorithms of the CSC trigger are expected to change substantially in the future. If the algorithms do not change, efficiency loss is expected due to high-occupancy effects, such as deadtime, that grow with luminosity. The loss of efficiency is worst in the ME1/1 chambers, which can lose as high as 15% efficiency, but this will be almost completely alleviated by updated firmware that allows simultaneous processing of trigger hits into LCTs in different parts of each CSC chamber. Additionally, the upgraded FPGAs used for the new electronics will allow finer granularity of trigger patterns; studies have shown that the cathode (bend direction) position and angle resolutions can be improved by factors of 1.87 and 1.35, respectively. Studies to determine whether the anode position and angle resolution can be improved as well are being done. Larger factors of improvement in the bend direction resolution are possible by combination of CSC trigger hits with GEM hits in the same station; these are described in the GEM section (Section 2.5.4).

### 2.5.4 Gas electron multipliers

New gas electron multiplier (GEM) chambers will be installed in the forward region $1.6 < |\eta| < 2.8$ [11]. The installation of GEM detectors will allow a precise measurement of the muon bending angle in the first and second stations to be performed and used as a handle to control the muon trigger rate. The added sensitive detecting layers will increase the trigger efficiency and improve the operational resilience of the system. GEM foils have been demonstrated to be a suitable technology for the CMS forward region. A single GEM chamber is made of three GEM foils. A stack of two or six GEM chambers forms a superchamber. These superchambers will be installed in three distinct locations in the forward region ($1.6 < |\eta| < 2.8$), dubbed GE1/1, GE2/1 and ME0. The GE1/1 station features 36 superchambers, each having two chambers, in front of the ME1/1 chambers. Each superchamber covers 10 degrees in $\phi$ and $1.6 < |\eta| < 2.15$ in pseudorapidity. The GE2/1 station is similar to the GE1/1 station, although GE2/1 chambers are much larger. Eighteen superchambers, each 20 degrees wide, cover the region $1.6 < |\eta| < 2.4$. CMS will also be equipped with a new ME0 station behind the HGCAL, closer to the interaction point than any other endcap muon detector. ME0 superchambers have six 20-degree wide chambers and cover the region $2.0 < |\eta| < 2.8$. 
The GEM detectors will deliver two types of trigger primitives to the L1T system. The double-layer GE1/1 and GE2/1 chambers will send identical trigger pad clusters onto two separate trigger paths: to the neighboring CSCs and to the EMTF via an ATCA card (GEM concentrator). By comparison, the ME0 superchambers will send multi-layer segments directly to the EMTF. Both types of trigger primitives will be constructed from trigger pads which have an angular resolution of 0.9 mrad in $\phi$. The trigger pads are built on-chamber in the VFAT3 chip [41] as an OR'ed combination of two neighboring strips. GE1/1 has 192 trigger pads per $\eta$ partition (8 $\eta$ partitions per chamber), whereas GE2/1 and ME0 each have 384 pads per $\eta$ partition. Single trigger pads are at least 97% efficient. The trigger pad data will be transmitted from the VFAT3 chip to the OptoHybrid (OH) board. The GE1/1 and GE2/1 OH boards will construct clusters from maximum 8 adjacent trigger pads in the chamber with a latency of 1.5 BX. Each pad cluster is 14 bits wide, as can be seen in Table 2.10 (left). The cluster data will be transmitted from the OH to the nearby CSC optical trigger motherboard (OTMB) and to the GEM concentrator. Per superchamber, up to 16(40) clusters will be transmitted from GE1/1 (GE2/1) OHs. The overflow rate of the cluster finder has been estimated to be $O(10^{-5})$ per BX.

Table 2.10: Bit assignment for the GEM trigger primitive data. Left: GE1/1 and GE2/1 trigger pad cluster. Right: ME0 track segment.

<table>
<thead>
<tr>
<th>Purpose</th>
<th>N bits</th>
<th>Purpose</th>
<th>N bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partition number</td>
<td>3</td>
<td>$\phi$ Coordinate</td>
<td>10</td>
</tr>
<tr>
<td>Pad number</td>
<td>6</td>
<td>$\eta$ Coordinate</td>
<td>5</td>
</tr>
<tr>
<td>Sector number</td>
<td>2</td>
<td>Pattern</td>
<td>7</td>
</tr>
<tr>
<td>Cluster width</td>
<td>3</td>
<td>Quality</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>14</strong></td>
<td><strong>Total</strong></td>
<td><strong>24</strong></td>
</tr>
</tbody>
</table>

In the GEM-CSC trigger path, GEM trigger clusters are combined with CSC Cathode and Anode LCTs to form highly efficient GEM-CSC LCTs. The requirement in the number of hit layers can be relaxed from ‘at least four CSC layers hit’ to ‘at least three CSC layers and at least one GEM layer hit’. The GEM-CSC LCT trigger objects will have a width of 32 bits as regular LCTs, but unused or redundant bits will be repurposed to indicate the presence of GEMs. In the GE1/1-EMTF trigger path, the GE1/1 concentrator multiplexes trigger pad cluster data from every three OH optical links running at 3.2 Gb/s into a link running at 9.6 Gb/s (with a latency of less than 0.5 BX). In the GE2/1-EMTF trigger path, eight 3.52 Gb/s links are multiplexed into a single 25 Gb/s link. Coincidences of trigger pad clusters (with an efficiency > 95%) will be used in the EMTF track-builder and momentum assignment. In contrast to GEM chambers, 24 ME0 OHs per superchamber (4 OH per chamber) will transmit all raw trigger data, without losses, directly to an off-chamber ATCA card every BX at 366 Gb/s. The ATCA cards will reconstruct track segments from multi-layer coincidences of trigger pads. Each ME0 ATCA card will transmit up to 16 track segments from two neighboring ME0 superchambers per BX onto a single 25 Gb/s fiberlink to the EMTF. A Level-1 ME0 segment builder has been developed in the context of the Phase-2 Muon TDR [42]. The algorithm is based on the offline segment reconstruction, but uses trigger pads with two times poorer spatial resolution than offline reconstructed hits. Simulations show that the segment builder is 99% efficient for prompt muons with $p_T > 3$ GeV and for displaced muons with $p_T > 5$ GeV and $d_{xy} < 20$ cm. The latency of the ME0 track segment builder is expected to be similar to the latency of the LCT builder in the OTMB. A draft data format for the ME0 track segment is presented in Table 2.10 (right). A simplified picture of the entire overlap and endcap trigger primitive architecture is shown in Fig. 2.20.
2.5. Muon endcap trigger primitives

2.5.5 Endcap muon trigger primitives performance

The performance of the LCT trigger primitives is described in what follows. Figure 2.21 shows that large losses in efficiency (up to 15%) for an average pileup level of 200 can be mitigated by upgrading the trigger algorithm (CSC Phase-2). By combining GEM hits with CSC hits in the trigger algorithms for ME1/1 and ME2/1 (GEM-CSC), additional inefficiency can be recovered for an average pileup level of 200 (between 1% and 20%). Figure 2.22 compares the directional resolution for CSC LCTs in ME1/1 and ME2/1 (left) and GEM-CSC LCTs (right) for displaced muons. By combining GEM with CSC information in the EMTF, a more precise measurement can be done of the muon direction in ME1/1 and ME2/1.

As described in Section 2.4.4, the improved time resolution is an important aspect of the upgrade program as it allows us to develop new trigger algorithms such as the HSCP algorithm described in Section 3.3.5. The finer time information in the muon endcap TPs comes from the RPC clusters. Their time resolution, shown in Fig. 2.23, is slightly better than the barrel muon TPs due to the fact that the endcap RPC chambers have shorter strips than the barrel ones and that the 2D readout of iRPC chambers completely suppress the uniform smearing originating from signal propagation along the strips. The timing simulation is performed under the conditions described in Section 2.4.4. Recent measurements performed on iRPC prototypes have shown time resolutions below 500 ps [43], when using an improved version of the frontend electronics that could not be included in the simulation.

2.5.6 Future improvements

The upgrades to the CSC trigger system discussed in Sections 2.5.3 and 2.5.4 are currently being studied for implementation in the firmware during LS2 or during Run-3 data-taking. The CSC backend system will be upgraded to read out GEM-CSC LCTs with CSC information partially missing. Research into possibilities to trigger on exotic signatures from long-lived particles...
Chapter 2. Inputs to the L1 trigger system: the trigger primitives generation

Figure 2.21: Comparison of the L1T track segment efficiency in the endcap for various scenarios: Run-2 algorithm at PU0 (red) and PU200 (orange), Phase-2 algorithm at PU200 (green), GEM-CSC algorithm at PU200 (black) and PU300 (blue). Left: ME1/1 station. Right: ME2/1 station. The efficiency loss in ME1/1 near $|\eta| \sim 2.1$ is caused by a discontinuity in the cathode strip readout. The efficiency losses in ME2/1, near $|\eta| \sim 1.8$ and 2.05, are due to missing wires at the edges of high voltage sectors.

Figure 2.22: Resolution of the muon direction in the EMTF for ME1/1 and ME2/1 stations. Muons are simulated with a transverse momentum $p_{T,\mu}$ between 9 and 11 GeV, absolute rapidity $\eta_\mu$ between 2.05 and 2.15 and an absolute value of the impact parameter $dxy_\mu$ between 10 and 50 cm. For each chamber type or chamber combination in the legend, the relevant $\eta$ range is given in brackets. Left: CSC without GEMs. Right: CSC with GEMs (GE1/1, GE2/1, ME0).

with more than 2 LCTs per chamber and per BX is also ongoing.

The possibility of matching the (i)RPC clusters with CSC TPs and complementing them with the sub-BX timing information will be studied. This would greatly improve the 25 ns time resolution of the CSC TPs and could be used to spot the ones that are not compatible with the expected time arrival of prompt muons, thus lowering the muon track finder rates. Another possible gain of combining iRPC with CSC TPs is to remove ghost signals in the latter arising if two or more hits occur in a given CSC chamber within $\pm 1$ bunch-crossing. Given its 2D readout, iRPC does not suffer from combinatorial ambiguity inherent to the two independent 1D readouts. Even though a common CSC+(i)RPC backend is not part of the baseline scenario, this combination could potentially be performed later in the hardware chain.

Following the design of the ME0 electronics scheme, an ME0 track segment builder algorithm...
2.6 External Triggers

The Global Trigger (Section 3.7) finds the trigger decision through the parallel evaluation of O(1000) trigger algorithms based on trigger objects reconstructed by the muon, calorimeter, track and particle-flow triggers. Additionally, the Global Trigger may receive a small number of simple "external" condition signals that may either be directly used as additional trigger algorithms or may be used as additional conditions in regular trigger algorithms.

The BPTX subsystem will send the BEAM1 and BEAM2 signals corresponding to BPTX detecting a bunch in beam 1 or beam 2 in a specific bunch crossing. Any derived signals such as BPTX-AND, BPTX-OR, etc., that were produced by the BPTX subsystem in Phase-1, will be derived from the BEAM1 and BEAM2 signals by the Global Trigger.

Additional sources of external triggers are to be defined. Examples of such trigger sources in the past were the CASTOR forward calorimeter and the ZDC "zero-degree calorimeter" and the PPS/Roman Pots.

Where possible, these external condition signals should be transmitted via the same type of 25 Gb/s optical fibers as the other signals mentioned above. To retain flexibility, the Global Trigger will continue to offer the alternative of receiving a limited number of external condition signals via galvanic LVDS lines as in the past. Such connections are easy and quick to establish in case of newly arising requirements. An additional advantage could be that signal transmission over galvanic lines can be slightly faster than over optical cables.

Figure 2.23: Time distribution of the endcap (i)RPC trigger primitives associated to a generated muon track with $p_T > 20$ GeV. The solid blue line, labelled as 'All RPC TP', represents the distribution obtained without separating RPC and iRPC TPs. The width of the distributions is an estimate of the time resolution.
2.7 Summary of trigger input bandwidth and latency

Table 2.11 summarizes the trigger primitives inputs from the Phase-2 sub-detectors. The time-multiplexing period for each system is provided. The number of output links, link speed and expected latency are also specified. The number of output links is given without including the (available) number of links to the 40 MHz scouting system as this is not expected to be a limiting factor. Table 2.12 provides the required bandwidth from each sub-detectors. A total input bandwidth of 62593 Gb/s is expected to be digested by the Phase-2 L1 trigger logic downstream.

A summary diagram displaying the links between the trigger primitives, the trigger objects, the Level-1 algorithms used in the menu and the physics channels, is shown in Fig. 2.24.

Table 2.11: Trigger primitive (TP) inputs expected from Phase-2 sub-detectors. This table summarizes the main features of the trigger backend system of each sub-detector sending TP to the Phase-2 L1 trigger. Reported here: the TMUX period, the number of output links, the link speed and the latency (defined as the time after the bunch crossing at which the first data are received by the trigger system). The numbers in parenthesis account for the links required to transmit data from overlapping sector regions. Note that for some sub-detectors (DT, RPC, iRPC and ME0) the latency would need to be estimated for the Phase-2 detector.

<table>
<thead>
<tr>
<th>Detector</th>
<th>TMUX period</th>
<th>Output links</th>
<th>Link speed (Gb/s)</th>
<th>Latency (µs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Track Finder</td>
<td>18</td>
<td>$9 \times 2 \times 18 = 324$</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>ECAL</td>
<td>1</td>
<td>3060</td>
<td>16</td>
<td>1.5</td>
</tr>
<tr>
<td>HCAL</td>
<td>1</td>
<td>144</td>
<td>6.4</td>
<td>1.5</td>
</tr>
<tr>
<td>HF</td>
<td>1</td>
<td>36</td>
<td>6.4</td>
<td>1.5</td>
</tr>
<tr>
<td>HGCAL</td>
<td>18</td>
<td>$2 \times 3 \times 4 \times 18 = 432$</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>DT+RPC to BMTF</td>
<td>18</td>
<td>$60 \times 18 = 1080$</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>DT+RPC to OMTF</td>
<td>1</td>
<td>72 (+18)</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>RPC(endcap) to OMTF</td>
<td>1</td>
<td>42 (+6)</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>RPC(endcap) to EMTF</td>
<td>1</td>
<td>48 (+12)</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>CSC to OMTF</td>
<td>1</td>
<td>360 (+30)</td>
<td>3.2</td>
<td>1.75</td>
</tr>
<tr>
<td>CSC to EMTF</td>
<td>1</td>
<td>480 (+108)</td>
<td>3.2</td>
<td>1.75</td>
</tr>
<tr>
<td>iRPC</td>
<td>1</td>
<td>24 (+12)</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>GEM (GE1/1)</td>
<td>1</td>
<td>96 (+12)</td>
<td>9.6</td>
<td>1.0</td>
</tr>
<tr>
<td>GEM (GE2/1)</td>
<td>1</td>
<td>36 (+12)</td>
<td>25</td>
<td>1.0</td>
</tr>
<tr>
<td>ME0</td>
<td>1</td>
<td>24 (+12)</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>
2.7. Summary of trigger input bandwidth and latency

Table 2.12: Summary of the logical input data to the Phase-2 L1 trigger. The type of objects transmitted by each detector and their corresponding format are provided. The total number of objects per event is estimated by looking at multiplicity distributions obtained from MC simulated events. The required bandwidth is specified in the last column and adds up to 62 593 Gb/s.

<table>
<thead>
<tr>
<th>Detector</th>
<th>Object</th>
<th>N bits/object</th>
<th>N objects</th>
<th>N bits/BX</th>
<th>Required BW (Gb/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRK</td>
<td>Track</td>
<td>96</td>
<td>1665</td>
<td>159 840</td>
<td>6 394</td>
</tr>
<tr>
<td>EB</td>
<td>Crystal</td>
<td>16</td>
<td>61 200</td>
<td>979 200</td>
<td>39 168</td>
</tr>
<tr>
<td>EB</td>
<td>Clusters</td>
<td>40</td>
<td>50</td>
<td>2 000</td>
<td>80</td>
</tr>
<tr>
<td>HB</td>
<td>Tower</td>
<td>16</td>
<td>2 304</td>
<td>36 864</td>
<td>1 475</td>
</tr>
<tr>
<td>HF</td>
<td>Tower</td>
<td>10</td>
<td>1 440</td>
<td>13 824</td>
<td>553</td>
</tr>
<tr>
<td>HGCAL</td>
<td>Cluster</td>
<td>250</td>
<td>416</td>
<td>104 000</td>
<td>4 160</td>
</tr>
<tr>
<td>HGCAL</td>
<td>Tower</td>
<td>16</td>
<td>2 600</td>
<td>41 600</td>
<td>1 664</td>
</tr>
<tr>
<td>MB DT+RPC (SP)</td>
<td>Stub</td>
<td>64</td>
<td>1 720</td>
<td>110 080</td>
<td>4 400</td>
</tr>
<tr>
<td>ME CSC</td>
<td>Stub</td>
<td>32</td>
<td>1 080</td>
<td>34 560</td>
<td>1 382</td>
</tr>
<tr>
<td>ME RPC</td>
<td>Cluster</td>
<td>16</td>
<td>2 304</td>
<td>36 864</td>
<td>1 475</td>
</tr>
<tr>
<td>ME iRPC</td>
<td>Cluster</td>
<td>24</td>
<td>288</td>
<td>6 912</td>
<td>276</td>
</tr>
<tr>
<td>ME GEM</td>
<td>Cluster</td>
<td>14</td>
<td>2 304</td>
<td>32 256</td>
<td>1 290</td>
</tr>
<tr>
<td>ME0 GEM</td>
<td>Stub</td>
<td>24</td>
<td>288</td>
<td>6 912</td>
<td>276</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>62 593</strong></td>
</tr>
</tbody>
</table>
Chapter 2. Inputs to the L1 trigger system: the trigger primitives generation

Figure 2.24: Summary diagram displaying the links between the trigger primitives (first column), the trigger objects (2nd column), the Level-1 algorithms used in the menu (3rd column), and the physics channels (4th column). The trigger primitives include crystals, towers and clusters from calorimeters (ECAL, HCAL, HF and HGCAL), stubs and clusters from the muon detectors (DT, RPC, CSC, GEM and iRPC), as well as L1 tracks from the track finder. The trigger objects types produced by the Phase-2 L1 trigger system are represented: standalone, track-matched, tracker-based and PF/PUPPI-based. The L1 algorithms column regroup classes of triggers as described in the simplified menu, described in Section 4.1. The physics channels listed here are further described in the introduction (see Section 1.4) and in the chapter presenting the L1 menu, Chapter 4. The links to the main L1 algorithms used for selecting these final states are represented.
Chapter 3

Trigger algorithms

3.1 Introduction: triggering on physics objects at the HL-LHC

This Chapter describes the development of trigger algorithms envisioned for the Phase-2 L1 upgrade system. Their development was guided by the requirements of the HL-LHC physics program, including heavy ion collisions. Although many approaches could be considered, the studies presented here aim primarily to demonstrate that baseline algorithms complying with both physics needs and data-taking conditions can be realized. The main characteristics of these algorithms and their performance are provided, as well as areas for potential future improvement. Performance results including efficiencies, as well as energy and position resolutions, are presented.

For this purpose, dedicated simulated signal process samples were produced choosing a center-of-mass energy of 14 TeV and an average of 200 pileup events, corresponding to the ultimate configuration expected for the HL-LHC accelerator; these conditions will be used throughout this document unless specified otherwise. The samples chosen range from full physics processes to single particle production with specific $p_T$ and $\eta$ intervals, which are stated in the text. Algorithm performance results are generally obtained by matching trigger objects with generator-level objects. Another essential quantity to evaluate is the algorithm trigger rate, which directly relates to its capacity to discriminate signal objects from backgrounds. Appropriate simulated background MC samples were also produced with equivalent machine conditions as the ones for signals. The rates are normalized assuming a bunch scheme with 2748 (See Table 1.1) out of 3564 LHC bunches filled. Resilience to pileup is a prerequisite feature for these algorithms, and optimization of identification and isolation variables to operate in the harsher pileup conditions is addressed. Most of the baseline algorithms presented have been implemented in firmware and their associated FPGA resource consumption and latency have been obtained; constraints on the algorithm structure imposed by the implementation are described as well.

Most of these algorithms are inspired by offline object reconstruction methods, but in some areas more advanced approaches based on machine learning techniques have been investigated. For example, identification of physics objects may rely on a set of discriminating variables for which an optimal working point can be obtained most efficiently through multivariate methods. Other pattern recognition methods may be enhanced significantly through the use of artificial neural networks. In order to assess the feasibility of these options, firmware implementations were investigated and reported here.
3.2 Triggering on electrons and photons

3.2.1 Introduction

The reconstruction and identification of electron and photon (e/γ) objects begins with the trigger primitives produced in the barrel (ECAL and HCAL) and endcap (HGCAL) calorimeters, covering the pseudorapidity region up to |η| < 3. Given the intrinsic difference of the barrel and endcap technologies and performance, the corresponding e/γ algorithms are described separately in Sections 3.2.2 and 3.2.3.

The online e/γ reconstruction aims at having high efficiency over the whole detector coverage, and a resolution as close as possible to that achieved offline for transverse energies above 10 GeV. In the current implementation of the e/γ trigger logic, the ECAL and HGCAL trigger primitives near the barrel-endcap transition region are treated independently. However, the architecture of the calorimeter trigger is designed to provide the required flexibility to combine primitives coming from different parts of the calorimeter.

In the pseudorapidity region where L1 tracks are available (|η| < 2.4), the calorimeter e/γ objects are combined with the track finder trigger primitives to build track-matched objects for the electrons, and to exploit track-based isolation. This matching stage, described in detail in Section 3.2.4, allows for a significant rate reduction with respect to the standalone calorimeter-based reconstruction.

A further extension of the pseudorapidity coverage of the e/γ triggers up to |η| < 5 requires the utilization of the hadronic forward calorimeter (HF). The high particle density, together with the limited granularity in the forward detector, pose important challenges in controlling the background rate in the high pileup environment of the HL-LHC. While this option is being explored, it is currently not considered as part of the baseline strategy described in this document.

3.2.2 Electron and photon reconstruction in the calorimeter barrel

Following the upgrade of both on-detector and off-detector electronics for the barrel calorimeters, the digitized response of every crystal of the barrel ECAL will provide energy measurements with a granularity of 0.0175 × 0.0175 in η × φ, which is 25 times higher than the input to the Phase-1 trigger, consisting of trigger towers with a granularity of 0.0875 × 0.0875. The HCAL tower size of 0.0875 × 0.0875 remains unchanged. The trigger algorithm under development should closely reproduce the algorithm used in the offline reconstruction, albeit with a number of simplifications required by trigger latency considerations.

The description of ECAL and HCAL trigger primitives is given in Sections 2.2.1 and 2.2.2. The ECAL crystals and HCAL towers comprising the trigger primitives used in the algorithm are required to have $p_T > 0.5$ GeV. Lowering this value increases the contribution from pileup, and increasing it degrades the energy resolution of the reconstructed clusters. The ECAL trigger primitives are ordered in $p_T$, and clustering starts from a seed trigger primitive with $p_T > 1$ GeV, corresponding to a local maximum. The seed $p_T$ value is chosen to minimize the contribution from pileup deposits without degrading the energy resolution. A core cluster is defined by a set of $η \times φ = 3 \times 5$ crystals around the seed crystal. The cluster position is determined as an energy weighted sum of the individual crystals within the cluster. The crystals used in the core are marked as used and excluded from the next iteration. The next cluster is built using the remaining unused highest-$p_T$ crystal as a seed, and this process continues until there are no remaining crystals with $p_T > 1$ GeV to be used as seed.
The angular distance, $\Delta R = \sqrt{\Delta \phi^2 + \Delta \eta^2}$, between the position of the simulated and reconstructed electron at the entrance of the ECAL is shown in Fig. 3.1 (left). A sample of single electron events with electron $p_T$ between 15 and 100 GeV is used. To simplify the reconstruction, and to speed up the algorithm in firmware, the position of the reconstructed electron cluster is assigned to the closest crystal position within the cluster. This has no significant effect on the precision and increases the operation speed significantly. The mean of the $\Delta R$ distribution is approximately 0.01 with a standard deviation of 0.007. This is very close to the offline resolution and allows tight requirements on the track-to-cluster matching when combining with tracks to be chosen, thus reducing the background rate. The energy of the electron cluster requires additional calibration, and this is performed using the same MC sample in bins of $p_T$ and $\eta$ of the reconstructed cluster. The calibrated $p_T$ resolution is shown in Fig. 3.1 (right).

An electron may lose significant energy to bremsstrahlung radiation. This effect leads to a spread of the electron footprint in the $\phi$-direction, which would be detected as additional energy outside of the $3 \times 5$ core region. To account for this effect, the core of the e/γ cluster in the barrel is increased by an additional $3 \times 5$ crystals above or below in $\phi$ with respect to the core region. If the energy in either of the adjacent $3 \times 5$ crystal regions in $\phi$ is more than 1/10 of the energy in the core, these regions are integrated into the energy and position of the cluster and labeled as used. The geometry of the reconstructed clusters and bremsstrahlung regions are presented in Fig. 3.2. Figure 3.3 shows the $p_T$ resolution of the electron reconstruction in the barrel, with and without the bremsstrahlung correction.

To improve the e/γ identification, and to reduce possible background contributions, identification and isolation algorithms are implemented. The identification is based on shower shape variables from the $3 \times 5$ core crystals. The electrons usually have a collimated footprint in the core since the size of the single crystal is close to the Molière radius, therefore different ratios of energy $1 \times 1/3 \times 5, 2 \times 2/3 \times 5$ were investigated for discriminating prompt leptons from misidentified jets. The best performance is achieved by using the ratio of the energies within $2 \times 5/3 \times 5$ crystal windows. The $2 \times 5$ energy is obtained by taking the largest of the $2 \times 5$ energies within the $3 \times 5$ core, as shown in Fig. 3.2. An additional photon ID uses $2 \times 2/2 \times 5$ crystals, where the maximum is taken out of four possible $2 \times 2$ combinations, and the seed crystal is always in one of the corners of the $2 \times 2$ area.

The ratio of ECAL to HCAL energies for each cluster is also stored. It was verified that using this ratio for the electron identification does not improve the ID performance after the shape
Chapter 3. Trigger algorithms

Figure 3.2: The geometry of the $e/\gamma$ clusters in the barrel with bremsstrahlung, shape, and isolation regions. Each square represents one ECAL crystal, and the core has a size of $3 \times 5$ in $\eta \times \phi$, which can be extended to recover deposits from bremsstrahlung radiation. The shape variables use the $2 \times 5$ areas around the seed cluster, while the isolation is calculated using a $27 \times 27$ area.

Figure 3.3: The $p_T$ resolution of the reconstructed $e/\gamma$ clusters in the barrel calorimeter, with and without bremsstrahlung correction. Extending the cluster core in the $\phi$ direction for events with bremsstrahlung significantly improves the $p_T$ reconstruction. It also slightly decreases the efficiency of the electron reconstruction at high $p_T$, where the number of electrons with relatively high fraction of energy measured in HCAL is increasing. This ratio will be used in future studies, including $p_T$-dependent cut, different geometry of the HCAL region, but is currently not used in the $e/\gamma$ ID.

The isolation of each cluster is calculated as a sum of the individual $p_T$ of all crystals in the $27 \times 27$ region around the seed crystal, excluding those crystals that are in the core cluster, divided by the $p_T$ of the entire cluster. The size of the region corresponds to approximately $\Delta R = 0.3$–0.4 isolation cone size in the offline reconstruction. The isolation uses only ECAL information, and since each crystal requires $p_T > 0.5$ GeV, it does not require any pileup correction. In the barrel region, the ID and isolation requirements are defined as a function of the $p_T$ of the reconstructed $e/\gamma$ cluster, such that for all $p_T$, the efficiency of the signal selection exceeds 99% for this requirement alone. Thus the efficiency with both selections applied should exceed 98% at the plateau of the efficiency curve. The corresponding $e/\gamma$ efficiencies and trigger rates for
3.2. Triggering on electrons and photons

the barrel are shown in Fig. 3.4 for the cluster only reconstruction and after the ID and isolation requirements are applied.

The ECAL trigger primitives also contain timing information that can be useful in the high pileup conditions of the HL-LHC. This information is not yet fully included in the simulation and is not discussed here, but will be utilized in the future. The timing information for each individual crystal can be used to calculate the timing of the clusters, aiming to distinguish between e/γ objects coming from the primary vertex or from pileup. It is expected that the usage of the crystal timing information will likely reduce the pileup contribution to the cluster energy.

![Figure 3.4: Left: the efficiency of reconstructing and identifying as e/γ a cluster in the barrel calorimeter as a function of the $p_T$ of the generated electron for a threshold of 25 GeV on the L1 object in single electron simulated events. Right: the trigger rate as a function of the $p_T$ threshold on the barrel L1 object at the ultimate HL-LHC luminosity.]

3.2.3 Electron and photon reconstruction in the endcap calorimeter

The reconstruction and identification of electron and photon candidates in the endcap calorimeter starts from clusters reconstructed in the backend electronics of the HGCAL calorimeter, as described in Section 2.3.2. Due to its high granularity, the HGCAL can resolve electromagnetic showers originating from single particles even in the very high occupancy environment of the HL-LHC at full luminosity. This capability represents an asset for the e/γ trigger algorithms in many different ways, by providing increased resiliency to pileup contamination, recovery of bremsstrahlung radiation for electrons, and precise matching with L1 tracks in the region covered by the L1 track-finder.

Despite the reduced readout granularity at L1, the HGCAL trigger primitive clusters carry detailed and rich information concerning the shower’s 3-dimensional development. This intrinsic feature of the calorimeter can be exploited to identify the clusters originating from electrons or photons. Several approaches to this problem are being evaluated, including advanced machine learning techniques exploiting the full detailed knowledge of cluster constituents. In the current baseline, the identification of electromagnetic-like deposits is implemented using a boosted decision tree (BDT) to distinguish signal candidates from clusters originating from pileup. Input features to the decision tree are five longitudinal and four lateral shower shape variables. The longitudinal variables include the cluster length, position of shower onset, and the energy weighted RMS of the z-coordinates of the cluster components, which is shown in Fig. 3.5.
(left). The lateral variables relate to various width parameters of the cluster in the $r$, $\eta$ and $\phi$ directions, calculated as the energy-weighted sum of widths in each layer of the HGCAL. Since observables in electron and photon identification evolve rapidly with $\eta$, two decision trees are implemented, one pertaining to a lower $\eta$ region, $1.5 < |\eta| \leq 2.7$, the other to a higher $\eta$ region, $2.7 < |\eta| \leq 3.0$. Separating the training in this way has been shown to improve background rejection in comparison to training a single discriminator inclusively in $\eta$. As the features of photon and electron showers in the HGCAL are similar, each BDT is trained only on electron clusters, simulated with 200 pileup interactions. Signal clusters are defined as those consistent with originating from a generator-level electron of $p_T > 20\text{ GeV}$ and passing a minimum $p_T$ threshold of $10\text{ GeV}$. The background sample comprises all clusters with $p_T > 20\text{ GeV}$, which are not matched to a generator level electron.

Figure 3.5: Left: energy weighted RMS of the $z$ coordinate (longitudinal dimension) of the HGCAL cluster components for signal and background. This is one input feature to the BDT used for the $e/\gamma$ ID in the endcaps showing a good discrimination power. Signal clusters are defined as those consistent with originating from a generator-level electron of $p_T > 20\text{ GeV}$, and passing a minimum $p_T$ threshold of $10\text{ GeV}$. The background sample comprises all clusters with $p_T > 20\text{ GeV}$, which are not matched to a generator level electron. Right: ROC curve showing the efficiency for generator-matched electron clusters (signal) against the fraction of rejected pileup clusters (background) evaluated at a given BDT output score.

The clusters passing this selection are promoted to calorimeter-only $e/\gamma$ candidates. Their energy is assigned using the “electromagnetic interpretation” provided with the HGCAL TPs (see Section 2.3.2). This is based on the energy deposited only in the CE-E layers of the calorimeter, with a smaller size with respect to the full cluster and with dedicated layer-by-layer calibrations. The chosen size has been optimized to get the best resolution on electrons in the $5–100\text{ GeV}$ $p_T$ range in the presence of 200 pileup events. This optimization requires the clusters to be large enough to contain as much of the electromagnetic shower as possible and small enough to exclude most of the energy deposited from soft pileup. At this stage, an average $\eta$-dependent correction to the energy is applied to subtract the residual average pileup contri-
bution to the cluster energy and equalizing the calorimeter response across the full HGCAL detector.

To recover part of the bremsstrahlung radiative loss for electrons, lower energy clusters passing the $e/\gamma$ identification selection are merged to the highest energetic one in a $\Delta\eta \times \Delta\phi = 0.02 \times 0.1$ window. The $p_T$ response for a sample of electrons with a flat $p_T$ spectrum between 15–100 GeV, with an average pileup of 200 events, is shown in Fig. 3.6, for both the pseudorapidity region inside and outside the acceptance of the L1 track-finder. The plots also illustrate the effect of the re-clustering for bremsstrahlung radiation, which allows part of the lower response tail to be recovered.

Figures 3.7 and 3.8 show the efficiency of reconstructing and identifying as an electromagnetic object a HGCAL cluster given a generated electron. The effect of the high pileup gets harsher for higher pseudorapidity values, affecting the resolution of the clusters and, as a consequence, yielding a slower efficiency turn-on curve. This effect becomes particularly important in the region beyond $|\eta| = 2.4$. The large reduction in background rate obtained applying the electromagnetic identification criteria is illustrated in Fig. 3.9. The distributions, obtained with a minimum-bias simulated sample, show the integrated rate as a function of the threshold applied on the $p_T$ of the $e/\gamma$ object.

3.2.4 Exploiting L1 track-finder primitives for $e/\gamma$ triggers

In the high pileup environment of the HL-LHC, it is essential to suppress background contributions originating from pileup collisions to maintain a viable trigger rate. In this respect, the upgraded L1T can exploit the L1 tracks as a further robust handle.

The role of tracks for the $e/\gamma$ algorithms is twofold:

- they are matched to the calorimeter deposits to help in the identification of the electron candidates;
Figure 3.7: Efficiency of finding a HGCAL cluster in the endcaps, passing the calorimeter-based electromagnetic ID criteria as a function of the matched generated electron momentum. The efficiency curves are obtained by applying a threshold of 25 GeV on the L1 reconstructed transverse momentum and refer to electrons falling within (left) and outside (right) the acceptance of the L1 track-finder.

Figure 3.8: Efficiency of finding a HGCAL cluster passing the calorimeter-based electromagnetic ID criteria and matching a generator-level electron with $p_T$ between 35–100 GeV, as a function of the pseudorapidity of the generator-level particle. The efficiency curve is obtained by applying a threshold of 25 GeV on the L1 reconstructed transverse momentum. The dashed lines indicates the barrel-endcap transition region ($|\eta| = 1.52$) and the maximum coverage of the L1 track-finder ($|\eta| = 2.4$).

- they are used to define an isolation variable for both photon and electron candidates.

The electron identification relies on matching ECAL and HGCAL clusters to prompt L1 tracks. Given a seed cluster, the electron candidate track trajectories with $p_T > 10$ GeV are propagated to the calorimeter surface, taking into account the CMS magnetic field and the longitudinal position of the track with respect to the detector origin.

The distance between the calorimeter clusters and the closest track in the $\eta - \phi$ plane, computed
3.2. Triggering on electrons and photons

Figure 3.9: Left: the integrated trigger rate as a function of the \( p_T \) threshold applied to the HGCAL clusters within the acceptance of the L1 track-finder, with and without requiring electromagnetic ID. Right: the integrated trigger rate as a function of the \( p_T \) threshold applied to the HGCAL clusters over the full endcap calorimeter acceptance.

Clustering at the calorimeter surface, is illustrated in Fig. 3.10, separately for the barrel and the endcap. We define an elliptic cut of the form \((\Delta \eta/\Delta \eta_{\text{max}})^2 + (\Delta \phi/\Delta \phi_{\text{max}})^2 < 1\) in this plane to create a cluster-track pair, where \(\Delta \eta_{\text{max}}\) and \(\Delta \phi_{\text{max}}\) are defined as:

\[
\Delta \eta_{\text{max}} = \begin{cases} 
0.025 & \text{for } |\eta| \leq 0.9 \\
0.015 & \text{for } 0.9 < |\eta| \leq 1.479 \\
0.0075 & \text{for } 1.479 < |\eta| \leq 2.4
\end{cases}
\]

\(\Delta \phi_{\text{max}} = 0.07\)

Clusters with at least one track satisfying this criterion are promoted to track-matched electron candidates.

Figure 3.10: \(\Delta \eta\) vs \(\Delta \phi\) distances between calorimeter clusters (L1EG) and the closest L1 track (L1Tk) in the \(\eta - \phi\) plane. The plots are computed for single electron events with an average pileup of 200. The tracks are requested to have \(p_T > 10 \text{ GeV}\). The elliptic cuts used to select electron candidates are shown. The left plot refers to the barrel region, while the right one refers to the endcaps.
The track-matched electron identification efficiency mostly depends on the track reconstruction efficiency illustrated in Fig. 2.5. The performance for barrel and endcap are illustrated in Figs. 3.11 and 3.12 respectively. Figure 3.13 shows the efficiency as a function of the electron pseudorapidity (left) and the corresponding trigger rate as a function of the selected \( p_T \) threshold (right). The sizable reduction in rate comes at the price of a small efficiency reduction with respect to the standalone electron identification. This rate reduction is of paramount importance for analyses relying on single electron triggers since it allows thresholds very close to those used during the LHC Run-2 to be set, as illustrated in Sections 1.4.1.1. Figure 3.14 shows the “effective RMS” (68% containment of the distribution) of the response function \( (p_L^{L1}/p_{gen}^{L1}) \) for track-matched electrons in the barrel and endcap, as a function of the generator-level \( p_T \).

Figure 3.11: Left: single electron efficiency as a function of the generated \( p_T \) for clusters passing the e/\( \gamma \) ID in the barrel, with and without requiring matching to L1 tracks. Right: trigger rate as a function of the \( p_T \) threshold for clusters passing the e/\( \gamma \) ID in the barrel, with and without requiring matching to L1 tracks.

Figure 3.12: Left: single electron efficiency as a function of the generated \( p_T \) for clusters passing the e/\( \gamma \) ID in the endcaps, with and without requiring matching to L1 tracks. Right: trigger rate as a function of the \( p_T \) threshold for clusters passing the e/\( \gamma \) ID in the endcaps, with and without requiring matching to L1 tracks.
3.2. Triggering on electrons and photons

Figure 3.13: Left: efficiency as a function of the pseudorapidity of the generated electron. Right: trigger rate as a function of the $p_T$ threshold for single-electron objects. The different curves correspond to the performance of calorimeter-only electrons (black), L1 track-matched electrons (red), and L1 track-matched electron objects satisfying additional charged isolation criteria (green).

Figure 3.14: The “effective RMS” (68% containment of the distribution) of the response function ($p_{T,1} / p_{T,\text{gen}}$) for track-matched electrons in the barrel and endcap, as a function of the generator-level $p_T$. The plot is derived from a simulated sample of single electrons with HL-LHC ultimate running conditions of 200 average pileup events.

The track-based isolation is computed using tracks with $p_T > 2$ GeV in an annulus around the calorimeter cluster. The annulus is defined by the inner and outer dimensions $\Delta R_{\text{min}} = 0.03$ and $\Delta R_{\text{max}} = 0.2$. The inner radius helps to reduce the radiation contamination from the electron, while the outer radius is determined to optimize the efficiency and background rejection. In the case of electron candidates, the tracks are required to be compatible with the vertex of the matched track, requiring $\Delta z < 0.6$ cm at the point of closest approach to the beamline. For photons, the same constraint is applied, assuming the primary vertex of the event as the origin of the $e/\gamma$ candidate. The relative isolation is defined as the ratio of the sum of the track $p_T$ within the isolation annulus and the calorimeter cluster $p_T$. The working point
is chosen based on the curves presented in Fig. 3.15. The plot shows the efficiency of signal and background as a function of the isolation variable. Selecting candidates with relative isolation smaller than 0.1 (0.125 in the endcaps) allows more than 99% of the signal events to be retained while reducing the background contribution by roughly a factor of 2. A similar optimization procedure is also applied for photons: the threshold on the relative charged isolation is set to 0.29 and 0.39 for barrel and endcap respectively, corresponding to a signal efficiency of about 95%.

The next step in the development of the isolation algorithm is to combine track and calorimeter-based isolations to consider contributions from both charged and neutral particles. In the current architecture proposal, the isolation is computed on the correlator trigger. At that stage of the processing, both PF candidates and PUPPI candidates are available, and could potentially be exploited also for the electron isolation (see Section 3.5.4.1), thus avoiding duplication of the matching procedures.

Figure 3.15: Efficiency of identifying an electron with charged isolation as a function of the cut on the relative charged isolation. The signal efficiency is computed on a single electron sample with an average pileup of 200 interactions. The background efficiency is defined as the fraction of accepted minimum bias events at the same luminosity. The two vertical dashed lines indicate the working points chosen for barrel (0.1) and endcap (0.125).

In the current implementation of the Phase-2 L1T menu, the $e/\gamma$ triggers (Section 4.1.2) are based on four kinds of objects:

- calorimeter-only electrons: standalone calorimeter $e/\gamma$ objects satisfying the calorimeter-based $e/\gamma$ ID for barrel and endcap ($|\eta| < 3$)
- track-matched electrons: track matched $e/\gamma$ objects satisfying the calorimeter-based $e/\gamma$ ID for barrel and endcap ($|\eta| < 2.4$)
- charged-isolated track-matched electrons: isolated track-matched $e/\gamma$ objects satisfying the calorimeter-based $e/\gamma$ ID for barrel and endcap and the track-based isolation cut ($|\eta| < 2.4$)
- charged-isolated photons: isolated standalone calorimeter $e/\gamma$ objects, satisfying the calorimeter-based $e/\gamma$ ID for barrel and endcap and the track-based isolation cut ($|\eta| < 2.4$)

In the rest of this Section, ongoing work is described to further optimize these selections, both
3.2. Triggering on electrons and photons

in terms of object and working point definitions.

Additional working points for the identification of track-matched electrons are under evaluation for deployment in the trigger menu.

The superior electron energy resolution of the calorimeter with respect to the L1 tracks in the presence of bremsstrahlung radiation can be exploited to define tighter matching criteria. Using the calorimeter measurement of the \( p_T \) for the track extrapolation allows for a more stringent elliptic cut in the \( \Delta \eta - \Delta \phi \) plane to be implemented, as illustrated in Fig. 3.16. While this tighter cut reduces the background rate by 10% at no expense in signal efficiency, it also allows lower thresholds on the \( p_T \) of tracks considered for matching. For instance, lowering the requirement on the track \( p_T \) from 10 to 5 GeV increases the electron efficiency by a few percent on the whole spectrum at a small cost in trigger rate, as illustrated in Fig. 3.17. In conclusion, the tighter matching window and the lower \( p_T \) cut on the track can be used to define a set of working points, optimized either for efficiency or trigger rate reduction, depending on the physics requirements of the experiment.

Figure 3.16: \( \Delta \eta \) vs \( \Delta \phi \) distances between calorimeter clusters and the closest L1 track in the \( \eta - \phi \) plane. The plots are computed for single electron events with an average pileup of 200 interactions. The tracks are required to have \( p_T > 5 \text{ GeV} \), and the e/\( \gamma \) cluster \( p_T \) is used for the extrapolation instead of the track \( p_T \). The tighter elliptical cuts used to select electron candidates are shown. The left plot refers to the barrel region, while the right one refers to the endcap region.

Furthermore, as discussed in Section 2.1, an “extended” L1 track reconstruction is explored to provide sensitivity to tracks originating from long-lived particles. Although originally developed to identify displaced objects, the extended tracks also have a strong case of application in the prompt electron reconstruction. The primary source of inefficiency in reconstructing electron tracks at L1 is induced by bremsstrahlung losses as the electrons traverse the tracker layers, which distorts their trajectories and reduces the measured \( p_T \). While electron tracks with distortions due to bremsstrahlung in the outer tracker layers are better accommodated by the higher degrees of freedom in the extended tracking algorithm, those tracks with distortions in the inner tracker layers have trajectories resembling those of displaced tracks, and thus are more easily reconstructed by the extended track producer. Figure 2.5 shows that the use of extended tracks yields up to 5% increase in absolute reconstruction efficiency for electron tracks at L1.

Finally, special working points optimized for the highest possible efficiency are being considered, such as using e/\( \gamma \) candidates without requiring any track matching but using the L1 tracks only for the isolation computation. These objects provide a significant rate reduction
with respect to the calorimeter-only ID criteria at almost no cost for the efficiency, as shown in Fig. 3.18 for the barrel.

Figure 3.17: Left: the $p_T$ dependence of the efficiency for track-matched electrons built using different matching strategies in the barrel. Right: the trigger rates of the track-matched electrons built using different matching strategies in the barrel. The green curves refer to the matching performed using the $p_T$ of the calorimeter object for the extrapolation and a looser 5 GeV cut on the L1 track momentum. The blue points refer to the matching performed using the track $p_T$ and requiring $p_T > 10$ GeV on the L1 tracks.

Figure 3.18: Left: efficiency for different working points in the barrel region with respect to generator-level $p_T$. Right: corresponding trigger rates. Four working points are illustrated: calorimeter-only $e/\gamma$ cluster (red), $e/\gamma$ cluster with charged isolation (green), $e/\gamma$ cluster matched to track (pink), and $e/\gamma$ matched to track and with charged isolation (brown).

### 3.2.5 Firmware implementation and resources

The firmware implementation of the $e/\gamma$ algorithm for the calorimeter barrel is discussed in Section 6.4.1. The ECAL crystal clustering is performed in the regional calorimeter trigger (RCT) boards, and it has been demonstrated to satisfy the required latency and resource requirements.

In the endcaps, the clustering is part of the HGCAL trigger primitive generation step, and runs on the sub-detector electronics backend boards, as described in Section 2.3.2. The BDT model used for identifying the $e/\gamma$ clusters has been implemented in firmware using the hls4ml library [44]. The resources needed for running the inference have been estimated assuming the VU9P FPGA chip as a reference. For a single instance, using a fully pipelined implementation,
they amount to 0.41% of LUTs, and 0.10% of flip flops, with a latency of 8 clock cycles, corresponding to 33 ns for a 240 MHz clock. Given these figures, the decision of where to fit this step in the overall trigger architecture needs to be driven mostly by considerations of bandwidth requirements for sending the shower-shape data from the HGCAL backend electronics to the calorimeter trigger, and about future flexibility in accessing other low-level features of the HGCAL trigger primitive clusters without modifying their data format.

In the architecture of the Phase-2 trigger, the natural place to perform the matching of calorimeter clusters with L1 tracks for the identification of track-matched electrons is the correlator trigger. This step is conceptually and algorithmically very close to what is done for building the PF candidates. The PF algorithms are described in detail in Section 3.5. Although a complete implementation of the electron matching algorithm in firmware is not yet available, the PF demonstrator can be used as an approximate evaluation of the needed resources. The correlator trigger processing of the input primitives is organized in regions: the approach where 27 regions are used for covering the barrel and 18 for the endcap within the L1 track-finder acceptance (see Sec. 3.5.1) is assumed in the following considerations. Simulated $t\bar{t}$ and Drell-Yan events with HL-LHC ultimate running conditions of 200 average pileup events have been used to benchmark the occupancy in the various correlator trigger regions. To prevent truncation in at least 95% of these events, the algorithms have to deal with a maximum of 2 clusters above 10 GeV and 5 tracks passing the quality and $p_T$ requirements for the matching. The maximum number of tracks above 2 GeV to be considered for the isolation is 23. Given these maximum occupancies per event, and assuming that resources scale as the product of the maximum numbers of track and clusters (see Section 3.5.5), the track-electron matching is expected to require approximately 3–5% of the resources of the correlator FPGA chip.
Chapter 3. Trigger algorithms

3.3 Triggering on muons

3.3.1 Introduction

The success of the Phase-2 CMS physics program relies significantly on the capacity to trigger efficiently on muons originating from many different physics processes. A number of challenges have to be faced, including the L1 reconstruction of new types of muon trigger objects defined by the properties of the muons produced in these processes; the ability to resiliently reconstruct those muons in the extreme pileup (PU) and detector aging conditions of the HL-LHC; and the mitigation of the expected very high trigger rates. To overcome these challenges CMS is expanding significantly its muon trigger capabilities through several upgrades that include additional muon detectors, for extended and more robust coverage, and improved back-end electronics. The latter provide more computational resources that allow for more accurate timing and more precise trigger primitives (TPs) to be produced. These will be used downstream in fast and efficient tracking algorithms that reconstruct the required muon trigger objects and provide substantially improved momentum and spatial resolution. The Level-1 (L1) muon momentum can be measured to be higher than the true one due to limited resolution, thus increasing the contribution to the trigger rate. The improvement of the L1 momentum resolution thus helps mitigate the challenge of high rates, allows for better alignment between the L1 and offline thresholds to be performed, and gives access to lower thresholds required by the physics program.

Very importantly, the CMS upgrades will also include tracks reconstructed at L1 from the new outer tracker detector that can be combined with other trigger objects, including the muons [3]. The use of the tracker for the L1 muon reconstruction can improve its $p_T$ resolution by an order of magnitude at low and moderate $p_T$, as is achieved in the offline reconstruction. An efficient combination of L1 tracker tracks and muon tracks will benefit from the improved spatial resolution of the L1 muons. The deployment of the track finder will extend the CMS reach to include signatures with soft muons, including low-threshold multi-muon triggers and track-based isolation requirements.

A number of new physics scenarios predict final states that contain muons originating from decays of sufficiently long-lived particles. Because of a significant impact parameter and/or the absence of stubs in the inner tracker layers, such muons may not be reconstructed by the L1 track finder algorithm. Maintaining standalone muon triggers, even with somewhat higher thresholds, will therefore remain important both for retaining physics sensitivity to displaced signatures, as well as providing an important calibration for the prompt muon triggers based on the matching of L1 tracker track and muons.

As mentioned above, the CMS physics program requires triggering on a variety of muon trigger types originating from many different physics processes. A summary of the required types is listed below.

- **Standalone prompt muons.** These are the type used the most during Phase-1. Important for the Phase-2 program is the extension of this type into more forward directions to include very forward signatures involving muons. The required extension will be achieved by the addition of new muon detectors in forward directions, as discussed briefly in what follows.

- **Standalone displaced muons.** Many well-motivated theories predict the existence of long-lived particles (LLPs) that can decay into final states including muons at distances up to several meters away from the interaction point, leading to distinct signatures involving one or more muons originating from at least one significantly
displaced vertex.

- **Track-correlated muons**, i.e. L1 muon tracks or segments matched to L1 tracker tracks. Benefiting from the integration of the tracker into the Phase-2 L1 trigger system, this entirely new muon type can be introduced. The new type aims to improve the L1 muon efficiency and resolution bringing it close to the level of the offline reconstruction provided by the tracker, as required by the Phase-2 physics program. The tracker will perform the actual L1 track reconstruction for this type, while the muon system will serve only for the L1 muon identification. This muon type will be produced in the Global Muon Trigger (GMT) unit of the Phase-2 L1 architecture, which is interfaced with the tracker finder, as described in Section 5.3.

- **Track-isolated muons**. These form a subset of the previous type that can be directly identified by the desired isolation requirement in the implementation of L1 tracker track to muon correlation. They are among the most commonly used probes of prompt heavy particle decays.

- **Slow “muons”**. Heavy charged LLPs with very long lifetimes, called heavy stable charged particles (HSCPs), will traverse the detector resembling a muon, except that their speed will be significantly smaller than the speed of light. Therefore, it may be possible to distinguish them from muons through their time-of-flight [8].

- **Topological muon triggers**. These involve various types of muons produced by special decays which are difficult to reconstruct at L1, such as very soft muons or muons in a dense environment. An example is the exotic (lepton flavour violating) decay $\tau \rightarrow 3\mu$ [45], for which one needs both high efficiency to identify all three muons and high resolution to reconstruct accurately the $\tau$ mass. Some types of topological muons can possibly be reconstructed with L1 global trigger algorithms (see Chapter 4).

All the above L1 muon types are new with respect to the Phase-1 L1 muon reconstruction, except for the standalone prompt muons. The objectives of the Phase-2 L1 muon algorithms, listed below, follow from the requirements to define the new types and to meet the high quality criteria for optimal performance of the new L1 muon trigger system:

- Provide the ability to trigger on each L1 muon type, making use of the improved trigger primitive quality through multi-chamber and multi-system algorithms developed by the Muon Upgrade Group.

- Maintain fast trigger decisions (within about 5 $\mu$s) and high L1 trigger efficiency ($> 95\%$) at reasonable rates, including $p_T$ thresholds that are as low as possible where needed.

- Improve momentum, spatial, and timing resolutions and ensure robustness of the L1 muon trigger system against noise, aging, dead detector regions, etc.

- Optimize performance vs. PU with safety margins, for example up to about 300 PU.

- Develop firmware to implement the L1 muon algorithms in platforms consistent with the overall Phase-2 L1 architecture and demonstrate the performance and compliance of the algorithms with the latency constraints and the resource specifications of those platforms.

The CMS muon detection system is illustrated in Fig. 3.19, including the Phase-2 upgrades. The main design of the system is essentially unchanged from Phase-1 to Phase-2, except for new layers of GEM and iRPC, which strengthen the detector in the forward directions. The system consists of drift tubes (DT), with $\sim 100\mu$m spatial resolution, resistive plate chambers (RPC),
with $\sim 1 \text{ cm } r - \phi$ resolution and $\sim 2 \text{ ns}$ timing resolution, and cathode strip chambers (CSC), with $\sim 75 - 150 \mu\text{m } r - \phi$ resolution. In the barrel region, fully covering the pseudorapidity range $|\eta| < 0.83$, the detectors are placed parallel to the CMS axis and consist of four DT+RPC layers arranged radially outwards in four stations MB/MB1,2,3,4 and longitudinally in five wheels labeled with $-2,-1,0,1,2$, with the sign corresponding to the sign of $\eta$. The four stations are separated by four layers of steel of the magnet yoke (see Fig. 3.19). In the region of the two endcaps, fully covering the range $1.24 < |\eta| < 2.4$, the detectors are placed perpendicular to the CMS axis and comprise four CSC+RPC layers arranged longitudinally in four stations ME/ME1,2,3,4 on each side and separated again by four layers of steel, as shown in Fig. 3.19. In Phase-2, the endcap muon detectors will be upgraded with three GEM chambers ME0, GE1/1, GE2/1 and two improved RPC (iRPC) chambers RE3/1 and RE4/1 on each side (see Fig. 3.19) for improved resolution in forward directions. The ME0 chamber, combined with the silicon tracker, will also extend the pseudorapidity range up to $|\eta| = 2.8$. The overlap region, defined by the pseudorapidity range of $0.83 < |\eta| < 1.24$, is covered partly by the barrel and partly by the endcap detectors.

The design of the CMS muon detector system naturally leads to consider three distinct $\eta$ regions, barrel, overlap, and endcap, featuring different detector technologies and geometries and thus posing different challenges to L1 muon reconstruction. Additional challenges to the detector design arise from the different profile of the magnetic field, turning smoothly from nearly uniform in the barrel to highly nonuniform in the endcaps, and from the particle occupancy, which increases rapidly in going from the barrel to the endcaps. Based on the experience of successful L1 muon tracking in the Phase-1 Upgrade, the same regional approach of the muon track finding is retained in Phase-2. Three baseline muon track finders are considered in the three $\eta$ ranges of the barrel, overlap, and endcap detector regions, aiming both to improve standalone prompt muon track finding relative to the Phase-1 track finders and to provide the new muon types required in Phase-2. Optimal algorithms are developed in each of the three regions. These developments do not preclude some future consolidation, and even newer algorithms to be developed, but given the Phase-2 challenges the diversity of algorithms is a strength at the present stage.
3.3. Triggering on muons

A new design of the L1 trigger architecture is developed for Phase-2, which is discussed in detail in Chapter 5. The muon trigger architecture is described in Sections 5.3 for the standalone muons and 5.5 for the track-matched muons. The design is optimized for synchronization of the subsystems and latency compatible with the HL-LHC beam structure, considering appropriate time-multiplexing where needed. In the new design, a global muon trigger (GMT) layer is devised, where the L1 track-muon matching algorithms are implemented (see Fig. 1.3). The outputs from the L1 track finder and from the muon track finders are received in the input of the GMT. The tracker and muon tracks are spatially matched to provide the correlated muons in the GMT output, which are then sent to the correlator trigger, to be integrated into the L1 particle-flow reconstruction (see Section 3.5), and to the Global Trigger for the final L1 trigger decision (see Section 3.7).

The L1 standalone muon and track finder to muon correlation algorithms are designed coherently with the Phase-2 L1 hardware specifications for optimal performance and cost (see Chapters 6 and 7). To verify the overall performance of the baseline algorithms and their firmware implementation, detailed demonstration tests (demonstrators) are foreseen at the mature stage of the development. The planning for these demonstrators is described in Section 6.4.

The Phase-1 L1 muon trigger system is described in detail in Ref. [46]. Its upgrade, in effect since 2016, is described in Ref. [47]. The performance of the Phase-2 L1 muon algorithms presented here is tested using Phase-1 TPs, described in detail in Ref. [47], in the barrel and overlap regions and using preliminary Phase-2 TPs in the endcap region, in order to accommodate the new Phase-2 muon detectors. The Phase-2 muon TPs are discussed in detail in Section 2.4 for the barrel and Section 2.5 for the endcaps. Performance tests using Phase-2 TPs in the barrel and in the overlap regions are envisioned in the near future. In both Phase-1 and Phase-2 TPs, the basic element for the reconstruction of an L1 muon track is the muon “stub”. Its usage is explained below in the description of the various standalone muon reconstruction algorithms. Depending on the station, the stub is extracted at every single muon station and contains local information on the position, bending angle, and timing of the track.

### 3.3.2 Standalone muon reconstruction in the barrel region

In the barrel region, the prompt and displaced standalone muons are provided by the same algorithm and thus discussed together.

In Phase-1, the BMTF algorithm [48] is responsible for the identification of tracks created by muons and for the measurement of their $p_T$ with a precision of approximately 10% in the barrel region. The TwinMux processor [49] sends detector data in the form of muon stubs, segmented in 12 $\phi$ sectors and in 5 $\theta$ wheels. By combining information from DTs and RPCs (super-primitives), each muon stub provides 22 bits of coordinate information in the bending $r - \phi$ plane (azimuthal angle $\phi$ and bending angle $\phi_b$), 7 bits of information in the longitudinal $r - \theta$ plane, and quality bits. The algorithm operates in three steps. First, the barrel stubs are matched in pairs and extrapolated to the interaction point. The extrapolation uses super-primitive $\phi$ and charge information to form an acceptance window for the next station. A track assembly step follows, where the stub pairs are combined to form a track. In the final assignment step look-up tables (LUTs) are used, implemented with Block RAM (BRAM), to assign parameters to the tracks. Due to the limited address space provided by BRAM, the momentum assignment is performed using information from only two stations. In addition, the LUTs are filled assuming that the track originates from the center of the detector. This beamspot constraint improves the momentum resolution, since it effectively adds one more point at the CMS center, exploiting the full bending power of the CMS solenoid.
In Phase-2, the DT backend electronics will be upgraded, thus providing better position and time resolution (see Section 2.4). This allows possible improvements in track reconstruction to be studied. The first goal is to improve momentum resolution by taking advantage of the better DT position resolution and by including information from more than two stations in the L1 muon reconstruction. This goal cannot be achieved by the LUT approach since the address space required for four stations is 88 bits with the current precision. The precision is expected to increase even more in Phase-2. The second goal is to implement momentum assignment without the beamspot constraint. This is motivated by searches for displaced particles. Since a displaced particle does not originate from the beamspot, the beamspot constraint results in mismeasured $p_T$ and inefficiency of the trigger for such physics signals. A better time resolution will help assign a significantly displaced muon to the correct bunch crossing (BX). To achieve both goals, a new tracking paradigm is implemented exploiting a Kalman Filter algorithm [50].

A Kalman filter for the muon track finder

Kalman Filtering is a technique used very widely in track reconstruction at hadron colliders. A Kalman Filter has been implemented for the CMS Phase-2 Track Finder [8] with excellent performance, but with a long latency, which is not suitable for the muon trigger. For this reason, a special KF implementation has been devised [51], tuned for the barrel muon trigger and optimized for low latency. The basic steps of this special implementation are described here. The full description of the implementation of a Kalman Filter for tracking can be found in Ref. [50].

The track parameters at each detector station are given by the state vector $x_n = (k, \phi, \phi_b)$, where $k = q/p_T$ is (proportional to) the signed curvature of the track, $q = \pm 1$ is the muon charge, $\phi$ is the azimuth angle of the track in the spherical coordinate system of the detector, and $\phi_b$ is the bending angle in the muon detector azimuth plane due to the motion of the muon in the magnetic field. A track is seeded by a muon stub in the outer available station, and the track parameters and their uncertainties are propagated inwards. When neglecting the energy loss in the muon system, the new state and the old state are related as follows:

$$x_{n+1} = Fx_n \quad \text{or} \quad \begin{pmatrix} k \\ \phi \\ \phi_b \end{pmatrix}_{n+1} = \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & b \\ c & 0 & 1 - b \end{pmatrix} \begin{pmatrix} k \\ \phi \\ \phi_b \end{pmatrix}_n , \quad (3.1)$$

where the propagation matrix $F$ containing the parameters $(a, b, c)$ is defined by the detector geometry and simulation. The state uncertainties expressed by a $3 \times 3$ covariance matrix $P$ are also propagated to the next station by the transformation $P_{n+1} = FP_nF^T + Q$, where $Q$ is an additional covariance matrix arising from multiple scattering in the return yoke between stations. After propagating the state to the new station, the closest stub is selected and the track parameters and their uncertainties are propagated inwards. When neglecting the energy loss in the muon system, the new state and the old state are related as follows:

$$x_{n}^{\text{upd}} = x_n + H^T Gr_n \quad \text{and} \quad P_{n}^{\text{upd}} = P_n - H^T GHP_n , \quad (3.2)$$
3.3. Triggering on muons

where \( r_n \) is the residual between the propagated state and the measurement:

\[
r_n = z_n - Hx_n = \begin{pmatrix} \phi_s^\text{stub} \\ \phi_b^\text{stub} \end{pmatrix}_n - \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}_n \begin{pmatrix} k \\ \phi \\ \phi_b \end{pmatrix}_n
\]  

(3.3)

and the Kalman gain

\[
G = HPH^T (HPH^T + R)^{-1}
\]

(3.4)

acts as the “ratio” between the prediction and the measurement. The 2 \( \times \) 3 matrix \( H \) reduces the dimension of the track parameter space to the dimension of the measurement space. The iterative process of station-to-station propagation and update is schematically illustrated in Fig. 3.20.

Figure 3.20: Sketch of a muon track traversing the full CMS detector. The Kalman Filter-based track finder is illustrated, starting from the outermost muon station, propagating inwards and updating the track parameters at each station. At the innermost station it provides a vertex-constrained and an unconstrained measurement, allowing for triggering on prompt and displaced muons.

To reduce the number of computations, the values of the Kalman gain for different tracks in simulation were studied and it was found that the gain matrix can be precalculated for different values of curvature (implying different multiple scattering) and different combinations of the stubs used at each station (implying different uncertainty of the already reconstructed track at each station). Therefore, an approximate Kalman Filter was implemented, propagating the track from station to station, and updating it using a precalculated Kalman gain that depends on a 4-bit stub pattern of the already used stubs and the value of the curvature at the given point of the track.

After reaching the first station, a measurement without the beamspot constraint is stored and then the track is propagated to the center of CMS (see Fig. 3.20). While the energy loss was neglected for the station-to-station propagation, for the propagation to the vertex it is taken into account since the material budget of the calorimeters, the magnet and the tracker is significant. After the vertex propagation, the impact parameter of the track to the beamspot is estimated and a Kalman update is performed, requiring that the track should pass through the origin. This final update provides a beamspot constrained measurement. This approximate KF algorithm, named henceforth KBMTF, exploits the measurements in all detector stations, provides a vertex unconstrained measurement for displaced muons, and implements a vertex constrained...
measurement for muons originating from the beamspot, satisfying all requirements for the improved muon trigger.

**Firmware implementation**

The algorithm consists of two major operations to be implemented in firmware: propagation and update. The track propagation is implemented as in Eq. 3.1. This matrix operation consists of many multiplications that are relatively expensive in firmware. The resource utilization is substantially reduced by exploiting the digital signal processor (DSP) slices present in modern FPGAs. Therefore all matrix operations are mapped to DSP, minimizing resource utilization.

A well defined track requires at least two stubs, therefore there are 11 possible tracks with two, three, or four stubs that can be reconstructed. All those combinations are implemented in parallel. Each track update in those 11 track chains corresponds to a different Kalman gain to be precalculated, which is mapped into one BRAM. Then, the state update in Eq. 3.2 is also implemented in DSPs. Given the maximum multiplicity of two stubs per station, the logic is duplicated to create 22 tracks for each detector sector. Subsequently, ghost cleaning and sorting are performed before forwarding the tracks to the GMT.

The firmware is implemented using Vivado high level synthesis (HLS). Vivado HLS compiles C code to HDL by optimizing the pipeline for the specific chip and the required clock frequency and by mapping calculations to DSP cores. As the clock frequency is increased, the latency is reduced as expected. However, for very high frequencies, the timing constraints require too many steps in the pipeline, resulting in increased latency. Therefore, there is an optimal clock frequency for each design and chip. For the KBMTF implementation integrated in the 2018 CMS data-taking, the optimal clock frequency was 200 MHz, however 160 MHz was used, in order to be compatible with the current BMTF algorithm.

![Figure 3.21: Schematic layout of the Phase-2 barrel muon track finder firmware.](image)

In Phase-2, the barrel muon trigger (BMT) architecture foresees two Layers (see Section 5.3): Layer-1 will receive DT+RPC data, segmented in 12 φ wedges and in 5 θ wheels (60 sectors in total), and create stubs with BX assignment. Based on the number of required links, Layer-1 can only receive data from at most one sector out of the 60. Therefore, no tracking is possible at Layer-1, only stub construction and BX assignment. Layer-1 needs 60 boards to receive data from all 60 sectors. The GMT will receive muon stubs from the 60 Layer-1 output fibers, and L1 tracks from two track finder output fibers per φ nonant (using a time-multiplexing period of 18), and will perform both standalone track finding and track matching. The deployment of the time-multiplexed track finding is based on the following three steps:

- A KBMTF central sector module (one instance of the algorithm) receives input from one sector and the five adjacent inner (towards the central wheel) sectors, and returns two tracks output.
- In the Phase-1 implementation, a KBMTF central wedge module (five instances of
3.3. Triggering on muons

the algorithm running in parallel) receives input from one wedge and two adjacent wedges, applies the KBMTF sector module to each central sector, finds 10 tracks, and cleans neighboring tracks to return the top 3 after bitonic $p_T$ sorting.

- In the Phase-2 implementation (see Fig. 3.21), the TMUX KBMTF payload receives input from all 12 wedges, uses the KBMTF sector module 60 times (instantiated once) and performs KBMTF wedge cleaning to find 3 tracks per wedge, and then the "ghost" cleaner removes overlap tracks between adjacent wedges, and the sorter returns the top 8-12 tracks bitonically sorted in $p_T$.

Studies with a ZYNQ Ultrascale+ (ZU19EG-2) FPGA show reasonable total resource utilization, with 17% used by LUT, 3% by FF, 11% by DSP, and 7% by BRAM. More details on the firmware implementation of the algorithm with the foreseen Phase-2 instrumentation are given in Section 6.5.9.

Integration in the 2018 CMS data-taking

The firmware was integrated in the CMS trigger system at the end of the 2018 data-taking period, running in parallel with the Run-2 BMTF algorithm. Both algorithms were implemented into the same XILINX Virtex 7-690T FPGA and both were taking the same data. The BMTF was used for trigger, while the KBMTF algorithm was read out in the DAQ for each collected event. This implementation exploited real data to study the algorithm with a plan to deploy it online as the default track finder in Run-3.

The performance of the algorithm is studied using CMS data. Data were collected by the BMTF and then the output of the KBMTF firmware was compared to the KBMTF emulator, demonstrating agreement better than 99% [51]. The total latency of the KBMTF algorithm is 9.5 bunch crossings, which fits within the latency budget of the current trigger. Therefore, the algorithm is already commissioned for Run-3. Details on the commissioning of the KBMTF muon track finder in 2018 can be found in Section 6.4.

Performance for Phase-2

Figure 3.22 shows the efficiency of the KBMTF algorithm as a function of the generated $p_T$. The efficiency is calculated from a high-PU Phase-2 muon sample with uniform $p_T$ and $\eta$ distributions in the ranges [2,100] GeV and [-3,3], respectively. The plot on the left shows the efficiency for prompt muons with vertex-constrained $L1 \ p_T > 20$ GeV compared to the corresponding efficiency of the Run-2 BMTF algorithm. The two algorithms have about the same efficiency. The plot on the right shows the KBMTF efficiency for various $L1 \ p_T$ thresholds. The algorithm maintains the same high efficiency down to very low thresholds, as required by the Phase-2 CMS physics program.

Figure 3.23 shows the efficiency of the KBMTF algorithm as a function of the generated $\eta$ compared to the efficiency of the Run-2 BMTF algorithm. The plot on the left shows the efficiency for muons with $L1 \ p_T > 20$ GeV (generated $p_T > 22$ GeV), and the one on the right for muons with $L1 \ p_T > 5$ GeV (generated $p_T > 7$ GeV). The efficiency of the two algorithms is about the same for muons with $L1 \ p_T > 20$ GeV and slightly better for KBMTF when including muons with $L1 \ p_T > 5$ GeV. The drops of the efficiency in the intervals of $0.2 < |\eta| < 0.3$ originate from the gaps between wheels 0 and $\pm1$ (see Fig. 3.19).

Figure 3.24 on the left shows the efficiency of the KBMTF algorithm for prompt (vertex-constrained) and displaced (unconstrained) muons compared with the BMTF efficiency as a func-
Figure 3.22: Left: Trigger efficiency for muons with L1 $p_T > 20$ GeV as a function of the generator-level $p_T$. The efficiency for muons reconstructed with the Phase-1 algorithm is also shown for comparison. Right: Trigger efficiency for muons reconstructed with various L1 $p_T$ thresholds.

Figure 3.23: Trigger efficiency for muons as a function of the generator-level $\eta$. Left: Efficiency with L1 $p_T > 20$ GeV. Right: Efficiency with L1 $p_T > 5$ GeV. The efficiency for muons reconstructed with the Phase-1 algorithm is also shown for comparison.

The maximum of the vertex-constrained and unconstrained L1 $p_T$ values is assigned to the displaced muons, whereas simply the vertex-constrained L1 $p_T$ is assigned to the prompt muons. The efficiency is calculated from a high pileup Phase-2 muon sample with uniform $p_T$, $\eta$, and $d_{xy}$ distributions in the ranges $[2,100]$ GeV, $[-3,3]$, and $[0,3]$ m, respectively. The KBMTF algorithm improves the efficiency by an order of magnitude for a track displacement of 80 cm from the vertex. The plot on the right shows the KBMTF efficiency as a function of the generated muon $p_T$, calculated from the same simulated sample of displaced muons for various intervals of the impact parameter. In this sample, the muon is not produced by another particle in a physical process and thus it is assigned the time of the primary vertex. This leads eventually to wrong BX assignment for large impact parameter.
Therefore, raising the impact parameter threshold results in the efficiency drop manifested in Fig. 3.24. Further studies with more realistic samples are planned for the future.

Figure 3.24: Trigger efficiency for displaced muons. Left: Efficiency as a function of the impact parameter. Right: Efficiency as a function of the generated muon $p_T$ in various intervals of the impact parameter.

Figure 3.25 on the left compares the L1 muon rates for the KBMTF and BMTF (Phase-1) algorithms as a function of the L1 $p_T$ threshold on the left and as a function of the average pileup on the right. The rates are measured in a high-PU minimum bias sample. The rates of the prompt KBMTF and BMTF triggers are similar. For a threshold of 20 GeV, the KBMTF rate is 10 kHz, while the BMTF rate is 15 kHz. The PU dependence of the KBMTF rate is shown on the right plot. The rate scales approximately linearly with PU, as expected.

Figure 3.25: Left: Trigger rate as a function of the L1 $p_T$ threshold for prompt and displaced muons. The rate for prompt muons reconstructed with the Phase-1 algorithm is also shown for comparison. Right: Trigger rate for prompt muons as a function of pileup.

Summary and outlook
Chapter 3. Trigger algorithms

The studies done so far have provided baseline algorithms for the Phase-2 standalone prompt and displaced L1 muons in the barrel region. The performance tests of the algorithm show that they fulfill the expectations of the Phase-2 physics program. Ongoing studies include the integration of Phase-2 TPs in the algorithm, tests of the resilience against detector aging, performance studies with TeV muons, and improvement of the displaced muon efficiency for large displacements. The results of these studies are expected to have no significant impact on the design of the Phase-2 L1 trigger architecture.

3.3.3 Standalone muon reconstruction in the overlap region

The overlap region covers $0.83 < |\eta| < 1.24$ and contains DT, RPC, and CSC detectors that are arranged both horizontally and vertically, and that experience a weak and very nonuniform magnetic field. The muon reconstruction algorithm used in Phase-1 and those proposed for Phase-2 are all based on a Bayesian probability method that uses muon hit information from the (up to 18) different stations that comprise this region.

Prompt muons

The algorithm used in Phase-1 to reconstruct standalone muons in the overlap region (OMTF) was based on a “Naive Bayes” classifier [52]. The algorithm identifies the muon track candidates and estimates their most likely transverse momenta by calculating the probabilities of matching the detector stubs found to different transverse momentum hypotheses. The same OMTF algorithm, with some small modifications, is proposed for Phase-2. In what follows, we discuss the algorithm in more detail and demonstrate its performance for Phase-2.

The magnetic field produced by the CMS solenoid, even if inhomogeneous in the overlap region, bends charged particles in the $r - \phi$ plane (i.e. perpendicular to the beam axis), so a precise measurement of the muon stub $\phi$ position data allows the $p_T$ to be determined. By examining the $\phi$ positions of stubs in the various detector layers, the OMTF algorithm performs both the muon candidate track identification and its transverse momentum measurement in essentially one step. The muon $p_T$ spectrum is divided into 52 bins (or classes) $k$, 26 for each charge. For a given set of stubs that comprise a track candidate, the probability for all classes is tested, and the track candidates are assigned to the $p_T^k$ class that is the most likely. The conditional Bayesian probability that a muon has a given $p_T^k$ given a set of $\phi_i$ stubs in $L$ detector layers is given by:

$$P(p_T^k | \phi_1, ..., \phi_L) = \frac{P(\phi_1, ..., \phi_L | p_T^k) \cdot P(p_T^k)}{P(\phi_1, ..., \phi_L)} \quad (3.5)$$

Since the denominator in the above equation is the same for all $p_T^k$ classes (if the same stubs are used), it is not relevant when selecting the $P(p_T^k | \phi_1, ..., \phi_L)$ with the largest probability and can be dropped. If the $\phi$ positions of the stubs were not correlated, the following trivial relation would hold: $P(p_T^k | \phi_1, ..., \phi_L) = \prod_{i=1}^{L} P(\phi_i | p_T^k)$. However the $\phi_i$ positions are correlated, and in order to weaken that correlation the algorithm uses instead $\phi_i^{\text{dist}}$, defined as $\phi_i^{\text{dist}} = \phi_i - \phi^{\text{ref}} - \Delta \phi_i^{\text{mean}}(p_T^k)$, where $\phi^{\text{ref}}$ is the position of a reference stub as illustrated in Fig. 3.26.

The $\Delta \phi_i^{\text{mean}}(p_T^k)$ term is the average $\phi$ distance for a given $p_T^k$ in layer $i$, and by adding this term the new $P(\phi_i^{\text{dist}} | p_T^k)$ is centered at 0 and its width is reduced — this reduces also the number of bits used in the LUT in the firmware implementation of the algorithm (see the end of this Section for more details).
3.3. Triggering on muons

The algorithm has additional features as follows. If more than one stub is found in a layer in a given pattern (due to noise, photon emission by a muon, or other effects) the algorithm selects the stub with the smallest $\phi_i^{\text{dist}}$. If no stubs are found in a layer (due to inefficiency or geometrical acceptance) the pattern probability is calculated only for layers with detector stubs; patterns with less than three layers with stubs are rejected, reducing the rate of fake triggers. Ultimately, taking also these cases into account, the pattern with the highest total probability is selected. The sum of logarithms of probability distribution functions (PDF) over all layers, is computed, omitting contributions from $\phi_i^{\text{dist}}$ outside a selected range. For each event, the track finding procedure is executed four times, with a different reference stub in each iteration (the reference stubs are chosen from 8 selected layers according to a predefined priority). This ensures that more than one muon can be reconstructed in the detector region handled by one processor. To remove duplicate muon candidates that may result from these iterations, the reconstructed $\phi$ coordinate of candidate pairs is examined, and if the difference in $\phi$ is within a predetermined window, the candidate with a lower number of stubs is removed.

Towards Phase-2 several minor modifications have been made to the algorithm, related to how stubs are selected in the DTs, what cluster size is allowed in the RPCs, and whether RPC stubs are considered at all in the presence of a DT stub in certain layers. In addition, the internal $p_T$ scale around the typical L1 threshold will be re-optimized. For Phase-2 the algorithm will benefit from the expected improvements in DT and CSC TP reconstruction. Another improvement to consider is to include the PDFs dependence on $\eta$ and to study the $\phi_i^{\text{dist}}$ layer-to-layer correlations. In addition, in case of significant aging of the DTs (especially in MB1, wheel±2) together with possible significant punch-through, further improvements to segment reconstruction could be implemented.

Figure 3.27 shows, for events with single in events with 200 average pileup, the Phase-2 OMTF trigger efficiency as a function of the generated muon $p_T$ for various $p_T$ thresholds, and as a function of $\eta$ (for generated muons with $p_T \geq 25$ GeV and with $7 \leq p_T \leq 15$ GeV).

Figure 3.28 shows the trigger rate in the overlap region as a function of $p_T$ threshold and the linearity of the rate as a function of pileup for $p_T$ thresholds of 20 and 25 GeV (for which the rate for two pileup points from Run-2 is shown), respectively. The rate is linear with pileup and the simulation matches the data very well for runs with average pileup of 50 and 110.

The OMTF algorithm is susceptible to the expected aging of the DT detectors, as described in 2.4.1. Figure 3.29 shows the effect on the efficiency and rates after a total integrated luminosity exposure of 1,000 and 3,000 fb$^{-1}$. For the worst case scenario, the overall efficiency drops about 5% for all $p_T$ and is dominated by the loss of efficiency in the MB1 station in the $0.8 < |\eta| < 1.2$ region. The rate is also affected, showing a small decrease due to the loss in efficiency.

Displaced muons

No dedicated algorithm for displaced muons in the overlap region has been finalized at the time of this writing. However there are ongoing studies that use the Bayesian approach. One solution is to use the existing patterns for prompt muons but to interpret the stubs as possibly coming from muons of different displacements. A rotation function, that transforms stubs assuming a given displacement into the prompt patterns can then be applied, and the most likely rotation would correspond to the most likely displacement. Another way to address displaced muons is to directly derive a separate set of patterns for different displacements and $p_T$ and obtain the most likely pattern for a given set of stubs. Finally, one could also use the method for displaced muons used in the endcap region, described below, where fiducial
Chapter 3. Trigger algorithms

Figure 3.26: OMTF algorithm $\phi$ coordinate definitions. The red stars represent stubs from the passage of muons in three detector layers. The $\phi$ positions of the stub in layer-1 is used as a reference. See the text for more details.

Figure 3.27: Left: Efficiency of the Phase-2 OMTF algorithm for a sample with an average of 200 pileup events as a function of muon $p_T$, for different L1 $p_T$ thresholds. Right: Efficiency as a function of $\eta$ for muons with generated $p_T$ above 25 GeV and within the range $7 < p_T < 15$ GeV.

patterns for displaced muons are obtained followed by an artificial neural network (NN) that outputs the preferred $p_T$ and displacement in one single step.

Firmware considerations

The Phase-1 algorithm has been running successfully with Phase-1 hardware in CMS during Run-2. Below are some details on the firmware implementation of this algorithm which runs on Xilinx Virtex-7 FPGA chips that reside on the MP7 $\mu$TCA boards. As described in Section 5.3, for Phase-2 larger FPGAs and LUTs will be available in the new ATCA boards and even better performance than has been demonstrated here can realistically be expected. For example, more finely grained patterns could be derived, more precise likelihood calculations could be made, or correlations could be added.
3.3. Triggering on muons

Figure 3.28: Trigger rates for the Phase-2 OMTF algorithm for a sample with an average of 200 pileup events. Left: As a function of \( p_T \) threshold. Middle: As a function of pileup for a \( p_T > 20 \text{ GeV} \) threshold. Right: As a function of pileup for a \( p_T > 25 \text{ GeV} \) threshold, comparing with two Run-2 data points at that threshold.

Figure 3.29: Performance deterioration of the Phase-2 OMTF algorithm when aging is included in the DT detectors. Left: trigger efficiency as a function of \( p_T \). Middle: efficiency as a function of \( \eta \). Right: trigger rate as a function of \( p_T \).

The calculation of the \( \prod_{i=1}^{L} P(\phi_i^{\text{dist}} | p_k^T)P(p_k^T) \) term described above requires many floating point multiplications, which in FPGA would be slow and consume excessive resources. Instead of multiplication, the sum of the log-likelihoods is used. The values of \( \log P(\phi_i^{\text{dist}} | p_k^T) \) are stored in LUTs implemented in the FPGA block-RAM (BRAM) modules. The BRAMs of the Xilinx Virtex-7 FPGA (which is used in the OMTF) can be used as look-up tables with 10 bits of address and 18 bits of output values. Just 7 bits are enough to encode the \( \phi_i^{\text{dist}} \) (which is used to address the LUT), and the remaining 3 bits are used to encode the number of the reference layer, since \( \log P(\phi_i^{\text{dist}} | p_k^T) \) depends on the layer of the reference stub in a given iteration.

For the output value, 6 bits are enough to encode the \( \log P(\phi_i^{\text{dist}} | p_k^T) \) — another advantage of using the logarithm, as it reduces significantly the required range of values. Thus it is possible to store in each BRAM cell up to three values of \( \log P(\phi_i^{\text{dist}} | p_k^T) \) coming from the same layer of different \( p_T \) classes with neighboring \( p_T \). The same address (i.e. \( \phi_i^{\text{dist}} \)) is then used for the three distributions. Additionally, the logic selecting the stubs with minimal \( \phi_i^{\text{dist}} \) can be common for these group of “golden patterns” (the same value of \( \Delta \phi_i^{\text{mean}}(p_k^T) \) is used for them). These optimizations reduce significantly the amount of the FPGA logic and BRAMs that are needed to implement the algorithm.

Summary and outlook

The studies described above demonstrate, for events with 200 average pileup, the feasibility
of triggering on prompt standalone muons in the overlap region with high efficiency (>95%) for $p_T > 3$ GeV, and a rate of $\sim 15$ kHz at a $p_T$ threshold of 20 GeV. The firmware resources expected for Phase-2 will allow a similar (Bayesian) method to be implemented for displaced muons. A 5% loss in trigger efficiency is expected when the aging of the DT stations is included in the studies. However this loss is expected to be mitigated when the Phase-2 DT TPs are used.

### 3.3.4 Standalone muon reconstruction in the endcap region

The endcap region covers $1.24 < |\eta| < 2.8$, and contains the Phase-1 CSC, and RPC detectors, which will be supplemented in Phase-2 with new iRPC and GEM detectors in the most forward directions. In this region all detector layers are arranged vertically and are subject to a strongly nonuniform magnetic field, to higher radiation, and to more copious punch-through interactions.

For many reasons, including the possible absence of the L1 track finder, and during commissioning, it is critical for CMS to maintain the ability to reconstruct standalone muons and measure their $p_T$ as precisely as possible, so that reasonable rates can be achieved for a low enough $p_T$ threshold. Moreover, for muons with large enough $p_T$, for which the rates are low, the experiment could trigger directly on these standalone muons without the need for further processing. Achieving this in the endcap region is not entirely straightforward, as with the current algorithm the muon trigger rate increases nonlinearly with luminosity, as can be seen in Fig. 3.30.

However, in Phase-2 the new ATCA platforms envisioned for the standalone endcap muon track finder, with more optical fiber inputs and larger FPGAs than the Phase-1 platforms, will allow for more sophisticated algorithms to be implemented. These algorithms benefit from using the information coming from the full set of Phase-2 muon detectors and are able to reconstruct both prompt and displaced muons that can be sent directly to the GT within a few microseconds. In what follows, we describe these new algorithms, their performance, and the possibility for further developments.

![Figure 3.30: The nonlinearity of the Phase-1 EMTF trigger rates for muons with $p_T > 22$ GeV and $|\eta| < 2.4$ as a function of pileup, as measured from data. The straight line is a fit to the data up to a pileup of 40.](image)

**Prompt muons**

The Phase-1 algorithm that reconstructs standalone prompt muons in the endcap (EMTF) operates in two stages. The first, pattern-recognition stage, identifies TPs in the CSC and RPC...
3.3. Triggering on muons

detectors that are consistent with muon trajectories. The second stage uses the combined information from these TPs as input to a boosted decision tree algorithm that determines the most likely value for the muon $p_T$.

Significant improvements to these algorithms have been studied in detail for Phase-2. These include (i) the extraction of more precise TP information from the existing detectors and new TP information from the Phase-2 systems (ii) a significant increase in the number of patterns in all detectors that are used to determine which stubs are consistent with muon trajectories, and (iii) the use of a neural network in which the correlations of the TP information from the stubs within a pattern is used to determine the best value of the muon candidate $p_T$. In what follows, we describe in more detail these items.

The information carried by the CSC and RPC TPs is improved as follows. The precision for both the position and local momentum measurements of the CSC TPs for ME1/1 and ME1/2 (where the magnetic field is stronger) has been improved considerably due to a new fit to the hits in the six layers that comprise the CSC chambers. The timing resolution for the RPC and iRPC detectors is reduced to 1.5 ns and the position resolution of the iRPC detectors is significantly better than the existing RPCs. The new GEM and ME0 detectors provide the position for muons traversing the higher rapidity regions, increasing robustness against random hits from punchthrough or noise when used in coincidence with the CSC detectors. In addition, the six layer ME0 also provides a bend angle. Table 3.1 lists the information available from the TPs of the different muon detectors.

The number of predefined patterns has been increased more than tenfold with respect to Phase-1. In the Phase-2 algorithm, the fiducial volume of the endcap region is divided in 0.5-degree segments in the azimuth view. These segments encompass a different number of detector layers, depending on $\eta$. Each 0.5-degree segment in $\phi$ on ME2 is defined as a unique key segment. For each 80-degree sector in azimuth (60-degree sector with 20-degree overlap with an adjacent sector), individual muons are shot through the detector simulation. They are distributed flat in 9 bins of $q/p_T$: $-0.5, -0.32, -0.22, -0.13, -0.085, 0.085, 0.13, 0.22, 0.32, 0.5$, and in 6 regions in $\eta$ per side: $1.24, 1.56, 1.7, 1.8, 2.0, 2.16, 2.15$. For a given $\eta$ region, muons of different $p_T$ will traverse specific key segments, but also other 0.5-degree segments on different layers. For each $q/p_T$ bin, integrating over the trajectories of many muons that traverse a specific key segment yields a well-defined pattern across all detector layers for that bin. Figure 3.31 (left) shows all 54 patterns for prompt muons for one specific key segment. The different zones represent the 6 different regions of pseudorapidity and the ”straightness” represent the 9 different bins of muon $q/p_T$. The total number of key segments is 160 per 80-degree sector and there are 12 such sectors (6 per side) being processed independently in the firmware, for a total of about 104k patterns. Practically, because of $\phi$ symmetry, it is enough to build in the simulation the 54 patterns for a key segment and then rotate them in steps of 0.5-degrees to form all needed patterns.

The next step in the algorithm is to estimate the $p_T$ for all patterns (muon candidates) that are satisfied in a collision, called roads. First, to remove possible fake muon tracks, the median $\theta$ of all stubs is found and stubs are required to be within a certain distance in $\theta$ from this median. Roads are discarded if less than 3 stubs remain after this requirement is applied. Next, the TP information from each detector layer is used as input to a regression NN. Table 3.2 shows the information used. After optimization of the NN structure, by training with prompt muons, together with pileup tracks from 200 PU events, a working point that yields very good rates and efficiency was reached using a NN with 36 input nodes and three hidden layers, each with 30/25/20 nodes. The results presented below are from this scheme.
Table 3.1: Information available in the various muon trigger primitives that serves as potential input to machine learning algorithms.

<table>
<thead>
<tr>
<th>Detector</th>
<th>num of bits</th>
<th>resolution</th>
<th>num of bits</th>
<th>resolution</th>
<th>num of bits</th>
<th>resolution</th>
<th>num of bits</th>
<th>resolution</th>
<th>num of bits</th>
<th>resolution</th>
<th>num of bits</th>
<th>resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSC</td>
<td>11</td>
<td>1/16-strip</td>
<td>7</td>
<td>wiregroup</td>
<td>5</td>
<td>1/16-strip</td>
<td>2</td>
<td>num layers</td>
<td>4</td>
<td>1/16-BX</td>
<td>4</td>
<td>1/16-BX</td>
</tr>
<tr>
<td>RPC</td>
<td>5</td>
<td>strip</td>
<td>2</td>
<td># partition</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>cluster width</td>
<td>3</td>
<td>cluster width</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>iRPC</td>
<td>7</td>
<td>strip</td>
<td>7</td>
<td>position along strip</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>cluster width</td>
<td>-</td>
<td>cluster width</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GEM</td>
<td>8</td>
<td>2-strip</td>
<td>3</td>
<td># partition</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>cluster width</td>
<td>-</td>
<td>cluster width</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ME0</td>
<td>10</td>
<td>1/2-strip</td>
<td>4</td>
<td>1/2-# partition</td>
<td>10</td>
<td>1/4-strip</td>
<td>4</td>
<td>num of layers</td>
<td>-</td>
<td>quality</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DT</td>
<td>12</td>
<td>1/4096 rad</td>
<td>3</td>
<td>bunch and track identifier</td>
<td>10</td>
<td>1/512 rad</td>
<td>3</td>
<td>quality</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 3.31: EMTF++ patterns for prompt (left) and displaced (right) muons. Within each dark square the different detector layers are represented by horizontal bars. In light blue are the regions in which stubs are allowed in that pattern. For prompt muons, the squares represent bins in $q/p_T$ (9 along the x axis) and in $\eta$ (6 along the y axis). For displaced muons the squares represent bins in $d_0$ (9 along the x axis) and in $\eta$ (6 along the y axis). See the text for more details.

Figure 3.32 compares the performance of the Phase-1 EMTF algorithm to the new EMTF++ algorithm for Phase-2 for the scenario with 200 average pileup events. For a trigger threshold of $p_T > 20$ GeV, a significant efficiency improvement as a function of muon $p_T$ is observed for EMTF++, with a plateau efficiency of $\sim 95\%$ and a sharper turn-on. Also shown are turn-on curves for the EMTF++ algorithm for different $p_T$ thresholds. As shown in the figure, the efficiency improvement is largest for the $|\eta| > 1.7$ region, with up to a 10% increase, for muons with $p_T > 20$ GeV. This holds true for muons with $5 < p_T < 20$ GeV.

Figure 3.33 shows the comparison of rates in the case of 200 average pileup events, as a function of L1 $p_T$ threshold, for the two algorithms. For $p_T = 20$ GeV, the EMTF++ rate is 18.4 kHz, a reduction factor of about 2.5 relative to the EMTF rate, with the largest reduction coming from the $|\eta| > 1.7$ region. Also shown are the trigger rates as a function of pileup, showing linear behavior up to 300 average pileup events for EMTF++, whereas the rates from the Phase-1 EMTF algorithm are significantly nonlinear.

In conclusion, the new Phase-2 EMTF++ algorithm, by taking advantage of the new muon detectors and by using more predetermined patterns together with a NN to determine the $p_T$

Table 3.2: Variables used as input to the EMTF++ neural network to determine the muon $p_T$.

<table>
<thead>
<tr>
<th>Variables</th>
<th>ME1/1</th>
<th>ME1/2</th>
<th>ME2</th>
<th>ME3</th>
<th>ME4</th>
<th>RE1</th>
<th>RE2</th>
<th>RE3</th>
<th>RE4</th>
<th>GE1/1</th>
<th>GE2/1</th>
<th>ME0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi$</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>$\theta$</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>bend</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>quality</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
of muon candidates, has been shown to provide excellent performance for standalone muons, both in terms of efficiency and rates. While several elements of the algorithm lend themselves to further optimization, the implementation presented here already achieves the desired performance. A discussion on firmware considerations regarding algorithm is provided below.

![Efficiency Graphs](image)

Figure 3.32: Efficiencies for endcap single muons ($|\eta| < 2.4$) with 200 average pileup events. As a function of generated muon $p_T$ for L1 $p_T > 20$ GeV (top left) and for L1 $p_T > 5, 10, 20$ GeV (top right). As a function of $\eta$ for muons with $p_T > 20$ GeV and a L1 $p_T > 20$ GeV (bottom left) and for muons with $5 < p_T < 20$ GeV and a L1 $p_T > 5$ GeV (bottom right).

**Displaced muons**

The algorithm studied for standalone displaced muons in the endcap region is similar to the one described above for prompt muons. However, given that the trajectories of displaced muons in the muon detectors deviate significantly from those of prompt muons, new displaced patterns are needed. Simulated samples of single muons that cover the relevant phase space are used, with generated flat spectra in $1/p_T$ for $2 < p_T < 200$ GeV, in $d_0$ transverse displacements from 0 to 120 cm, in $1.2 < |\eta| < 2.4$, and in $|z_0| < 30$ cm.

The number of patterns tested as a possible working point is the same as that of the prompt muons, but the pattern strategy is somewhat different, with the requirement that $p_T > 14$ GeV and the binning is done in 9 bins of impact parameter $d_0$, from $-120$ to 120 cm and not in $p_T$. 
Figure 3.33: Left: endcap trigger rate comparison of the Phase-1 EMTF and the Phase-2 EMTF++ algorithms as a function of $p_T$ threshold for events with 200 average pileup. Right: Trigger rate comparison as a function of PU for a $p_T > 20$ GeV threshold.

The same 6 $\eta$ zones are retained for a total of 54 patterns per 0.5-degree in $\phi$, as for the prompt muon patterns. Figure 3.31 (right) shows these patterns.

The TP information in the stations from stubs that satisfy a displaced pattern are input to a NN that in this case has been trained to perform a regression that returns simultaneously values for $1/p_T$ and $d_0$ of displaced muons. The NN configuration used is the same as that for prompt muons, using 3 hidden layers with 30/25/20 nodes each. Batch normalization is inserted after each layer, including the input layer. A total of 23 inputs are used in the NN, these are:

- 6 $\Delta\phi$ quantities between stations: S1-S2, S1-S3, S1-S4, S2-S3, S2-S4, S3-S4
- 6 $\Delta\theta$ quantities between stations: S1-S2, S1-S3, S1-S4, S2-S3, S2-S4, S3-S4
- 4 bend angles: set to zero if no CSC stub is found and only RPC stub is used
- For ME1 only: 1 bit for front or back chambers and 1 bit for inner or outer $\eta$ ring
- 1 track $q$ taken from stub coordinate in ME2, ME3, ME4 (in this priority)
- 4 RPC bits indicating if ME or RE stub was used in each station (S1, S2, S3, S4)

At the time of this writing, information from the new Phase-2 detectors (GE1/1, GE2/1, ME0, iRPC) has not been incorporated into the study, and neither has the more precise CSC bend information described above. As such, this study is geared towards possible implementation of this algorithm during Run-3. An update to incorporate new Phase-2 detector information is in progress. The already positive conclusions on triggering on standalone displaced muons in the endcap with only the Phase-1 detectors, as shown below, is expected to improve significantly when all Phase-2 information is included.

Figure 3.34 shows, for events with single muons and no pileup, the $q/p_T$ and the $d_0$ resolutions as determined by the NN estimate of these quantities. The $p_T$ resolution is about 60%, which is large compared to the 20% resolution obtained from EMTF++ for prompt muons. A bias towards underestimating the $p_T$ can be observed. However, the $d_0$ resolution is very good, $\sim 5$ cm. Figure 3.35 shows the trigger rates of the displaced muon algorithm for PU 200 events. In order to keep the rates at approximately the same 10 kHz level as those from prompt muons, reasonable L1 thresholds of, for example, $p_T > 20$ GeV and $|d_0| > 20$ cm can be applied.
The figure also shows the efficiencies as a function of $p_T$, $\eta$, and $d_0$ for displaced muons in events with no pileup, when these $p_T$ and $d_0$ thresholds are applied. Efficiencies of $\sim 50\%$ are obtained for $d_0 > 30$ cm and up to 100 cm. The structure on the efficiency as a function of $\eta$ results from the gaps in transitions between detector stations. Significantly higher efficiencies can be achieved by relaxing these thresholds while also incurring a significant rate increase. Also shown in the figure is the efficiency as a function of $d_0$ for the Phase-1 EMTF algorithm for prompt muons, which intrinsically has a high efficiency for muons with small displacements $|d_0| < 10$ cm.

![Figure 3.34: Resolutions for the EMTF++ displaced muon algorithm for single muons in the endcap. Left: $q/p_T$. Right: $d_0$.](image)

**Firmware Considerations**

The Phase-1 EMTF algorithm has been running successfully in CMS in Run-2 conditions. As shown above, the new EMTF++ algorithm is needed for Phase-2. The status of the implementation so far is described in detail in Section 6. As shown in this Section, the large FPGA resources anticipated in the new ATCA platforms will be able to accommodate not only the algorithm for prompt muons but also the one for displaced muons.

**Summary and outlook**

We have demonstrated excellent performance for the new Phase-2 algorithms developed for standalone muons in the endcap for the harsher conditions expected at the HL-LHC. For prompt muons high reconstruction efficiencies of about 95% have been reached with very reasonable trigger rates of $\sim 20$ kHz for a 20 GeV $p_T$ threshold; about a factor of three lower than what the Phase-1 algorithm can yield. Moreover, the rate dependence with pileup is linear. We have also demonstrated the feasibility of triggering on displaced muons, with low rates, 20 kHz for $p_T = 20$ GeV, and efficiencies of 50% for displacements between 25 and 100 cm. This is a new feature for Phase-2 that will expand the CMS physics program.

### 3.3.5 Global muon trigger algorithms

Both in the High Level Trigger and in the offline software the CMS muon reconstruction algorithms combine the tracker information with that of the muon detectors. With the Phase-2 L1 muon trigger, in addition to the standalone muons, new types of muon objects can be reconstructed by taking advantage of the availability of tracks from the track finder [8]. To recon-
Figure 3.35: Rates and efficiencies for single displaced muons from the displaced EMTF++
algorithm in the endcap. Trigger rates as a function of $p_T$ for different $d_0$ selection criteria
(upper left). Trigger efficiency as a function of $p_T$, for L1 $d_0 > 20$ cm, for different $p_T$ thresholds
(upper right), as a function of $\eta$ for L1 $p_T > 20$ GeV (lower left), and as a function of $d_0$, also for
L1 $p_T > 20$ GeV, including, in red, the efficiency for displaced muons for the prompt algorithm
(lower right).

struct these various types of muons, the GMT receives tracks from the track finder, standalone
muon tracks from the regional muon track finders, and all stubs from the muon detectors. This
allows for great flexibility and creativity in the design of Phase-2 muon trigger algorithms.

The objects reconstructed by the GMT are sent directly to the Global Trigger, where they can be
used either as a single-object trigger, or in conjunction with other trigger objects. In addition,
GMT objects can also be sent to the Layer-2 of the Correlator Trigger, implementing the particle-
flow algorithm. Figure 5.4 displays the interconnections between the GMT and the regional
muon track finders, the Correlator Trigger, and the GT.

Matching L1 tracker tracks to standalone muons from the regional track finders ("TkMu" trig-
ger objects) or to selected combinations of individual stubs ("TkMuStub" trigger objects), al-
ows those L1 tracker tracks to be tagged as muons. The L1 tracker track $p_T$ measurement,
which (as shown below) has significantly better resolution than the measurement for stan-
3.3. Triggering on muons

dalone muons, can be assigned to those muon trigger objects.

As demonstrated in the following Sections, the TkMu matching is performed in windows that vary in $\Delta \phi$ and $\Delta \eta$, depending on the L1 tracker track $p_T$ and $\eta$, and results in considerably sharper trigger efficiency curves. Consequently, lower trigger rates are achieved at any given $p_T$ threshold, significantly enhancing the physics capabilities of the CMS experiment.

Matching L1 tracker tracks to individual muon stubs, as opposed to reconstructed muon tracks, can have some advantages and some disadvantages. The advantages include the possibility of tagging low $p_T$ muons that only leave hits in the muon stations closest to the interaction region, and would not be reconstructed by the muon track finding algorithms. This can be particularly important in the case of physics signatures with low $p_T$ muons such as $\tau \rightarrow 3 \mu$ (described in more detail below). Also, when muons traverse the fiducial regions that do not have complete coverage by more than one or two muon stations (where again the muon track finding algorithms fail to reconstruct them as muon tracks), matching L1 tracker tracks with only stubs will help cover these gaps and increase the overall muon acceptance. On the other hand, the different TkMuStub algorithms might be more susceptible to punch-throughs from pileup, to dead chambers, to aging, and to detector or electronics noise, resulting in inefficiencies and/or increased trigger rates. Ultimately the detector's performance and occupancy during real HL-LHC running conditions will help determine how to best define a trigger strategy that works optimally, with the interplay between different TkMu and TkMuStub algorithms. It is therefore paramount for the CMS muon trigger to maintain the ability to match L1 tracker tracks to fully reconstructed muons in the regional muon track finders, and to selected stubs not necessarily associated to these muons. These cases are described and compared in the Sections below.

Other important algorithms that can use L1 tracker track matching to muons and that can be implemented in the GMT include: 1) the reconstruction of HSCPs, where both the position and the timing as measured by the improved muon detectors can be matched with that of the L1 tracker track, 2) the determination that certain muons might be isolated from other charged tracks in an event, 3) the identification of displaced muons by reconstructing segments from stubs in the muon detectors that do not point to the interaction region and demanding that prompt L1 tracker tracks do not point to these segments, 4) the identification of clusters of stubs in the muon detectors from new interesting physics processes such as displaced jets, and 5) the identification of well-defined states that decay to multi-muons such as $\tau \rightarrow 3 \mu$. Some of these cases are described in the following sections. For all these algorithms to be implemented in the GMT, while also leaving room for new ideas, it is important to ensure that enough resources and flexibility are built into the GMT architecture.

1. L1 tracker tracks and standalone muon correlations

The muon track finders in the barrel, overlap, and endcap regions determine which sets of stubs are used to reconstruct standalone muon candidates, and then transmit this information to the GMT. The availability of L1 tracker tracks at the GMT level allows for algorithms that can match L1 tracker tracks to those regional muon candidates and form TkMu trigger objects. Figure 3.36 shows, for muons reconstructed in the endcap region and with samples in which no additional pileup interactions are included, the spatial and momentum resolution from the EMTF++ algorithm and from the track finder. The track finder spatial resolution is approximately one order of magnitude better than the EMTF++ for both the $\theta$ and $\phi$ coordinates. The muon $p_T$ is determined with a resolution of $\sim 3\%$ for muons in the tracker and 15 to 30$\%$ for EMTF++ reconstructed muons, depending on the momentum considered.

The track finder to muon correlation takes into account the L1 tracker track’s propagation from
the track to the muon system and the multiple scattering that occurs on its path. Both of these effects scale as $1/p_T$, although the precise dependence is more complex because of the nonuniform magnetic field in the endcap regions and its strength dependence on $\eta$. The amount of material traversed also varies with $\eta$, the bulk being from the calorimeters and the magnet’s iron return yokes. Consequently, the multiple scattering depends significantly on $\eta$.

In the TkMu algorithm, a L1 tracker track and a standalone muon are considered as matched if their angular separations ($\Delta \theta$ and $\Delta \phi$) at a chosen station in the muon detectors are within precomputed intervals, called “matching windows”. The center and width of such intervals varies as a function of the track $p_T$ and $\eta$, and can be tuned to achieve a desired efficiency. Single-muon samples generated with no additional pileup are used to unambiguously identify the associated L1 tracker track and muon objects, and study their $\Delta \theta$ and $\Delta \phi$ separation as a function of $\eta$. Similar procedures have been derived for the barrel, overlap, and endcap regions. Therefore, in what follows, as an example, we provide an example of such algorithm for the endcap region, while performance figures are provided for all regions.

The simulated muons in the endcap region are divided into 10 bins of 0.12 units in $|\eta|$, and 200 $p_T$ bins of 1 GeV width. For each of these two-dimensional bins, the $(1 - \alpha)/2$ and $(1 + \alpha)/2$ quantiles are computed to give the interval, where the fraction $\alpha$ of the candidates can be found. The resulting intervals are denoted as $[\Delta \theta_{\text{low}}, \Delta \theta_{\text{high}}]$ and $[\Delta \phi_{\text{low}}, \Delta \phi_{\text{high}}]$ in what follows. The value of $\alpha$ can be chosen in the procedure to achieve a desired efficiency. Statistical fluctuations may significantly impact the determination of the matching window boundaries. The values obtained in each $|\eta|$ range are thus interpolated as a function of $p_T$ using the functional formula:

**Figure 3.36**: Resolution comparisons between the the EMTF++ algorithm (top row) and the track finder (bottom row). Shown are the difference between the measured $\theta$ and $\phi$ coordinates and the true value from the generated muons, and the ratio of the measured $p_T$ to the true muon $p_T$. The different colors indicate the different generator level $p_T$ intervals considered.
\[ \Delta \theta_{\text{low/high}}(p_T) = \text{const.} + A \times p_T^B \]
\[ \Delta \phi_{\text{low/high}}(p_T) = \text{const.} + A' \times p_T^B. \]  

Equation 3.6 is used to determine the values of \( \Delta \theta_{\text{low/high}} \) and \( \Delta \phi_{\text{low/high}} \) as a function of the track momentum.

Figure 3.37 shows the values obtained for the endcap matching windows targeting 99% efficiency and the corresponding fitted function for several of the pseudorapidity bins. The windows can be tuned to achieve any desired efficiency, where larger windows result in larger trigger rates.

The efficiency as a function of the generated muon \( p_T \) and \( \eta \), for 200 average pileup events, is shown in Fig. 3.38 for the endcap, and it is compared to the standalone EMTF++ algorithm performance. As expected the \( p_T \) turn-on is significantly sharper, with a corresponding slightly higher efficiency for muons above the selected \( p_T \) threshold. Figure 3.39 shows a significant rate reduction with respect to EMTF++, which had already achieved a very considerable reduction with respect to the Run-2 EMTF algorithm. The expected rate at a \( p_T = 20 \text{ GeV} \) threshold is about 10 kHz. Also shown in the figure, are the fractions of the rate that originate from true muons and from fakes, which are roughly the same at that threshold. Finally, the figure shows the rate dependence as a function of pileup, for a \( p_T \) threshold of 20 GeV, compared to the rate from EMTF++. Reasonable rates are achieved up to an average pileup of 300.

To associate L1 tracker tracks with standalone muons reconstructed in the barrel (KBMTF) and overlap (OMTF) regions, we use a similar correlation algorithm to the one used in the endcap, with matching windows depending on \( p_T \) and \( \eta \).

Figure 3.40 shows the trigger efficiencies in the barrel region, as a function of generated muon \( p_T \) and \( \eta \), comparing the L1 tracker track plus KBMTF correlator performance to the KBMTF
algorithm. Again, the turn on of the matching algorithm is very sharp, with a small increase in efficiency.

Figure 3.41 shows the trigger rate in the barrel region as a function of L1 $p_T$ for PU 200 events, comparing the BMTF, KBMTF, and the track matching algorithm which, as expected, reduces the rate very significantly, by about a factor of 4. Also shown in the figure is the dependence on pileup for the KBMTF and the track+muon matching algorithm for L1 $p_T > 20$ GeV. Up to average pileup of 300, reasonable rates of 10–15 kHz are achieved.

Figure 3.42 shows the trigger efficiencies in the overlap region as a function of generated muon $p_T$ and $\eta$, comparing the L1 tracker track plus OMTF correlator performance to the OMTF algorithm. The efficiencies are shown as a function of $p_T$ and $\eta$. Also in this region, the turn on of the matching algorithm is very sharp, with a small increase in efficiency.

Figure 3.43 shows the trigger rate in the overlap region, as a function of L1 $p_T$ for PU 200 events, comparing the OMTF, and the track matching algorithm (TkMu) which, as expected, reduces the rate very significantly by about a factor of 5. The figure also shows the pileup
dependence for values up to 300, for the OMTF and the track+muon matching algorithm, for a L1 $p_T$ threshold of 20 GeV. It can be seen that reasonable rates, of 6–16 kHz, are achieved by these algorithms.

2. L1 tracker tracks and muon-stub correlations

Given the diversity and complexity of the muon detectors, there are many ways to select individual stubs, and their attributes, to match to L1 tracker tracks and form the TkMuStub trigger objects. Different matching techniques and stub selection criteria have been studied in each of the different detector regions, all demonstrating very good performance. Ultimately, the detector’s performance and occupancy under real data-taking conditions will indicate how to best adjust and utilize these algorithms in the different detector regions. Building flexibility into these algorithms, as done here, will also help adapt to changing conditions as the run progresses.

2.1 The TPS algorithm in the barrel region

The track plus stubs (TPS) algorithm in the barrel region matches L1 tracker tracks with muon stubs in the $|\eta| < 0.83$ region. The TPS algorithm operates in four steps: track propagation, track-stub association, overlap cleaning, and $p_T$ sorting. The track propagation is approximately linear with respect to the signed curvature $k = q / p_T$ of the track, for the position angle $\phi_{i,j}(k) = \phi + c_i j k$ and the bending angle $\phi_{b,i}(k) = c_{b,i} k$, where $q = \pm 1$ is the charge of the particle and $\phi$ is the azimuth angle of the track at the interaction point. The propagation coefficient $c_i$ for station $j$ is determined from the mean values of Gaussian fits of $\Delta \phi = \phi_1 - \phi$ histograms made as functions of $k$. Energy loss from ionization in the material budget is included by using the corrected curvature $k^* = k / (1 - ek) \approx k + ek^2$ for the small and nearly constant energy loss $e$. Figure 3.44 (left) shows the derivation of $c_i$ for the barrel wheel-0.

The Gaussian fits also provide the resolution $\sigma_j = \sqrt{\alpha_j k^2 + \beta_j}$, which is approximated by $\sigma_j = \alpha_j' |k| + \beta_j'$ to simplify the firmware implementation. The $\alpha_j$ (or $\alpha_j'$) term describes the contribution from multiple scattering and the $\beta_j$ (or $\beta_j'$) term describes the contribution from the chamber resolution. Figure 3.44 (right) shows the resolution for RPC and DT chambers within the same wheel and station. The track-stub association is performed by matching the track and stub angles within their respective resolutions. This is achieved using the pull distribution $P_j = \left| \phi_{i,j}(\text{stub}) - \phi_{i,j}(\text{prop}) \right| / \sigma_j$, by cutting on $|P_j(\phi_i)| < a_j$ and $|P_j(\phi_b)| < b_j$.
Figure 3.40: Barrel correlation performance. Single muon efficiency as a function of $p_T$ compared to the KBMTF algorithm (top left), and for different L1 $p_T$ thresholds (top right). Single muon efficiency as a function of $\eta$ for L1 $p_T$ thresholds of 20 GeV (bottom left) and 5 GeV (bottom right).

where $a_j$ and $b_j$ are appropriate limits, depending on the muon station. The pull is formed for each sub-detector and the resolution calculation is verified, checking if it follows a normal distribution with mean 0 and $\sigma = 1$.

Overlapping muon candidates are cleaned with respect to shared stubs. If candidates share stubs, the candidate with the most stubs is selected while the other one is discarded. If the two candidates have the same number of stubs, then the one with the smallest sum of $|\phi_j(\text{stub}) - \phi_j(\text{prop})|$ values is kept and the other is removed. The dominant background originates from matching a high-$p_T$ track with stubs generated by a low momentum hadron decay in flight, or a nuclear interaction of a punch-through hadron in the detector. In the case of punch-through hadrons, the same track is matched to stubs of the hadronic shower remnants in the muon system, while in the case of hadron decays in flight, a close-by energetic track is matched to a stub of a low momentum decay in flight. These backgrounds are rejected by examining the stub pattern of the track in different detector regions and rejecting patterns that are not compatible.
3.3. Triggering on muons

Figure 3.41: Barrel correlation performance. Left: As a function of L1 $p_T$ threshold, compared to the standalone BMTF and KBMTF algorithms. Right: As a function of pileup, for a L1 $p_T$ threshold of 20 GeV.

with a real muon. Significant improvement in performance is possible when the pattern selection is performed through a configurable pattern memory, which is the preferred option for future operation.

Firmware implementation

The TPS algorithm is translated to firmware using Vivado HLS targeting a ZYNQ Ultrascale+ (ZU19EG-2) FPGA running with an internal clock of 200 MHz. The algorithm is split into three modules:

- The Muon Associator propagates up to 18 L1 tracker tracks, each one coming from one fiber, and associates them to stubs coming from the 60 Layer-1 barrel sectors.
- The Muon Accumulator stores in memory the output tracks associated with at least one stub from the Associator, until 36 such tracks are accumulated, and then returns them as output.
- The Muon Sorter receives the associated tracks from the Accumulator, it sorts them bitonically in $p_T$ and returns 8 muons with the highest $p_T$.

Preliminary studies show reasonable FPGA total utilization, with 18% used by LUT, 6% by FF, 15% by DSP, and 4% by BRAM. Implementation studies are still ongoing to optimize both resource utilization and latency, including a demonstration on hardware. More details are discussed in Chapter 6.

Performance

The performance of the algorithm is examined by measuring the L1 efficiency and the trigger rate. The efficiency is measured in a high pileup muon sample with uniform $p_T$ and $\eta$ distributions. For the performance shown here, generator-level muons are selected with $|\eta| < 0.83$, corresponding to the barrel acceptance. The rates are measured in a high pileup minimum bias sample. The efficiency in a given bin of $p_T$ or $\eta$ is the ratio of the generated muons matched
Figure 3.42: Overlap correlation performance: As a function of $p_T$, for single muons with L1 $p_T > 20$ GeV, compared to the OMTF algorithm (top left); for L1 $p_T > 3$ GeV (top right); and as a function of $\eta$ for muons with L1 $p_T > 25$ GeV (bottom left) and $7 < p_T < 15$ GeV (bottom right).

(within a cone of $\Delta R < 0.1$) to a TPS L1 muon that passes the threshold, divided by the number of generated muons in the bin.

Figure 3.45 shows the TPS efficiency as a function of the generated muon $p_T$. The efficiency is shown for various L1 $p_T$ thresholds (left) and compared with the efficiencies of the Phase-2 standalone and L1 tracker track + muon track (“global muon”) algorithms for a L1 threshold of 20 GeV (right). The improved $p_T$ resolution of the TPS algorithm is demonstrated by the sharp efficiency turn-ons. The TPS algorithm efficiency at the plateau is above 99%, demonstrating a substantial improvement in reducing the inefficiency by about a factor of 5 compared to the standalone trigger, where the plateau approaches 95%. The improvement comes from regaining muon tracks in the gaps between the barrel muon detector wheels which are lost in the standalone muon trigger.

Figure 3.46 shows the TPS efficiency compared with the standalone and track + muon efficiencies as a function of the generated muon $\eta$ for a L1 threshold of 5 GeV in the left plot and
3.3. Triggering on muons

20 GeV in the right plot. By requiring only one matched stub in TPS the standalone and track + muon efficiency lost between barrel wheels, due to missing stubs to reconstruct muon tracks in the detector gaps, is regained to the level of 99%.

Figure 3.47 shows the TPS trigger rate compared with the standalone and track + muon rates as a function of the L1 \( p_T \) threshold in the left plot and of the average pileup in the right plot. For a typical single muon threshold of 20 GeV, the expected rate at 200 average pileup events is about 4.5 kHz, which is less than 1% of the available Phase-2 L1 budget.

2.2 The Bayesian tracker track and muon-stub correlator algorithm in the overlap region

The task of tagging the L1 tracker tracks as muons can be considered as a classification problem: each tracker track has to be classified either as a muon or not. To perform this classification, the Bayesian track plus muon-stub correlator algorithm determines the probability that the tracker track is a muon given a set of muon stubs and the tracker track parameters:

\[
P(T_\mu | \eta, \phi, p_T, q, \text{muonStubs}) = \frac{P(\text{muonStubs} | T_\mu, \eta, \phi, p_T, q)P(T_\mu | \eta, \phi, p_T, q)}{P(\text{muonStubs} | \eta, \phi, p_T, q)} \quad (3.7)
\]

In equation 3.7, \( T_\mu \) denotes the hypothesis that a given tracker track is a muon, and \( \eta, \phi, p_T, q \) are the pseudorapidity, the azimuth angle, the transverse momentum, and the charge of the tracker track, respectively. Following the Bayes rule, the probability can be re-written as shown on the right side of equation.

If symmetry in \( \phi \) is assumed, then the likelihood \( P(\text{muonStubs} | T_\mu, \eta, \phi, p_T, q) \) depends only on the difference between the muon stub position in a given \( \phi \) layer, \( \phi^i_\mu \), and the extrapolated track position in the same layer, \( \phi^i_\ell \). In case of \( \eta \) layers, if assumed that the deflection in \( \eta \) can be neglected, the likelihood depends on the differences between the muon stub \( \eta^i_\mu \) in a given \( \eta \) layer and track \( \eta_\ell \). Furthermore, when it is assumed that approximately the \( (\phi^i_\mu - \phi^i_\ell) \), \( \phi^i_{B\mu} \) (the stub’s bending angle) and \( (\eta^i_\mu - \eta_\ell) \) of the muon stubs in different layers are independent from
Chapter 3. Trigger algorithms

Figure 3.44: Left: Derivation of the propagation coefficient in two stations of wheel-0 in the barrel muon detector. The difference in slope demonstrates the dependence of the coefficient on the detector depth. Right: Fitted Gaussian resolution for the difference between the propagated angle and the closest stub for RPC and DT detectors in the same station of wheel-0, demonstrating the different offset due to the different position resolution and the same curvature due to the same multiple scattering.

each other, then the likelihood can be expressed as the product of the likelihoods for muon stubs in every layer, as expressed in equation 3.8. This is the same “Naive Bayes” classifier approach used for the OMTF standalone muon reconstruction.

\[
P(\text{muonStubs } | T_\mu, \eta_t, \phi_t, p_T, q) \approx \prod_{i\in\text{layers}} P(\phi_{\mu i} - \phi_{t i} | T_\mu, \eta_t, p_T, q) \times \prod_{i\in\text{phi layers}} P(\phi_{Bi} | T_\mu, \eta_t, p_T, q) \times \prod_{i\in\text{eta layers}} P(\eta_{\mu i} - \eta_i | T_\mu, \eta_t, p_T, q) \tag{3.8}
\]

In equation 3.8, the bending angles \(\phi_{Bi}\) of the DT stubs are treated as separate “layers” of measurements. In the case of the DT and CSC chambers, the measurement of the azimuth angle and pseudorapidity are provided as distinct stubs, thus are also treated as separate layers.

To facilitate the implementation of the likelihood computation in the firmware, the log-likelihood form is used so that the multiplication of likelihoods is replaced by a sum of log-likelihoods. Exploiting the charge symmetry and assuming that the likelihood for a given layer can be approximated by a Gaussian distribution, the log-likelihood for a given phi layer can be expressed as a second order polynomial:

\[
\log P^i = \log P(q \cdot (\phi_{\mu i} - \phi_{t i}) | T_\mu, \eta_t, p_T) = a^i (\phi_{\mu i} - \phi_t - m^i \cdot q)^2 + c^i \tag{3.9}
\]

In equation 3.9, \(m_i\) is the average deflection of a muon with a given momentum from the interaction point (where the track \(\phi_t\) is measured) relative to a given chamber layer \(i\), (or an average \(\phi_{Bi}\) in case of the bending layer), and \(\phi_{\mu i}\) is the phi of the muon stub. The coefficients \(a^i\) and \(c^i\) depend on the standard deviation of the likelihood distribution. The \(\log P^i\) is linearly transformed such that it is convex and positive in a chosen range.
3.3. Triggering on muons

Figure 3.45: Trigger efficiency of the TPS algorithm. Left: Efficiency as a function of the generated muon $p_T$ for various L1 $p_T$ thresholds. Right: Comparison with the standalone and track + muon (“global muon”) efficiencies for a L1 threshold of $p_T = 20$ GeV.

Figure 3.46: Trigger efficiency of the TPS algorithm as function of the generated muon $\eta$ compared with the standalone and track + muon (“global muon”) algorithm efficiencies. Left: Efficiency for a L1 threshold of 5 GeV. Right: Efficiency for a L1 threshold of 20 GeV.

Both $m_i'$ and the standard deviation —and thus the coefficients $a_i'$ and $c_i'$— depend on the $p_T$ and $\eta_i$ of the tracker track. Therefore, they are stored in look-up tables: $\text{LUT}(\eta_i, p_T) = (a_i', m_i', c_i')$. To be used as the LUT addresses, the track $p_T$ and $\eta_i$ are binned in the following way:

- $p_T$ is divided into 64 bins in a nonlinear scale,
- $\eta_i$ is divided into 16 bins for the endcap layers, and one eta bin in case of the barrel layers (in the barrel the variation of the $a_i$, $m_i$ and $c_i$ versus $\eta_i$ can be neglected).

In a given muon layer there can be many muon stubs, therefore for a given tracker track only one stub is chosen, which is the one closest to the extrapolation of the track to the muon layer, i.e. the stub originating from a chamber with $\eta$ compatible with the track $\eta_i$, and with the
smallest $\phi_\mu - \phi_t - m^t \cdot q$. The stub is considered to be matched if its log $P_t$ is positive.

More than one track can be matched to the same muon stub. To select a unique matching, the stub is assigned to the track for which the log $P_t$ is a maximum. Only those tracks that match to a minimal number of muon stubs are considered as muon candidates. In the current implementation, the minimal number of stubs is two, but a track that is matched to only two stubs in the first station (DT/CSC and RPC layers) is discarded. If additional rate reduction is needed, a cut on the sum of the log $P_t$ can be applied (depending on the number of stubs matched to a given track). This cut is equivalent to rejecting tracks with low probability of being a muon.

### Performance

Figures 3.48 and 3.49 show the performance for the overlap region. Efficiency and trigger rate between the standalone OMTF algorithm, the algorithm matching L1 tracker tracks to OMTF muons, and the Bayesian correlator of L1 tracker tracks and muon stubs are compared. As expected, the efficiency, shown as a function of generated muon $p_T$ and $\eta$, increases with respect to both the OMTF standalone reconstruction and the L1 tracker track matching to OMTF muons. The largest increase occurs at the edges of the overlap region. Also as expected, the use of L1 tracker tracks reduces the rate considerably, more than a factor of 5 at a L1 threshold of 20 GeV. Moreover, the rate remains low (at 6 kHz) for an average pileup up to 300.

### Firmware considerations

At the time of this writing, the algorithm is not yet implemented in firmware. However, the implementation design concept is developed and the software emulator reflecting this design is available. The algorithm is in many aspects similar to the OMTF algorithm, which has been running in CMS throughout Run-2. Therefore, no issues are foreseen for it to be implemented in the newer FPGAs to be used in Phase-2. The implementation consists of the following blocks:

- The matching block is implemented for every tracker input link, it contains a sep-
3.3. Triggering on muons

Figure 3.48: Efficiency comparison for the Bayesian muon-stub correlator: As a function of $p_T$ for muons with $L1 \ p_T \geq 20 \text{ GeV}$ (upper left) and with $L1 \ p_T \geq 3 \text{ GeV}$ (upper right); as a function of $\eta$ for muons with $p_T \geq 20 \text{ GeV}$ (lower left) and $p_T \geq 5 \text{ GeV}$ (lower right).

rate module for matching L1 tracker tracks with the stubs in each muon layer. This logic first selects the best-matching stub, and then calculates the $\log P^i$.

- The L1 tracker track candidates that have a sufficient number of matched stubs are stored in a buffer, together with the identifiers of the matched stubs and $\log P^i$.
- The next track candidates are checked against those stored in the buffer: if they have the same matched stubs, the $\log P^i$ is compared, and the stub is removed from the candidate with lower log $P^i$.
- If the number of remained stubs in a given candidate is less than the minimum required, the candidate is removed.
- For the remaining candidates the sum of the $\log P^i$ is calculated.

2.3 The Dynamic Window tracker track and muon-stub correlator in the endcap region

A similar TkMuStub method to the one used for matching L1 tracker tracks to reconstructed muon candidates in the regional track finders can be implemented with single muon stubs or
combination of stubs. Given the diversity of detectors in the endcap region, many options to select stubs for such matching can be considered.

For the results presented here we use stubs from the CSC and RPC detectors found in stations 1 or 2, within the 1.2–2.4 \( \eta \) range. Matching windows were derived for muons in different \( p_T \) and \( \eta \) bins. The location and size (in \( \Delta \theta \) and \( \Delta \phi \)) of the matching windows were tuned independently for each station to achieve a 99% matching efficiency.

For each L1 tracker track, all stubs located in the matching windows in both stations 1 and 2 are examined, and the stub closest in \( \phi \) to the projected \( \phi \) of the track at the corresponding station is considered as the matched stub. If more than one tracker track is matched to the same stub, the track with the larger \( p_T \) takes precedence and the lower \( p_T \) track is matched to the next closest unmatched stub. With this approach, no pair of tracks are allowed to share a matching stub.

**Performance**

Figure 3.50 shows the efficiency of the TkMuStub algorithm described here, as a function of muon \( p_T \) and \( \eta \), in comparison with the efficiency from the standalone EMTF++ and L1 tracker track plus EMTF++ algorithms. As expected, the TkMuStub efficiency is very high, about 99%, and a few percent higher than that of the other algorithms. The ability to match tracks to individual stubs in multiple stations mitigates the intrinsic inefficiency of the EMTF++ algorithm, which reconstructs muon tracks that require stubs in at least three stations, including stubs in both stations 1 and 2. This inefficiency is present in the TkMu algorithm as well, for the same reason. The improvements are larger, of about 5%, in the higher \( p_T \) range and in \( \eta \) regions with gaps between detector layers. In the lower, 2 to 5 GeV, \( p_T \) range the improvement is of the order of 20% and is discussed further in Section 4.3.2.

Figure 3.51 shows the trigger rate (left) and purity (middle) obtained for this TkMuStub matching algorithm. For muons with \( 5 < p_T < 20 \) GeV, due to a slightly lower purity, the rate is about 10% higher than the TkMu rate. In the \( 2 < p_T < 5 \) GeV range, the TkMuStub rate is a factor 2 to 4 higher than the TkMu rate, coming predominantly from charged pions and \( K \)-mesons (60%), with a smaller fraction being from prompt muons (30%). At a L1 \( p_T \) threshold of 20 GeV,
about 50% of the rate is from prompt muons, labeled as true in the middle figure. The remaining fraction, labeled as fake, is from punch-through and decays in flight of charged pions and $K$-mesons (25% and 10%, respectively), and from fake L1 tracker tracks matched to spurious muon stubs (15%). The TkMuStub rate is reasonable up to a pileup of 250 and increases significantly at 300. At that point, the rate could be mitigated by tightening the matching windows or, if by doing so the efficiency gets compromised too much, by switching to the TkMu algorithm, which maintains a reasonably low rate for an average pileup of 300.

Figure 3.50: Endcap correlation performance: Single muon trigger efficiency for the standalone EMTF++ (blue), the track finder to EMTF++ correlator (red), and the track finder to muon-stub correlator (green) algorithms. Top row: as a function of generated muon $p_T$ for L1 $p_T$ thresholds of 20 and 5 GeV. Bottom row: as a function $\eta$ for L1 $p_T > 20$ GeV and $5 < p_T < 20$ GeV.

**Firmware considerations**

A firmware solution for matching L1 tracker tracks to reconstructed muons and muon stubs is described in some detail in Chapter 6. The algorithm will run in the GMT and the resources needed are modest and dominated by the look-up tables.

**3. Track-isolated muons**
Figure 3.51: Endcap correlation performance: Trigger rates as a function of L1 $p_T$ threshold comparing the standalone EMTF++, the L1 tracker track to EMTF++, and the L1 tracker track muon-stub matching algorithms (left). Also shown, are the fractions originating from true muons and from L1 tracker track matches that are not true muons for the L1 tracker track to muon-stub matching algorithm (middle), and the trigger rate dependence on pileup comparison of all three algorithms, for a $p_T$ threshold of 20 GeV (right).

As mentioned above, the presence of L1 tracker tracks in the GMT allows the identification of candidate muons isolated from other tracks to be performed with acceptable efficiency. Results obtained with the TPS algorithm are presented here, bearing in mind that any of the other algorithms presented above would perform very similarly, given that once a track is tagged as a muon candidate the isolation algorithm is purely based on L1 tracker tracks. The isolation $p_T$ is calculated as the scalar sum of $p_T$'s of all tracker tracks in a cone of radius $\Delta R = \sqrt{(\Delta \eta)^2 + (\Delta \phi)^2} = 0.3$ centered around the matched track, excluding the $p_T$ of the matched track. The tracks are also required to originate from vertices within a maximum distance $|dz|$ from the detector center. An absolute isolation cut of $\sum p_T < 5$ GeV for $|dz| < 1$ cm maintains the same or slightly lower efficiency and rate with respect to the non-isolated muons, as shown in Fig. 3.52. In this figure, the efficiency is defined as the number of generated muons matched within $\Delta R = 0.3$ to a L1 track with $p_T > 10$ GeV having at least two associated muon stubs. The rate is measured with a high pileup minimum bias sample.

Figure 3.52: Performance for isolated and nonisolated TT-correlated muons in a $W \rightarrow l\nu$ sample with high pileup. Left: Efficiency as a function of the generated muon $p_T$. Middle: Efficiency as a function of the generated muon $\eta$. Right: Rate as a function of the L1 muon $p_T$ threshold.
As mentioned in the introduction, the identification of HSCPs by the L1 muon trigger is a new feature that can be exploited during the Phase-2 era, expanding significantly the CMS physics program. The HSCPs considered here look like muons originating from the interaction point, but their speed ($\beta \cdot c$) can be significantly smaller than the speed-of-light. Thus, the associated muon stubs can occur a few bunch crossings later than stubs originating from ordinary muons. Therefore, the muon trigger architecture design needs to allow for enough bandwidth so that all stubs from these additional bunch crossings can be transmitted to the GMT.

To reconstruct such slow particles, the algorithm has to match a L1 tracker track with muon stubs not only in the bunch crossing that corresponds to the L1 tracker track under consideration, but also in several later bunch crossings. However, the probability of tagging wrongly the L1 tracker track as a muon or HSCP candidate increases, leading to a significantly higher trigger rate. Therefore, the proposed algorithm checks if the timing of the muon stubs matched by the L1 tracker track plus muon-stubs algorithm could correspond to a specific hypothesis of the particle velocity. For each L1 tracker track, 16 $\beta$ hypotheses are tested in parallel. A similar approach as used in the Bayesian correlator algorithm presented above is used. The log-likelihood of the velocity $\beta$ given the observed set of the times $t_i$ (relative to the L1 tracker track bunch crossing) registered in $n$ muon layers for a L1 tracker track with pseudorapidity $\eta$, is:

$$\log P(t_1, ..., t_n \mid \beta, \eta) \cong \log(P(t_1 \mid \beta, \eta) \times ...P(t_n \mid \beta, \eta))$$

(3.10)

This log-likelihood is calculated for several bins of $\beta$ and the bin with the largest log-likelihood is chosen. An appropriate cut on the obtained likelihood can be used to reduce the rate of the fake candidates. The version implemented here uses 16 bins in $b = (\frac{1}{b} - 1) \cdot 4$. The values of the log $P(t_i \mid b, \eta)$ are found from simulation and are stored in look-up tables. The muon stubs from 3 additional BX’s are used, which allows to find the HSCPs with $\beta$ down to 0.2. The muon candidates are assigned to one of two exclusive types: (1) “muon”, if it has at least two matched muon stubs originating from the same BX as the track, and (2) “HSCP”, otherwise. With this approach, the candidates in the “muon” type are the same as those generated using the L1 tracker track matching to muon-stubs algorithm, with the muon stubs coming only from the BX’s corresponding to the L1 tracker track, and can be used as ordinary muons in the Global Trigger.

Figure 3.53 shows the relationship between $(\frac{1}{b} - 1) \cdot 4$ and the timing measured in the RPC chambers. Also shown, are the trigger rates as a function of HSCP $p_T$ for different selection criteria and the corresponding algorithm efficiencies as a function of $\beta$. The magenta curve on the rate figure shows the rate of the HSCP type candidates without applying any additional cuts, for the Bayesian L1 tracker track plus muon-stub correlator algorithm described above. To reduce this rate, cuts on the tracker track $\chi^2$, the sum of the log $P_0$ and the $\beta$ log-likelihood can be applied, resulting in the behavior shown by the blue curve. The obtained rates for HSCP candidates are very reasonable in both cases, below 10 kHz for $p_T > 7 \text{ (20) GeV}$ when the cuts are (not) applied. As shown by the blue curve in the middle of Fig. 3.53, efficiencies above 40% and up to 100% are obtained for 0.2 < $\beta$ < 0.6 for HSCPs (from SUSY $\tilde{\tau}$) that satisfy these cuts. The red curve is for the Bayesian track plus muon-stub correlator algorithm. The combination of both algorithms results in very high efficiency, near 100%, for $\beta > 0.3$.

5. Muons in topological triggers - muon jets

The three muon reconstruction algorithms described in this Chapter, the standalone, TkMuon,
and TkMuStub, all have an intrinsic minimum \( p_T \) requirement. To reconstruct standalone muons, stubs need to be present in several different stations, which only muons above a certain \( p_T \) can generate. For the TkMu and TkMuStub algorithms a matched tracker track is required, which is reconstructed only with \( p_T \) above 2 GeV at L1.

Soft muons with \( p_T < 2 \) GeV and muons in the \( 2.4 < |\eta| < 2.8 \) region (where L1 tracker tracks are non-existent) are not reconstructed by any of these algorithms. These low \( p_T \) or forward muons can only be reconstructed at the HLT or offline stages by matching muon stubs to L1 tracker tracks that go as low as a few hundred MeV in \( p_T \) and populate the full rapidity range of the muon detectors. This type of reconstruction significantly benefits processes such as \( \tau \to \mu \mu \mu \), where tau leptons originate from \( D_s \) meson decays. In this Lepton-Flavor-Violating (LFV) process, the three muons are collimated, and produced predominantly in the high \( |\eta| \) endcap region, and with low \( p_T \) (below 2 GeV).

A possible way to relax the minimum \( p_T \) requirement at Level-1 for these very soft muons would be to reconstruct them as individual stubs. However, this would result in an unacceptably high trigger rate. To target very low-\( p_T \) and topologically correlated muons, as in \( \tau \to \mu \mu \mu \), we define a new trigger object type, referred to as “muon-jet”. These objects are formed in the Global Muon Trigger and, in this case, are defined as a group of three muon stubs found in the first station of the muon detector and within a cone of a specified size. To further control the muon jet trigger rate, one can require one or more of the individual stubs to be matched to a tracker track and form a part of a TkMuStub object.

The properties of a muon jet are derived from the properties of its constituent muon objects and their spatial and kinematical correlations. For muons that are reconstructed as standalone stubs, the stub bending angle is used to estimate both the muon \( p_T \) (as shown in Fig. 3.54) and the value of the azimuth angle of the muon at the vertex. For the TkMuStub muons these quantities are taken from the matched tracker track. With all three constituent objects characterized by curvature, \( p_T \), \( \phi \) and \( |\eta| \), we can, for example, reconstruct the muon jet total charge, its invariant mass, its width in azimuth, its pseudorapidity, and its \( p_T \). In case two TkMuStub objects are contained in the muon jet, we can also reconstruct the separation of the tracks along the \( z \)-coordinate at the collision vertex. These correlation variables can then be used to discriminate against the uncorrelated background.

In Section 4.3.2 we present an analysis of the \( \tau \to \mu \mu \mu \) process originating from \( D_s \)-meson decay at the L1 trigger level with muon jets consisting of individual stubs and TkMuStub objects. There are ongoing studies with muon jets that use different combinations of TkMu and
3.3. Triggering on muons

Figure 3.54: Profile of generator-level muon $p_T$ versus bending angle in simulated stubs reconstructed in the ME1/1 station of the CSC muon detector. The color lines are two empirically derived functions used to describe the relationship of $p_T$ and bending angle in the low-$p_T$ range: 1st-order rational function (in red), exponential times 2nd-order polynomial function (in blue).

TkMuStub objects as a possible future extension of this work.

Similarly, L1 muon jet objects are being studied for $\tau \to \mu \mu \mu$ decays, where the tau lepton originates from W-boson decays. These decays have different event kinematics to that of the $\tau \to \mu \mu \mu$ process in $D_s$ meson decays. The tau leptons are produced with higher momentum and nearly uniform pseudorapidity, allowing the use of muon trigger objects from all muon detector regions. The higher tau $p_T$ results in a cleaner signal relative to the $\tau \to \mu \mu \mu$ search in $D_s$ meson decays. Preliminary studies have shown that there is significant acceptance times cross section for this process, as well as good reconstruction efficiency and manageable rate. These studies are expected to provide additional sensitivity to the search for this LFV process.

GMT summary and outlook

We have demonstrated that the new capability to combine L1 tracker tracks with reconstructed muons and/or individual muon stubs significantly enhances the possibilities for efficient, robust, and innovative triggers in the challenging HL-LHC conditions. The muon $p_T$ resolution improves significantly resulting in sharp efficiency turn-on curves that plateau at very high efficiency (> 95%), and in single muon trigger rates that are very manageable (of order 10 kHz) for muons with $p_T > 20$ GeV. Moreover, matching L1 tracker tracks to individual muon stubs improves significantly the efficiency in fiducial regions with gaps between muon detectors, while also maintaining reasonable trigger rates. We are also able to generate single muon triggers, in which the muon is isolated from other tracker tracks, with > 95% efficiency and rates of a few kHz at $p_T$ of 20 GeV. The possibility of correlating the timing of muon stubs over several bunch crossings at the GMT will provide CMS with a trigger for particles that travel at velocities considerably less than $c$, such as HSCPs. Overall, we obtained efficiencies near 100% for $0.3 < \beta < 1.0$ with a trigger rate of about 10 kHz for $p_T > 8$ GeV. Finally, we demonstrated the possibility of combining the various trigger objects that can be formed at the GMT to topologically identify, and trigger on, interesting physics processes such as lepton-flavor violation in $\tau \to \mu \mu \mu$ from $D_s$ meson or W-boson decays. Many other such triggers, such as multi-muon jets, muon-stub clusters from displaced jets, $B_s \to \mu \mu$ etc. could also be readily implemented.
at the GMT. It is therefore important to ensure a flexible architecture and large FPGA resources in this new promising element of the CMS L1 trigger for Phase-2.
3.4 Vertex reconstruction

Within the L1 trigger, the identification and reconstruction of the primary vertex from tracks is crucial for the mitigation of pileup and the efficient use of resources. The goal here is not to identify every interaction within the event, as it would be during offline event reconstruction, but to identify the hard scatter “primary” vertex and its associated tracks or particle-flow (PF) candidates. The reconstructed vertex with the largest sum of track $p_T$ is taken to be the primary vertex. By selecting for the tracks and/or candidates associated to the primary vertex, the resources for the downstream algorithms will be targeted toward the interactions resulting from the hard scatter. To this end, many vertexing algorithms, including those described here, can be broken into three stages: (i) selection of input tracks, (ii) clustering of the tracks that originate from the same interaction, and (iii) a fitting procedure for extracting the vertex parameters.

The vertex reconstruction begins as soon as the L1 tracks are available. The algorithm executes in parallel to the PF algorithm (see Section 3.5.3.1), but on separate hardware. Once its reconstruction is complete, the primary vertex is used by several downstream algorithms. This naturally leads to algorithmic and architectural constraints which are discussed in Section 5.4, which describes the architecture of the Global Track Trigger (GTT), the system which contains the vertexing. In addition, the GTT contains several of the vertex consuming algorithms. Other systems which make use of the reconstructed primary vertex will be described in Sections 3.5.3.2 and 3.7.

3.4.1 Histogram-based vertexing algorithm

One of the simplest vertex reconstruction techniques is a histogram-based algorithm dubbed "FastHisto", which was first reported in the Technical Proposal [7]. This algorithm is based on two key principles: (i) that the tracks associated to each vertex will have a separation along the $z$ axis while the individual vertices will have a comparatively large separation and (ii) that the primary vertex will be the interaction with the highest scalar sum of the track $p_T$. Furthermore, the FastHisto procedure will follow the three basic stages introduced at the start of this Section.

In principle, the tracks used as input to this algorithm will first pass through a selection process to remove fake tracks and select for prompt tracks. However, the current studies make use of all tracks sent out by the track finding system in order to maintain a high vertex finding efficiency. Once selected, the tracks are sorted into bins based on their position along the $z$ axis, a procedure analogous to forming a histogram. Each entry can be weighted by the track $p_T$ (or possibly other quantities). The algorithms being considered use the $p_T$ weighting scheme and thus the primary vertex position is then determined by the $p_T$ weighted average of the $z$ position of the $b$ bins with the maximum scalar sum $p_T$, where $b$ is a configurable parameter.

Fig. 3.55 shows how the FastHisto algorithm would work in a simulated $t\bar{t}$ event with 200 pileup interactions. The density of tracks from the primary vertex (PV) and their relatively high $p_T$ make the bins around the primary vertex stand out from the background of tracks from pileup.

An alternative algorithm is the unseeded, inherently iterative density-based spatial clustering of applications with noise (DBSCAN). However, it was found to be unsuitable for use within the Level 1 Trigger due to its long latency (at best several tens of microseconds). Further discussion of this algorithm can be found in Ref. [53].
Figure 3.55: A $p_T$ weighted histogram of the track positions along the $z$ axis. The gray markers show the scalar sum of the track $p_T$ within each bin, while the colored crosses indicate the actual position of each track. The different colors for the crosses indicate the source of each track: blue for tracks originating from a pileup interaction, red for fake tracks, and green for a track from the primary vertex. The black dashed line shows the vertex position found by the FastHisto algorithm, while the green dashed line shows the true vertex position.

3.4.2 Track to vertex association

The FastHisto algorithm only coarsely clusters the tracks during the histogram forming step within fixed bins. After the vertex position is computed, as described in Section 3.4.1, tracks can be assigned to the vertex more precisely based on the fitted PV $z_0$, and the $z_0$ positions of the tracks. The most simple procedure for this assignment is to associate each track within a fixed distance from the reconstructed PV position. A more advanced, multivariate technique was also developed to associate tracks to the PV on a per-track basis. The $\chi^2$, $|\eta|$, and $p_T$ of the track, as well as $|dz|$, the difference between the track $z_0$ and reconstructed PV $z_0$, were used as input variables to a boosted decision tree (BDT). The BDT was trained to classify the association, or not, of a track to the PV. By introducing the additional track parameters, which are known to correlate with the track $z_0$ resolution, more selective association of tracks to the PV is enabled. The performance of these association techniques is presented in Section 3.4.3.

3.4.3 Vertex finding performance

Algorithms for vertex reconstruction are evaluated on two aspects: firstly, on the quality of the regression (the ability to precisely locate the $z_0$ of the primary vertex), and secondly, on the correct association of tracks to the primary vertex. Figure 3.56 shows the distribution of the difference between the reconstructed primary vertex and the true vertex of the hard interaction in $t\bar{t}$ and in $Z \rightarrow \mu\mu$ events. A resolution of a few mm is achieved in $t\bar{t}$ events, with the resolution being a bit worse in $Z \rightarrow \mu\mu$ events.

Track-to-vertex association performance is measured with the efficiency and purity of the reconstructed primary vertex. In this context, the primary vertex efficiency is the fraction of reconstructable tracks from the Monte Carlo primary vertex which is successfully attributed to the reconstructed primary vertex by the algorithm. An efficiency of 1 is achieved when all of
the tracks originating from the PV in Monte Carlo are found in the reconstructed PV, otherwise it is less than 1. The purity is then the fraction of the tracks associated to the PV by the algorithm which truly originate from the Monte Carlo PV. A reconstructed PV with only PV tracks from the Monte Carlo will have a purity of 1, whereas the presence of pileup, fake tracks, or tracks from another vertex will result in a purity less than 1.

An efficiency vs. purity ROC curve is shown in Fig. 3.57 for several combinations of regression and association algorithms. The ROC curves are produced by reconstructing the PV $z_0$ with the labelled regression algorithm, followed by scanning the tunable parameter of the association algorithm. For the “Fixed Window” (FW) association, this is done by changing the size of the window around the $z_0$ inside which all tracks are associated to the PV. For the BDT association, the probability threshold at which tracks are assigned to the PV is adjusted. When using the BDT association, a BDT was trained separately for each regression algorithm. In all cases, as the matching efficiency increases, the purity is degraded as more non-PV tracks are associated to the reconstructed PV. The BDT association performs much better than the FW algorithm. Application of the vertexing algorithm, described in Section 3.4.5, that makes use of a convolutional neural network (CNN) would improve the performance even further. The improvement to matching performance brought by the BDT compared to a Fixed Window can be understood by its ability to select tracks with a $z_0$ far from the reconstructed PV, but with other properties making it likely to originate from the PV, and simultaneously to deselect tracks which are close to the reconstructed PV in $z_0$, but with other properties making it less likely to originate from the PV.

Differences in vertex finding performance across different samples, as seen in Fig. 3.56 and in Fig. 3.57, are understood to arise from the different topologies and corresponding quality of track reconstruction. Events of $t\bar{t}$ production, for which the vertexing is seen to perform the best, are characterised by central jets of high multiplicity. The resolution of track parameters, especially $z_0$, is best for central tracks. Together with the high multiplicity, this yields sufficient
Figure 3.57: Track-to-vertex association performance ROC curves in $t\bar{t}$ (left) and $Z \rightarrow \mu\mu$ events (right), both with 200 PU. Each curve corresponds to a different combination of regression and association algorithm. The performance obtained with the FastHisto algorithm is shown by the blue curves, when the association is performed with a fixed window (FW) or a BDT. For reference, the performance obtained when the association uses the true (“Gen”) PV position is also shown. The orange curves (“CNN”) illustrate the performance of a potential algorithm using a convolutional neural network, that will be described in Section 3.4.5.

$p_T$ density across a small region of the $z$ axis that the simple FastHisto can generally reconstruct these vertices well. $Z \rightarrow \mu\mu$ events, on the other hand, have relatively fewer tracks originating from the primary vertex. Such events are harder to reconstruct well with FastHisto, which relies on these tracks’ reconstructed $z_0$ falling within a narrow range of bins. If one of the muons is reconstructed with a large $z_0$ residual this vertex may be lost due to the large separation of the reconstructed muons along the $z$ axis compared to the histogram bin size.

Figure 3.58 shows the ranking of vertices containing muons found in $Z \rightarrow \mu\mu$ events with 200 pileup (the vertex with rank = 0 being the one reconstructed with the largest total $p_T$). In 58% of events, both muons are associated to the same vertex. As can be seen in the left panel of the figure, in the vast majority of these events, that reconstructed vertex was the one with the highest $p_T$, and hence the primary vertex. For the 34% of events where both muons were reconstructed to vertices, but not the same vertex, the right panel of the figure shows the distribution of rank of the vertex in which the second (lower $p_T$) muon was found. In the remainder of events, either one or both muons were not reconstructed. This was either because they were not within the tracker acceptance, or were not reconstructed by the track finder, which reconstructs muons with 98% efficiency [8]. Matching performance in these types of events is seen to be greatly improved with the use of the BDT track-vertex association in Fig. 3.57.

Another possibility for recovering vertexing performance in topologies that are difficult to reconstruct is to consider multiple primary vertex hypotheses and compute PUPPI weights (see Section 3.5.3.2) for each vertex candidate. From the right side of Fig. 3.58 it can be seen that keeping just one or two vertices in addition to the PV would capture the vast majority of PV particles in these low multiplicity events. The added latency required to acquire multiple vertices is minimal.
Figure 3.58: The rank, when ordered by total $p_T$, of reconstructed vertices containing either both muons (left), or just one muon (right) in $Z \rightarrow \mu\mu$ events with 200 pileup. In 58% of events, both muons are reconstructed in the same vertex, contributing to the left plot. In 34% of events, both muons are reconstructed, but to different vertices.

### 3.4.4 Firmware implementation of vertexing algorithms

Two implementations of the FastHisto algorithm were studied in hardware. One implementation forms separate histograms for each stream of input data and then uses a merge stage to combine the histograms. This allows the histograms in the first stage to act as buffers to handle the large influx of tracks and to prevent address collisions when tracks fall within the same bin. The algorithm can find the single highest $p_T$ sum using a sequential reduction stage or it can return the top N vertices using parallel sorting stages plus a linear search. The other implementation uses custom routing logic to perform the track clustering and to form a highly granular histogram. It also makes use of a sequential reduction algorithm for finding the bin with the highest $\sum p_T$. This implementation is extremely resource efficient. Sections 3.4.4.1 describes the first implementation, while 3.4.4.2 describes the second.

#### 3.4.4.1 FastHisto with track clustering using sequential merge logic

A schematic view of the FastHisto architecture that uses a sequential merge logic to cluster the tracks is illustrated in Fig. 3.59. The firmware takes as input two streams of 96-bit tracks from each of the nine track finding regions (18 streams in total). The tracks will come unordered in both $z_0$ and $p_T$. Because of this, no preselection based on track $p_T$ can be made and all of the tracks will need to be accepted to remove the risk of dropping a high-$p_T$ track at the end of the TMUX period. Each track will need to be placed in a histogram according to its $z_0$ position. The histogram will consist of 72 bins, which corresponds to a width of 0.4 cm per bin, spanning a region of $-14.4 < z < 14.4$ cm around the center of the detector.

Each clock cycle a new track will be read in, which means that all of the operations needed to place a track in a bin and compute the sum of its $p_T$ and the already accumulated value must take place within one clock cycle. In order to accommodate this time requirement, a histogram is formed for each stream and the accumulate operations are handled in parallel. Once all of the tracks for a given event have arrived, the 18 histograms are merged into a single histogram.
Figure 3.59: Diagram of the FastHisto implementation which makes use of the sequential merge logic to do the track clustering. The tracks arrive from the track finding boards at the left and the vertex, or vertices, are output at the right. Several parameters are output for each vertex. This process is illustrated by the “Track Clustering” section of Fig. 3.59.

The next part of the FastHisto algorithm is to compute the bin, or set of bins, with the highest combined $p_T$. This “Peak Finding” section is sometimes referred to as a “sliding window algorithm.” Given the window parameter $b$, there will be $72 - b$ sums to compute. Once that is accomplished, the maximum value of this unordered list must be determined as this will be the bin containing the primary vertex.

To find the maximum value, a sequential reduction algorithm is used. This algorithm does pairwise comparisons of every two entries in the list, keeping only the highest value. At each stage the number of comparisons is cut in half. The number of stages needed to find the highest value goes as $\log_2(N_{\text{entries}})$. In order to simplify the logic, the list of sums is always padded by zeros such that its size is a power of two. In the case of 72 bins and a window size of 3, hence of 69 sums, this means that the initial list will have 128 entries (the first integer larger than 69 that is a power of two) and it will take 7 clock cycles to return the maximum value. A diagram of this procedure is shown in Fig. 3.60.

While this sort of sequential reduction algorithm is ideal for returning a maximum or minimum value, it is not suited for returning the top $N$ values. For this, a different algorithm was implemented, starting from the list of entries which has a size equal to a power of 2. This list is broken into four lists of equal size. The number and size of each list is again a variable parameter of the algorithm. Each list is then sorted using a nonrecursive, bitonic sorting algorithm. From there, the top four values are compared, the highest being returned and then removed from its list. This comparison, return, and remove loop is repeated until the desired number of vertices is returned. An example of the second to last stage of the sort and select algorithm is shown in Fig. 3.61.

The implementation used for the “Peak Finding” section of the FastHisto algorithm depends on the number of vertices which will be returned. The resource usage and timing of the two different implementations were studied. Figure 3.62 shows the resource and latency comparison between the two peak finding algorithms. For one vertex, the sequential reduction algorithm, here called the “ParallelFindMax” algorithm, is preferable due to its low latency and vanishingly small resource usage. However, for anything over two vertices the “SortAndSelect” al-
3.4. Vertex reconstruction

Figure 3.60: A sequential reduction algorithm where at each stage the number of entries in a list is reduced by half based on a comparison criteria, in this case keeping the largest of two values.

Figure 3.61: An example of a sort and select algorithm which can find the top N highest, or lowest, values.

gorithm, described in the previous paragraph, is preferable. For both algorithms, the original histogram of 72 bins was copied into an array of 128 entries for using in the algorithms as this corresponds to the next highest power of two. For the SortAndSelect algorithm, the smaller arrays used to sort the entries had a size of 32. This was tuned based on the resource usage and latency of the sort.

Overall, the FastHisto firmware is very lightweight in resource usage, as seen in Table 3.3. Infrastructure resources used by the GTT demonstrator framework are not included. The LUTs and Flip-Flops are mostly used for making and merging the histograms from each input link, with a small amount used by the vertexing portion of the algorithm. The BRAM and DSP are used by the vertexing. The resource utilization is low enough that multiple copies of the algorithm could be used in parallel in one FPGA.

Table 3.3: Resource usage for the implementation of the FastHisto algorithm with sequential merge logic, not including board infrastructure.

<table>
<thead>
<tr>
<th>Resource</th>
<th>LUTs</th>
<th>Flip Flops</th>
<th>BRAMs</th>
<th>DSPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Used</td>
<td>113566</td>
<td>32350</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Percentage of VU9P</td>
<td>9.6</td>
<td>1.4</td>
<td>~0</td>
<td>~0</td>
</tr>
</tbody>
</table>
Figure 3.62: The resource usage (left) and timing (right) of the two peak finding algorithms. The ParallelFindMax algorithm (red) is preferable for single vertex selections; the SortAndSelect algorithm (blue) is used to return multiple vertices.

Because this algorithm makes use of every track in the event, the firmware must accept tracks for the entire time-multiplexing period of 18 bunch crossings, or 450 ns. After that the peak finding section of the algorithm adds 10 clock cycles. For demonstration purposes this was clocked at 180 MHz, yielding 56 ns.

### 3.4.4.2 FastHisto with track clustering using routing logic

The FastHisto algorithm was also implemented in a firmware that uses a routing logic to cluster the tracks, with an architecture illustrated in Fig. 3.63. The firmware is designed to accept 27 input streams of tracks from the track finder (3 per nonant), which are initially unordered in $z_0$, and distribute them within the chip according to their $z_0$ to 256 histogram bins. This distribution is performed with a purpose-built ‘router’, which can also be found in the track finder and HGCAL firmware. The 256 bins corresponds to a width of 1.17 mm per bin, spanning the $-150 < z < 150$ mm luminous region.

Figure 3.63: Schematic architecture of one implementation of the FastHisto algorithm for vertex finding, with a routing logic to cluster the tracks. Tracks arrive from the track finder at the left, and the vertex $z_0$ is output at the right.

The purpose of the router is to perform an ‘any-to-any’ distribution of track inputs to $z_0$ bin outputs in several sequential steps to minimize the distance any data has to move in a single clock cycle. BRAMs are used to buffer data while control logic determines the data routing and...
3.4. Vertex reconstruction

performs arbitration when multiple tracks are destined for the same output. The use of the router enables a high clock frequency to be targeted, by avoiding the complicated routing that might otherwise be required.

Histogram bin logic is placed at each output of the router. These bins accumulate the $p_T$ of each arriving track. At the end of the time-multiplex period the accumulated $p_T$ in each bin is its weight in the histogram. The “Peak Finder” logic then compares pairs of bins and propagates the higher $p_T$ bin, with a “tree reduce” structure, such that the highest $p_T$ bin of the histogram is output.

Just as for the previous implementation, the FastHisto firmware is very lightweight in resource usage, as seen in Table 3.4. Infrastructure resources used by the extensible, modular data processor (EMP) framework, which can be found in Section 6.3.2.2, are not included. BRAMs are used exclusively by the router, while LUTs and Flip Flops are used for the router control logic, the bin accumulation, and peak finding. Given the low utilization, it would be possible to operate multiple copies of the algorithm in parallel in one FPGA.

Table 3.4: Resource usage of one implementation of the FastHisto algorithm for vertex finding, with a routing logic, not including board infrastructure.

<table>
<thead>
<tr>
<th>Resource</th>
<th>LUTs</th>
<th>Flip Flops</th>
<th>BRAMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Used</td>
<td>26683</td>
<td>32128</td>
<td>42</td>
</tr>
<tr>
<td>Percentage of KU115</td>
<td>4.0</td>
<td>2.4</td>
<td>1.9</td>
</tr>
<tr>
<td>Percentage of VU9P</td>
<td>2.4</td>
<td>1.4</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Since every track contributes to the histogram, the firmware must accumulate tracks for the entire time-multiplex period before the peak bin can be known. With tracks arriving with a TM Period of 18, this is 450 ns. In addition to this accumulation, the pipelined algorithm logic adds 20 clock cycles. For the demonstration this was clocked at 240 MHz, yielding 83 ns. The algorithm is sufficiently pipelined to meet the timing constraints for up to 480 MHz, which would result in a 42 ns latency.

3.4.5 End-to-end neural network vertex reconstruction model

The concepts of filtering input tracks, binning them along the z axis, performing regression on the histogram to reconstruct the vertex $z_0$ position, and finally to associate tracks to the vertex, are combined in an approach using end-to-end machine learning as a potential improvement to the vertexing performance. This end-to-end model, illustrated in Fig. 3.64, uses a 1D convolutional neural network (CNN) to assign weights to each input track, accumulate those tracks in a histogram and use another set of convolutional layers to perform peak finding within the histogram. The training of such an end-to-end model required the implementation of a histogram that is differentiable, which enabled the optimization of the weight generation CNN against the overall objective of minimal separation between the true and reconstructed vertex position. The model can also be extended with a set of neural layers that also perform the track association which would allow for the training against the combined objective of minimal vertex-position residual and best track association performance.

The first functional block of the end-to-end model is the CNN used for generating weights for each input track before it is stored in the histogram. The network was trained using up to 250 tracks per event with track $p_T$, $\eta$ and $\chi^2$ used as inputs. The input volume was zero padded if less than the maximum number of tracks were available and the track parameters were encoded as “color” channels in the input.

The end-to-end training of the model was enabled by the construction of a custom histogram
module for Tensorflow [54] whose contents can be differentiated with respect to variation in
the network weights so as to enable back-propagation up to the inputs of the network. The
histogram is configured to have 256 bins of 1 mm width and represent the \( z_0 \) position of tracks.
The contribution of each track was determined using a kernel density estimation (KDE) with a
triangular kernel with very narrow bandwidth.

The peak finding layers of the CNN scan over the bins of the histogram with 4 bin wide kernels
and produce 16 channels for each layer. The resulting regression performance of the model can
be seen in Fig. 3.56, where over 30% improvement relative to FastHisto can be seen as measured
by the half-width of the 68% region in the \( t\bar{t} \) sample. In the more challenging \( Z \rightarrow \mu \mu \) events,
the resolution of the PV position does not improve compared to FastHisto, however, the track-
to-vertex matching performance is improved by the CNN, as seen in Fig. 3.57.

### 3.4.5.1 End-to-end ML algorithm firmware

The full end-to-end model has not yet been implemented in firmware, however it was designed
with future FPGA implementation in mind. The track weight generation model is small, and
can be used to generate weights for tracks in parallel as they arrive from the track finder. The
choice of 256 bins and histogram accumulation means that these track weights can be directly
input into the FastHisto implementation described in Section 3.4.4.2. Following this, the im-
plementation of the larger CNN which performs the position regression can be tuned with
the tradeoff between latency and resources enabled by the hls4ml package [44]. Since this re-
gression need only be performed once per event, a large “Initiation Interval” with significant
reuse of FPGA resources should be achievable, as long as the latency can be controlled. The
implementation of this full algorithm in firmware will be pursued in further work.
3.5 Particle-flow reconstruction

The particle-flow approach has the goal of separately identifying and reconstructing all individual particles produced in each event, with an optimal combination of information from all sub-detectors. This approach has been adopted for the CMS event reconstruction offline and in the HLT already since LHC Run-1, yielding substantial improvement in the physics performance, especially for the reconstruction of hadronic jets, missing transverse momentum and hadronically decaying tau leptons, and as input to pileup mitigation strategies [55].

Two necessary ingredients for the particle-flow algorithm are an efficient reconstruction of charged particles in the tracker and a fine granularity calorimetry to resolve the contributions from neighboring particles. The Phase-2 upgrade of CMS will for the first time make these two ingredients available at the L1 trigger, allowing the algorithm to be implemented in the Correlator Trigger. For the implementation, a single grouping of boards, forming the Correlator Layer-1, will be responsible for the particle-flow reconstruction along with the application of pileup mitigating algorithms coming from vertexing and Pile Up Per Particle Identification (PUPPI) [56]. A second layer of the correlator, Correlator Layer-2, is intended to be used for algorithm reconstruction based on particle-flow candidate inputs. A functional diagram of the design is shown in Fig. 1.3. To fit latency requirements, the Correlator Layer-1 must receive detector inputs, reconstruct particles, and transmit the resulting candidates to the Correlator Layer-2 within 1 µs.

An implementation of the particle-flow algorithm within the latency and resource constraints of the L1 trigger system is possible due to the local nature of the algorithm. The identification and reconstruction of a particle can be performed relying only on detector input objects in the vicinity of the particle, allowing the detector to be subdivided in regions that can all be processed separately. The limited size of the regions reduces the multiplicity of the input objects, and thus the combinatorial complexity of the algorithm. The possibility of processing multiple regions in parallel, also on different boards, reduces the latency for the processing of each event, and the resources needed on each individual board.

To perform the particle-flow reconstruction inputs are streamed into a buffer on the FPGA. As part of the streaming process, the inputs are rearranged to correspond to specific rectangular regions in $\eta$ and $\phi$. These regions are then fed into the particle-flow algorithm, which performs the linking of the detector objects iteratively on a region by region basis. The local nature of the particle-flow algorithm allows for this process to be parallelized separately onto many FPGAs, with each FPGA processing a different set of geometric regions. Full details about the linking procedure are discussed later in Section 3.5.3.1.

3.5.1 Calorimeter inputs

As described in Section 5.5.1, the full detector is divided into 7 regions in $\eta$. These regions are further subdivided into sub-regions as shown in Table 3.5. Additional overlaps of size 0.25 in $\eta$ and $\phi$, are defined such that the total size of each sub-region is larger than an equal partition of the corresponding region, as illustrated in Fig. 3.65. The overlaps are necessary to make sure that for each input object in the sub-region proper (i.e., not in the overlap area), any other input object it may have to be linked to in the particle-flow reconstruction is available in the extended sub-region (i.e., including the overlap area).

In order to determine the number of input objects to consider in each region for the PF algorithm, we study a high-multiplicity process, namely $t\bar{t}$ events with 200 pileup. We define the input capacity for each type of object as the minimum number necessary to prevent truncation
of the objects in at least 95% of these events. In this way, we ensure that we consider a sufficient number of input objects per region for the vast majority of events and regions for typical processes.

![Image of detector regions partitioned in sub-regions with overlaps]

Figure 3.65: Illustration of how detector regions are partitioned in sub-regions, with overlaps. First, the region is subdivided in nonoverlapping sub-regions in $\eta, \phi$ (left). Then, each sub-region is enlarged by adding a boundary in $\eta, \phi$ that overlaps with the other sub-regions (right); to preserve readability of the figure, the overlaps are shown only for three of the sub-regions.

Table 3.5: Summary of the input regions to the particle-flow algorithm. Additional overlaps of size 0.25 in $\eta$ and $\phi$ are included in the sub-region sizes.

<table>
<thead>
<tr>
<th>Region</th>
<th>Coverage</th>
<th>Number of regions $(\eta \times \phi)$</th>
<th>sub-region size $(\eta \times \phi)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barrel (regions 1, 2, 3)</td>
<td>$</td>
<td>\eta</td>
<td>&lt; 1.5$</td>
</tr>
<tr>
<td>Endcap with tracker (regions 4, 5)</td>
<td>$1.5 &lt;</td>
<td>\eta</td>
<td>&lt; 2.5$</td>
</tr>
<tr>
<td>Endcap without tracker (region 6)</td>
<td>$2.5 &lt;</td>
<td>\eta</td>
<td>&lt; 3.0$</td>
</tr>
<tr>
<td>Forward (region 7)</td>
<td>$3.0 &lt;</td>
<td>\eta</td>
<td>&lt; 5.0$</td>
</tr>
</tbody>
</table>

3.5.1.1 Barrel clusters

Input ECAL clusters are taken from the global calorimeter trigger without any additional identification applied. A calibration is applied to the cluster energy, as a function of $p_T$ and $\eta$, derived from simulated photons and neutral pions. The calibrated clusters are used as input to the particle-flow algorithm, provided that they have $p_T > 0.5$ GeV. The energy response and resolution for photons before and after the calibration is shown in Fig. 3.66.

Input calorimeter towers are constructed from HCAL energy and ECAL energy not included in ECAL clusters. Input towers are clustered by first finding local maxima and then adding the energy in the eight surrounding towers to form $3 \times 3$ clusters. These clusters are then linked to the pure ECAL clusters, and the resulting combined clusters are calibrated using charged pions as a function of $p_T$, $\eta$, and $E_{EM}/E_{\text{cluster}}$ where $E_{EM}$ is the electromagnetic energy of the cluster and $E_{\text{cluster}}$ is the total cluster energy. These combined clusters are used as input to the particle-flow algorithm, provided that they contain sufficient energy (greater than 1 GeV). The calibration with $\pi^\pm$ is illustrated in Fig. 3.67. The multiplicities of ECAL clusters and combined clusters in $t\bar{t}$ events with 200 pileup is shown in Fig. 3.68 as function of $|\eta|$, for different $p_T$ thresholds. The distributions of the maximum number of clusters located in a single $1.5|\eta| \times 1.2\phi$ sub-region of the barrel per event is shown in Fig. 3.69. In order to prevent truncation of the clusters in at least 95% of these events, the PF algorithm requires a minimum input capacity of 15 ECAL clusters and 19 HCAL clusters.

An alternative design using only the combined clusters as input to the particle-flow algorithm has also been studied. This allows a single uniform firmware to be deployed across the full
3.5. Particle-flow reconstruction

Figure 3.66: Raw (green) and corrected (blue) response for photons in zero pileup as a function of $p_T$ for $|\eta| < 1.3$ (left) and $\eta$ for $p_T > 10$ GeV (middle) and the corresponding resolution as a function of $p_T$ for $|\eta| < 1.3$ (right). The vertical dashed line indicates the barrel-endcap transition.

Figure 3.67: Raw (red), electromagnetic-only (purple) and fully corrected (blue) response for $\pi^\pm$ in zero pileup as a function of $p_T$ for $|\eta| < 1.3$ (left) and $\eta$ for $p_T > 10$ GeV (middle) and the corresponding resolution as a function of $p_T$ for $|\eta| < 1.3$ (right). The vertical dashed line indicates the barrel-endcap transition.

Figure 3.68: $dN/d\eta d\phi$ in $t\bar{t}$ events with 200 pileup as a function of $\eta$ for ECAL (left) and combined (right) clusters used as input to particle-flow. Different colors represent different thresholds on the minimum $p_T$ required of the clusters.
Chapter 3. Trigger algorithms

Figure 3.69: Distributions of the maximum number of ECAL clusters (left) and combined clusters (right) located in a single sub-region of the barrel per event in $t\bar{t}$ events with 200 pileup. The maximum is evaluated over all sub-regions of the barrel event-by-event. Based on this event-by-event maximum, the PF algorithm requires a minimum input capacity of 15 ECAL clusters and 19 combined clusters in order to prevent truncation in at least 95% of these events.

The major difference in this design is that the hadronic towers are not clustered independently. Rather, the $1 \times 3 (\eta \times \phi)$ towers behind each ECAL cluster are added to the ECAL cluster, and then if the $H/E$ for the resulting cluster is above a threshold then the adjacent $1 \times 3$ towers are also added to create a $3 \times 3$ cluster. This method of clustering provides similar position resolution and momentum response and resolution as the previous method and has the added benefit of mirroring the input setup of HGCAL. This would allow for a common firmware and algorithm block across both calorimeter detectors.

A simple neural network is used to classify the combined clusters into three categories: those originating from pileup, $e/\gamma$, and $\pi^\pm$. The inputs to the network are those related to the ECAL cluster shape and energy, the $H/E$, and the cluster position and momentum. The network has two outputs and is trained to discriminate between clusters originating from pileup, from electromagnetic objects (electrons and photons), or from charged pions. One output node is designed to separate pileup from electrons, photons, and charged pions, and the other output node is designed to separate electrons and photons from charged pions. A combination of simulated $t\bar{t}$ and single particle events with 200 pileup is used to train the network. The ROC curves for the two nodes are shown in Fig. 3.70. The efficiency of the network to identify clusters as a function of momentum is shown in Fig. 3.71.

3.5.1.2 HGCAL clusters

Input HGCAL clusters are calibrated using photons, neutral kaons, and charged and neutral pions as a function of $p_T$, $\eta$, and $E_{EM}/E_{\text{cluster}}$, where $E_{EM}$ is the electromagnetic energy of the cluster and $E_{\text{cluster}}$ is the total cluster energy. The distribution of the maximum number of HGCAL clusters located in a single sub-region of the HGCAL per event in $t\bar{t}$ events with 200 pileup is shown in Fig. 3.72, separately for the region inside and outside of the track finder coverage. No additional cuts are placed on the cluster energy, yielding a minimum $p_T$ of roughly 1 GeV varying with $\eta$. The maximum is evaluated over all sub-regions of the endcap region event-by-event. Based on this event-by-event maximum, in order to prevent truncation of the clusters in at least 95% of these events, the PF algorithm requires a minimum input capacity of 15 HGC-
3.5. Particle-flow reconstruction

Figure 3.70: ROC curves for the neural network trained to discriminate between clusters from pileup, clusters from electrons and photons, and clusters from pions. The left plot is obtained from the output node designed to separate pileup from electrons, photons, and charged pions. The right plot is obtained using the output node designed to separate electrons and photons from charged pions. The black stars represent the chosen working points.

Figure 3.71: Efficiency to classify clusters in the barrel using a neural network as a function of $p_T$ in $t\bar{t}$ events with 200 pileup. Red (blue) points represent the efficiency when selecting on a discriminator between pileup and signal clusters from electrons (charged pions). Orange (purple) points represent the electron (charged pion) efficiency when using an additional selection on a discriminator between electrons and charged pions. In all cases the chosen working points are designed to achieve signal efficiencies of 90%.
CAL clusters per sub-region for $1.5 < |\eta| < 2.5$, and 10 HGCAL clusters per sub-region for $2.5 < |\eta| < 3$.

![Figure 3.72: Distribution of the maximum number of HGCAL clusters located in a single sub-region of the endcap region with (left) and without (right) tracker overlap per event in $t\bar{t}$ events with 200 pileup. The maximum is evaluated over all sub-regions of the endcap region event-by-event.](image)

Similar to the barrel clusters, machine learning algorithms have been designed to classify clusters in the HGCAL as originating from photons, pions or pileup. The classification is comprised of two BDT algorithms, that use variables related to the shower shape and profile of the 3D HGCAL clusters. One BDT algorithm is used to discriminate between those clusters originating from particles from the leading vertex and those from PU. The motivation behind this classification is to reduce the cluster multiplicity entering the particle-flow algorithm computations, which proves to be crucial for reducing the resource usage of the algorithm in the hardware. The reduction of the number of particles is achieved by rejecting the PU clusters and this has the added effect of reducing the workload of PUPPI in later computations. The other BDT is used to tag whether a cluster is of electromagnetic or hadronic origin, and can improve the linking of the HGCAL clusters to tracks done in PF algorithm. The photon efficiency and ROC curve for these BDTs are shown in Fig 3.73.

Taking advantage of the longitudinal information that the HGCAL provides, the first algorithm is capable of rejecting 90% of pileup while maintaining 95% signal efficiency. The efficiency of the algorithm when it comes to discriminating electromagnetic from hadronic clusters is up to 95% with 10% fake rate. The use of this combined classification in the particle-flow algorithm itself provides an improvement in the pileup rejection performance and leads to slightly decreased thresholds for $H_T$ and $E_T^{\text{miss}}$ triggers.

### 3.5.1.3 HF clusters

Input HF clusters are calibrated using charged and neutral pions as a function of $p_T$ and $\eta$. The calibration and resulting resolution is illustrated with charged pions in Fig. 3.74. Figure 3.75 shows the maximum number of HF clusters with $p_T > 15$ GeV located in a single sub-region of $1.5 \times 1.2$ in $\eta \times \phi$ in the HF per event in $t\bar{t}$ and VBF $H \rightarrow b\bar{b}$ events with 200 pileup. In order to prevent truncation of the clusters in at least 95% of these events, the PF algorithm requires a minimum input capacity of 14 HF clusters per sub-region.
3.5. Particle-flow reconstruction

Figure 3.73: The photon identification efficiency (left) and the ROC curve for both BDTs (right) trained to perform identification on clusters in the HGCAL, with example working points corresponding to 90% efficiency.

Figure 3.74: Raw (red) and corrected (blue) response for $\pi^\pm$ in zero pileup as a function of $p_T$ for $3.5 < |\eta| < 4.5$ (left) and $p_T > 10$ GeV (middle) and the corresponding resolution as a function of $p_T$ for $3.5 < |\eta| < 4.5$ (right). The reduction in raw response and increase in resolution at high $p_T$ is caused by tower saturation, which occurs at energies greater than 1024 GeV.

Figure 3.75: Distribution of the maximum number of HF clusters located in a single sub-region of the HF per event in events with 200 pileup, for $t\bar{t}$ (left) and $VBF H \rightarrow b\bar{b}$ (right). The maximum is evaluated over all sub-regions of the HF event-by-event.
Unlike in the barrel and in the HGCAL, information on shower shape and electromagnetic energy fraction is not available for the HF clusters, so they not classified using a machine learning algorithm. Instead, their $p_T$ is used as the main handle in rejecting pileup and reducing rates.

### 3.5.2 Tracking and muon inputs

The PF correlator receives in its input “standalone” muons (i.e., tracks reconstructed by the CMS muon chambers alone), as well as L1 tracks (i.e., tracks reconstructed by the CMS silicon tracker). A PF muon is then formed by associating the two. This is done by matching them geometrically, in $\eta$ and $\phi$, as well as in $p_T$. Once the matching is achieved, the L1 track is promoted to a PF muon and is not reused during the linking of calorimeter clusters with L1 tracks. The four momenta of the PF muon is inherited from the associated L1 track. The Global Muon Trigger is be used in this scenario since the latency of the muon trigger is too long and the increase in firmware complexity to do muon linking within the particle-flow is small.

The correlation of L1 tracks with standalone muons needs to take place with minimal latency, in the beginning of the PF. This is crucial to be done for two reasons. Firstly, high $p_T$ tracks, above the configurable threshold of 20 GeV, will be suppressed and classified as non-genuine (“fakes”) if they are found to remain unmatched in the end of the PF correlation. Secondly, in the case of muons within jets, significant disorientation of the measured energy flow will occur if calorimeter clusters are accidentally associated with a L1 track stemming from a muon. This is because the calorimeter cluster that would have otherwise given rise to the creation of a neutral PF candidate, instead accidentally forms a charged PF candidate. In this case, the energy absorbed by the calorimeter will be ignored in PF (in favor of the assumed four momenta of the associated L1 track) and that will result in an underestimation of the jet energy in the event. Studies with heavy flavor jets using HH → b$b\bar{b}$b simulation showed that without standalone muon information in the PF there would be a flat and constant 4–6% degradation of the energy resolution for all $p_T^{b\text{-jet}} > 30$ GeV.

All L1 tracks received from the track finder are used in PF, provided they meet the minimum requirements of having $p_T > 2.0$ GeV, $|\eta| < 2.5$ and at least 4 stubs with $\chi^2 < 15$. These cuts are a subset of those used for the L1 track-based jets and $E_T^{\text{miss}}$ computation (Table. 3.7). It is worth to emphasize that in PF, one can afford looser selection on the track quality criteria (and thus gain in efficiency and resolution) compared to those needed for track-only physics objects described later in Section 3.6. Protection against non-genuine tracks naturally arises by virtue of the PF matching of the L1 tracks with calorimeter clusters, which provides independent confirmation for the presence of a genuine charged particle. For the purpose of matching, the tracks are extrapolated to the calorimeter surface, a propagation that could be envisioned to take place in a dedicated LUT and further specialized per physics object (i.e., PF electrons, PF charged hadrons and PF muons) in the future. For the standalone muons, the same $p_T$ and $\eta$ restrictions apply and no further quality criteria are imposed (or relaxed) on top of those required by the Global Muon Trigger. It is in principle possible to correlate the corresponding PF muon track with a calorimeter cluster having energy consistent with that of a minimum ionizing particle, but this is not as of yet exploited in the present system and is left open as a possibility for the future.

The geometrical matching of the L1 track with the standalone muon is done in a rectangle having sides of $|\Delta \eta| < 0.2$ and $|\Delta \phi| < 0.2$, with $|\Delta \eta|$ and $|\Delta \phi|$ representing the corresponding pseudorapidity and azimuthal differences of the two. In addition to the geometrical matching, consistency on the measured $p_T$ of the two tracks is required. This is done by demanding that the $p_T$ ratio, having in the numerator (denominator) the largest (smallest) of the two measured
transverse momenta, be smaller than 4. While the aforementioned matching procedure will probably be further optimized in the future, the present version of it already gives sufficiently good performance as shown in Fig. 3.76.

![Figure 3.76: PF muon reconstruction efficiency as a function of $\eta$. Efficiency is determined for cases where the generator muon has $p_T > 30$ GeV. The varying efficiency for $|\eta| < 1$ is due to gaps between drift tube chambers.](image)

3.5.3 Particle-flow and PUPPI algorithms

3.5.3.1 Particle-flow algorithm

The PF algorithm for the L1 trigger was redesigned from first principles, as the algorithm used offline and at the HLT was not suitable for the Level-1 environment: the offline algorithm is designed to operate on a CPU and processes individual objects sequentially, with complex interdependencies and branching structures. The execution time may vary significantly depending
on the event occupancy. The L1 algorithm has instead been designed to be simpler and able to process all input objects in parallel for pipelined execution on FPGAs with a short and fixed latency. An overview of the whole algorithm is shown in Fig. 3.77.

Figure 3.77: Overview of the PF+PUPPI algorithm. Time is indicated roughly by motion from left to right and the solid black boxes indicate the grouping of the algorithm across different boards (Correlator Layer-1, Correlator Layer-2 and Global Track Trigger). Each shaded box, regionizer, linker, and clustering denote firmware blocks needed for the algorithm. PF it the top green box denotes particle-flow. Both the forward PUPPI (from HF) and the clustering can be run before Correlator Layer-1 in which case the inputs to the regionizer will be the clusters. Correlator Layer-2, the algorithmic layer constitutes a broad class of different types of algorithms that can be run in parallel on separate boards. While for Correlator Layer-2, PUPPI particle inputs are considered the default. There is the possibility of PF candidate inputs as well, this is indicated by the green dashed line.

Outside of the $|\eta|$ acceptance of the L1 track finder, only calorimeter information is available. The PF algorithm simply promotes every calorimeter cluster into either a neutral hadron or a photon depending on the identification information described in the previous section.

As discussed in the previous section, inside the track finder coverage, the first step of the particle-flow algorithm is to identify muon tracks. For each input standalone muon object, the best-matching track in the inner tracker by $\Delta R$ and $p_T$ is tagged as a muon and excluded from further processing in the PF algorithm.

Remaining tracks are then linked to the calorimeter clusters. Given the different nature of the calorimeters used as input in the barrel and in the endcap, the algorithm is implemented differently in the two regions.

In the barrel, where electromagnetic energy deposits can be reconstructed with the finer granularity provided by the ECAL while hadronic deposits cannot, a three step procedure is used:

1. Each track is linked to the closest electromagnetic cluster (EM) if any is found within
3.5. Particle-flow reconstruction

\[ \Delta R < 0.04 \] and the association is used for all EM clusters to compute the scalar sum of the \( p_T \) of all associated tracks.

Clusters with no associated tracks are tagged as photons, while if \( p_T^{\text{cluster}} \geq \sum p_T^{\text{track}} - 2\sigma \) the tracks are tagged as electrons and furthermore if \( p_T^{\text{cluster}} - p_T^{\text{track}} \geq 1\sigma \) then the excess is promoted to a photon, with the \( \eta, \phi \) coordinates of the cluster.

If instead \( p_T^{\text{cluster}} < \sum p_T^{\text{track}} - 2\sigma \), the cluster is discarded as it is likely originating from a hadronic shower starting in the electromagnetic calorimeter.

In this comparison, the tolerance \( \sigma \) is the largest between the expected momentum resolution for the cluster and for the sum of the tracks, in both cases extracted from look-up tables as function of the \( p_T \) and \( |\eta| \) (in case of multiple tracks, their resolutions are added in quadrature).

2. Surviving EM clusters are linked to the closest hadronic cluster, and the energy of the hadronic clusters is updated subtracting that of all linked EM clusters. The subtraction is needed since the input hadronic clusters contain both hadronic and EM calorimeter energy. Hadronic clusters whose energy is reduced below about 10% of their original energy after this subtraction are discarded entirely.

3. Tracks can be linked to a hadronic cluster if they are within \( \Delta R < \Delta R^{\text{max}} \), and \( p_T^{\text{calo}} \geq p_T^{\text{track}} - 2\sigma_{\text{calo}}(p_T^{\text{track}}) \). For each track, if there are multiple compatible hadronic clusters, the algorithm selects the one that minimizes the combined angular and \( p_T \) distance

\[
q^2 = \left( \frac{\Delta R}{\Delta R^{\text{max}}} \right)^2 + \left( \frac{\max(p_T^{\text{track}} - p_T^{\text{calo}}, 0)}{\sigma_{\text{calo}}(p_T^{\text{track}})} \right)^2.
\]  

(3.11)

The maximum distance \( \Delta R^{\text{max}} \) is 0.15, larger than the one used for electromagnetic clusters to account for the less regular development of the hadron showers and, in the barrel, of the worse granularity of the hadron calorimeter. The resolution \( \sigma_{\text{calo}}(p_T^{\text{track}}) \) is the expected transverse energy resolution for the calorimeter cluster of a charged pion of \( p_T \) equal to \( p_T^{\text{track}} \), extracted from a look-up table also dependent on \( |\eta| \).

The asymmetry in the \( p_T \) part of the linking distance is to not penalize cases in which the calorimetric deposits of multiple charged particles are merged in a single cluster, as expected in the core of energetic hadronic jets.

The bound requirement \( p_T^{\text{calo}} \geq p_T^{\text{track}} - 2\sigma_{\text{calo}}(p_T^{\text{track}}) \) is instead imposed to reject “fake” tracks (i.e., not corresponding to a physical particle), or those with largely overestimated momentum.

Tracks are promoted to charged hadrons if linked to a cluster, or if \( p_T^{\text{track}} < 10 \) \((20)\) GeV if the track passes loose (tight) quality criteria.

Conversely, calorimetric clusters can give rise to a neutral hadron or photon if \( (p_T^{\text{calo}})^2 \) exceeds \( (\sum p_T^{\text{track}})^2 \) by more than \( \sum(\sigma_{\text{calo}}(p_T^{\text{track}}))^2 \), depending on whether the excess is mainly hadronic or electromagnetic; otherwise, the cluster is discarded.

In the HGCAL, a single linking step is performed. Similar to the barrel case, tracks are linked to the closest cluster in \( \Delta R \) and \( p_T \), with the additional requirement \( p_T^{\text{calo}} \geq p_T^{\text{track}} - 2\sigma(p_T^{\text{calo}}) \).

Tracks linked to a cluster are promoted to charged hadrons or electrons depending on whether the cluster has been identified as hadronic or electromagnetic on the basis of calorimetric information, while tracks not linked to clusters are subject to the same \( p_T \) and quality criteria as in the barrel region. Outside the tracker volume, the particle-flow algorithm is trivial and amounts simply to promoting calorimeter clusters to photons or neutral hadrons.
3.5.3.2 PUPPI algorithm

PUPPI relies on the use of vertexing to reduce the contribution of pileup at the particle level. Through vertex identification, charged particles can be definitively removed and neutral particles can be either removed or downweighted to reflect the probability the particle originates from pileup.

Application of PUPPI in the trigger can be performed through the use of the fact that, outside of vertexing, PUPPI is an inherently local algorithm. To compute the likelihood that a particle is from pileup, a metric is used that takes into account neighboring particles within a cone. Within the tracker volume, the vertexing can be used to effectively separate charged pileup particles from those originating from the leading primary vertex. This separation, in turn, can be applied directly to neutral particles yielding an effective association of neutral particles to the primary vertex. Outside the tracker volume the algorithm needed cannot rely on vertexing; consequently, PUPPI relies on other mechanisms to discriminate pileup including particle $p_T$ and the density of neighboring particles.

For each of the PUPPI algorithms, a firmware implementation exists and is discussed later. The final choice of algorithm does not need to be exactly the same provided that the changes performed fit within the current firmware restrictions. A critical component of PUPPI is the selection of the primary vertex. Firmware algorithms already exist which are capable of vertexing within the latency constraints of particle-flow, and which can also select more than a single primary vertex so that recovery of particles from multiple primary vertices is possible, as discussed in Section 3.4.

Within the coverage of the L1 track finder, the particle-flow reconstruction delivers charged hadrons, neutral hadrons, charged electromagnetic, neutral electromagnetic, and muon particles. Additionally the excellent position resolution of vertices allows for the assignment of charged particle-flow candidates to the primary vertex. As a consequence, the algorithm consists of a computation of the metric $\alpha_C$ ($C$ denotes central) for each neutral defined as

$$
\alpha_C = \log \sum_{i \in PV, \Delta R < 0.3} \frac{\min(p_T^i, p_T^{max})^2}{\max(\Delta R, \Delta R^{min})^2},
$$

where the sum is over all tracks from the primary vertex that are within a cone of $\Delta R < 0.3$ from the neutral particle candidate (if no tracks are found, $\alpha_C$ is set to zero). $p_T^{max}$ and $\Delta R^{min}$ are constants to avoid individual tracks that contribute disproportionately to the sum, and are set to 50 GeV and 0.07 (0.04) in the barrel (endcap), respectively. This algorithm is inherently local and thus can be parallelized over regions and particles.

The summation is performed using tracks, rather than the reconstructed PF candidates, in order to preserve efficiency for particles close to the transition between the barrel and endcap calorimeters: the separate barrel and endcap correlator instances have access only to clusters from the barrel or endcap calorimeters, and only thus to the PF candidates in that region, but have access to tracks also in a wider $\eta$ range.

When the particles are in the forward region, usage of tracks to associate PF candidates to a vertex is not possible. In this scenario, we repeat the same procedure as that of PUPPI performed in the tracker volume but the value of $\alpha_C$ is now replaced by a value $\alpha_F$ ($F$ denotes forward) defined as

$$
\alpha_F = \log \sum_{i \in PF, \Delta R < 0.3} \frac{\min(p_T^i, p_T^{max})^2}{\max(\Delta R, \Delta R^{min})^2},
$$

(3.13)
where the sum is over all particle-flow candidates within a cone of $\Delta R < 0.3$. When considering HF the $p_T^{\text{max}}$ and $\Delta R^{\text{min}}$ are changed to 100 GeV and 0.1 to account for the different granularity. As with $a_C$, the computation is local and can be performed in parallel with all other particles at the same time. However, now that the sum is over all particles within a cone, the resource usage is considerably larger.

In the original implementation of PUPPI used for the offline reconstruction, the per-particle weight is derived on the basis of $a$, and on quantities derived event-by-event from the tracks not associated to the primary vertex. In the L1 trigger, instead, the algorithm combines information from $a$ with that on the particle $p_T$ and type, but does not rely on other per-event quantities except for an overall estimate of the level of pileup $N_{\text{PU}}$.

The $a_C$ value is combined with the particle $p_T$ and the per-event pileup estimate with coefficients depending on the detector region ($|\eta|$) and particle type in order to compute the per-particle weight $w_i$

$$w_i = \frac{1}{1 + e^{-x_{\text{tot}}}}$$

$$x_{\text{tot}} = x_a + x_{p_T} - x_{\text{PU}}$$

$$x_a = \min(\max(c_a \cdot (a - a^0), -x_{\text{max}}^a), +x_{\text{max}}^a)$$

$$x_{p_T} = c_{p_T} \cdot (p_T - p_T^0)$$

$$x_{\text{PU}} = \log(N_{\text{PU}}/200) + c_0.$$ 

The parameters $a^0$ and $p_T^0$ are representative values of $a_C$ and $p_T$ for particles from pileup, and the remaining parameters $c_a$, $x_{\text{max}}^a$, $c_{p_T}$, $c_0$ are tuned to ensure good response and resolution for hadronic jets and missing transverse momentum.

The weight $w_i$ is the final weight applied to the $p_T$ of the neutral particle, yielding a corrected $p_T$ given by $p_T' = w_i p_T$. All the coefficients are taken from look-up tables, and also the function to compute $w_i$ from $x_{\text{tot}}$ is implemented as a look-up table.

The dependency on $p_T$ causes energetic collimated energy deposits to be accepted as signal even in the absence of associated tracks from the primary vertex, which in particular also ensures efficiency for high energy photons.

As the dependency on $N_{\text{PU}}$ is logarithmic, a precise estimate is not critical for the performance of the algorithm. It is expected that such an estimate can be computed in the vertex finding algorithm, e.g., by counting the number of tracks not associated to the selected primary vertex.

The values of the parameters for the algorithm are shown in Table 3.6. The physics performance of the algorithm is stable for variations of the parameters around the chosen values.

### 3.5.4 Particle-flow and PUPPI algorithm performance

The performance of the particle-flow algorithm is first assessed on simulated events without pileup, both on individual particles and on hadronic jets in $t\bar{t}$ events. The response and resolution for hadrons are shown in Fig. 3.78. The response is defined as the median of the ratio between the sum of the $p_T$ of all reconstructed PF candidates in a $\Delta R < 0.2$ cone around the simulated hadron and the hadron energy. The resolution is defined as the Gaussian core of the above ratio. The worse resolution for charged pions in the endcap compared to the barrel arises both from the degradation of the momentum resolution of tracks with increasing $|\eta|$, and the possibility of partial double-counting of energy if a pion produces multiple clusters in the calorimeter.
Table 3.6: Parameters for the PUPPI algorithm.

| Parameters       | $|\eta| < 1.5$ | $1.5 < |\eta| < 2.0$ | $2.0 < |\eta| < 2.4$ | $2.4 < |\eta| < 3.0$ | $3.0 < |\eta|$ |
|------------------|--------------|-----------------|-----------------|-----------------|-----------------|
| Neutral $p_T$ cut| 1.0          | 1.0             | 2.0             | 4.0             | 10.0            |
| $c_0$           | 0.7          | 1.5             | 1.5             | 2.2             | 0.6             |
| $\alpha^0$      | 6.0          | 6.0             | 6.0             | 9.0             | 9.0             |
| $x_{\text{min}}$| 4            | 3               | 3               | 4               | 4               |
| $c_{p_T}$       | 0.3          | 0.4             | 0.4             | 0.4             | 0.25            |
| $p_T^0$(Had)    | 4.0          | 5.0             | 7.0             | 9.0             | 14.0            |
| $p_T^0$(EM)     | 2.5          | 3.0             | 4.0             | 5.0             | 14.0            |
| $c_0$(Had)      | 5.0          | 5.0             | 5.0             | 7.0             | 6.0             |
| $c_0$(EM)       | 1.0          | 3.5             | 3.5             | 5.0             | 6.0             |

Figure 3.78: Particle-flow response (left) and resolution (right) for charged pions, neutral pions, and long-lived neutral kaons, in two detector regions separated by $|\eta|$ corresponding to barrel, filled in points, and the endcap HGCAL region, unfilled points.

The response for multiparticle final states is evaluated by looking at hadronic jets in $t\bar{t}$ events. Generator-level jets are clustered from stable generator particles except neutrinos with the anti-$k_T$ algorithm [17, 57] with a distance parameter of 0.4. For each jet, a response is computed from the scalar sum of the $p_T$ of all trigger objects in a $\Delta R < 0.4$ cone around the jet axis. The response is computed for PF candidates, and for comparison also separately for the input calorimeter clusters and tracks. In order to estimate the impact of the momentum acceptance, a response is also computed for generator particles with $p_T$ above 2 GeV (1 GeV) for charged (neutral) particles. The median response is shown in Fig. 3.79. Compared to calorimeter clusters alone, PF candidates recover some response, especially in the transition region between the barrel and the endcaps.

For each set of objects, a resolution is also estimated from the Gaussian core of the response distribution, after applying $p_T$-dependent corrections to the response in order to bring the median to unity as shown in Fig. 3.80. Especially at low $p_T$, the resolution is substantially improved by the use of the particle flow. The gain is smaller for energetic jets in the endcap, where the calorimeter response is better and the track momentum resolution is worse.
3.5. Particle-flow reconstruction

Figure 3.79: Median response for jets in $\bar{t}t$ events computed from the scalar sum of $p_T$ of trigger objects in a $\Delta R < 0.4$ cone around the generated jet axis. The different colors correspond to the stages of reconstruction. Purple shows the response of just the track components within the cone. Red shows the response of the calorimeter inputs alone. Blue shows the response taking the sum of all particle-flow candidates. Finally, gray shows the sum of the gen particles correcting for cone acceptance effects.

Figure 3.80: Transverse momentum resolution for jets in $\bar{t}t$ events computed from different trigger objects. The scalar sum of $p_T$ of the trigger objects a $\Delta R < 0.4$ cone around the generated jet axis is computed, and a calibration is applied as function of $p_T$ to ensure unit response for all sets of objects. The different colors correspond to the stages of reconstruction. Purple shows the response of just the track components within the cone. Red shows the response of the calorimeter inputs alone. Blue shows the response taking the sum of all particle-flow candidates. Finally, gray shows the sum of the gen particles correcting for cone acceptance effects.
3.5.4.1 Isolation

Particle-flow and PUPPI-based isolation of candidates can further extenuate the performance of muon and electron identification. Its usage follows the behavior of $\tau_{h}$ lepton identification (Section 3.6.2.3). In this scenario, isolation is computed by taking the sum of candidates within a cone. For muons, in the global muon trigger, track isolation can be used to measure the muon isolation. With particle-flow, it is possible to use both charged and neutral PF candidates for the computation of the isolation. This further extends to electrons, where either PF or PF+PUPPI candidates can be used to compute the isolation around a lepton. This allows for an improved trigger acceptance, particularly for muons at very low $p_T$, where isolation is needed to reduce the rate of background events.

We define isolation as

$$\text{Iso} = \sum_{i, \Delta R < 0.4} p_T^i,$$

where the sum over $i$ is either particle-flow candidates for PF isolation or PUPPI candidates for PUPPI isolation. We define isolation under two working points:

- **loose**: Isolation efficiency of 99% for muons originating from a Z boson with $p_T^{\text{gen}} > 20$ GeV and $|\eta^{\text{gen}}| < 2.5$
- **tight**: Isolation efficiency of 95% for muons originating from a Z boson with $p_T^{\text{gen}} > 20$ GeV and $|\eta^{\text{gen}}| < 2.5$

The choice of these working points is to show relatively extreme working points that are capable of probing very low $p_T$ muons, where the impact of isolation will be largest. To be clear, this is just for the choice of isolation cut. It does not involve the $p_T$ cut choice, which is varied in the ensuing studies.

The rates for PUPPI seeded muons with an isolation on PUPPI candidates and PF seeded muons with PF candidate based isolation are shown in Fig. 3.81. When no isolation cut is applied PF and PUPPI muons perform very similarly. When an isolation cut is applied, a cut on PUPPI isolation is found to have a higher rate with respect to a cut on PF isolation for low-$p_T$ muons. For high-$p_T$ muons, the rates are reversed with PF isolated muons having a higher rate than PUPPI muons. The higher rate for low-$p_T$ muons is a result of the fact that PUPPI rejects too many particles in low-$p_T$ jets, causing nonisolated muons to appear as isolated.

For PUPPI muons, a scan of the bit precision is performed in Fig. 3.82 on PF muons. A total of 3 bits below the decimal point is sufficient to obtain an optimal $S/B$ for PF muons. Lastly, a comparison of the lepton efficiency is shown in Fig. 3.83 and an optimization of $H \rightarrow ZZ \rightarrow 4\mu$ events is shown in Fig. 3.84.

3.5.5 Firmware implementation

Before inputs can be provided to PF and PUPPI, they must be sorted into appropriately sized small regions. The regionizer algorithm is responsible for receiving all inputs, determining which small regions they belong to, and then passing them to the PF and PUPPI firmware. The implementation is designed to minimize both latency and overall resource usage, specifically taking into account the available resources not used by PF and PUPPI and the firmware shell necessary for managing the links and clock on the VU9P FPGA.

The regionizer algorithm is divided into three main steps. In the first, inputs from the same bunch crossing are streamed from the links into BRAM to buffer them. Next, the inputs are
3.5. Particle-flow reconstruction

Figure 3.81: Comparison of the event rate with 200 average pileup interactions for PF muons, and PF+PUPPI muons for (left) no isolation cut, (center) a loose working point, and (right) a tight working point. The dashed line corresponds to a rate of 50 kHz.

Figure 3.82: Scan of the bit precision for PF muon isolation. The different colored points correspond to a variation in the precision after the decimal point. The x axis shows the relative isolation, the y axis shows the ratio of signal efficiency, computed from $Z \rightarrow \mu\mu$ events, over background rate at pileup 200, normalized to have an inclusive $S/B$ ratio of 0.5 when no isolation criteria is applied. Increasing the bit precision below the decimal allows for more total points and allowing for 3 bits below the decimal point yields the possibility for a cut that is maximal in $S/B$ to be placed on the isolation. Application is performed on a per-muon basis to reflect the effectiveness of this selection when included in trigger with one or more muons. In this plot, two bits are used to describe the isolation above the decimal point. However, one bit is sufficient from this study.
Figure 3.83: Comparison of the single muon efficiencies as a function of generated muon $p_T$ after requiring the single muon trigger $p_T$ threshold to be above a fixed value such that the total trigger rate is at 50 kHz. This selection is performed for both for PF muons (red) and PF+PUPPI muons (black) for (left) the case when no isolation cut is applied, (center) a loose isolation cut is applied corresponding to a plateau efficiency of 99% relative to no isolation, and (right) a tight isolation cut corresponding to a plateau working point efficiency of 95% relative to no isolation applied.

Figure 3.84: Comparison of the $H \rightarrow ZZ \rightarrow 4\mu$ signal efficiency requiring all 4 muons having $|\eta_{gen}| < 2.5$ as a function of collision rate for (left) a single muon trigger, (center) di-muon trigger, and (right) a tri-muon trigger for these plots instantaneous luminosity of $7.5 \times 10^{-34}$ cm$^{-2}$ s$^{-1}$ is utilized. The dotted line corresponds to a rate of 50 kHz for the single muon trigger (left), 12 kHz for the di-muon trigger (center), and 6 kHz for the tri-muon trigger (right).
passed through multi-tap shift register arrays that determine of which small regions they are a member. The multi-tap shift register arrays route the inputs to additional buffers, each designed to hold a small number of inputs of a particular type for a particular small region. In the final step, all inputs for each small region are successively extracted from the corresponding buffers and input to the PF and PUPPI firmware block. The buffers in the second step are constructed using ultra RAMs, allowing the inputs to be read at very low latencies. A diagram of the regionizer logic is shown in Fig. 3.85.

![Diagram of the regionizer logic](image)

Figure 3.85: Data organizer for PF

Each board in the Correlator Layer-1 is responsible for processing specific small regions per event. Groups of boards are divided into three types, "central", "endcap", and "endcap, no track", defined by $|\eta| < 1.5$, $1.5 < |\eta| < 2.5$, and $2.5 < |\eta| < 3$, respectively. The boards in each group run the specific PF and PUPPI algorithms designed for the types of inputs arriving in each region. These divisions correspond to the total number of small regions for each group of 27, 18, and 18, respectively. The region defined by $|\eta| > 3$ uses only HF clusters as input, and therefore is processed in the Global Calorimeter Trigger boards where this information is already available. It is divided into 36 small regions for each side of the detector.

The regionizer firmware has been validated against an independent implementation of the algorithm in software considering a common set of inputs. A sample of simulated $t\bar{t}$ events is considered for this purpose, overlaid with an average of 200 simultaneous pp collisions due to pileup. For each species of input object entering a central board, regional boundaries are verified to be consistently enforced. The classification of each object entering this large region is further verified to be consistently sorted into one or more of the small regions. Perfect agreement is observed for all events considered, including the multiplicity and identity of small regions to which each object is assigned. Figure 3.86 demonstrates the number of each type of input that enters each of the 18 small regions for processing by the PF + PUPPI block, as determined by the firmware and software implementations of the regionizer.

In order to design a system capable of processing the full detector at 40 MHz, multiple possible initiation intervals and region sizes were considered. For each possible combination, the number of PF and PUPPI implementations per VU9P chip necessary to process all small regions was
Figure 3.86: Comparison of small region assignments for each PF+PUPPI input object for a sample of ten $t\tau$ events. Perfect agreement is achieved between the regionizer firmware and software emulation.

determined, and the resources needed for these implementations was computed. For a Layer-1 design with 36 boards, the values from this scan are shown in Fig. 3.87. This calculation represents a rough metric for the ease of successfully placing and routing the PF and PUPPI firmware in the current design. Two configurations were identified as optimal: 54 regions in the barrel running at an initiation interval of 2, and 27 regions in the barrel running at an initiation interval of 4. Both of these configurations require a single PF and PUPPI implementation per board, and produce the minimum usage of LUTs/registers and DSPs, respectively. A diagram of the board layout for Correlator Layer-1, assuming the latter configuration, is shown in Fig. 3.88. From the firmware implementation, we also have a preliminary estimate of the latency of the Regionizer + PF + PUPPI firmware demonstration. This is shown in Fig. 3.89. The total latency of the algorithm including the streaming in of the presumed last input (the tracks), the regionizer firmware, and the PF+PUPPI algorithm is approximately 1.5 $\mu$s.

The particle-flow and PUPPI firmware assumes that each input object is encoded in 64 bits. Different information is associated with each input type. Every input encodes $p_T$ using 16 bits, $\eta$ using 10 bits, and $\phi$ using 10 bits. The additional information for each input type and the number of bits used to store it is given below.

- **Tracks**
  - $p_T$ - 16 bits
  - $z_0$ - 10 bits
  - *Passes tight quality* - 1 bit

- **EM clusters**
  - $p_T$ - 16 bits

- **Combined clusters**
Figure 3.87: The fraction of various resources needed per VU9P FPGA to process events in the barrel with particle-flow and PUPPI at 40 MHz. LUTs used as logic is abbreviated to \textit{LUT as L} and LUTs used as shift registers is abbreviated to \textit{LUT as SR}. Different colors represent different scenarios for dividing the barrel into small regions. These regions are described in the text, and the necessary number of input tracks, electromagnetic calorimeter clusters, and combined calorimeter clusters are given in brackets in the legend. Different sub-columns represent different initiation intervals (II) in clock cycles used for particle-flow and PUPPI. Solid points represent scenarios where the optimal number of particle-flow and PUPPI implementations per VU9P is a whole number; empty points represent scenarios where the optimal number of implementations is fractional; fractional points use much more resources than are actually needed since only integer numbers of firmware implementations can be added to the FPGA.
Figure 3.88: Flow of particle-flow processing IP. A box corresponds to a specific board that covers the designated geometrical region. The multiplicity factor $\times x$ denotes the number of respective boards. The $y$ axis denotes the total time for processing the inputs for a design at a time-multiplex factor of 6; recall the latency of the overall algorithm can be longer provided the throughput is maintained. Within this latency the algorithm must processes 9 total regions with an initiation interval of 4 clocks between regions; this is indicated by the labels within the diagram.
3.5. Particle-flow reconstruction

The resulting particle-flow and PUPPI particles are also allotted 64 bits, and the information is stored with the same precision as for the input objects. The PUPPI weight $a$ is stored using 10 bits. This scheme leaves at least 10 bits unused for output objects. This allows additional information to be included later and passed with each object. Given that the used bandwidth due to output objects is not too large, these currently unused bits do not strain the system and can be reserved without issue.

Studies of the effects of different granularities for encoding the $p_T$ have also been performed. These studies suggest that while the jet response is modified for granularities larger than 0.25 GeV, the jet resolution is unaffected for granularities as large as 1 GeV. Given that $H_T$ and $E_T^{\text{miss}}$ are both affected for larger granularities, 0.25 GeV is chosen as the optimal granularity.

In the nominal design, the central version of the PUPPI algorithm is run using the primary vertex $z$ position as input. However, the algorithm itself can be run using any number of vertices, not just the primary vertex. Running the algorithm using alternative vertices allows the recovery of inefficiencies due to the misidentification of the primary vertex. As discussed in Section 3.4.4, alternative vertices can be computed by the GTT and passed to the PUPPI algorithm. To make use of these vertices in the PUPPI algorithm, the algorithm is modified such that a separate PUPPI weight $w$ is computed for each vertex. These weights can then be used in multiple ways. Taking the maximal weight for each particle allows the recovery of inefficiencies due to misidentification of the primary vertex, and is an entirely transparent change to all other algorithms using PUPPI particles. Alternatively, the weights (or a subset of the weights) for each particle can instead be used individually in dedicated algorithms. Modifying the PUPPI firmware to enable it to compute multiple weights results in an increase in the
resources necessary for implementation. Each additional vertex considered requires the use of an additional 0.5%, 1.5%, and 4.5% of available DSPs, registers, and LUTs, respectively, in a VU9P. The additional latency necessary is very minimal and amounts to one clock cycle per two additional vertices.

In the Correlator Layer-2, the regionizing is performed by taking the inputs from each of the boards used for Layer-1. For 36 boards, at a time multiplex factor of 6, this equates to inputs from each of six different Layer-1 boards. For these inputs, we assume the final PUPPI candidates are separated by particle type. Namely charged hadrons, neutral hadrons, and muons each come on a separate set of fibers. With six boards as input, we assume that 3 regions enter the detector per clock. Thus, for each fiber input, one 64 bit PUPPI candidate can be accepted every clock. When 16 PUPPI candidates are accepted per Correlator Layer-1 region, a total of 48 fibers are used as input into Correlator Layer-2. Such a scheme would allow for a time-multiplex factor of 3 as well since in that case a total of 96 fibers would be used as inputs. With a time-multiplex factor of six and a clock frequency of 240 MHz, the resulting algorithms must work with an initiation interval of 36 clock or smaller.

For the first implementation of the input and routing for the Correlator Layer-2, an HLS IP was constructed that performed the streaming and sorting of the inputs into a global event region. For this setup, a pair of regions are read per clock taking in as many as 36 input fibers per region. The region is then directly inserted into the block RAM. The Block RAM is partitioned to correspond to the leading candidate, sub-leading, and so on, split by particle-flow type (charged or neutral). As a consequence, each block RAM contains $i$-th PF candidate of a certain type for all regions. Each region is located in a specific place within the block RAM. Additionally, the inputs for each particle type are assumed to be $p_T$ sorted allowing for the direct placement into the block RAM. The sorting is currently not implemented in Layer-1, but it is expected to not take a considerable amount of resources.

To perform the seeding for all algorithms, we rely on a grid structure of the sub-regions. This structure is shown in Fig. 3.90. A simple seeding algorithm has been implemented in firmware that takes the highest $p_T$ particle per region. For the current PF tau algorithm, this is the highest $p_T$ track per region. For the muon isolation algorithm, this is the highest $p_T$ muon per region. A more sophisticated seeding has been developed whereby the highest $p_T$ particle in a large region is used as a seed. The large regions correspond to one fourth of the total detector volume. Once this seed is chosen, the second highest $p_T$ particle in the region is selected as a second seed, and so on. Following the choice of the seed for the two seeding scenarios, a look-up table is used to select the 4 neighboring regions adjacent and covering the chosen seed (Fig. 3.90). For each of these neighboring regions all candidates are loaded, and a subset of these candidates are chosen as inputs into an algorithm. In most cases, the first choice is to require the candidates to be within $\Delta R < 0.4$ of the chosen seed. For this requirement candidates with $\Delta R > 0.4$ are set to have a null 4-vector, which is discarded later in the algorithm.

For the tau algorithm, discussed in Section 3.6.2.3, the 10 highest candidates per region, per PF type are chosen. These are then merged into a larger array. The combination of 4 regions with 32 candidates (3 particle types and two muons), gives a maximum of 128 PUPPI candidates for each tau. This collection of a potential 128 candidates is then sorted and the top 10 candidates are taken and applied to the neural network. The resulting procedure is repeated on the 6 highest nonoverlapping seeds.

For the muon algorithm, the seeding is performed with PUPPI muons and then the isolation is computed by summing the particle $p_T$ of the candidates.
For the jet algorithm, the seeding is performed using the full iterative procedure. The resulting jet axis is taken to be the seed, and the jet four-vector is taken as the vector sum of all the four-vectors of PUPPI candidates within the cone.
3.6 Triggering on jets, hadronic tau decays, and energy sums

This section describes triggering on jets, hadronic tau decays, and energy sums. A range of algorithms with different levels of sophistication are explored to take maximum advantage of the information available to the L1 trigger during HL-LHC. These algorithms aim to provide the best rate versus efficiency working points as a function of object transverse momentum threshold, as well as to provide robust methods for triggering, estimating efficiencies, and monitoring. Some algorithms discussed are based on simulation studies only, while for others full firmware implementation and test-bench studies are described.

Algorithms of different complexity are explored for identifying jets, hadronic taus, and energy sums. The types of algorithms considered are based on either calorimeter information only, tracker information only, track-matched calorimeter objects, or use full particle-flow / PUPPI-based inputs. Simpler algorithms, such as identification of jets and hadronic taus based on fixed-size windows are generally easier to implement to firmware, and provide robustness to the system. The Phase-1 upgrade of the CMS trigger system has demonstrated the superiority of sophisticated algorithms adapted to higher pileup and luminosity conditions. The computing power available in Phase-2 should allow these approaches to be implemented even more comfortably, and to remain within the latency budget of 3–4 $\mu$s.

The implementation of both complex, high performance particle-flow based algorithms and complimentary robust, standalone algorithms allows low thresholds to be maintained, providing similar physics acceptance to LHC data-taking, whilst providing maximum efficiency at high thresholds and redundancy during changing detector conditions.

3.6.1 Jet finding algorithms

Multiple jet finding algorithms are developed and described below. These include a jet algorithm based only on calorimeter information from the barrel ECAL and HCAL, HGCAL, and the forward HF system. This algorithm is similar to existing jet-clustering algorithms used in the current L1 trigger. Taking advantage of the addition of tracking at L1, tracker-only jets are defined using a rapid firmware-implemented track clustering algorithm. Finally, jet clustering is performed using particle-flow objects as input. This algorithm has been implemented in HLS and a small-scale version is running on a demonstrator platform. Currently, calorimeter jets and particle-flow jets are used in the menu.

3.6.1.1 Calorimeter-based jet finding

The calorimeter-only jet algorithm uses calorimeter tower information from the barrel, HGCAL, and HF. The towers are provided as trigger primitives from the HGCAL and HF readouts directly. For the barrel, the towers are built in Layer-1 of the barrel calorimeter trigger after the clustering of $e/\gamma$ objects is finished, as described in Section 3.2.2. After this clustering is complete, the Layer-1 contains three types of barrel objects: $e/\gamma$ clusters based on ECAL crystals, ECAL towers that contain the remaining unclustered ECAL energy, and HCAL towers. This information is combined into trigger towers, re-using the $e/\gamma$ clustering, where the tower size corresponds to that of an HCAL tower (or equivalently, $5\times5$ ECAL crystals). The towers are ordered by $p_T$ and the algorithm starts from the highest-$p_T$ tower, which serves as a seed for a jet. After building a jet, all the towers used in this jet are removed and the algorithm continues with the next remaining highest-$p_T$ tower. The algorithm continues until there is no tower with $p_T > 2.5$ GeV to seed a new jet.

The jet size at the trigger level should be as similar as possible to the offline jet size. This requirement allows better adjustment of the L1T decision to the HLT decision, thus improving the
3.6. Triggering on jets, hadronic tau decays, and energy sums

overall efficiency of the trigger chain. The offline reconstruction uses the anti-$k_T$ algorithm with a distance parameter of 0.4. For the high PU expected at the HL-LHC, this may be reduced to 0.3. Throughout the barrel, endcap and forward calorimeter systems, it is more practical to use the simple geometry of a cone algorithm. The size of one tower is approximately $0.087 \times 0.087$ in $\eta$–$\phi$ phase space, therefore, a few different configurations were studied: $11 \times 11$, $9 \times 9$, $7 \times 7$, and $5 \times 5$ towers, centered at the seed tower. Both a simple square geometry and a square with corner towers removed to mimic a cone geometry were considered. The $7 \times 7$ simple square geometry was chosen, corresponding to a cone size of approximately $0.3$–$0.4$ in $\eta$–$\phi$. An increase in the jet size leads to a higher contribution from PU, while decreasing the jet size leads to a worsening of the jet energy resolution.

The jet is corrected for energy reconstruction and PU, as described below. The jet energy correction is performed using simulated samples with no PU, and can therefore be done independently of the PU correction, which is explained first below.

The presence of additional interactions within the same bunch crossing leads to additional energy flow that increases the energy of reconstructed calorimeter jets. This contribution must be corrected for, since it cannot be separated from the energy of prompt particles coming from the hard interaction in the calorimeter trigger, where no tracking information is present. To reduce the contribution due to PU, the jet size is restricted to $7 \times 7$ trigger towers, where only towers with an energy greater than 0.5 GeV are used to reconstruct jets. Study of different cutoffs 0.5–2.0 GeV demonstrated that the 0.5 GeV requirement does not affect the jet energy resolution, but significantly reduces the PU contribution. The jet PU correction is based on an assumption that the PU initiated energy flow can be estimated on average using first MC, later on measured in data, as a function of $\eta$. That is, for a given PU one can estimate an average PU energy contribution in each trigger tower and subtract it. The correction is applied on an event-by-event and tower-by-tower basis from the average expectation of the PU energy flow for a given number of PU vertices. The estimated corrections are loaded in the look-up table of calorimeter cards and for each tower this energy is subtracted from the measured tower energy, truncated at 0. While the PU energy is estimated from simulation, one needs to estimate the number of PU vertices from the data. This is done using the total number of towers with $0.5 < p_T < 2$ GeV in each part of the calorimeter: ECAL barrel, HCAL barrel, HGCAL electromagnetic, HGCAL hadronic, and HF. The energy region is chosen based on MC studies showing that towers with this energy are mainly populated by PU. Figure 3.91 shows the resolution of the reconstruction of the number of PU vertices relative to all generator vertices using this method. The best measurement is achieved by the HGCAL since it has the highest total PU contribution and therefore the combined measurement is also driven by the HGCAL measurement. The PU energy per tower, estimated from simulation, varies between 0.1 GeV in the barrel, 2 GeV in the endcap, and up to 5 GeV in HF. The correction works well in the HGCAL and HF. It has almost no effect in the barrel since the average PU energy there is small, so only a small amount of energy is subtracted. To be sure that possible PU fluctuations in the barrel are taken into account, the energy subtracted per tower is increased by a factor of four. The change in jet energy response in the barrel due to this pileup correction is corrected back by the jet energy calibration.

The jet energy correction is performed in steps. First the ECAL energy is corrected based on simulated single electron samples that provide scaling factors based on the ratio between the reconstructed and the generated energy of electrons as a function of the reconstructed electron $p_T$ and $\eta$. After the ECAL correction is applied, the HCAL energy is corrected in a similar way using simulated QCD samples with single charged pions and jets. In the HF, the energy correction is performed in a single step. The results of the energy corrections are shown in
Figure 3.91: Resolution of the number of PU vertices using different calorimeter systems, ECAL, HCAL, HF and the electromagnetic and hadronic HGCAL in $t\bar{t}$ events at PU200.

Figure 3.92: Jet energy response for jets in $t\bar{t}$ events with 200 average pileup interactions for the barrel calorimeter (left) and HGCAL (right) after PU and energy corrections are applied. Fig. 3.92 for jets in the barrel and HGCAL for different energy regions.

In the barrel, the HCAL corrections are slightly different for jets with high and low contribution of ECAL in the total jet energy, which allows the energy resolution of reconstructed jets to be improved.

Figure 3.93 shows the resulting efficiency of the jet reconstruction as a function of generated $p_T$ for reconstructed jet $p_T > 100$ GeV (left) and as a function of $\eta$ for jets with reconstructed $p_T > 80$ GeV and generated $p_T > 100$ GeV (right), with jet energy corrections applied. The jet trigger turn-on curve has a steep rising edge at the threshold value, which is well aligned with the reconstructed value. Except for regions around detector boundaries, the efficiency as a function of $\eta$ is almost flat across the full acceptance.
3.6. Triggering on jets, hadronic tau decays, and energy sums

Figure 3.93: Jet finding efficiency as a function of jet $p_T$ (left) and $\eta$ (right) in $t\bar{t}$ events with 200 average pileup interactions.

Figure 3.94: L1 trigger rates for the calorimeter-only based jet finding algorithm, shown separately for different parts of the calorimeter system, barrel, endcap (labeled as HGCAL) and HF, for single neutrino events with 200 average pileup interactions.

The resulting trigger rates are shown in Fig. 3.94 separately for different parts of the calorimeter.

3.6.1.2 Track-based jet finding

A rapid firmware-implemented track clustering algorithm enables the combination of a collection of tracks into an output collection of track-based jets. The track-based jets can then be used for global trigger quantities like $H_T$ and $H_T^{miss}$. A set of track purity requirements are applied to the input selection of tracks to keep the L1 trigger objects resilient to pileup. In this section, we describe the simple track clustering algorithm that is implemented in firmware. The track clustering is done in parallel for ranges of track $z_0$, so makes no explicit use of the L1 primary vertex. However, the algorithm becomes more robust against pileup when considering tracks
from only a narrow z-range from the primary vertex, \( \Delta z(z_{PV}, z_{trk}) < 1 \) cm, which would require the L1 primary vertex for preselecting the L1 Tracks to be clustered, which is reconstructed in the same board and can be made available to the jet algorithm.

The clustering of tracks in the \( \eta-\phi \) plane is performed using a nearest-neighbor approach in two one-dimensional (1D) steps. The first phase of the algorithm parses tracks into 27 \( \phi \) regions and 24 \( \eta \) regions, which divides the region of tracking acceptance into cells that are approximately 0.2 x 0.2. The track \( z_0 \) is also used to assign the track to a z-cell along the beam axis. The clustering is done in \( \eta-\phi \) in z-regions that span the beam axis. In between every two z-cells is a third cell to cover the overlapping region and eliminate edge cases where the primary vertex is close to the edge of a given bin. This gives a maximal jet size that is approximately a 3x3 square of cells with a half-side equal to \( \Delta R = 0.3 \). The width of the z-cell used for this study is 1 cm, which means the clustering is run in 60 z-cells including the overlapping z-regions. The number of cells would be limited by FPGA resource usage in firmware, but emulates the performance when preselecting input tracks to be in a narrow window around the primary vertex before track clustering.

Only tracks fulfilling the track purity requirements listed in Table 3.7 are clustered. In the first layer of the track clustering in \( \phi \), the sum of track \( p_T \) in the \( i^{th} \) cell, \( \sum p_{T}^{ij} \), is compared to its two neighbors in \( \phi \) and the \( p_T \) is summed into the local maximum cell. The result of this step is to create a list of cells centered on the local maxima in the \( \phi \) dimension. The next step is to check the \( j^{th} \sum p_{T}^{ij} \) clusters for neighbors in \( \eta \), which are merged into the \( \phi \)-cluster with larger \( \sum p_{T}^{ij} \). In these two steps on the firmware implementation, nearest neighbors in \( \phi \) and \( \eta \) cells are merged based on \( p_T \) comparison. The track-based jets are then the list of cells centered on the local \( p_T \) maximum in the \( \eta-\phi \) plane and including the \( \sum_{ij} p_{T}^{ij} = p_{T}^{jet} \) from its neighbors. This procedure is done in parallel z-bins, and the primary interaction z-window is found by summing the \( p_{T}^{ij} \), in each z-cell, \( \sum p_{T}^{ij} = H^{ij} \), and finding the z-cell with the largest \( H^{ij} \).

In addition to the \( \sum p_{T}^{ij} \), the track multiplicity is also determined using the same procedure to provide an additional handle for the output collection of track-based jets. This approach gives a list of track-based jet positions in \( \eta, \phi, z_0, p_T, \) and \( N_{tracks} \).

### Table 3.7: L1 track purity requirements for tracks input to both the L1 track-based jet clustering and the L1 \( E_T^{miss} \) computation

<table>
<thead>
<tr>
<th>Track Variable</th>
<th>Cut</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N_{stubs} ) per track layer</td>
<td>( \geq 4 )</td>
</tr>
<tr>
<td>( p_T )</td>
<td>( \geq 2 ) GeV</td>
</tr>
<tr>
<td>( \chi^2/\text{dof} )</td>
<td>( &lt; 40 )</td>
</tr>
<tr>
<td>( \chi^2_{\text{bend}} )</td>
<td>( &lt; 2.4 )</td>
</tr>
</tbody>
</table>

Figure 3.95 shows the reconstruction efficiency for this firmware implementable track clustering compared to a full anti-\( k_T \) clustering with FASTJET. The same set of input L1 tracks that fulfill the track purity requirements is used. Tracks within a \( |\Delta z| < 1 \) cm of the L1 primary vertex are then input to FASTJET anti-\( k_T \) clustering while L1 track jets are reconstructed as described above. This difference in the z-window of tracks gives the L1 track jets a slightly larger efficiency in the forward region. The efficiencies are generally similar across the tracking acceptance and range of generator level \( p_T \). The approach of 1D nearest-neighbor clustering in two steps gives a similar efficiency to a full 2D jet reconstruction with an algorithm similar to the offline jet reconstruction. The largest inefficiency is seen towards the edge of the tracking.
acceptance, $|\eta| > 2.0$, and for lower $p_T$ generator level jets, $p_T < 50$ GeV, which typically have lower track multiplicities.

![Efficiency plots](image)

Figure 3.95: Comparison of the L1 track jet clustering and anti-$k_T$ clustering using FASTJET with $R = 0.3$ with the same input collection of L1 reconstructed tracks. The efficiency in fully hadronic $t\bar{t}$ events with no pileup (top) and 200 pileup (bottom) as a function of generator $\eta$ (left) and $p_T$ (right). The efficiency is computed by matching the generator-level jet to the reconstructed track jet within $\Delta R < 0.4$. The two plots show similar efficiency in events with no overlaying pileup interactions and at 200 pileup interactions. The SUSY model sample and the $t\bar{t}$ sample have a similar jet $p_T$ distribution and so can be compared.

To keep global variables like L1 track-based $H_T$ and $H_T^{miss}$ resilient against pileup, we optimize the quality criteria on the input L1 tracks to preserve a low trigger threshold while maintaining low data-taking rates. The same criteria is applied for the track-based MET (Table 3.7). When clustering tracks to form jets, the track multiplicity in each jet gives an additional handle, compared to what is used for track-based MET, in rejecting tracks from bad combinations of track hits that result in fake high-$p_T$ tracks. Specifically, jets with $p_T > 50(100)$ GeV are required to have at least 2 (3) tracks. Track-based $H_T$ is computed as the scalar sum of $p_T$ for track jets with
To study the performance of the L1 track-based $H_T$ and $H_T^{\text{miss}}$ triggers, a signal model is used that simulates a supersymmetric top-quark (stop) decay to its super-partner, the top-quark, and the stable lightest supersymmetric particle, $\tilde{\chi}_1^0$. The mass difference between the stop and $\tilde{\chi}_1^0$ is near the top-quark mass so that the missing transverse energy is small. To measure the trigger turn-on, a collection of generator-level jets is used to emulate offline $H_T$ and $H_T^{\text{miss}}$ variables. The generator level collection of jets is made by clustering generator-level particles, excluding neutrinos, using the anti-$k_T$ algorithm. The generator-level jets considered are those with $p_T > 30$ GeV and $|\eta| < 2.4$.

Figure 3.96 shows the improvement in the L1 trigger rate, the signal efficiency, and the L1 turn-on when reducing the impact of tracks from bad track hit combinations through stringent track quality criteria (solid vs dashed lines). The improvement is observed for both the L1 track-based $H_T$ and $H_T^{\text{miss}}$ triggers. For a fixed L1 track-based $H_T$ rate of 25 kHz, the 95% offline efficiency turn-on for the stop signal is lowered from 675 GeV to 450 GeV. For the track-based $H_T^{\text{miss}}$ trigger, the threshold is lowered from 675 GeV to 290 GeV. In both cases, the signal efficiency for a fixed rate is greatly increased for triggering on the BSM stop phase-space when including the track purity criteria.

A firmware implementation for track jets is available, and has been successfully tested in a hardware demonstrator. This is presented in Section 6.4.3.2.

### 3.6.1.3 Particle-flow based jet finding

The histogrammed PF jet algorithm builds jets from PUPPI candidates received from Correlator Layer-1 by binning them into pseudo trigger towers, and clustering using a $7 \times 7$ window centered on a local maximum, similar to the L1 jet algorithm used during Run-2.

In order for the algorithm to use PUPPI candidates as input, pseudo trigger towers are created by binning the inputs into a 2D histogram in $h-j$, with the PUPPI candidate $p_T$ assigned to the weight. The full calorimeter is divided into 120 $\times$ 72 bins, corresponding to a bin size of 0.083 $\times$ 0.087 in $h-j$.

Clustering is performed around a jet seed that satisfies the inequality mask shown in Fig. 3.97 to avoid double counting and self vetoing. The $7 \times 7$ window corresponds to a jet size of 0.6 $\times$ 0.6. The jet momentum is computed as the sum of the momentum of the $7 \times 7$ window, and the jet position is taken as the seed position in both $\eta$ and $\varphi$.

A jet energy correction factor is applied based on the jet momentum and position in the detector. These factors are computed by matching generator-level anti-$k_T$ jets with distance parameter of 0.4 (AK4) to L1 jets in a combination of QCD and $t\bar{t}$ MC samples simulated with PU200. Matching is done by looking for the closest L1 jet within $\Delta R < 0.5$ for each generator jet. To prevent mismatches, it is checked whether the generator jet is also closest to the matched L1 jet. Correction factors are obtained by computing and fitting the distribution of the jet response $p_{T,\text{gen}}/p_{T,\text{trig}}$ in bins of $p_{T,\text{trig}}$ and $\eta$. The PUPPI candidates used as inputs have their $p_T$ weighted according to the probability they originate from pileup interactions, and no further pileup correction is applied to the jets. Jet energy calibrations are applied to all jets except those in HF, due to limited statistical precision in the QCD and $t\bar{t}$ MC sample. A VBF MC sample can eventually be used to derive HF jet calibrations.

The jet performance has been studied for different jet sizes. When using the particle-flow calorimeter clusters as input, a $5 \times 5$ window provided the best compromise between including
3.6. Triggering on jets, hadronic tau decays, and energy sums

Figure 3.96: Signal efficiency versus L1 data-taking rates (top) for track $H_T$ (left) and track $H_T^{miss}$ (right). The signal efficiency is computed as the fraction of events in the signal stop sample that pass the L1 threshold at a given L1 rate. The track quality cuts allow these L1 thresholds to be greatly lowered for a fixed rate by rejecting fake high-$p_T$ tracks while preserving the signal efficiency. The trigger turn-on efficiency (bottom) is measured as the efficiency for events with generator level $H_T$ (left) and $H_T^{miss}$ (right) to pass a given L1 trigger threshold. The generator level jets used to compute these quantities have $p_T > 30$ GeV and $|\eta| < 2.4$ to emulate offline jet selection. A 25 kHz L1 threshold is chosen for track $H_T$ and 35 kHz L1 threshold for $H_T^{miss}$. The L1 thresholds applied are for the online track $H_T$ and track $H_T^{miss}$ at the selected L1 rate. The efficiency turn-on for the BSM stop signal at 95% is greatly lowered with the track purity requirements.

most of the jet energy and integrating as little pileup energy as possible, in agreement with a preliminary study presented in the HGCAL Technical Design Report [10]. However, when using PUPPI candidates, a $9 \times 9$ jet gives the best performance, since the larger jet size includes more energy deposits from the jet, and the PUPPI algorithm suppresses the contributions from pileup. The demonstration of the algorithm in hardware, described later in the section, uses a $7 \times 7$ jet size, and this is used for the menu studies in Chapter 4. Both will soon be updated to use $9 \times 9$ jet size.

The jet energy response in a combination of QCD and $t\bar{t}$ MC samples with 200 average pileup
Figure 3.97: A jet seed tower (green) must have energy greater than, or greater than or equal to, all other towers in the $7 \times 7$ window.

Figure 3.98: Jet response for $7 \times 7$ histogrammed jets in QCD and $t\bar{t}$ MC at PU200 in bins of $p_T$ for the full calorimeter (left), and for the different $\eta$ regions used in the PF algorithm (right).

interactions for the $7 \times 7$ histogrammed PF jet algorithm, using PUPPI candidates as inputs, in bins of $p_T$ for the full calorimeter, and for the different $\eta$ regions used in the PF algorithm with $p_T > 30$ GeV, is shown in Fig. 3.98. The single jet and $H_T$ trigger efficiencies for $7 \times 7$ histogrammed PF jets and AK4 PF jets are compared in Fig. 3.99. The performance of the histogrammed PF jet algorithm is close to that of the offline AK4 algorithm when PUPPI inputs are used for both algorithms. The efficiencies for $9 \times 9$ jets are shown in Fig. 3.100, and show very close agreement to the offline algorithm. The single jet and $H_T$ trigger rates for $7 \times 7$ and $9 \times 9$ histogrammed PF jets and AK4 PF jets using MC samples with 200 average pileup interactions are shown in Fig. 3.101. The histogrammed algorithm produces similar rates to the AK4 algorithm.

The algorithm has been implemented in HLS with the assumption that time-multiplexing will be used with a period of 6. The implementation is split into three modules running in sequence. The first module receives inputs from regions of the detector and returns histograms of these inputs. The second buffers histograms from these regions until all bins in $\eta$ with the same $\phi$ coordinate have been received. Bins corresponding to detector slices sorted in $\phi$ are then sent to the jet clustering module, which builds the jets. The algorithm is designed to be streaming with limited data buffering, which greatly reduces FPGA resource usage. As a proof-of-concept, a scaled-down version of this sequence for $7 \times 7$ jets has been successfully built and loaded onto a Serenity development board carrying a Xilinx KU15P FPGA. The demonstrator finds jets in two regions of the barrel that are received over four consecutive clock cycles at 360 MHz. Resource
3.6. Triggering on jets, hadronic tau decays, and energy sums

Figure 3.99: Single jet (left) and $H_T$ (right) trigger efficiencies in QCD and $t\bar{t}$ MC with 200 average pileup interactions, for $7\times7$ histogrammed jets and AK4 jets, using PUPPI inputs.

Figure 3.100: Single jet (left) and $H_T$ (right) trigger efficiencies in QCD and $t\bar{t}$ MC with 200 average pileup interactions, for $9\times9$ histogrammed jets and AK4 jets, using PUPPI inputs.

Usage for this chain is around 38% of LUTs, 16% of FFs, and 4% of DSPs of a KU15P FPGA. The latency, defined as the number of clocks between the first insertion of data and the first output, is 23 clocks at 360 MHz. The initiation interval (the number of clock cycles before the algorithm is ready to receive the next event) is 17 clock cycles.

To test the firmware implementation in hardware, $t\bar{t}$ MC data was run through the Serenity demonstrator card and firmware output was compared to the output of an independently written emulator implemented in CMSSW. The demonstrator results are presented in Section 6.4.4.2. Firmware and emulator output agrees perfectly for 98% of events. Resource usage estimates for a chain covering the whole detector have been computed based on the results from this demonstrator. 42% of LUTs, 7% of FFs, and 4.2% of DSPs of a VU9P are expected to be required for the $7\times7$ jets, with a latency of 15–20 clock cycles at 360 MHz. The $9\times9$ jet finder is expected to require 58% of LUTs, 11% of FFs, and 4.2% of DSPs of a VU9P at 360 MHz. These numbers have been obtained by producing designs covering successively larger areas, using HLS to estimate their resource usage, and extrapolating to the final design. Optimisations that reduce the LUT usage down to 47% for the full algorithm have been found, and it is expected that additional improvements can be made to the design and implementation to further improve the resource usage and latency. Taking those into account, the $9\times9$ jet finder is expected
3.6.1.4 Jet algorithms performance summary

To provide a direct comparison of the jet algorithms performance, the single jet matching and turn-on combined trigger efficiency for track jets, calorimeter only jets, and histogrammed PUPPI jets for thresholds that provide the same fixed rate, in $\bar{t}t$ events with 200 average pileup interactions, is shown in Fig 3.102, for barrel, endcap and forward jets. Also shown, is the fixed rate efficiency comparison for $H_T$.

3.6.2 Hadronic $\tau$ algorithms

The reconstruction of hadronically decaying tau leptons ($\tau_h$) is explored with a variety of algorithms to maximally take advantage of the detector systems and provide robustness. Different algorithms are developed based on calorimeter inputs only, tracker-only or tracks matched to e/\gamma clusters, and using particle-flow candidate inputs. Two implementations of the $\tau_h$ reconstruction based on particle-flow candidates are explored; both have first versions of firmware implementations.

3.6.2.1 Calorimeter-based hadronic $\tau$ reconstruction

Two approaches to reconstructing hadronic $\tau$ decays relying solely on calorimeter information are explored. The first, providing the default calorimeter-based hadronic $\tau$ reconstruction, is heavily based on the calorimeter e/\gamma and jet reconstruction described in Sections 3.2.2 and 3.6.1.1, respectively, and details of the e/\gamma cluster and tower reconstruction are described there. The second is developed specifically for the HGCAL, taking maximal advantage of the available trigger information provided by this system, and benchmarks the possible future improvements to calorimeter-based $\tau$ reconstruction using more sophisticated techniques.

Tower-based algorithm

The calorimeter-based $\tau_h$ reconstruction algorithm uses the same towers as the calorimeter jet algorithm described in Section 3.6.1.1. For the barrel these are calculated by the barrel calorimeter trigger Layer-1. For HGCAL and HF they are delivered directly as trigger primitives (see Section 2). All corrections are done in the same way as for jets. Since the $\tau$ leptons decaying hadronically can be also reconstructed as a jet, the efficiency of its reconstruction can
be as high as the efficiency of the jet reconstruction. The main challenge is to distinguish between the $\tau$-initiated jet and those from background processes with much higher production cross section. The products of $\tau$ decays are collimated and less spread out than background jets, so in the offline reconstruction a dynamic cone of around 0.05–0.1 is used to define the $\tau_{h}$. At L1, we restrict the core size of the $\tau_{h}$ jet to $3\times5$ ($0.26\times0.44$ for barrel towers) in $\eta-\phi$ towers within the $7\times7$ towers for a reconstructed jet. For the barrel, this corresponds to $0.26\times0.44$ and $0.61\times0.61$ in $\eta-\phi$ for the $3\times5$ and $7\times7$ respectively. Thus the $\tau_{h}$ are seeded in the same way as jets, but only the central $3\times5$ towers are used to define the $p_T$ of the $\tau_{h}$ jet, while the rest of the $7\times7$ jet area is used to calculate the $\tau_{h}$ isolation as shown in Fig. 3.103. The $3\times5$ size of the core is chosen to maximize the $p_T$ resolution of the reconstructed $\tau_{h}$ and to decrease possible PU contribution, and is used in the barrel, endcap and forward calorimeters. The $p_T$
resolution for different geometries of the $\tau_h$-jet core is shown in Fig. 3.104 for the barrel for $\tau_h$ with $p_T > 20$ GeV. It can be seen that the $3 \times 5$ geometry provides a response that is closer to the truth value with resolution slightly better than for the other configurations.

Figure 3.103: The $3 \times 5$ towers in $\eta-\phi$ around the seed tower defines the core of the calorimeter-based $\tau_h$ reconstruction. The core defines the $p_T$ and position of the $\tau_h$. The $7 \times 7$ area is used for isolation.

Figure 3.104: The $\tau_h$ $p_T$ resolution for different sizes of the $\tau_h$ core. The $3 \times 5$ configuration provides better agreement with the generated $\tau_h$ $p_T$ value and also has slightly better resolution than other geometries.

The $\tau_h$ algorithm uses the same inputs and calibration method as the calorimeter jet algorithm, except the calibration is modified by considering the different $\tau_h$ decays. The decay modes considered are single $\pi^\pm$, single $\pi^\pm + \pi^0$, and three $\pi^\pm$ possibly with additional $\pi^0$'s, where $\pi^0$'s decay into two photons. All reconstructed $\tau_h$ can therefore be subdivided into one of three categories according to the number of $e/\gamma$ clusters contained in the $\tau_h$ core, which should reproduce the decay of neutral pions to zero, one, or two $e/\gamma$ clusters. The calibration of the $\tau_h$ energy response therefore is split separately for three categories based on the number of $e/\gamma$ clusters. This improves the energy resolution of the reconstructed $\tau_h$.

The $\tau_h$ isolation cut is calculated by taking the ratio of the energy contained within the $7 \times 7$ jet that is outside of the core $3 \times 5$ region, with the energy in the core region, $(E_{7,7} - E_{3,5})/E_{3,5}$. It is adjusted separately for the barrel and HGCAL as a function of $p_T$ of the reconstructed $\tau_h$, and is defined to keep approximately 99% of reconstructed $\tau_h$ for $p_T$ above 100 GeV and to give a rate reduction at the trigger threshold (32 GeV) of approximately a factor of 2 compared to the non-isolated case. Figure 3.105 shows the $\tau_h$ reconstruction and isolation efficiency as a function of generated $\tau_h$ $p_T$ in the barrel (left) and endcap (right), for a 32 GeV threshold, corresponding to the online $\tau$ threshold for Run-2. Figure 3.106 shows the efficiency as a function of $\eta$ (left), and the corresponding rates (right).

**Optimized HGCAL-based algorithm**
3.6. Triggering on jets, hadronic tau decays, and energy sums

A further optimized \( \tau_h \) identification algorithm for the HGCAL is discussed in this section. It relies solely on the inputs from the endcap calorimeter, covering \( 1.9 < |\eta| < 2.9 \), without making use of any tracker information, and may provide an alternative to the tower-based approach for calorimeter taus in the endcap. The front-end electronics take as input the EE, FH and BH energy deposits and transform them into \( 2 \times 2 \) or \( 3 \times 3 \) trigger cells, typically around \( 0.17 \times 0.17 \) or \( 0.26 \times 0.26 \) in \( \eta - \phi \) respectively. These are then dynamically clustered by the back-end electronics in both the transverse and longitudinal directions, forming the so-called 3D-clusters. The algorithm discussed in this section, referred to as ‘ECHTT’ (Endcap Calorimeter Hadronic Tau Trigger), takes as inputs these 3D-clusters to build the \( \tau_h \) candidates. This is done in three separate steps, which are designed to reject the pileup contributions and collect the relevant energy deposits, estimate the \( \tau_h \) energy, and identify its decay mode. Algorithm
efficiency is measured using di-tau MC samples with 200 average pileup interactions.

The first step of the algorithm consists of rejecting the PU contributions to the $\tau_h$ candidates. This is done by setting a cutoff on the output score of a Boosted Decision Tree (BDT) trained for that purpose. The BDT exploits features related to the transverse shape and the compactness of the 3D-clusters, which show distinct separation between the PU footprint (more diffuse) and $\tau_h$ footprint (more localized). The algorithm sets a working point of 99% signal efficiency, corresponding to a 16% background efficiency, applied on individual 3D-clusters before building the $\tau_h$ candidate in order to avoid potential bias in the calibration.

The second step of the algorithm aims to identify and calibrate the energy deposits in the calorimeter due to the $\tau_h$ interaction. The creation of a L1 tau candidate is initiated, or ‘seeded’, by a local maximum of energy with $E_T \geq 4\text{ GeV}$ which passes the 99% efficiency point of the PU BDT score. The algorithm takes this single 3D-cluster as the tau candidate. Surrounding lower-energy 3D-clusters are not considered, since they mostly arise from residual PU contributions, and the full tau deposit is contained in the leading cluster in 90% of the cases. The position assigned to the L1 $\tau_h$ candidate is taken as the position of the 3D-cluster, which corresponds to a mean difference with respect to the position of the generated visible $\tau_h$ of approximately 0.02 in both $\eta$ and $\phi$ directions, as shown in Fig. 3.107.

![Figure 3.107: Difference in $\eta$ (left) and $\phi$ (right) positions between the generated visible $\tau_h$ and the $\tau_h$ L1 candidate, obtained with di-tau MC samples with 200 average pileup interactions.](image)

After the creation of the L1 $\tau_h$ candidate, a calibration procedure is performed to correct the response and improve the resolution in three different steps. First, an $\eta$-dependent pileup subtraction obtained from a linear regressor is applied. Second, a multiplicative correction based on a BDT regression is applied to improve the resolution. Third, a $p_T$-dependent multiplicative factor is applied to correct for the response as a function of $p_T$. This results in a calibrated energy response centered at 1 with a relative resolution of 19% for 200 average pileup interactions. The resolution before and after calibration in bins of $p_T$ of the generated visible $\tau_h$ can be found in Fig. 3.108.

The third step of the algorithm identifies the decay mode of the $\tau_h$ as a starting point for a better isolation procedure to discriminate against jets. The decay modes considered are single $\pi^\pm$, single $\pi^\pm + \pi^0$s, and three $\pi^\pm$ possibly with additional $\pi^0$s. The decays containing $\pi^0$s have showers that start in the electromagnetic calorimeter, where the energy deposit also tends
3.6. Triggering on jets, hadronic tau decays, and energy sums

Figure 3.108: Resolution in bins of $p_T$ of the L1 $\tau_h$ candidates with respect to the generated $\tau_h$ $p_T$, obtained with di-tau MC samples with 200 average pileup interactions.

to accumulate. In addition, the showers tend to be longer in the 3-prong and 3-prong+$\pi^0$'s decay modes. The decay mode identification is performed with a multi-class Random Forest classifier which exploits the relevant variables related to these different features. The Random Forest provides for each L1 $\tau_h$ candidate a probability associated to each decay mode. The mode with the highest probability is taken as the predicted decay mode. The normalized confusion matrix between predicted and generated decay modes shows a diagonal assignment, as can be seen in Fig. 3.109.

Figure 3.109: Confusion matrix between predicted and generated $\tau_h$ decay modes, obtained with di-tau MC samples with 200 average pileup interactions.

The $\tau_h$ efficiency as a function of the generated visible $\tau_h$ $p_T$ for typical L1 thresholds for di-tau MC samples with 200 average pileup interactions can be found in Fig. 3.110, together with the efficiencies for the different decay modes for a L1 threshold of 50 GeV. The figures show sharp turn-ons and a plateau efficiency of 100%. The single-$\tau_h$ rates corresponding to different L1 thresholds can be can be found in Fig. 3.111.
Figure 3.110: L1 $\tau_h$ efficiency as a function of the generated visible $\tau_h$ $p_T$ for L1 thresholds of 40 GeV, 50 GeV and 60 GeV (left), and for different decay modes for a L1 threshold of 50 GeV, both obtained with di-tau MC samples with 200 average pileup interactions.

Figure 3.111: Single L1 $\tau_h$ rates as a function of the L1 $\tau_h$ $p_T$ threshold, obtained with a neutrino MC sample with 200 average pileup interactions.

3.6.2.2 Track+e/$\gamma$ $\tau$ reconstruction

The $\tau$ reconstruction algorithm presented in this section provides a potential alternative option to the other track-based approaches described so far. This tracks +e/$\gamma$ algorithm exploits the presence of the charged products of a hadronically decaying $\tau$ lepton ($\tau_h$) and uses tracks as its base, covering the full tracking acceptance ($|\eta| \leq 2.5$). The first step of the tracks +e/$\gamma$ algorithm is to construct a tracks-only object by only employing tracking information for the identification of charged hadrons originating from $\tau$ lepton decays. A single track, found as the highest-$p_T$ track in a cone in $\Delta R$, is used as the seed for constructing a $\tau_h$ candidate. The criteria used to select seed tracks were optimised using a multivariate analysis with the rectangular cut optimisation method of the TMVA toolkit [58]. The signal collection consisted of tracks matched to the leading charged daughter of the generator $\tau_h$ ($\tau_h^{vis}$) using $\Delta R$ criteria.
The background collection consists of tracks that fail the matching criteria, and are the highest \( p_T \) track within a cone of \( \Delta R = 0.15 \). The track parameters with the best discrimination in descending order of discriminating power were found to be the reduced chi-squared, \( \chi_v = \frac{\chi^2}{N_{\text{bins}}} \), the bend chi-squared (\( \chi_{\text{bend}} \), defined in Section 3.6.3.1), and the stub multiplicity (\( N_{\text{stubs}} \)). The optimized values of these variables were found to be \( \chi_v \leq 64.6, \chi_{\text{bend}} \leq 6.2 \) for \( p_T < 15 \text{ GeV} \) and \( \chi_v \leq 9.2, \chi_{\text{bend}} \leq 6.2 \) for \( p_T \geq 15 \text{ GeV} \). The stub multiplicity was found to be independent of the track \( p_T \) with the optimal value being \( N_{\text{stubs}} \geq 5 \).

The high-quality and high-\( p_T \) (> 2 GeV) seed tracks are then used to define a dynamic signal cone whose size is inversely proportional to the seed track \( p_T \). Any track found inside the signal cone that satisfies the preselection criteria is considered as a signal track, and its momentum is vectorially added to the seed track. Signal cone tracks are clustered together with the seed track, provided that the track-based invariant mass \( (m_{\text{tks}}) \) is less than 1.5 GeV, and the \( z_0 \) separation of signal cone tracks from the seed track is less than 1.0 cm. Once all signal tracks are clustered, an isolation annulus is constructed starting from the edge of the signal cone and extending to \( \Delta R_{\text{iso}} \leq 0.3 \). Any quality track found inside the isolation annulus is considered in the calculation of the isolation variables. The vertex isolation \( (vtxIso) \) requires that no isolation track is found with a \( z_0 \) value closer than a specified distance from the \( z_0 \) of the seed track. Two vertex isolation working points were studied; 0.25 cm (loose) and 1.0 cm (tight). The relative isolation \( (relIso) \) is defined as the \( p_T \) sum of the isolation tracks consistent with coming from the same vertex as the seed track (\(|\Delta z_0| \leq 0.5 \text{ cm}\)), divided by the \( p_T \) of the seed track. Two relative isolation working points are also studied; 0.25 (loose) and 0.01 (tight).

<table>
<thead>
<tr>
<th>Variable</th>
<th>tracks +e/γ</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_T^{\text{seed}} \geq 2 \text{ GeV} )</td>
<td>transverse momentum of seed track</td>
<td></td>
</tr>
<tr>
<td>( p_T^{\text{sig}} \geq 2 \text{ GeV} )</td>
<td>transverse momentum of tracks inside signal cone</td>
<td></td>
</tr>
<tr>
<td>( E_T^{\gamma} \geq 1.5 \text{ GeV} )</td>
<td>transverse energy of clustered e/γ</td>
<td></td>
</tr>
<tr>
<td>(</td>
<td>\eta^{\gamma/e}</td>
<td>\leq 2.5 )</td>
</tr>
<tr>
<td>( \Delta R_{\text{sig}} \leq 2.5 \text{ GeV} ) ( p_T^{\text{seed}} )</td>
<td>shrinking signal cone parametrisation</td>
<td></td>
</tr>
<tr>
<td>( \Delta R_{\text{iso}}^{\text{max}} = 0.15 )</td>
<td>maximum signal cone opening (cut-off)</td>
<td></td>
</tr>
<tr>
<td>(</td>
<td>z_0^{\text{tks}} - z_0^{\text{tk}i}</td>
<td>\leq 1.0 \text{ cm} )</td>
</tr>
<tr>
<td>( m_{\text{tks}e/\gamma} \leq 1.8 \text{ GeV} )</td>
<td>invariant mass of signal cone (tracks + e/γ)</td>
<td></td>
</tr>
<tr>
<td>( \Delta R_{\text{iso}} \geq \Delta R_{\text{sig}} )</td>
<td>minimum opening of isolation annulus</td>
<td></td>
</tr>
<tr>
<td>( \Delta R_{\text{iso}} \leq 0.30 )</td>
<td>maximum opening of isolation annulus</td>
<td></td>
</tr>
</tbody>
</table>

The second step of the \( +e/\gamma \) algorithm is the addition of neutral energy to the tracks-only object by employing calorimeter information for the identification of neutral hadrons originating from \( \tau \) lepton decays. These are typically neutral pions \( (\pi^0) \), which are amply produced in \( \tau \)-lepton decays and which almost exclusively decay to photon pairs through the decay \( \pi^0 \rightarrow \gamma \gamma \). The algorithm associates nearby e/γ clusters to the tracks +e/γ candidate if they are found inside the signal cone and they also satisfy predefined quality criteria. Starting from the e/γ cluster that is leading in \( E_T \), more e/γ clusters are added as long as the invariant mass of the tracks +e/γ candidate remains below the \( \tau \)-lepton mass, so that \( m_{\text{tks}+e/\gamma} \leq 1.77 \text{ GeV} \). If at least one e/γ cluster is associated with the tracks +e/γ object, the relative isolation variable is also modified to allow the consideration of the neutral energy found inside the isolation annulus. It is defined as the scalar sum of the \( p_T \) of isolation annulus tracks plus the scalar sum of the \( E_T \) of the isolation annulus \( e/\gamma \), divided by the signal cone \( p_T \) of the tracks +e/γ candidate. The full set of parameters that defines the tracks +e/γ algorithm are tabulated in Table 3.8.
The expected rate of the tracks \( +e/\gamma \) algorithm for a single-\( \tau \) selection is shown in Fig. 3.112 (left). A 50 kHz trigger rate can be maintained at 200 pileup with an \( E_T \) threshold of 62 GeV for non-isolated tracks \( +e/\gamma \) candidates. The introduction of track-based isolation results in a noticeable reduction in the online trigger thresholds down to 50 GeV. The corresponding rates for the double-\( \tau \) selection are shown in Fig. 3.112 (right). The seed tracks of the two tracks \( +e/\gamma \) objects are required to be consistent with originating from the same vertex by requiring that their \( z_0 \) values are similar (\( |\Delta z_0| < 1.0 \text{ cm} \)). The 12 kHz trigger rate can be achieved with \( E_T \) thresholds of 40 GeV and 26 GeV for non-isolated and isolated tracks \( +e/\gamma \) candidates, respectively. For both the single-\( \tau \) and double-\( \tau \) selections the tracks-only step is also shown, using tight vertex isolation criteria. The relatively lower thresholds of 40 GeV and 22 GeV for the single-\( \tau \) and double-\( \tau \) selections, respectively, can be understood by the higher seed track \( p_T \) requirement, and from the absence of \( e/\gamma \) clusters which can be susceptible to pileup contamination, especially in the forward regions of the detector.

In Fig. 3.113, the trigger rate versus the corresponding signal efficiency of the tracks \( +e/\gamma \) algorithm is plotted by scanning different \( E_T \) trigger thresholds. The signal efficiency is defined as \( \frac{n}{N} \), where \( N \) is the number of events with all \( \tau^\text{mc} \) within the acceptance and \( n \) is the subset of those events with at least one MC-matched trigger object (\( \Delta R < 0.1 \)). The efficiency is measured with respect to events where the \( \tau \) leptons from the signal process decay hadronically, are within the tracking volume (\( |\eta| < 2.4 \)), and have a generated visible transverse energy (\( E_T^{\text{vis}} \)) larger than 20 GeV. With a 50 kHz trigger rate, the single-\( \tau \) selection efficiency is 40% for non-isolated and 51% for isolated candidates. This represents a notable improvement when compared to the corresponding efficiencies of the tracks-only step (28% and 42%, respectively). For the double-\( \tau \) selection, the same (symmetric) threshold is required for both tracks \( +e/\gamma \) objects, while the seed tracks of the two tracks \( +e/\gamma \) objects are also required to satisfy \( |\Delta z_0| < 1.0 \text{ cm} \). The trigger efficiency for a 12 kHz trigger rate is 18% for non-isolated and 30% for isolated candidates, which also represents a marked improvement from the corresponding values achieved with the tracks-only step (9% and 22%, respectively).
3.6. Triggering on jets, hadronic tau decays, and energy sums

Figure 3.113: Trigger rate as a function of signal efficiency for the tracks $+e/\gamma$ algorithm for the single-$\tau$ (left) and double-$\tau$ (right) selection. For the double-$\tau$ configuration, symmetric thresholds are used and the two tracks $+e/\gamma$ objects are required to satisfy $|\Delta z_0| < 1.0$ cm. Two vertex isolation ($vtxIso$) and two relative isolation ($relIso$) working points are used: 0.25 cm (loose) and 1.0 cm (tight), and 0.25 (loose) and 0.01 (tight), respectively.

Figure 3.114: Energy resolution for the full tracks $+e/\gamma$ algorithm for $\tau$ particles from simulated VBF $H^0 \rightarrow \tau^+\tau^-$ events with an average pileup of 200. The central (0.0 $\leq |\eta| < 0.8$), intermediate (0.8 $\leq |\eta| < 1.6$), and forward (1.6 $\leq |\eta| < 2.5$) pseudorapidity regions (left) are shown separately, and for various $\tau$ decay modes (right). The energy resolution is defined as $\delta E_T/E_{T^{vis}} = (E_T^{vis} - E_{T^{vis}})/E_{T^{vis}}$, where "vis" denotes the visible part of the $\tau$-lepton decay.

In order to further assess the performance of the tracks $+e/\gamma$ algorithm, we define the energy resolution as $\delta E_T/E_{T^{vis}} = (E_T^{vis} - E_{T^{vis}})/E_{T^{vis}}$, where "vis" denotes the visible part of the $\tau$-lepton...
To study the turn-on, three different definitions are used to quantify the per-object efficiency. The matching efficiency is defined as \( \frac{n}{N} \), where \( N \) is the number of \( \tau_h^{\text{mc}} \) objects with \( |\eta^{\text{vis}}| \leq 2.4 \) and \( n \) is the inclusive subset matched to a tau trigger object nearby (\( \Delta R \leq 0.1 \)). The threshold efficiency is defined as \( \frac{m_X}{M} \), where \( M \) is the number of \( \tau_h^{\text{mc}} \) objects with \( |\eta^{\text{vis}}| \leq 2.4 \) and which are matched to a tau trigger candidate with \( \Delta R \leq 0.1 \). The term \( m_X \) is the subset of the \( M \) objects with \( E_T > X \) GeV. Finally, the matching + threshold (turn-on) efficiency is defined as \( \frac{n_X}{N} \), where again \( N \) is the number of \( \tau_h^{\text{mc}} \) objects with \( |\eta^{\text{vis}}| \leq 2.4 \). The numerator \( n_X \) is the subset of \( N \) objects which are matched to a tau trigger candidate (\( \Delta R \leq 0.1 \)) and also satisfy \( E_T \geq X \) GeV. The efficiency curves of the tracks +e/γ algorithm are shown in Fig. 3.115 (left) using tight vertex isolation criteria, and with all three efficiency types overlaid. The matching efficiency, which describes how good the tracks +e/γ algorithm is at reconstructing a \( \tau_h \) without considering energy thresholds, reaches a plateau of 85% at \( E_T^{\text{vis}} \approx 40 \) GeV. The threshold efficiency describes
the effect of the energy resolution on the per-object efficiency, by using tracks $+e/\gamma$ objects that have already been matched to a $\tau_h^{\text{mc}}$. In this case, the step region is offset to about 10 GeV and rises fast before reaching a plateau at $E_{\text{vis}}\simeq 40$ GeV, with efficiency extending well over 95%. The combined matching plus threshold, or turn-on curve, which includes both the resolution and the matching effects, has a step region that extends to $E_{\text{vis}}\simeq 60$ GeV, and a plateau reaching an efficiency of 80-85%. In Fig. 3.115 (right) a breakdown of the turn-on efficiency curve is shown for various $\tau_h$ decay modes, demonstrating that the overall performance of the tracks $+e/\gamma$ algorithm is largely independent of the $\tau_h$ decay mode.

From the above results it can be concluded that the tracks-only step of the tracks $+e/\gamma$ algorithm is inferior in terms of resolution and overall performance. Nevertheless, there are reasons why such algorithm might be explored with the capabilities of the L1 trigger architecture. The motivations for a standalone tracks-only trigger algorithm are numerous and extend beyond complementarity, with a prime example being the study of the calorimeter response to positively and negatively charged particles. The uncertainty of the jet energy response can be extracted by propagating the calorimeter response for single isolated charged hadrons in data and comparing with MC simulations. The results presented in this section also show that overall, the tracks $+e/\gamma$ algorithm has a simple yet effective design that delivers very good efficiency that is comparable to other complex algorithms. It provides good resolution and rate control at high pileup, with the robustness and flexibility to cope with changes to the LHC running conditions, but also the simplicity to easily maintain optimal performance.

### 3.6.2.3 Particle-flow based $\tau$ reconstruction

Currently, two implementations of tau reconstruction are available that utilize particle-flow candidate inputs. The main neural network based algorithm, which has been implemented and demonstrated in firmware, and an algorithmic approach to particle-flow reconstruction. Both of these algorithms are implemented with inputs that consider either the option of PUPPI particle candidates as inputs or particle-flow particle candidates.

### 3.6.2.4 Neural network tau reconstruction

The neural network $\tau_h$ algorithm is summarised in the following steps:

- Iteratively seed from the highest-$p_T$ charged particles that have $\Delta R > 0.4$ from each other.
- For each particle, take all particle-flow or PUPPI candidates within $\Delta R < 0.4$.
- Of these candidates, take the 10 highest-$p_T$ candidates within the cone, compute $p_T$, $\Delta \eta_{\text{seed}}$, $\Delta \phi_{\text{seed}}$, and particle ID (40 inputs total), and input these variables into a dense neural network.
- For all candidates within $\Delta R < 0.1$ of the seed, assign them to the $\tau$ candidate to compute its four-vector.

The neural network approach relies on using event level quantities. The advantage of using event level quantities is that it improves the overall efficiency of the $\tau_h$ selection at the cost of more complicated firmware that incorporates full event level information. Additionally, the use of neural networks within the L1 trigger is a relatively new concept. Despite these added complications, a firmware implementation already exists for this algorithm. Within the current NN $\tau_h$ algorithm firmware, the final output collection is sorted and the top 10 candidates are considered for further processing. Currently, all $\tau_h$ candidates are required to have $|\eta| < 2.5$.

Figure 3.116 shows the reconstructed $\tau_h$ resolution. For single $\pi^{\pm}$ and three $\pi^{\pm}$ tau candidates,
the resolution is narrow and is dominated by the track resolution. When additional $\pi^0$ are present, the resolution is worse with asymmetric tails corresponding to the addition of pileup.

To reduce the rate for lower-$p_T$ particles, a cut can be placed on the neural network output. This reduction is shown in Fig. 3.117. As example working points, we consider the optimized cuts defined as:

- **Tight PF/PUPPI NN $\tau$ identification:** $NN \times (0.1 + 0.2 \times (\min(p_T, 100.))) / 20.1 > 0.25$
- **Loose PF/PUPPI NN $\tau$ identification:** $NN \times (0.1 + 0.2 \times (\min(p_T, 100.))) / 20.1 > 0.05$

For each scenario, the respective inputs (i.e. PUPPI or PF) are used in either the computation of the neural network or in the isolation computation. It is also possible for PF/PUPPI to use track isolation or PF/PUPPI candidates. The performance of track isolation was found to be worse than summing over all candidate inputs, which is in turn worse than the neural network discriminator. For the NN $\tau$ ID, a sliding cut is used to ensure that the efficiency plateaus to a high value when the $\tau_h$ candidate has a high energy. This is shown in Fig. 3.116, where the PF+PUPPI loose ID is found to plateau at roughly 88% efficiency.

The energy response of the different $\tau_h$ reconstruction algorithms are shown in Fig. 3.116. The use of PUPPI candidates combine to yield the closest resolution to the visible $\tau_h$.

The efficiency is shown for various levels of the PF/PUPPI $\tau$ reconstruction in Fig. 3.116. Importantly, the efficiency of the NN+PUPPI combination gives the best performance. The slow turn-on is a result of the sample choice (gluon fusion $h \rightarrow $tt). When more jets are present, such as for di-Higgs production in the $h \rightarrow \tau \tau$ and $h \rightarrow b\bar{b}$ final state, the collision vertex is easier to identify and the turn-on with PUPPI-based $\tau_h$ identification is improved with both a sharper turn-on and a higher plateau efficiency.

![Figure 3.116](image_url)

Figure 3.116: (Left) Energy scale for $\tau_h$ for NN-based $\tau_h$ using PUPPI or particle-flow candidates as inputs. (Right) Efficiency for the loose working point of the PF and PUPPI based NN Tau ID. For the scale selection the loose working point is also applied. In both cases the Standard Model gluon fusion produced $h \rightarrow $tt sample is used.

To understand the performance in the critical region of low $\tau_h$ $p_T$, we consider the signal efficiency for $h \rightarrow \tau \tau$ decays, requiring $\tau_h^{\text{gen}} p_T > 40$ GeV. To compute the rate, the $p_T$ cut on reconstructed $\tau_h$ is adjusted and the corresponding efficiency for a given $p_T$ cut is computed. The performance of the single-$\tau_h$ trigger is shown in Fig. 3.118. The gain is largest for the NN+PUPPI combined $\tau_h$ algorithm. The impact of PUPPI is larger for the di-$\tau_h$ trigger when
3.6. Triggering on jets, hadronic tau decays, and energy sums

η_{\text{gen}} > 60 \text{ GeV})

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.117}
\caption{(Left) Efficiency for a single $\tau_h$ trigger to pass the loose working point at 50 kHz for $p_T > 60$ GeV generator level $\tau_h$. The efficiency is computed with a gluon fusion Higgs boson to $\tau_h$ sample. (Right) Rate for single $\tau_h$ selection using gluon fusion $h \rightarrow \tau_h \tau_h$ selection and applying the loose $\tau_h$ identification. In both cases the NN ID is used with either PF or PUPPI candidates as input.}
\end{figure}

compared with the single $\tau_h$ trigger, as shown in Fig. 3.119. Here, the gain in performance is shown for the loose $\tau_h$ identification for visible generator-level $\tau_h$ $p_T^{\text{gen}} > 20 \text{ GeV}$. When using PUPPI, the performance can be viewed as a worst case scenario, in terms of physics selection, since analyses with low $p_T \tau_h$ usually require additional objects in the final state, such as jets. The additional objects help to improve the primary vertex definition and thus improve the selection efficiency. As a reference for the level of improvement, the NN $\tau_h$ identification efficiency without any selection based on the NN discriminator is found to increase from a plateau efficiency of 89% in the barrel to 94%, when considering the $\tau_h$ in gluon fusion $h \rightarrow \tau \tau + h \rightarrow b\bar{b}$ events.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.118}
\caption{Efficiency compared to rate for a di-$\tau_h$ trigger selection requiring $|\eta| < 2.5$. Single-$\tau_h$ efficiency is defined as the selection of at least one $\tau_h$ decay from the Higgs boson where the lowest $\tau_h^{\text{gen}}$ $p_T^{\text{gen}} > 40 \text{ GeV}$ and the $\tau_h$ is found with $|\eta^{\text{gen}}| < 2.5$.}
\end{figure}

Hardware implementations of the particle-flow and NN $\tau_h$ algorithms exist. For the PF and PUPPI $\tau_h$ algorithms the algorithm implementation has been performed by reading in the whole event on a region-by-region basis as is planned with the current implementation of particle-flow in Layer-1 of the Correlator Trigger, described in Section 5.5. The resource es-
Figure 3.119: Rate (left) and efficiency compared to rate (right) for a di-τ_h selection on gluon fusion H→τ_hτ_h. In this scenario, loose identification criteria is applied for the NN and τ_h are required to have |η| < 2.5. Finally, di-τ_h efficiency is defined requiring with a preselection of a di-τ_h decaying Higgs boson where the lowest τ_h^{gen} p_T^{gen} > 20 GeV and both τ_h are found with |η^{gen}| < 2.5.

3.6.2.5 Hadron plus strips tau reconstruction

An alternative approach to τ reconstruction at L1 using particle-flow candidates, the “hadron plus strip” (HPS) algorithm, uses parallel regions separated in η and φ, that form a grid across the detector, from η = -3.5 to η = 3.5. This yields 92 regions, each with a width in η-φ space of 0.7×0.7, corresponding to 92 possible τ_h candidates. For each region, the algorithm performs several steps to obtain the candidate:
3.6. Triggering on jets, hadronic tau decays, and energy sums

Table 3.9: HLS estimate for resources with various \( \tau_h \) algorithms. The parenthesis indicate the FPGA utilized.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Clock(MHz)</th>
<th>Latency(clocks)</th>
<th>II</th>
<th>DSP</th>
<th>FF</th>
<th>LUTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF Tau(VU9P)</td>
<td>240</td>
<td>125</td>
<td>6</td>
<td>0</td>
<td>1228</td>
<td>48168</td>
</tr>
<tr>
<td>NN Tau(KU115)</td>
<td>240</td>
<td>154</td>
<td>1</td>
<td>1514</td>
<td>90893</td>
<td>96382</td>
</tr>
<tr>
<td>NN Tau+Inputs(KU115)</td>
<td>240</td>
<td>285</td>
<td>30</td>
<td>1514</td>
<td>161187</td>
<td>198965</td>
</tr>
<tr>
<td>NN Tau(VU9P)</td>
<td>240</td>
<td>51</td>
<td>1</td>
<td>1387</td>
<td>154230</td>
<td>89809</td>
</tr>
<tr>
<td>NN Tau+Inputs(VU9P)</td>
<td>240</td>
<td>169</td>
<td>30</td>
<td>1467</td>
<td>189207</td>
<td>177810</td>
</tr>
</tbody>
</table>

- A seed is determined by looking for the highest-\( p_T \) charged particle within each of the 92 detector regions.
- A grid of \( 5 \times 5 \) smaller regions, each with \( \Delta \eta = 0.0874 \) and \( \Delta \phi = 0.0874 \), is placed about the seed candidate. These smaller regions are aligned to match the size of hadron calorimeter towers.
- The highest sum of electron & photon candidates in \( \eta \) and spanning the full \( \phi \) range within this grid, is taken as the electromagnetic object (referred to as a strip); the \( \eta \) and \( \phi \) position is determined as the energy-weighted location of the strip.
- All additional tracks that are found within a \( \Delta R \) cone of varying size, from \( \Delta R < 0.18 \) for \( 3 < p_T < 5 \text{ GeV} \) ranging to \( \Delta R < 0.03 \) for \( p_T > 25 \text{ GeV} \), are added to the \( \tau_h \) candidate.
- The remaining strips as well as all tracks within \( \Delta R < 0.4 \) are summed and computed into isolation variables.

These steps aim to match the current offline HPS algorithm. The limitation here is that restricted \( \eta-\phi \) region limits the algorithm to one candidate per region, and additionally does not deal effectively with candidates on the edge of the regions. However, this approach is easily parallelizable and can be implemented on an FPGA with a low amount of resources.

Once the candidates are constructed in each separate region, they are sorted in \( p_T \) and the highest 12 candidates are selected. These candidates can further be discriminated through the use of the isolation variables. Currently, all \( \tau_h \) candidates are required to have \( |\eta| < 2.5 \).

Similar to the offline \( \tau \) identification at CMS, for each L1 PF reconstructed tau an MVA tau ID is used. The MVA provides three working points: loose (90% efficiency), medium (85% efficiency), and tight (70% efficiency). The input to this MVA is the Level-1 \( p_T, \eta, \) strip \( p_T, \) decay mode, \( dz \) between the \( \tau z_0 \) and the L1 primary vertex \( z_0 \), H/E ratio for the leading charged hadron, the \( \tau z_0 \) and the charged isolation sum in a cone of \( \Delta R < 0.4 \). The \( \tau z_0 \) is defined as the \( z_0 \) of the leading hadron in the \( \tau \). We expect a reduction of rate by approximately a factor of 2 for the medium working point.

Reconstructing \( \tau_h \) candidates with the HPS algorithm without using the approach of dividing the detector into a grid has also been studied, to try and improve the reconstruction efficiency for \( \tau_h \) at the edges of the grid. This “non-grid” HPS algorithm is seeded by the combination of charged candidates of \( p_T > 5 \text{ GeV} \) and jets of \( p_T > 30 \text{ GeV} \). At the highest luminosity expected at the HL-LHC, corresponding to 200 pileup interactions per bunch crossing, only 5 such seeds are expected per bunch crossing, allowing parallelization of the HPS algorithm by seed instead of by detector region.

A few further modifications are made to either increase its efficiency or reduce its rate, typically
Chapter 3. Trigger algorithms

at the cost of a moderate increase in the computational complexity of the algorithm. The size of
the signal cone used to select charged particles produced in the $\tau$ decay and considered in the
reconstruction of the $\tau_h$ momentum is adjusted as a function of the scalar sum in $p_T$, denoted
by the symbol $p_S^\tau$, of the charged hadrons, photons and electrons that are within a cone of size
$\Delta R = 0.4$, centered on the direction of the leading track. The latter refers to the charged particle
that seeds the $\tau_h$ reconstruction or, for $\tau_h$ candidates built from jet seeds, to the charged particle
of highest $p_T$ within the jet. The non-grid HPS algorithm uses a signal cone of size inversely
proportional to $p_S^\tau$, $2.8/p_S^\tau$, taking into account that the collimation of the particles produced
in $\tau$ decays typically increases with the $p_T$ of the $\tau$, due to the Lorentz boost in the direction of
the $\tau$. In case the size of the signal cone thus computed is $< 0.05$ or $< 0.10$, it is set to 0.05 or
0.10 respectively.

The size of the strip used by the non-grid HPS algorithm amounts to $0.05 \times 0.20$ in $\eta \times \phi$.
The signal cone and strip are both centered on the direction of the leading track. Charged
particles within an isolation cone of size 0.4, centered on the direction of the leading track, and
neither included in the signal cone nor in the strip, are used for the purpose of computing the
isolation of the $\tau_h$. The scalar sum in $p_T$ of those charged particles is denoted by the symbol $I_{ch}$
and constitutes the main handle to separate genuine $\tau_h$ from the large background of quarks
and gluon jets and hence the main handle to reduce the trigger rate. In the non-grid HPS
algorithm, only those charged particles that are compatible with originating from the same
vertex as the leading track are considered for the purpose of computing the $\tau_h$ momentum
as well as for computing the isolation of the $\tau_h$. The condition for charged particles to be
considered as originating from the same vertex as the leading track is that their track cross the
beam axis within a distance of $\Delta z < 0.4$ cm with respect to the crossing of the leading track.

For $\tau_h$ reconstructed by the non-grid HPS algorithm to pass the $\tau_h$ trigger, the leading track
of both $\tau_h$ are further demanded to originate from the same vertex. In order to satisfy this
condition, the leading tracks must cross the beam axis within a distance of less than 0.4 cm
from each other.

Using isolation thresholds which are relaxed in proportion to the $p_T$ of the $\tau_h$ can increase the
trigger efficiency for $\tau_h$ of high $p_T$, taking advantage of the fact that the background from quark
and gluon jets is mainly concentrated at low $p_T$.

Feasibility studies for implementing the non-grid HPS algorithm in FPGA firmware are cur-
rently in progress. It should be noted however that since the algorithm uses jet candidates to
seed the tau candidates, latency requirements and the final trigger architecture are particularly
important when considering this algorithm.

3.6.2.6 Tau algorithms performance summary

To provide a direct comparison of the algorithms performance, the single (left) and di-$\tau$ (right)
combined matching and turn-on efficiencies for the various algorithms, in SM HH $\rightarrow b\bar{b}t^+\tau^-$
events at 200 average pileup events, for thresholds that provide the same fixed rate is shown in
Fig 3.120.

3.6.3 Missing transverse energy algorithms

The reconstruction of the missing transverse energy in an event is an important trigger signa-
ture. With the high pileup expected at the HL-LHC, utilizing L1 track information is critical in
achieving acceptable trigger rates at thresholds that will enable the physics goals of the experi-
ment. Two different algorithms are explored. The first algorithm is based on using tracks-only
3.6. Triggering on jets, hadronic tau decays, and energy sums

3.6.3.1 Tracker-based algorithm

The vector sum of the transverse momenta ($E_{\text{miss}}$) of all particles produced in the primary interaction is a key input for triggering on BSM signatures at L1. For the track-based algorithm described in this section, one of the main challenges is to exclude tracks from incorrect hit combinations, “fake tracks”, that have high transverse momentum. Although these tracks are rare in pileup interactions after requiring tracks to originate from within a tight window around the primary vertex $z$ position, events containing these high-$p_T$ tracks would dominate a L1 $E_{\text{miss}}$ trigger. The algorithm makes use of track purity requirements, in addition to the requirement on the $\Delta z$ between the primary vertex and the track, $\Delta z(z_{\text{PV}}, z_{\text{trk}})$, to reduce the number of events with poorly measured momentum balance.

The track purity selection is based mainly on the confines of the detector, the track reconstruction algorithm, and the available track fit quality parameters. The track $p_T$ and $\eta$ requirements shown in Table 3.7 are based on the minimal $p_T$ of tracks that can be reconstructed reliably and the tracking acceptance. The minimal number of track layers is also based on the minimal requirement for the track fit. There is a further requirement that 4 of the track layers must also have stubs to remove any 4-stub tracks that were created with only hits in 3 layers. Requirements on the track fit quality measured as $\chi^2/\text{dof}$ is set to a maximum of 40 to reduce tracks that are poorly reconstructed. To further reject fake tracks, a new variable, bend consistency, $\chi^2_{\text{bend}}$ is measured. This measure of the bend of the track is uncorrelated to the $\chi^2/\text{dof}$ from the track fit. $\chi^2_{\text{bend}}$ is calculated based on the horizontal distance between the two consecutive track hits.
hits in a track module – the stubs. The bend consistency quantifies how compatible the bend of the track hits measured in the detector is with the reconstructed track $p_T$. The bend consistency variable is defined as:

$$
\chi^2_{\text{bend}} = \frac{1}{N_{\text{stubs}}} \sum_{i=1}^{N_{\text{stubs}}} \left( \frac{\beta_i - \beta_i^{\text{exp}}}{\sigma_i} \right)^2
$$  \hspace{1cm} (3.20)

where $N_{\text{stubs}}$ is the total number of stubs comprising the track, $\beta_i$ is the measured bend angle for stub $i$, and $\beta_i^{\text{exp}}$ is the expected bend angle based on the track. A bad combination of track hits tends to have a large value of bend consistency compared to well-reconstructed tracks. The cut on $\chi^2_{\text{bend}}$ is optimized to lower the fraction of poorly reconstructed tracks that are included in the $E_T^{\text{miss}}$ calculation, as can be seen in Fig. 3.121 which shows the reduction in the fake rate from these optimized cuts, as well as maintain a reasonable L1 data-taking rate for a track-based $E_T^{\text{miss}}$ trigger. The fake rate is defined as the percentage of tracks produced by incorrect combinations of stubs that cannot be matched to truth-information. The largest suppression of fakes comes from cutting on the $\chi^2_{\text{bend}}$ variable. Reducing the fake rate is critical to the track $E_T^{\text{miss}}$ algorithm since fake tracks can result in a significant overestimate of momentum in the event, and these events can then dominate the rate of the track $E_T^{\text{miss}}$ trigger.

![Figure 3.121: The full track selection greatly reduces the fake rate.](image)

The L1 vertex described in the GTT firmware algorithm in Section 3.4.1 is input to the L1 track-based $E_T^{\text{miss}}$ algorithm. A window of $\Delta z (z_{PV}, z_{\text{trk}})$ is required for pseudo-rapidity regions within the tracking acceptance. Table 3.10 shows the range of values for the allowed range of $z_{\text{trk}}$ around the measured primary vertex $z_0$. Central tracks allow for a tight window of 0.4 cm while more forward tracks need a wider $z$-window (as large as 2.2 cm) to include tracks within $3\sigma$ of the Gaussian $z_{\text{trk}}$-resolution.

To gauge the performance of the track-based $E_T^{\text{miss}}$ we compare the reconstructed track-based $E_T^{\text{miss}}$ to simulated tracking particles, which are the simulation-level trajectories of charged particles in an event. The tracking particles are used to compute the simulated $E_T^{\text{miss}}$ using the same algorithm as the reconstructed L1 tracks. This gives the expected L1 track $E_T^{\text{miss}}$ in the case where no fake tracks are included in the $E_T^{\text{miss}}$ computation. Figure 3.122 shows the data-taking
3.6. Triggering on jets, hadronic tau decays, and energy sums

Table 3.10: Minimum $\Delta z$ requirements between the primary vertex and $z_{\text{trk}}$ in each pseudorapidity region for the track selection for the track-based $E_{\text{T}}^{\text{miss}}$ algorithm.

| $\eta$ range | $|\Delta z (z_{\text{PV}}, z_{\text{trk}})|$ (cm) |
|--------------|---------------------------------------------|
| $0 \leq |\eta| < 0.7$ | 0.4 |
| $0.7 \leq |\eta| < 1.0$ | 0.6 |
| $1.0 \leq |\eta| < 1.2$ | 0.76 |
| $1.2 \leq |\eta| < 1.6$ | 1.0 |
| $1.6 \leq |\eta| < 2.0$ | 1.7 |
| $2.0 \leq |\eta| < 2.4$ | 2.2 |

L1 rate for the track $E_{\text{T}}^{\text{miss}}$ from the reconstructed tracks and the tracking particles. Applying only the $|\Delta z (z_{\text{PV}}, z_{\text{trk}})|$ cut in Table 3.10 gives an L1 rate of 35 kHz for a track-based $E_{\text{T}}^{\text{miss}}$ threshold at 200 GeV. Applying the full track selection in Table 3.7, this threshold is lowered to 70 GeV, which is the same threshold as when only simulated charged particle trajectories are used for the $E_{\text{T}}^{\text{miss}}$ computation. This allows a low trigger threshold for offline-$E_{\text{T}}^{\text{miss}}$ for BSM models like a BSM stop signal. The trigger turn-on at 95% efficiency is lowered from 700 GeV applying only the primary vertex constraint to 350 GeV when including the full selection of tracks.

A firmware implementation for track $E_{\text{T}}^{\text{miss}}$ is available, and has been tested in a hardware demonstrator. This is presented in Section 6.4.3.2.

![Graph](image)

Figure 3.122: Track $E_{\text{T}}^{\text{miss}}$ rate from single neutrino MC events at 200 average pileup interactions (left). The full track selection reduces the threshold from 200 GeV to 70 GeV, which is closely aligned with the threshold from the simulated tracks. Track $E_{\text{T}}^{\text{miss}}$ efficiency for SUSY stop MC (right), with the stop mass at 1000 GeV and the LSP mass at 775 GeV, at 200 average pileup interactions.

3.6.3.2 Particle-flow based algorithm

An estimate of the missing transverse momentum based on particle-flow candidates incorporates information from all sub-systems of the CMS detector. The PUPPI algorithm may be applied to the resulting candidates to mitigate contributions due to pileup before higher-level
quantities such as $E_T^{\text{miss}}$ are calculated (see Section 3.5 for description about the particle-flow and PUPPI algorithms).

In general, the determination of $E_T^{\text{miss}}$ requires a largely uniform response to objects throughout the full $\eta$ coverage of the detector. Thus, together with di-jet triggers targeting VBF topologies, the $E_T^{\text{miss}}$ trigger is a driver of the requirements on forward hadronic objects. In the construction of PF and PUPPI candidates, the performance of each trigger is therefore sensitive to the set of input objects, and in particular the threshold applied to calorimeter clusters identified within the HF system. Low-energy HF deposits are dominated by pileup contributions and cannot be removed with the aid of tracking information at L1, so cluster $p_T$ is the dominant handle to remove noise contributions to the $E_T^{\text{miss}}$ calculation. The reduction in trigger rate and signal efficiency is thus studied as a function of HF threshold in two standard signal topologies: $t\bar{t}$ production and VBF production of an invisibly-decaying Higgs boson.

Figure 3.123 (left) demonstrates the clearest impact of the HF threshold by considering a pure di-jet trigger targeting the VBF topology. The effect is assessed by considering VBF Higgs to invisible events that enter the high signal-purity phase space selected by offline analyses: events which have a leading jet $j_1$ with $p_T > 70$ GeV and a sub-leading jet $j_2$ with $p_T > 40$ GeV, which must satisfy both $\Delta \phi(j_1, j_2) < 2$ and $\Delta \eta(j_1, j_2) > 4$. This selection and associated motivation is discussed further in Section 3.7 on Global Trigger strategies for the invisible Higgs scenario. Further, it is required that signal events have one of the two leading jets well-contained within the HF ($|\eta| > 3.4$) for the purpose of this study.

Figure 3.123 (right) shows the impact of the varied HF threshold on the $E_T^{\text{miss}}$ trigger build from PF+PUPPI inputs. The efficiency for VBF Higgs to invisible events and trigger rate are scanned as a function of the requirement on PF+PUPPI $E_T^{\text{miss}}$. Both $m_{jj}$ and $E_T^{\text{miss}}$ triggers are found to be insensitive to changes in the HF threshold up to 15 GeV, beyond which a noticeable degradation in performance is seen.

The performance of the PUPPI $E_T^{\text{miss}}$ calculation at L1 is compared for both benchmark signals considered in Fig. 3.124, and shows comparable performance in each scenario.

The $E_T^{\text{miss}}$ algorithm targeting PF+PUPPI inputs has been implemented in firmware using high-level synthesis (HLS) tools. The number of expected PF candidates after PUPPI to be included in the $E_T^{\text{miss}}$ calculation is expected to be less than 90 in 95% of events based on simulation of $t\bar{t}$ events with 200 average pileup interactions per proton-proton bunch crossing. The predetermined $p_T$ and $\phi$ for each candidate is transformed to $x$- and $y$-components using look-up tables and accumulated to determine the total $E_T^{\text{miss}}$ in each direction. Components are squared and added to determine the magnitude, while the $\phi$ coordinate is obtained based on the ratio of components, using look-up tables to perform the division and trigonometric functions. Calculation precision is maintained so that differences with respect to the floating-point calculation are at the sub-% level. The algorithm is implemented with a 240 MHz clock frequency and accepts new inputs at 25 ns intervals. The resource utilization estimated from HLS is presented in Table 3.11 for a range of input particle multiplicities.

### 3.6.3.3 Missing transverse energy algorithms performance summary

To provide a direct comparison of the $E_T^{\text{miss}}$ algorithms performance, the combined matching and turn-on trigger efficiency for track $E_T^{\text{miss}}$ and PUPPI $E_T^{\text{miss}}$, for thresholds that provide the same fixed rate, in $t\bar{t}$ events at 200 average pileup interactions, are shown in Fig 3.125.
3.6. Triggering on jets, hadronic tau decays, and energy sums

Figure 3.123: Signal efficiency for VBF Higgs to invisible events is shown (left) for trigger selections with a fixed rate of 20 kHz. The trigger requires two high-$p_T$ jets, aligned in azimuth, with large rapidity gap, and large di-jet mass. The efficiency is evaluated on signal events passing a VBF selection and with a jet well-contained in the HF. Performance is compared for various PF candidate thresholds in the HF. The addition of a 15 GeV threshold shows negligible impact on the performance for the VBF signal, while a 50 GeV requirement significantly reduces efficiency. Rates and efficiencies for all VBF Higgs to invisible events are compared for $E_T^{\text{miss}}$ triggers constructed from PUPPI inputs (right). Performance is compared for various candidate thresholds in the forward calorimeter (HF). In each case, the addition of a 15 GeV threshold shows negligible impact on the performance for the VBF signal, while a 50 GeV requirement significantly reduces efficiency.

Figure 3.124: Rate and efficiency are shown as a function of the requirement on $E_T^{\text{miss}}$ constructed from PF candidates with the PUPPI algorithm applied. The efficiency is shown for two signals: tt and VBF production of an invisibly-decaying Higgs boson.
Table 3.11: Summary of the latency and resources required for the PF+PUPPI$_{E_T}^\text{miss}$ algorithm for a range of input particle multiplicities. Results use a 240 MHz clock and accept a new set of inputs every 25 ns. Resource utilization estimates are presented in absolute units and as a percentage of one VU9P Super Logic Region (SLR).

<table>
<thead>
<tr>
<th>Input PF+PUPPI multiplicity</th>
<th>50</th>
<th>100</th>
<th>150</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRAM</td>
<td>33 (2%)</td>
<td>49 (3%)</td>
<td>65 (4%)</td>
</tr>
<tr>
<td>DSP</td>
<td>21 (&lt;1%)</td>
<td>37 (1%)</td>
<td>53 (2%)</td>
</tr>
<tr>
<td>FF</td>
<td>3 400 (&lt;1%)</td>
<td>7 900 (&lt;1%)</td>
<td>15 000 (1%)</td>
</tr>
<tr>
<td>LUT</td>
<td>6 300 (1%)</td>
<td>12 000 (3%)</td>
<td>18 000 (4%)</td>
</tr>
<tr>
<td>Latency</td>
<td>33 ns</td>
<td>75 ns</td>
<td>96 ns</td>
</tr>
</tbody>
</table>

Figure 3.125: Comparison of combined matching and turn-on efficiency for $E_T^\text{miss}$ algorithms using thresholds that provide a fixed online rate of 18.1 kHz.
3.7 Global trigger algorithms

The Global Trigger is responsible for implementing the trigger menu (described in Section 4) by parallel evaluation of $\mathcal{O}(1000)$ trigger algorithms with each one selecting a specific event signature. Trigger algorithms are based on one or multiple physics objects identified by the upstream trigger systems (see Section 1.3). They may place constraints on the variables of single objects, multiple objects or on topological correlations between objects, including the invariant mass computed from any pair of objects in the same collection, their transverse mass or the transverse momentum of the hypothetical mother particle (two-body $p_T$). Classical cut-based trigger algorithms apply thresholds on the objects variables and correlation variables [6]. New MVA-based algorithms may use these variables as an input to an MVA classifier, such as a neural network. In order to facilitate searches for long-lived particles with a lifetime longer than a bunch crossing (BX), the Global Trigger will also be able to compute so-called inter-BX conditions, correlating objects with each other, that are several (up to $\pm 3$) BX apart. The Global Trigger provides functionality to prescale\textsuperscript{1} individual algorithms, to enable individual algorithms only in certain bunch crossings of the orbit (applying a bunch mask\textsuperscript{2}) and to monitor their rate, both before and after taking into account dead time. It determines the trigger type based on what algorithms fired and sends it to the Phase-2 Timing and Control Distribution System (TCDS-2) [59].

The inputs to the Global Trigger algorithms are collections of objects identified by the upstream trigger systems. In general, di-object correlations such as $\Delta \eta$, $\Delta \phi$, $\Delta R$, invariant mass are calculated in the Global Trigger from the single objects, as an upstream calculation would lead to excessive bandwidth requirements due to combinatorics. One exception may be multi-track objects such as $f$ or $B$ mesons which will potentially be computed in the Global Track Trigger and for which an invariant mass window would be applied already upstream. Similarly, special muon trigger conditions requiring muons and muon stubs may be evaluated in the Global Muon Trigger (where stubs are available) and only the result may be sent on to the Global Trigger. More details on the inputs are given in Section 5.6

3.7.1 Cut-based trigger algorithms

Cut-based trigger algorithms are defined either as the result of a single trigger condition or as the result of a simple logical combination of multiple trigger conditions. The following types of cut-based trigger conditions can be executed in the Global Trigger:

- A single-object trigger condition is fulfilled, when one or more objects in a single collection fulfill a single-object comparison. A single-object comparison is fulfilled when one or more criteria on the object’s variables are satisfied. Different types of cuts exist for different variables. A greater or equal cut may be placed on the object’s $p_T$ or $E_T$. A range-cut, specifying a maximum and minimum value, may be placed on the object’s $\eta$ and $\phi$ (used for example in a muon trigger for cosmic rays in collisions, in order to select only downwards pointing muons), the distance in $z$ between the object’s distance of closest approach and the primary vertex $Dz$ ($PV_z, \text{obj}_z$) (for algorithms using PUPPI objects), the impact parameter $d_{xy}$, the displacement time (for objects with precision timing, if available). Look-up table based cuts, fulfilled when the variable has certain values, may be placed on the object’s quality, sign of charge and isolation variable. Each single-object trigger condition works on a single collection of

\textsuperscript{1}Prescaling can be used to reduce the trigger rate of an algorithm by allowing only 1 in $n$ triggers to be forwarded.

\textsuperscript{2}A bunch mask defines for each BX in the LHC orbit, whether a certain trigger is prevented from firing in this BX.
objects. Separate conditions need to be instantiated in order to deal with different collections corresponding to different flavors of the same objects (e.g. calorimeter / PF / tracker jets). The single-object trigger conditions are used to implement single-object paths in the trigger menu.

- A multi-object trigger condition works on a single collection of objects. Multiple distinct objects in the collection each need to fulfill a single object comparison as defined above. The single object comparisons can be the same or different for each of the objects in the multi-object trigger condition. Multi-object conditions are used to implement di-object, tri-object or four-object paths with symmetric or asymmetric cuts, that do not require inter-object correlation. Multi-object trigger conditions require considerable amount of combinatorial logic, since all permutations of objects need to be considered. For example, in a quad-jet condition based on a collection of 12 jets, about 12,000 combinations of objects need to be considered.

- A di-object correlational trigger condition works on a single or on two distinct collections of objects. It considers distinct objects of a single collection or any combination of one object of the first with one object from the second collection. Each of the two objects considered needs to fulfill a single-object comparison as defined above. Different single-object comparisons may be required for the first and second object. Additionally, cuts on correlations of the two objects may be applied. Range cuts may be applied on $\Delta\eta$, $\Delta\phi$, $\Delta R$, $\Delta z$, and $\Delta t$ (for objects with precision timing). In case of a single collection, additional range cuts may be applied on the invariant mass, the transverse mass or the transverse momentum of a hypothetical mother particle. Furthermore, in case of a single collection, a condition on the correlation of the sign of charge (same sign, opposite sign) may be applied. Di-object correlational trigger conditions are used to implement di-object and cross-triggers with topological constraints.

- A tri-object correlational trigger condition works on a single collection or on two collections of objects. It considers either three distinct objects from the single collection or one object from the first collection and two distinct objects from the second collection. Different single-object comparisons may be required for the first, second and third object. A cut on the compatibility of the $z$ coordinates of the three objects may be applied. Tri-object correlational trigger conditions are needed for tri-lepton triggers, that require all 3 leptons to come from the same vertex (e.g. L1_TripleTkMu, L1_TkMu_DoubleTkEle and L1_DoubleTkMu_TkEle).

Cut-based trigger algorithms are evaluated in four steps:

- **Conversions and calculations step:** In this step, variables of physics objects are converted to common scales if needed and di-object correlations such as $\Delta\eta$, $\Delta\phi$, $\Delta R$, invariant mass, transverse mass and two-body $p_T$ are calculated. The outputs of the first step may be used by one or more comparisons in the following step.

- **Comparison step:** In the second step, the variables of physics objects or results of correlations are compared to thresholds. The results are booleans, vectors of booleans or matrices of booleans. Outputs of the second step serve as input to the condition step. They may be used directly or after combining multiple comparison outputs with boolean operations. Outputs of this step may be used in multiple condition steps.

- **Condition Step:** In this step, the outputs of the comparisons are checked to determine if a condition from the list above is met. The condition step will also support taking into account comparisons for overlap removal in order to avoid the same particle,
3.7. Global trigger algorithms

identified in multiple object collections, being considered multiple times. This can be important between jet and tau collections, for example. The output of the condition step may already constitute the result of a simple algorithm or be combined with other conditions in the final step.

- **Algorithm step:** In the final step, (more complex) algorithms are computed by combining conditions with boolean operations. External triggers (discussed in Sections 2.6 and 5.6.1.1) may also be used in the final step.

With the above conditions, trigger algorithms ranging from simple single or multi-object triggers to cross-triggers and triggers for specific physics processes may be implemented. For example, a trigger on VBF Higgs production can be tailored to trigger on two jets with a high invariant mass (*di-object correlational condition*) in addition to objects created by the decay of the Higgs boson, such as a $\tau$, (single-object condition), a muon (single-object condition) or two muons (multi-object condition).

The MIP Timing Detector may in the future be up-scoped to send objects with precision timing information to the GT.

*Inter-BX conditions*, correlating objects across multiple bunch crossings, may be useful for particular physics channels. Such conditions have been used in Phase-1 for triggering on cosmic-ray muons during LHC runs (when timed in for collisions, cosmic-ray muons show up as two muons in consecutive bunch crossings). In Phase-2, such conditions may be used to trigger on long-lived particles. The GT system will support such algorithms for up to $\pm 3$ bunch crossings: *di-object correlational conditions* may be evaluated with one of the object collections from an earlier or future bunch crossing. Similarly, in the algorithm step, conditions from earlier or future bunch crossings may be combined with conditions from the current bunch crossing.

3.7.1.1 Cut-based trigger algorithm implementation

Cut-based multi-object trigger conditions need considerable resources in a Field Programmable Gate Array (FPGA) due to combinatorics. Correlation variables, such as $\Delta R$ or the invariant mass, need to be calculated for any possible combination of objects. In the Run-2 Global Trigger, which is implemented in 40 MHz logic, individual algorithms with correlation conditions use up to 33% of available Digital Signal Processor (DSP) resources in a Xilinx Virtex-7 XC7VX690T FPGA. The full menu implemented at the end of Run-2, comprising 345 algorithms, needs 6 Xilinx Virtex-7 FPGAs. Since in Phase-2 more types of object collections and different flavors of object collections will be available, the number of algorithms is expected to rise. Higher precision will increase the logic consumption per algorithm. The Virtex Ultrascale+ VU9P FPGA targeted for Phase-2 provides more resources than the Xilinx Virtex-7 690, but only a factor of 2–4 more (depending on the type of resource: LUTs, flip-flops or DSPs). It is therefore important to reduce the resource consumption of cut-based algorithms. One way to achieve this is to run the logic at a higher clock frequency. Modern FPGAs can run their logic including DSPs at several 100 MHz. Running the GT algorithms at 240 or 480 MHz reduces logic consumption and at the same time facilitates routing as combinatorial paths between flip-flops will be shorter. Collections of objects will be presented to the algorithm steps as a stream of objects with a new object arriving on every clock cycle. At 480 MHz, up to 12 objects can be processed within an LHC bunch crossing, while at 240 MHz two instances of the algorithm have to be used in a time multiplexed way. In this way, comparisons can be applied to a collection of $N$ objects using a single comparator over $N$ clock cycles. In a similar way, calculations that have to be performed for $N \times M$ objects may be performed with $N$ calculation units over $M$ clock cycles. The DSP usage for calculations of correlation variables can thus be reduced by a factor
of 6 (at 240 MHz) or 12 (at 480 MHz).

Many of the above algorithm steps can be applied on the streams of objects as they flow, with each algorithm step needing one or a few clock cycles at 240 or 480 MHz. However, resource-optimized calculations of correlations between different collections may require one collection to be deserialized so that the calculations can happen on every clock cycle, while the other collection is streamed. Such a deserialization step would need 2 BX at 240 MHz (1 BX at 480 MHz), part of which may be in the shadow of other operations.

Taking into account the above considerations, we estimate that cut-based algorithms will fill a similar number of FPGAs as used at the end of LHC Run-2.

### 3.7.1.2 Latency

While processing at 240 / 480 MHz will greatly reduce resource usage, a slight penalty on latency with respect to a highly optimized 40 MHz implementation is to be expected. The latency estimate for the above implementation of algorithms is 6 BX if run at 240 MHz (or 4 BX at 480 MHz). Assuming inter-BX conditions that are able to match signals from ±3 BX, three additional BX are needed. This brings the total latency for cut-based trigger algorithms to 9 BX (or 7 BX at 480 MHz). Whether the logic will be clocked at 480 or 240 MHz will be decided later, depending on the effort required to achieve timing closure, especially when the FPGA design gets full.

### 3.7.2 Machine learning based trigger algorithms

Machine learning methods, such as Boosted Decision Trees (BDTs) and Neural Networks (NNs) have a wide range of applications in event processing at the LHC, from lower level energy cluster calibration and regression to high level physics object classification, and physics analyses. In particular, they have shown to provide significant improvements over traditional cut-and-count methods in analyses that target a small signal-to-background ratio, such as in the recent observation of Higgs boson decays to bottom quark-antiquark pairs [23]. Such successes motivate the development of machine learning algorithms to be implemented in the Global Trigger system. A neural-network based classifier could increase the trigger efficiency, while reducing the trigger rate.

The attempts to implement neural networks on FPGAs started in the 1980s for hardware neurocomputers [60], but were only gaining attention recently as a practical solution for NN-inference due to the improvements in capacity and performance of FPGAs. Implementing an analysis-like signal background discriminator is now feasible in the Global Trigger system. Two examples of NN based triggers targeting VBF $H \rightarrow b\bar{b}$ signal separately, as described in Section 4, could be used in the Global Trigger to select against the minimum bias background. In Section 3.7.2.1, we introduce a model-independent trigger algorithm, based on autoencoders, as a novel method for detecting new physics with the Global Trigger.

#### 3.7.2.1 Autoencoder for New Physics detection

With CMS storing $\mathcal{O}(10^4)$ out of the $\mathcal{O}(10^7)$ collision events produced by the LHC per second, it is possible that new physics events, if any, are systematically rejected by the CMS trigger system. This could happen in many exotic scenarios in which the experimental signature of the new physics topology differs from those usually considered (e.g., low mass resonance). In order to extend the CMS sensitivity to otherwise uncovered new physics models, one would have to depart from the usual fully supervised search scheme, in which a signal hypothesis is made
upfront and searched for, first in the trigger systems and then in an offline data analysis. A possible strategy would be to formalize the problem as an anomaly detection task in real time. One would study the main features of standard events (those coming from known physics) and define a measure of typicality. Using the probability density function of such a quantity, one could then define a one-sided threshold test to isolate the most atypical events, e.g., the one-in-a-million event that would fall in the far tail of such a distribution.

Unsupervised data mining techniques based on deep learning offer practical ways to put such a strategy in practice. In particular, one could take a sample of data (e.g., prescaled L1 pass-through events) as input to train an autoencoder (AE) [61]. An AE is a sequence of an encoder and a decoder network. The former is trained to project a given set of inputs into some abstract latent space. The latter takes a point from this latent space to generate an output of the same kind as the input. The loss function is usually chosen as a function of the differences between the input and the output. Once trained, the AE with minimized loss function should return an output as close as possible to the input. The selected loss function can then be used as a metric to measure the distance between the input and the output. Accurate training would result in small loss values for typical inputs, while inputs other than those provided at training time might result in larger loss values. The loss values of the AE can then be used to quantify the typicality of a given event.

As a proof of principle, we studied an autoencoder that would take as input a subset of particles fed into the Global Trigger system. The input features consist of \((p_T, \phi)\) of PUPPI \(E_T^{\text{miss}}\), PUPPI \(H_T\), and \((p_T, \eta, \phi)\) of 5 PUPPI jets, for a total of 18 input variables. The AE is constructed as a multilayer perceptron (MLP) consisting of 8 layers. The encoder of the AE starts with an input layer with 18 nodes, followed by one layer with 12 nodes and two layers with 6 nodes each. The decoder of the AE inverts the structure of the encoder, consisting of two layers with 6 nodes, one layer with 12 nodes, and an output layer with 18 nodes. Each layer uses ReLU [62] as activation function. The objective function to train the AE and to measure the distance between inputs and outputs is the \(L_1\) loss function [63], which computes the sum of absolute differences between each feature. Both ReLU and \(L_1\) loss functions are chosen for their simplicity in FPGA implementation.

Input values of the AE require preprocessing during training and inference. All \(p_T\) values are divided by 40 to reduce the required range, preventing the AE from becoming over-sensitive to large differences in \(p_T\). All \(\eta\) and \(\phi\) values are made positive values by adding an offset of 5, which avoids the truncation by the ReLU function of negative values. In the events with less than five PUPPI jets, a traditional zero padding method is applied, setting the \(p_T, \eta, \phi\) of the extra jets to zero.

The AE was trained with simulated events at 200 average pileup interactions, mimicking the minimum bias data to be collected at the HL-LHC. We used simulated VBF \(H \rightarrow \text{invisible}, VBF H \rightarrow b\bar{b}\) and \(HH \rightarrow b\bar{b}b\bar{b}\) signals to test the AE performance. Using the loss value between input and reconstructed output as a discriminator, the performance of the AE is shown in Fig. 3.126. We observed sensitivity of the AE to these signals, even though they are unknown during the AE training. This demonstrates the anomaly detection power of the AE, which could provide guidance to trigger unexplored new physics events at the HL-LHC.

During data-taking, trigger rates can vary due to detector noise or channel masking. For typical noise in the detector, classical trigger algorithms, like single-jet or MET have been used as indicators. The L1 menu has been designed to tolerate certain rate changes online. We found that the AE is not sensitive to small variances of individual input features. In the scenario of small variances in a large number of input features, the AE will detect the differences (generating a
Chapter 3. Trigger algorithms

Figure 3.126: Acceptance and rate of VBF $H \rightarrow$ invisible, VBF $H \rightarrow b\bar{b}$ and HH $b\bar{b}b\bar{b}$ signals from the autoencoder. The autoencoder was trained with minimum bias samples only.

larger trigger rate), whereas classical triggers remain insensitive. This feature of AE could be utilized by online monitoring to spot changes in data-taking conditions. If needed, the AE trigger rate may be prevented from rising above a fixed budget by additional logic in the firmware that would dynamically adapt the autoencoder threshold. Changes in data-taking conditions could then be spotted by monitoring the AE threshold.

3.7.2.2 NN Implementation

We use the hls4ml package [44], tag v1.0.5, for implementing the NNs into firmware. It translates the NN model into FPGA firmware using the Vivado High-Level Synthesis from Xilinx. We use the Keras [64] interface of the package.

While floating-point precision is used during training, values are represented as fixed-point numbers in the FPGA. Numerous studies have shown that NNs can be implemented with fixed-point precision without significant performance degradation [44]. In testing, we use fixed-point precision with 5 bits for the integer part, and 10 bits for the fractional part.

The results presented below are synthesized and implemented for a Xilinx Virtex Ultrascale+ FPGA with part number xcvu9p-flgb2104-2-i (VU9P in the following), as a “placed and routed” implementation. We use a “bare” firmware design that uses minimal resources beyond those required by the neural network. This bare implementation consists of a simple VHDL wrapper that connects the NN firmware block directly to the FPGA’s general purpose input/output pins to prevent Vivado from removing the neural network logic during optimization.

It was assumed that Global Trigger algorithms for CMS Phase-2 are foreseen to be implemented at a clock frequency of 240 MHz. When performing the implementation, the clock period targeted by HLS was not initially achieved. This was resolved by requesting a shorter clock period of $\sim 2$ ns during the synthesis step, which allowed the timing constraints of 4.167 ns to be met in the final FPGA implementation.

**Latency** The latency of the NN trigger has two components: the latency for calculating the input features; and the NN inference. For the VBF NNs, some of the 24 input features are inputs to the Global Trigger system, requiring no preprocessing. Some features, such as invariant mass, $\Delta R$, and $\Delta \eta$ of the jet pairs built from three leading jets, require preprocessing
3.7. Global trigger algorithms

in the Global Trigger. All of the preprocessing calculations are typical correlations among jets, which are well developed for the CMS Phase-1 Global Trigger. While the Phase-1 Global Trigger has a latency budget of 3 BX for the evaluation of algorithms (except inter-BX algorithms), we conservatively assume a 4 BX latency for possible preprocessing for NN inputs at 240 MHz. The preprocessing of the input values for the AE is expected to be easily achieved within this 4 BX latency.

To minimize the latency of the NN inference, we fully unroll the network in the FPGA, which assigns one DSP per multiplication-and-addition (MAC) operation. The VBF NN performs inference within 21 clock cycles at 240 MHz (≈ 4 BX) with an initiation interval of 1 clock. As a result, a total latency of 9 BX can be achieved, which equals that of cut-based trigger algorithms with inter-BX conditions.

Similarly, the autoencoder can perform inference and the additional reconstruction loss calculation, within 23 clock cycles at 240 MHz (≈ 4 BX) with an initiation interval of 1 clock.

**Resource Usage**

The NNs for VBF physics described above consume 2775 DSPs each, corresponding to 40% of the DSP resources of the VU9P chip. The hls4ml package allows users to fine-tune the number of DSPs used in the NN inference by reusing the same DSP for additional MAC operations, at the cost of additional latency and with a larger initiation interval. For a NN running at 240 MHz, any initiation interval \( \leq 6 \) would be able to perform inference with events arriving at 40 MHz. Such fine-tuning can be done in the final implementation. In addition, 22k Flip Flops and 42k LUTs are utilized for routing and mimicking activation functions, which corresponds to only 1% and 4% of the respective resources of the VU9P.

The autoencoder will consume 678 DSPs, corresponding to 9% of the DSPs in the VU9P. In addition, 5k Flip Flops and 13k LUTs are used for routing, mimicking activation functions, and applying the reconstruction loss function.

### 3.7.3 Overall Global Trigger resource usage

The overall resource usage of the GT will depend on the number of cut-based algorithms and NN-based algorithms in the trigger menu. As the global trigger menu is expected to evolve continuously, it is not possible to give a firm resource usage estimate. It is therefore important that the architecture of the Global Trigger system (discussed in Section 5.6) allows for easily scaling the system size to provide more resources.

As discussed in Section 3.7.1.1 we estimate cut-based algorithms to fill about 6 VU9P FPGAs. Machine learning based algorithms, discussed above, will use a significant fraction of resources of a VU9P FPGA per algorithm. While this resource usage may be reduced at the cost of a slight increase in latency, we expect research in this area to continue and more algorithms to be proposed.

In order to allow for future developments, especially in the area of NN-based algorithms, an architecture allowing for up to 12 VU9P FPGAs or similar is pursued.
3.8 Trigger algorithms for heavy ion collisions

3.8.1 Introduction

Future opportunities for physics with Heavy Ion (HI) beams at the LHC after LS3 have been documented in Ref. [65]. They include collecting lead-lead (PbPb) at $p_{\text{NN}} = 5.5$ TeV and proton-lead (pPb) at $p_{\text{NN}} = 8.8$ TeV collision data, together with reference pp collisions (i.e., pp collisions at the same collision energies as the PbPb and pPb ones).

The desired/expected luminosities for the ion runs in LHC Run-4 are only about twice higher than the $\sim 1.7 \text{nb}^{-1}$ and $\sim 170 \text{nb}^{-1}$ recorded during the 2018 PbPb and 2016 pPb runs, respectively. Given the already very good performance of the trigger system (both at the L1 and HLT) during the LHC Run-2, the main goal for Run-4 (and Run-3) is, at a minimum, to maintain the Run-2 performance, and targeting to improve on it. In the following, the physics program envisioned with Run-4 data is summarized, together with the L1 trigger requirements needed for reaching it.

3.8.2 Physics program

In order to achieve a deeper understanding of the properties of strongly-interacting matter, baseline collisions systems (like pp and pA) also have to be studied and understood in parallel to AA collisions. Large samples of $W^\pm$ bosons, Z bosons, top quarks, and di-jets measured in pPb collisions, or of coherently produced quarkonia measured in ultraperipheral PbPb collisions (UPC), are needed to study the modifications (e.g., shadowing, anti-shadowing, the EMC effect [65]) of the parton densities inside a nucleus compared to those of a free nucleon. The uncertainties in the theoretical nuclear parton distribution functions (nPDFs) of gluons and quarks can be drastically improved by the availability of high-precision measurements, which in turn will propagate into reduced uncertainties in the perturbative QCD calculations for PbPb observables. Such measurements can be optimized by having a wide pseudorapidity coverage for leptons and jets, together with a high trigger efficiency and readout rate, which the Phase-2 trigger upgrade will provide.

The physics potential of the open and hidden heavy flavor measurements is driven by the pre-
cision reached across the widest possible $p_T$ range (ideally starting from $p_T = 0$ GeV), and the finest possible centrality (the variable reflecting the impact parameter of the two colliding Pb ions) intervals. A large sample of peripheral PbPb collisions, with multiplicities similar to those typical of PbPb and pp collisions, is essential to identify the onset of “hot/dense QCD effect”, and to study the transition from a standard “vacuum” QCD regime to another one driven by collective parton properties. These measurements depend on the capability of efficiently accumulating high-purity minimum bias (for low-$p_T$ measurements), and track/muon/jet-triggered samples (for high-$p_T$ measurements) samples. The Phase-2 trigger upgrade will help meeting these requirements.

The studies of jet-quenching related phenomena have entered a new era with the Run-2 results. Tomographic studies of the parton shower using different methods (e.g., jet grooming) have been initiated, the radial distributions of particles inside jets have been examined, results from first measurements of fully reconstructed heavy-flavor jets (charm in PbPb, and beauty in both PbPb and PbPb) have been carried to investigate the flavor-dependence of the energy loss. In addition, reference measurements using jets tagged by back-to-back photon and Z bosons, which give insights to the absolute amount of energy loss, have been performed. However, the amount of data available in Run-2 (and Run-1) has allowed so far to extract with limited precision the properties of the high-density QCD matter produced in PbPb collisions. To improve the situation, the preparations for the Run-4 (and Run-3) have as main goal to accumulate the largest data sample possible, with high-efficiency and purity for the photon, jet, and meson triggered samples. The Phase-2 trigger upgrade helps meeting these requirements.

Figure 3.127 shows the expected performance expected with HL-LHC data, for a few of the measurements mentioned above.

### 3.8.3 Trigger considerations

To set the stage for the L1 trigger preparations, we mention that while the HL-LHC pp events are expected to reach 200 average pileup interactions, in PbPb collisions, single events producing a multiplicity equivalent to more than 300 simultaneous pp collisions are the norm. On the other hand, the maximum collision rate that the LHC can deliver with PbPb beams is considered to be $\sim 50$ kHz, even for HL-LHC, which is to be compared to the maximum output bandwidth of 750 kHz expected to be reached for the pp physics program. So while the HI environment can be much more “dense” than what the LHC will deliver for the pp program during the high-luminosity era, the output rate requirements for the PbPb program are smaller. In this sense, storing all events on tape is one of the possible goals. However, the events being “heavier” in general, areas of potential bottlenecks have to be identified and problems addressed at L1, or, if needed, at the HLT. Given the similarities expected between the 2018 PbPb run and any subsequent PbPb runs, the goal is to build on the performance from the 2018 run, in order to reach the physics goals (some of which were mentioned above).

Because of the small collision rate, triggering on low-$p_T$ muons (for quarkonia and beauty physics) at L1 is technically possible, and has been done in both LHC Run-1 and Run-2. A larger $\eta$ coverage for muons (up to 2.8) will be beneficial for the electroweak boson program (e.g., nPDFs) and quarkonia studies (where an overlap with that of ALICE, whose muon acceptance spans $2.5 < \eta < 4$, will be achieved). An open area for improvement is to increase the efficiency of the L1 muon triggers and/or reduce their fake-rates. To compensate for the loss of efficiency in some corners of the phase space (very peripheral collisions, or very forward regions of the muon detectors), cross-triggers with information from other detectors (e.g., the tracker, the forward calorimeters, etc.) will be investigated to be deployed to a larger extent than what
was used for Run-2. No special tuning of the muon detectors is needed during the PbPb runs compared to the pp runs. Similar considerations can be made for the photon L1 triggers. The small collision rate allows triggering on very low energy deposits (the minimum was 3 GeV for Run-2), and, in addition, using looser ECAL settings compared to those used during the pp runs (e.g., the readout thresholds). The PbPb settings are chosen so as to improve the efficiency and fake rate of the L1 trigger.

Substantial changes to L1 trigger algorithms, and to the way the detector information is retrieved, have to be made during the PbPb runs, because of the higher multiplicities in these collisions. To start with, dedicated front end driver (FED) firmware was deployed in Run-1 and Run-2 for the tracker, for the readout of the zero suppression hybrid data, as well as pixel. While the 200 pileup expected for HL-LHC in pp collisions is getting closer to the PbPb environment, tests and careful checks have to be made to ensure that for Run-4 no such FED modifications are needed during the PbPb runs. The high multiplicity also has a direct impact on the L1 jet triggers; their associated collection needs to be “cleaned”. Therefore, jet-subtraction algorithms (i.e., algorithms that separate real jets from underlying background fluctuations) need to continue to be deployed at the Level-1 trigger, to ensure a low-enough rate to make the jet L1 triggers useful. In addition, similar to what has been done in Run-2 at the HLT level, the parameters of the tracking and particle-flow algorithms will have to be adjusted for the PbPb runs, in order to cope, time and memory wise, with the much higher combinatorial backgrounds.

While most of the L1 triggers can be derived rather easily from the pp menu, a few triggers are specific to the ion runs. These are the triggers that perform a selection based on the characteristics of the event: centrality/multiplicity (using HF and Zero Degree Calorimeter (ZDC) information), $\phi$ shape of the event (signals from the Reaction Plane Detector, which started collecting data during the 2018 PbPb run), or energy asymmetry (using calorimeter information). Such triggers have been available throughout Run-2; the goal for Run-4 is to improve their efficiency and purity.

One of the most crucial triggers in the HI L1 menus is the minimum bias trigger, which, up to the end of Run-2 was based mainly on information from the HF calorimeters alone, or in combination with information from the tracker or ZDC (with loss of efficiency and decrease of purity, for different combinations). As the performance of the HF calorimeters deteriorates with time (already seen between 2015 and 2018 PbPb runs), the expectation is that the region of inefficiency will increase to more central collisions (from just the centrality range 90–100% during the 2018 run) and will become noisier. Alternative algorithms (or settings) for HF, or mixed triggers using additional information from multiple detectors, have to be developed. In this spirit, several combinations were tried during the 2018 run, using CASTOR (which will not be available starting from Run-3) and the ZDC detector, which helped increase the purity of the HF-based L1 trigger.

While there is a good chance to pass-through all hadronic PbPb collisions, this is not true for ultra peripheral collision (UPC) events, which have a much larger cross section. Such events, which have a very small event activity in the detectors, require careful triggering to suppress backgrounds. The Phase-2 L1, with its enhanced bandwidth, provides new opportunities in this area. For example, lowering the $p_T$ threshold below 2 GeV (the design threshold for the nominal 40 MHz pp bunch crossing rate) in a new L1 trigger that uses tracks, during ion runs, would enable measuring new channels such as UPC $\rho, \phi$, or open heavy-flavor mesons, that are not possible before Run-4.

For pPb collisions, the total interaction rate will be of the order of tens of MHz. The rate increase
from Run-2 is likely to be larger than for PbPb due to the higher proton beam intensity. Even with the designed $750 \text{kHz}$ bandwidth, a reduction factor of $\sim 10$ is needed at L1 in Run-4. While most high-$p_T$ object triggers are standard pp ones, lower threshold calorimeter triggers are important to the pPb program, but they are also the ones more sensitive to noise. The high-multiplicity pPb triggers, meant to select events with multiplicities similar to those in some PbPb collisions, based on L1 $E_T$ sum or tower count, can be significantly improved with the Phase-2 L1. Similarly to the PbPb case, the L1 track finder is a new capability that is to be explored, and will benefit the pPb program for certain.

In summary, the strategy for Run-4 is to build on the strategy (and performance) of Run-2, exploiting the CMS detector capabilities following the Phase-2 upgrades, and the particularities of the ion runs (low collision rate), trying to increase the purity and overall quality of the selected events.
Chapter 4

Level-1 trigger menu

The Phase-2 Level-1 trigger uses information from the muon, calorimeter and tracking detectors to select collision events that are potentially interesting for physics analysis. The selection is performed using a set of algorithms (known as “seeds”), collectively called the “menu”, that check events against predetermined criteria. Any event that satisfies at least one seed in the menu proceeds further in the trigger chain. This initiates the readout of the complete detector information by the data acquisition system, and the data are sent to the HLT. During the LHC Run-2, a broad range of L1 seeds were deployed online, reflecting the variety of the CMS physics program. The Level-1 menu evolves with shifting CMS physics priorities and adapts to changes in beam conditions or detector performance.

The HL-LHC physics program, briefly summarized in Section 1.1, includes measurements of the properties of the Higgs boson and precision measurements of Standard Model properties in the electroweak, top-quark, and QCD sectors with special attention given to the physics of bottom quarks, where trigger objects often have low $p_T$. It also includes a large number of searches for supersymmetric or exotic particles and candidates for dark matter. Heavy-ion collisions are studied to expand our knowledge of quark-gluon plasma dynamics. The menu described in this Chapter demonstrates the capability of the Phase-2 L1 trigger to fully exploit the CMS discovery potential for such a physics program. To reach this goal, the trigger algorithms designed to collect events for investigation in all of these areas (see Sections 1.4.1 and 1.4.2) are included in the Phase-2 Level-1 trigger menu.

In this Chapter, the strategy used to design the Phase-2 L1 trigger menu is presented. As mentioned in Section 1.4, a typical menu is composed of a set of criteria applied to single or combinations of trigger objects. Additional triggers tailored for the selection of events with specific physics final states (VBF topologies, $b$-hadron decays) also enter the menu. The design of the envisaged selection strategy for CMS Phase-2 will exploit a richer range of complementary high-resolution trigger objects than those available for Phase-1 (see Chapter 3).

The design of this menu is inspired by the CMS Run-2 menu for an instantaneous luminosity of $2 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$, as discussed in Section 4.2. The trigger menu presented in this Section is a simplified menu that does not include specific seeds used for measuring trigger efficiencies, diagnostics, and calibrations. During the LHC Run-2, these seeds accounted for 30% of the total rate available at Level-1. The Phase-2 L1 menu features 41 seeds which in Run-2 comprised 70% of the total Level-1 bandwidth. The study reported here shows that offline thresholds similar to those used during Run-2 can be maintained even with the harsh environment of 200 average pileup events. The total L1 menu rate computed (accounting also for the additional 30% rate of a fully operational menu) remains under the upper limit of 500 kHz, ensuring a 50% safety margin. This safety margin is used to account for uncertainties in extrapolating rates to higher energy and instantaneous luminosity and in detector performance. Specific triggers for heavy-
ion collisions, which would be part of a dedicated menu, are not considered at the time of this writing.

The Phase-2 selection strategy is expanded beyond maintaining the Run-2 performance and it includes seeds used to collect events originating from physics processes previously unreachable and only available today as a result of the advanced triggering techniques proposed. The novel seeds are fashioned from the sophisticated trigger algorithms for which the system’s architecture was modeled. With the inclusion of these seeds, the Phase-2 L1 trigger menu allows for a significant improvement of the physics program as introduced in Section 1.4.2. The extensions as well as the evaluation of the gain achieved for targeted physics analyses are provided in Section 4.3.

4.1 Inputs and design strategy of the Phase-2 Level-1 trigger menu

The Level-1 trigger menu is implemented within the global trigger (GT) visible on the functional diagram provided in Section 1.3. The input objects received by the GT are listed in Section 3.7. In contrast with the current L1 trigger, the Phase-2 system reconstructs objects with multiple trigger algorithms. This includes the following categories of objects: standalone (including tracker-only objects as defined in Section 3.6), track-matched and objects reconstructed from particle-flow and PUPPI particle candidates. Different versions of the same trigger object can be used to target specific physics needs, while providing robustness to the selection strategy. The range of available trigger algorithms allows the menu to provide similar thresholds as those implemented during Phase-1.

Following the studies reported in the interim document [3], the total Level-1 menu targets a rate of 500 kHz which provides a 50% safety margin with respect to the maximum output rate of 750 kHz. A simplified menu was designed by comparing the performance of the various physics objects and by choosing the adequate algorithms for different seeds to cover the physics program detailed earlier, while keeping the overall bandwidth within the constraint imposed by the system. As for the current L1 menu, single- and multiple-object trigger seeds are implemented. The following steps are required to obtain this menu:

- **Evaluating the performance of the trigger algorithms:** The trigger objects reconstructed with the various algorithms are compared on the basis of signal efficiency (as presented in Chapter 3) and rate for a given target “offline” threshold. The best performing algorithms are used to build the menu. The selection thresholds applied to online objects correspond to different thresholds for objects reconstructed offline. The target thresholds chosen for the Phase-2 L1 menu development are equivalent to the Phase-1 L1 menu offline values, which are used as a reference.

- **Offline and online threshold scaling:** In order to evaluate the trigger rate of an algorithm for a given arbitrary offline threshold, an “offline-to-online” scaling function is derived. A set of online thresholds are chosen and the object’s trigger efficiency curve, also referred as “turn-on”, is evaluated for each threshold. The turn-on curves are derived with respect to generator level information as an approximation of offline quantity. The point where the object’s efficiency reaches a chosen percentage of the turn-on plateau efficiency is found and used to extract the corresponding offline threshold from the x-axis. For objects with slower turn-on behavior, such as $E_{\text{miss}}^T$ or $H_T$, a lower efficiency point is chosen. This procedure is repeated for each threshold. A linear fit is then applied to that set such that a scaling function is derived, which allows one to interpolate to any potential threshold as a function of offline threshold.
These steps are illustrated in Fig. 4.1 for the electron trigger algorithms that will be summarized in Section 4.1.2.

![Figure 4.1: Left: Turn-on curves for different electron trigger algorithms at three typical thresholds, in the barrel region. Right: Linear fits and offline-to-online scaling functions extracted using the 95% efficiency point of the turn-on curves in the right plot.](image_url)

- **Estimation of the trigger rate**: The scaling function derived in the previous step for each trigger algorithm is used to compute the associated single-object trigger rate as a function of the offline threshold. These rate distributions are therefore driven by the resolution and purity of each of the reconstruction algorithms. As already introduced in Chapter 3, the trigger rates are estimated with a sample of simulated events including an average of 200 pileup collisions.

In the following, the details of this strategy are described for each physics object considered in the menu.

### 4.1.1 Muon trigger algorithm inputs

The three main reconstruction algorithms for prompt muons described in Section 3.3 are considered here: standalone, TkMu and TkMuStub. The reconstruction efficiency for the standalone and TkMu algorithms is broadly similar (95% at plateau), while higher efficiencies are reached by the TkMuStub for all values of $p_T$, as shown in Fig. 3.45 and Fig. 3.46 for the barrel region, Fig. 3.48 for the overlap region and Fig. 3.50 for the endcap region. This is especially the case at very low $p_T$ values ($p_T < 5$ GeV) as discussed in Section 3.3.1 and shown in Figure 4.7.

For very soft muons, with $p_T$ in the range of 2 to 5 GeV, the TkMuStub reconstruction efficiency in the endcap is substantially higher than for the TkMu algorithm. However, in the barrel region, the current reconstruction algorithms have not yet been optimized for muons in this $p_T$ range. Studies to improve these efficiencies are in progress.

The comparison between the performance of muon reconstruction algorithms, in terms of the combined effect of background reduction and momentum resolution, is shown in the left side of Fig. 4.2, where the trigger rate as a function of the offline threshold is shown for each algorithm. The trigger rate is calculated for the range of $|\eta| < 2.4$. The offline-to-online scaling is derived using prompt muons from a simulated sample of Drell–Yan events, extracting turn-on curves...
for several $p_T$ requirements separately in the barrel, overlap and endcap regions of the muon
detectors. As already discussed in Section 3.3, the use of track-matched muons significantly
reduces the rate, by about a factor 5 at the typical Phase-1 thresholds, making it possible to
maintain them in the simplified menu (see Section 4.2). Nevertheless, standalone muons will
be used with higher $p_T$ thresholds to provide robustness for the overall trigger system, given
that they do not rely on L1 tracks. Despite better reconstruction efficiencies provided by the
TkMuStub algorithm, the TkMu algorithm is used instead as default in the simplified menu
since the rate of the former is too high at very low $p_T$ values to allow for thresholds similar
to those used in Phase-1. Low $p_T$ muons are extensively used in the menu to target BSM final states with soft leptons and $E_T^{\text{miss}}/H_T$.

Studies to improve the menu performance for low $p_T$ muons are in progress.

The use of track-matched objects in the menu allows for further rate reductions when multi-
object seeds are designed. The spatial resolution of the L1 tracks allows for an efficient selection
requirement to be placed on the $z$-distance at the point of closest approach to the beamline, to
ensure the L1 track candidates come from the same interaction vertex. The requirement of
$\Delta z < 1$ cm is used, corresponding to greater than 2 $\sigma$ confidence level, as shown in Fig 2.2.
This requirement drastically reduces the pileup contribution, and therefore the trigger rate,
while maintaining high signal efficiency for muons originating from the same vertex. The rate
reduction is obtained without requiring the leptons to come from the reconstructed primary
vertex thus avoiding loss in efficiency due to the vertex finding performance (see Section 3.4.3).

### 4.1.2 Electron and photon trigger algorithm inputs

As discussed in Section 3.2, the GT receives calorimeter-only electrons and photons, track-
matched electrons, as well as track-matched isolated electrons and isolated photons when the
charged isolation provided by the L1 tracks around the electron or photon candidates is used.
Figures 3.11, 3.12 and 3.13 show that track-matched electrons provide a significant rate reduc-
tion (factor 5 at the typical Phase-1 thresholds) at the price of some efficiency loss with respect
to the calorimeter-only objects (93% with respect to 99% at plateau) mostly due to the track-
reconstruction efficiency. The trigger rate for track-matched electron objects is further reduced
by a factor of 2 by imposing a charged isolation requirement to define isolated track-matched
electrons, without loss of efficiency on prompt isolated electrons.

For photons, calorimeter-only objects are used to seed the reconstruction of isolated photons
using a charged isolation working point that ensures an isolation efficiency for prompt iso-
lated photons above 95% at plateau, with respect to the calorimeter-only object reconstruction
efficiency (see Fig. 3.18).

The comparison between the performance of these objects, in terms of the combined effect of
background reduction and momentum resolution, is shown in the right side of Fig. 4.2, where
the trigger rate as a function of the offline threshold is shown for each algorithm. The offline-
to-online scaling is derived for prompt electrons using a simulated sample of Drell–Yan events,
and for prompt photons using a simulated sample of Higgs boson events decaying into a pair
of photons. Turn-on curves for several $p_T$ requirements are evaluated separately for the barrel
and endcap regions of the calorimeter. The trigger rate is calculated for a coverage of $|\eta| < 2.4$.

Track-matched (isolated) electrons and isolated photons are the main objects used in the simpli-
ified menu, allowing the total menu rate with Phase-1 thresholds to stay within the bandwidth
requirements; calorimeter-only electrons are used in dedicated seeds at higher thresholds to
recover the lower reconstruction efficiency of the track-matched ones. Figure 4.2 also shows
the rate for the calorimeter-only electrons with the pseudorapidity range extended from 2.4 to
3, the coverage of the upgraded endcap calorimeter. The use of calorimeter-only electrons up to $|\eta| < 3$ will be discussed in Section 4.3.1.

Moreover, as discussed above, the use of track-matched electrons further reduces the trigger rate for multi-object seeds using the $z$ coordinate of the matched L1 track, as in the muon case.

**Figure 4.2:** Left: Trigger rate as a function of the offline $p_T$ thresholds for standalone muon, TkMu and TkMuStub algorithms for the pseudorapidity range $|\eta| < 2.4$. Right: Trigger rate as a function of the offline $p_T$ thresholds for calorimeter-only electrons, track-matched electrons, track-matched isolated electrons and isolated photons for the pseudorapidity range $|\eta| < 2.4$; the rate for calorimeter-only electrons extended to $|\eta| < 3$ is also shown.

### 4.1.3 Jets and jet-sum trigger algorithm inputs

As discussed in Section 3.6 there are two standalone jet algorithms, track jets and calorimeter-based jets, which use only L1 tracks and calorimeter clusters, respectively. In addition, as discussed in Section 3.5, PUPPI jets are reconstructed using the particle-flow algorithm.

While track jets are only reconstructed in the $|\eta| < 2.4$ range, where L1 tracks are available, calorimeter-based jets and PUPPI jets are reconstructed up to $|\eta| < 5$, making use of all of the calorimeters. This study considers the $7 \times 7$ histogrammed PUPPI jets as described in Section 3.6.1.3. For all three algorithms, the reconstruction efficiency reaches 100% at plateau, as shown in Figs. 3.93, 3.95, and 3.99.

A summary of the jet trigger rates can be found in Fig. 4.3, where the trigger rate as a function of the offline threshold is shown for each of the algorithms. The offline-to-online scaling is derived using jets from a simulated sample of $t\bar{t}$ events and is used to extract turn-on curves for several $p_T$ requirements separately in the barrel, endcap, and forward (outside the tracker acceptance) regions.

For $p_T$ thresholds lower than 150 GeV, sustainable rates are provided only by the PUPPI jet algorithm; the calorimeter-based jet algorithm can be used for higher thresholds when its rate becomes comparable or even lower than the PUPPI jet algorithm rate. Moreover, the two standalone jet algorithms provide robustness for the overall trigger system, given they rely only on the calorimeter sub-detectors or only on the tracker detector.

Three different algorithms have been considered for the reconstruction of $H_T$ at trigger level:
calorimeter-based $H_T$, track $H_T$, and PUPPI $H_T$, obtained by summing the transverse energy of calorimeter-based jets, track jets, and PUPPI jets, respectively. The performance of all three algorithms is shown in Fig. 4.4, defining the generator level $H_T$ as the scalar sum of all the generator level jets with $p_T > 30$ GeV and $|\eta| < 2.4$. In this case, for all thresholds, PUPPI $H_T$ provides the best results.

![Graph of Trigger Rate vs. offline $p_T$ threshold for track jets, calorimeter-based jets, and PUPPI jets for the pseudorapidity range $|\eta| < 2.4$.](image)

Figure 4.3: Left: Trigger rate as a function of the offline $p_T$ thresholds for track jets, calorimeter-based jets, and PUPPI jets for the pseudorapidity range $|\eta| < 2.4$. Right: For calorimeter-based jets and PUPPI jets the rate for the pseudorapidity range $|\eta| < 5$ is shown.

### 4.1.4 Missing transverse energy trigger algorithm inputs

As discussed in Sections 3.6 and 3.5 the two algorithms for the computation of $E_T^{\text{miss}}$ both require use of L1 tracks to be able to cope with high pileup: one makes use of L1 tracks only, track $E_T^{\text{miss}}$, and one exploits the particle flow reconstruction and PUPPI selection, PUPPI $E_T^{\text{miss}}$. The offline-to-online scaling is derived using a simulated sample of $t\bar{t}$ events, with the generator level $E_T^{\text{miss}}$ calculated as the vector sum of the momenta of all the generator level particles (excluding the undetectable neutrinos). For the typical Phase-1 thresholds targeted in the simplified menu (see Section 4.2), PUPPI $E_T^{\text{miss}}$ provides a trigger rate which is a factor 20 lower than the corresponding one from track $E_T^{\text{miss}}$ as shown in Fig 4.4 right, and it is therefore used in the menu design.

### 4.1.5 Tau trigger algorithm inputs

As discussed in Section 3.6, prompt isolated taus that decay hadronically can be reconstructed using several newly designed algorithms: the standalone calorimeter-based taus, the tracks+$e/\gamma$ taus, and more complex algorithms which make use of particle-flow. In particular PUPPI taus come from an algorithm that uses PUPPI candidates as input to a neural network (NN) which provides different working points for object identification efficiency.

A comparison between the reconstruction efficiency for some of these algorithms is shown in Fig. 3.120. At plateau, calorimeter-based taus reach the highest efficiency (99%), followed by the loose working point of the PUPPI tau algorithm (90%), and by the tracks+$e/\gamma$ tau algorithm (80%). The comparison between the performance of these objects, in terms of the combined effect of background rejection and momentum resolution, is shown in Fig. 4.5, where the single
4.2 Simplified menu with Phase-1 thresholds

Towards the end of the LHC Run-2, the total number of seeds used in the CMS L1 menu was between 350 and 400 [6]. With the intention of defining the simplest L1 menu for Phase-2 that would cover most of the CMS physics program, only all the unprescaled\(^1\) triggers (~150 in 2018) seeding HLT paths which were used for physics analysis were retained for this exercise. The L1 seeds of the baseline Phase-1 L1 menu targeting an instantaneous luminosity of \(2 \times 10^{34} \text{cm}^{-2}\text{s}^{-1}\), that come out of this selection, were used to seed the HLT paths that recorded 75% of the whole CMS dataset. A further simplification was achieved by selecting only the less restrictive unprescaled trigger conditions of each “family” of seeds: less restrictive refers here to the lowest \(p_T\) threshold applied on trigger objects. From the remaining list, the triggers for which new techniques have been developed, such as displaced muon or displaced jet triggers (described in Section 4.3) were removed as well. A simplified list of 41 triggers is ob-

\(^1\)The trigger rate of an algorithm can be reduced by applying a “prescale” that determines what fraction of events selected by the seed will pass the trigger. The baseline Phase-1 L1 menu targeting an instantaneous luminosity of \(2 \times 10^{34} \text{cm}^{-2}\text{s}^{-1}\) contained approximately 200 to 250 such prescaled seeds.
Chapter 4. Level-1 trigger menu

Figure 4.5: Trigger rate as a function of the offline visible tau $p_T$ thresholds for different tau identification algorithms.

...tained through this process, which cover 70% of the total L1 trigger rate. This list also contains duplicated paths targeting higher $p_T$ objects with algorithms that have higher reconstruction efficiencies as discussed in Section 4.1.

This simplified list of trigger conditions, containing single-object and cross-object triggers, was used as a starting point for deriving the Phase-2 L1 trigger menu. The importance of maintaining Phase-1 thresholds in a 200 average pileup environment for single-object triggers has been extensively discussed in Section 1.4. The simplified menu also contains the cross-object triggers designed during Phase-1 to further reduce the thresholds on individual objects of the trigger seeds, targeting final states with multiple objects with soft $p_T$.

Table 4.1 describes the content of the L1 simplified menu for CMS Phase-2; all the selection details for each of the 41 seeds are listed, together with the Phase-1 offline $p_T$ and $E_T$ thresholds and their respective individual trigger rates. As described in Section 4.1 the corresponding online thresholds are obtained with the scaling procedure that chooses the trigger efficiency point at 95% of the plateau for $\mu$, $e/\gamma$, jets and at 90% for $\tau$, $H_T$ and $E_T^{miss}$. For triggers that use $\tau$, $H_T$ and $E_T^{miss}$, the corresponding offline thresholds at 50% of the trigger efficiency plateau are also quoted; physics analysis can increase their acceptance selecting the events with these offline thresholds far from the trigger plateau. For each trigger the reconstruction efficiency plateau for the objects used in the seed implementation is also quoted. In any given collision event, several seeds may be "fired". The overlaps (correlations) among different algorithms are accounted for when estimating the menu rate, reported at the end of the table.

The total rate of this set of seeds meets the bandwidth requirements, with a 50% margin; it therefore shows that a physics reach similar to the one of the Phase-1 L1 menu can be ensured even under the harsh HL-LHC luminosity conditions. Moreover, it should be noted that a number of these seeds using upgraded algorithms achieve low rates, even at 200 pileup. Therefore, this simplified menu baseline has available bandwidth which can be spent to reduce the thresholds on a number of seeds in order to increase the acceptance for key physics signals. These include reducing the Single TkMuon seed $p_T$ threshold from 22 GeV to 15 GeV with a rate increase of 30 kHz and the Single TkIsoElectron seed $p_T$ threshold from 28 GeV to 22 GeV with a 25 kHz rate increase. These reduced thresholds would help to increase the acceptance in a col-
lection of channels that trigger using leptons, for example $H \rightarrow \tau\tau \rightarrow \tau_\ell\tau_\ell$ and $HH \rightarrow bb\tau\tau \rightarrow bb\tau_\ell\tau_\ell$ ($\ell = e, \mu$), where the lepton tends to be quite soft (see Fig 1.5). Other triggers that are generally used by analyses limited by the available thresholds include: the PUPPI $E_T^{\text{miss}}$ seed whose offline threshold at 90%(50%) could be lowered from 200(128) GeV to 170(100) GeV, with a rate increase of 42 kHz, benefiting the search for the Higgs decay channel $ZH \rightarrow \nu\nu bb$ and the searches for BSM in compressed spectra as shown in Fig 1.6; or the Double PUPPI Tau seed whose thresholds on both legs could be lowered from 52/52(36/36) GeV to 40/40(30/30) GeV for an additional 9 kHz rate, to increase the acceptance in the $HH \rightarrow bb\tau\tau \rightarrow \tau_\ell\tau_\ell$ search (see Fig 1.6), for example. Lowering these thresholds all together would increase the total menu rate by just 70 kHz.
Table 4.1: Simplified menu for Phase-2: 41 trigger seeds with Phase-1 offline thresholds. In the first column the trigger seeds are listed: the names refer to the types of object algorithms described in Section 4.1. In the second column the Phase-1 offline $p_T$ and $E_T$ thresholds are reported, in the third column their total rates. In the fourth column the additional requirements of each seed selection are listed: they include pseudorapidity ranges, the $\Delta z$ cuts for multiple track-matched object seeds, requirements on spatial distance between objects ($\Delta \eta$ and $\Delta R$) and requirements on the invariant mass of two objects. The fifth column contains the trigger efficiency plateau for the objects used in the seed implementation.

<table>
<thead>
<tr>
<th>L1 Trigger seeds</th>
<th>Offline Threshold(s) at 90% or 95% (50%) [GeV]</th>
<th>Rate (PU) = 200 [kHz]</th>
<th>Additional Requirement(s)</th>
<th>Objects plateau efficiency [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single/TkMuon</td>
<td>$22$</td>
<td>$12$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Double/TkMuon</td>
<td>$15,7$</td>
<td>$1$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Triple/TkMuon</td>
<td>$5,3,3$</td>
<td>$16$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Single/TkElectron</td>
<td>$36$</td>
<td>$24$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Single/TkIsoElectron</td>
<td>$28$</td>
<td>$28$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Double/TkElectron</td>
<td>$22, 12$</td>
<td>$36$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Single/StatEG</td>
<td>$51$</td>
<td>$25$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Double/StatEG</td>
<td>$37,24$</td>
<td>$5$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Photons seeds</td>
<td>$36$</td>
<td>$43$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Single/TkIsoPhoton</td>
<td>$22, 12$</td>
<td>$50$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Tau seeds</td>
<td>$150(119)$</td>
<td>$21$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Double/CaloTau</td>
<td>$90,90(69,69)$</td>
<td>$25$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Double/PuppiTau</td>
<td>$52,52(36,36)$</td>
<td>$7$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Hadronic seeds</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single/PuppiJet</td>
<td>$180$</td>
<td>$70$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Double/PuppiJet</td>
<td>$112,112$</td>
<td>$71$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Puppi$H_T$</td>
<td>$480(377)$</td>
<td>$11$</td>
<td>jets: $</td>
<td>\eta</td>
</tr>
<tr>
<td>QuadPuppiJets-Puppi$H_T$</td>
<td>$70,55,40,40,400(328)$</td>
<td>$9$</td>
<td>jets: $</td>
<td>\eta</td>
</tr>
<tr>
<td>$E^{miss}_T$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Puppi$E^{miss}_T$</td>
<td>$200(128)$</td>
<td>$18$</td>
<td></td>
<td>$100$</td>
</tr>
<tr>
<td>Cross Lepton seeds</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TkMuon-TkIsoElectron</td>
<td>$7,20$</td>
<td>$1$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>TkMuon-TkElectron</td>
<td>$7,23$</td>
<td>$3$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>TkElectron-TkMuon</td>
<td>$10,20$</td>
<td>$1$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>TkMuon-DoubleTkElectron</td>
<td>$6,17,17$</td>
<td>$0.1$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>DoubleTkMuon-TkElectron</td>
<td>$5,5,9$</td>
<td>$4$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>PuppiTau-TkMuon</td>
<td>$36(27),18$</td>
<td>$2$</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>TkIsoElectron-PuppiTau</td>
<td>$22,39(29)$</td>
<td>$13$</td>
<td>$</td>
<td>\eta</td>
</tr>
</tbody>
</table>
## 4.2. Simplified menu with Phase-1 thresholds

<table>
<thead>
<tr>
<th>L1 Trigger seeds</th>
<th>Offline Threshold(s) at 90% or 95% (50%) [GeV]</th>
<th>Rate (PU) = 200 [kHz]</th>
<th>Additional Requirement(s) [cm, GeV]</th>
<th>Objects plateau efficiency [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross Hadronic-Lepton seeds</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TkMuon-PuppiHT</td>
<td>6,320(250)</td>
<td>4</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>TkMuon-DoublePuppiJet</td>
<td>12,40,40</td>
<td>10</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>TkMuon-PuppiJet-$\text{Puppi}_{\text{miss}}$</td>
<td>3,100,120(55)</td>
<td>14</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Double TkMuon-PuppiJet-$\text{Puppi}_{\text{miss}}$</td>
<td>3,3,60,130(64)</td>
<td>4</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>TkElectron-PuppiHT</td>
<td>3,3,300(231)</td>
<td>2</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Double TkMuon-$\text{Puppi}_{\text{miss}}$</td>
<td>10,10,400(328)</td>
<td>0.9</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>TkisoElectron-PuppiHT</td>
<td>26,190(124)</td>
<td>9</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>TkElectron-PuppiJet</td>
<td>28,40</td>
<td>34</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>PuppiTau-Puppi-$\text{Puppi}_{\text{miss}}$</td>
<td>55(38),190(118)</td>
<td>4</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td><strong>VBF seeds</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double PuppiJets</td>
<td>160,35</td>
<td>40</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td><strong>B-physics seeds</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double TkMuon</td>
<td>2,2</td>
<td>12</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Double TkMuon</td>
<td>4,4</td>
<td>21</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Double TkMuon</td>
<td>4,5,4</td>
<td>10</td>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>Triple TkMuon</td>
<td>5,3,2</td>
<td>7</td>
<td>$0 &lt; m_{\mu_3\mu_4}, q1 \times q2 &lt; 0,</td>
<td>\eta</td>
</tr>
<tr>
<td>Triple TkMuon</td>
<td>5,3,2,5</td>
<td>6</td>
<td>$5 &lt; m_{\mu_5\mu_2\mu_3}, q1 \times q2 &lt; 17,</td>
<td>\eta</td>
</tr>
</tbody>
</table>

Rate for above Trigger seeds      | 346                                          |
**Total Level-1 Menu Rate (+30%)** | 450                                          |
4.3 New L1 trigger algorithms to extend Phase-1 physics acceptance

The upgraded CMS detector provides the L1 trigger with high-resolution information utilized by sophisticated trigger algorithms. These algorithms can be used to define event selection strategies that could not be envisaged or sustained in terms of additional rate by the Phase-1 system. Beneficiaries of the new triggering techniques include searches for physics beyond the Standard Model and precision measurement analyses. These analyses rely on new triggers to reach an enlarged acceptance and cover an extended phase-space. In the present Section, analyses of selected benchmark signals are conducted to quantify the gain in acceptance for these final states and the benefits brought to the offline analysis strategies are discussed. In some cases, the evaluation of the impact of the enlarged acceptance on the final signal sensitivity remains to be assessed; a complete analysis performed with offline reconstructed objects is still under development at the time of this writing.

This Section highlights the improvements on benchmark physics analyses granted by the key features of the Phase-2 L1 trigger design summarized in Section 1.3. These key features include the availability of charged particle tracks and the computation of the event primary vertex, the extension of muon and electron acceptance covered by the upgraded detectors, the reconstruction of higher-level trigger objects with particle-flow, the reconstruction of displaced objects, the inclusion of timing information and the possibility of implementing machine learning approaches to enhance the signal extraction with kinematic multi-object discriminants. The following categories of physics signals were identified to illustrate the expected improvements.

- Physics with leptons at high pseudorapidity: extended coverage of leptonic triggers.
- Physics with soft and correlated muons (muon jets): use of L1 track finder and extended detector coverage.
- Physics with light mesons: use of L1 track finder.
- Physics with displaced muons: reconstruction of displaced standalone muon tracks.
- Physics with displaced jets: usage of displaced standalone L1 tracks and timing information.
- Higgs physics via the VBF production mechanism: use of particle-flow objects and of machine learning techniques.

For each category one or more physics analyses benefiting from the enlarged acceptance are described, and the technical implementation of the new trigger seeds is discussed. For each new seed, the additional rate associated with its deployment, on top of the simplified menu allocated bandwidth described in the previous section, is reported.

4.3.1 Physics with the extended pseudorapidity coverage of the L1 lepton triggers

As shown in Table 4.1 the Phase-1 lepton trigger seeds provide coverage up to $|\eta| = 2.4$, corresponding to the fiducial volume of the Phase-1 tracking detector, used to reconstruct offline electrons and muons. In Phase-2, the L1 track-matched muon and electron reconstruction, limited to $|\eta| < 2.4$, can be supplemented with standalone object reconstruction within a larger pseudorapidity region. This can be achieved through the extended coverage of the muon spectrometer (up to $|\eta| = 2.8$) and the coverage of the endcap calorimeter (up to $|\eta| = 3$). Given the Phase-2 tracking detector coverage expands up to $|\eta| < 4$, the extension of the lepton triggers coverage becomes even more profitable than in Phase-1. In the following, two examples of
4.3. New L1 trigger algorithms to extend Phase-1 physics acceptance

physics benchmarks, which would benefit from the extended coverage of the L1 lepton trigger seeds, are presented.

- Experimental studies of Double Parton Scattering (DPS) would provide useful information on transverse and longitudinal parton correlations within the proton. It is accessible at the LHC through the channel where two W-bosons of same charge are produced and decay into muons or electrons. This process is also a background to many BSM physics searches with same-sign light leptons, which could therefore benefit from improved experimental measurements of DPS. Understanding the parton correlations is best achieved by studying the slope of the DPS differential cross section $d\sigma/d(\eta_1\eta_2)$ whose precision is enhanced if leptons can be reconstructed in the high rapidity region. The most efficient triggering strategy to capture same-sign W events is to deploy single or double-lepton triggers with enlarged acceptance up to $|\eta| = 2.8$ or 3. The increased acceptance in pseudorapidity gives the opportunity to improve the precision on the $d\sigma/d(\eta_1\eta_2)$ measurement by a factor of $\sim 1.5$, as described in Ref. [11].

- The differential $t\bar{t}$ cross section measurement is used to constrain the parton distribution functions which are a necessary ingredient to any physics data analysis at the LHC. To maximize the signal acceptance, final states where the top-quarks decay semileptonically are retained with the use of single lepton triggers. The sensitivity of the analysis is significant at high rapidity of the $t\bar{t}$ system. Extending the trigger $\eta$ coverage from 2.4 to 2.8 or 3 increases the event yields in this important forward region remarkably. More information on the analysis with projections to the full Phase-2 dataset can be found in [69]; a demonstration that the increased acceptance of the single electron and muon triggers reduces the statistical uncertainties at high rapidities of the $t\bar{t}$ system is provided in Fig. 4.6. Although the results of the analysis performed in the high rapidity region are still dominated by the systematic uncertainties when these are reduced by a factor of $\approx 2$ [69], a reduction by a factor of $\approx 4$ would suffice to have equal contributions from statistical and systematic uncertainties. Given that with respect to [69] a factor of 2 could already be achieved today, as in Ref. [70], a factor of 4 is not unrealistic for HL-LHC.

As shown in Fig 3.13, calorimeter-only electrons feature a reconstruction efficiency above 90% up to $|\eta| = 3$; the corresponding trigger rate for this extended pseudorapidity coverage is shown in Fig. 4.2. Making use of these objects, a dedicated single-electron seed that makes use of track-matched electrons up to $|\eta| = 2.4$ and calorimeter-only electrons for $2.4 < |\eta| < 3$, using the offline threshold for the single-electron trigger of the Phase-2 simplified menu, 36 GeV (see Table 4.1), have therefore been designed. This new seed adds a pure rate of 12 kHz to the total menu bandwidth. Developments are ongoing at the time of this writing to provide the possibility to extend in pseudorapidity also the single muon trigger, exploiting the coverage up to $|\eta| = 2.8$ of the new muon spectrometer.

4.3.2 Physics with L1 soft and correlated muons

Given the large production rate at LHC, $b$-hadron decays can be used to test many BSM theories through enhancement of rare decays. Selecting these signal events was usually performed via low-$p_T$ double-muon triggers, which have become a challenge for CMS already during the LHC Run-2. These originally loose triggers had to become more restrictive due to the increasing average pileup: a requirement of maximal separation of muons in pseudorapidity had to be imposed, together with either an increase of the $p_T$ threshold for the leading muon or a restriction for both muons to be in the central rapidity region.
Figure 4.6: Demonstration of the reduction of statistical uncertainties in measurements of differential top-quark pair production cross sections when extending the single lepton triggers to higher pseudorapidity values.

The search for $\tau \rightarrow \nu \mu \mu$ is one of the flagship analyses of the HL-LHC program and considered as a golden channel in which new sensitivity can be realized for lepton flavor violation. The current best limit on the BR of $\tau \rightarrow \mu \mu$, $2.1 \times 10^{-8}$ (90% CL), obtained by the Belle experiment [71], has so far been out of reach for the LHC experiments. Many BSM theories predict a BR as high as $10^{-9}$ [72], which makes this signal an attractive target for the HL-LHC era.

Through proton-proton collisions, most of the $\tau$ leptons are produced in decays of $D$ and $B$-mesons (75% and 17% respectively). Since the mass of these light mesons is close to that of the $\tau$, the decaying muons appear forward and collimated in the detector. During LHC Phase-1, the triggering strategy at Level-1 was based on the requirement of two standalone muons with $|\eta| < 1.6$ and $|\Delta(\eta^{\mu_{1}}\eta^{\mu_{2}})| < 1.8$, which were subsequently matched at HLT to tracks with $p_T > 3$ GeV. After combining those objects with an additional track with a minimum $p_T$ of 1.2 GeV also reconstructed at HLT, an invariant mass of three objects was computed and required to be $\sim 170$ MeV around the $\tau$ lepton mass. Because of the thresholds at Level-1 and HLT, the offline analysis was restricted to two muons of $p_T > 3$ GeV and one muon of $p_T > 2$ GeV, two of which are required to be in the central barrel region of $|\eta| < 1.6$.

The Phase-2 L1 trigger upgrade provides solutions to the limitations of the Phase-1 trigger by exploiting the information from muon detectors combined with the information from the tracker. This allows for moving most of the trigger selection cuts from HLT to Level-1. The L1 tracks have an excellent resolution in $p_T$ and position ($\eta - \phi$ space) which can be successfully exploited to reconstruct the event kinematics with objects decaying to charged particles. The improved $p_T$ resolution allows to lower the trigger thresholds while the improved position resolution allows for a higher muon identification purity and precise computation of correlated quantities such as the invariant mass and the geometrical correlation between decaying objects in azimuthal angle and rapidity. Correlation variables can then be used to discriminate against the background, which displays predominantly non-correlated object topologies. In this case, the reduction of the $p_T$ thresholds from 3 to 2 GeV and the enlargement of acceptance from $|\eta| < 1.6$ to $|\eta| < 2.4$ is possible by using the track-matched muons at Level-1. The extended coverage of muon detectors ($|\eta| < 2.8$), where standalone muon stubs are used, provides a further increase of acceptance to this signal, as described below. The two improvements com-
4.3. New L1 trigger algorithms to extend Phase-1 physics acceptance

New L1 trigger algorithms are responsible for an increase of signal acceptance by about a factor 10 compared to the Phase-1 trigger.

The Phase-2 offline analysis strategy is to separate events into two categories according to the tri-muon mass resolution and search for a peak. The triggering strategy developed follows the same as the offline analysis, separating targeted events according to the expected $p_T$ resolution of the objects used in the trigger. The trigger object used is a muon jet, introduced in Part 5 of Section 3.3.5. The muon jets, reconstructed in the Global Muon Trigger, consist of three muon stubs in a cone of $\Delta R < 1$, with at least one matched to a L1 track forming a TkMuStub. The use of muon stubs without the track-matching requirement allows to circumvent the $p_T > 2$ GeV threshold imposed by the L1 tracks.

We consider muon stubs reconstructed in the first stations of the muon endcap detectors: the stubs in CSC chambers, and the stubs in ME0 chambers, reconstructed in regions $0.82 < \eta < 2.4$ and $2.0 < \eta < 2.8$, respectively. In the future, the use of CSC-GEM tandem stubs is foreseen. The stubs are characterized by their position, bending angle, and quality. The bending angle is used to estimate the particle’s $p_T$, with an expected resolution of 20% and 40% for CSC and ME0 stubs respectively. In turn, the estimated $p_T$ is then used to propagate the particle from the detector position back to the vertex, and estimate its azimuthal angle at that vertex. The quality is an integer which represents the number of layers with hits within a chamber used to reconstruct the stub position and the bending angle.

As discussed in Sections 3.3 and 4.1, track-matched muons are reconstructed in the fiducial volume of the track finder and have an implicit cut $p_T > 2$ GeV (the minimum L1 track $p_T$) with an expected $p_T$ resolution below 3%. The current choice of track-matched muon algorithm is TkMuStub which is shown to have 25% higher efficiency than the TkMu one for low $p_T$ muons, as can be seen in Fig. 4.7.

For the $\tau \rightarrow \mu\mu\mu$ analysis, two categories of muon jet triggers are designed according to the expected mass resolution of the muon jet object, labeled Category-1 and -2. An auxiliary trigger, Category-0, which can be constructed in the GT, is also considered. This category is almost completely a subset of Category-1, and is intended to be used as a back-up trigger or to collect events missed by Category-1 and -2.

- Category-1 muon jet trigger objects consist of one unmatched CSC muon stub and two TkMuStub, where all three constituent objects are defined in the region $|\eta| < 2.4$.
- Category-2 muon jet trigger objects consist of two unmatched muon stubs, at least one of which is in ME0 with $|\eta| > 2.4$, and one TkMuStub. This trigger extends the pseudorapidity reach to the maximum coverage of the Phase-2 muon detectors.
- Category-0 trigger objects do not contain unmatched muon stubs, and consist of three TkMuStub objects. This tri-muon object has the best invariant mass resolution, but due to the implicit requirement of its constituents ($p_T > 2$ GeV), its signal phase space is significantly smaller than the one of Category-1.

The tri-object invariant mass for the events in Categories 0 and 1 is shown in Fig. 4.8. In addition to the preselection $\Delta R$ cut, more selection requirements optimized for high signal efficiency are applied on the objects in each trigger category to further suppress the background. The total electric charge of the three constituents is required to be equal to unity, $|\Sigma q_i| = 1$. Additional cuts on maximum $\Delta \eta$ and $\Delta \phi < 0.4$ among the three constituents is required for Category-0 and -1. For Category-2 this cut threshold is relaxed to a value of 0.5. In the case of Category-1 and Category-0, which contain two and three TkMuStub objects respectively, a cut on the $z$-
Figure 4.7: The object reconstruction efficiency of TkMu and TkMuStub in the low $p_T$ region (2 to 5 GeV) as a function of pseudorapidity. In the endcap, the reconstruction of TkMu object shows lower efficiency because the standalone muon-tracks require hits in multiple stations of the muon detector, and very soft muons often do not reach stations further from the interaction point. In contrast, for the reconstruction of TkMuStub the existence of a single hit in the muon station which is matched to the L1 track suffices for the object to be reconstructed. In the barrel, the current reconstruction algorithms have not yet been optimized for muons in this $p_T$ range.

In all three categories, a cut on the invariant mass of two specific legs is required to be less than that of the tau lepton. In Category-1 the two track-matched muons are used and their invariant mass is reported in Figure 4.9, which clearly shows the power of this cut to discriminate against a good fraction of the non-correlated background. These three cuts have around 90% signal efficiency. To control the rate in Category-2 and still profit from the acceptance gain due to extended pseudorapidity coverage, only the muon jets with at least one stub in the ME0 detector, with $|\eta| > 2.4$, are considered. This improves the signal acceptance by about 15% efficiency. In Category-1, only the stubs with the two highest quality flags are considered, which reduces the trigger rates significantly while retaining 80% of the signal. In Category-2, this selection criterion is more stringent and only the stubs with the highest quality are used, with a selection efficiency of 15%.

Finally, in each of the categories, the invariant mass of the three-muon object is reconstructed. The events in the mass windows of 1.5 – 2.1 GeV, 1.0 – 2.5 GeV, and 0.5 – 3.0 GeV are considered to pass the final selection in Category-0, -1, and -2 respectively.

Category-1 has the largest trigger acceptance. It amounts to about 85% of the total acceptance provided by the three categories together. Most of the additional acceptance is provided by the Category-2. While the trigger rate of Category-1 is on the order of 25 kHz, the rate of the trigger in Category-2 is currently around 60 kHz, which would take too much of the available bandwidth. Imposing a $p_T$ requirement on the unmatched-stub would significantly reduce the trigger rate in this category as seen in Fig. 4.10, but with a large cost in efficiency. Nevertheless, Category-2 is included in this study to illustrate the use of the muon detector in the rapidity region 2.4–2.8, and as a starting point for future development. The work is ongoing to develop a
4.3. New L1 trigger algorithms to extend Phase-1 physics acceptance

Figure 4.8: The invariant mass of tri-object tau candidates in the $\tau \rightarrow \mu\mu\mu$ signal, after the pre-selection cuts on $\Delta R$ and $\Delta z$. Left: Category-0. Right: Category-1. Histograms are normalized to unit area. For Category-1, muon jets in events also triggered by Category-0 are shown and labeled as "Cat_1 | Cat_0". A MC signal sample with no simulated pileup is used.

Figure 4.9: Left: The invariant mass of the two track-matched muons of the Category-1 muon jets, after the preselection cuts on $\Delta R$, $\Delta z$ and $\Sigma q_i$; the 200 pileup MC background sample is shown in blue, the MC signal samples simulated with no pileup and with 200 pileup are shown in red. Histograms are normalized to unit area. Right: The invariant mass of the tri-object of the Category-1 muon jets, after the preselection cuts on $\Delta R$, $\Delta z$ and $\Sigma q_i$; muon jets in events also triggered by Category-0 are shown and labeled as Cat_1 | Cat_0. A MC signal sample with an average pileup scenario of 200 is used.
trigger strategy which would further increase the signal acceptance by relaxing the stub quality requirement in the full pseudorapidity coverage of the muon detector. For the Category-2 trigger, which currently has a high rate, one subject of study is a muon-jet trigger which uses TkMu instead of TkMuStub objects. Being more pure, a TkMu-based trigger would suppress the rate by factor of two while causing a loss of signal efficiency of about 30%.

Nevertheless, the present triggers are a considerable improvement with respect to Phase-1. The enlarged signal acceptance of the new triggers is a factor 5 to 10 higher than that of the Phase-1 Level-1+HLT trigger selection. Assuming HLT efficiency close to 1 and $B(\tau \rightarrow \mu\mu\mu) = 2.1 \times 10^{-8}$, the expected number of events to be selected in the HL-LHC dataset with this version of L1 triggers is about 16000. This surpasses by a factor two the number of events expected for the offline Phase-2 analysis reported in Ref. [11]. The signal efficiency, the trigger rates, and the number of expected events in the full HL-LHC dataset for the newly proposed triggers are shown in Table 4.2.

The excellent $p_T$ resolution of L1 tracker tracks will translate into no or only a very minor increase of the corresponding threshold on track-matched legs of the muon jet at HLT. Instead, the rate reduction at HLT will be obtained by imposing requirements on the stub-only leg to be matched to low-$p_T$ L1 tracks, reconstructed at HLT well below 2 GeV (i.e. 0.6–0.9 GeV) and in the full rapidity coverage. This is expected to reduce the rate of Category-1 to below 1 kHz, similar to that of the Category-0, while maintaining an improved signal efficiency for the offline analysis compared to that of Phase-1.

Another novel idea to further increase the signal acceptance at Level-1 is to relax the track-match requirement on any muon objects and allow muon jets consisting of reconstructed muon stubs only to be triggered on. In this trigger category, the targeted phase space is the signal with very low-$p_T$ or very forward decaying muons, which don’t have reconstructed L1 tracks. The expected increase in signal acceptance would be another factor of two compared to what is discussed here. However, the trigger rate of this category is unsustainable at the moment, a factor 10 higher than that in Category-2, and is a subject of ongoing studies.
4.3. New L1 trigger algorithms to extend Phase-1 physics acceptance

Table 4.2: Signal efficiency, trigger event rate, and number of expected events for $\tau \rightarrow \mu\mu\mu$ trigger Category-0, -1, and -2 for 200 average pileup interactions.

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Level-1 Expected number of events</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Efficiency (%)</td>
</tr>
<tr>
<td>Cat 0</td>
<td>4.6</td>
</tr>
<tr>
<td>Cat 1</td>
<td>21.1</td>
</tr>
<tr>
<td>Cat 2</td>
<td>2.6</td>
</tr>
<tr>
<td>Total</td>
<td>24.7</td>
</tr>
</tbody>
</table>

In parallel, work is ongoing to develop a triggering strategy targeting tau leptons from W-boson decays. In these types of events, $\tau$-leptons are produced mostly centrally, with muon decay products leaving signatures in all muon detectors and with higher $p_T$. Dedicated triggers, different from those described above, are therefore needed.

4.3.3 Physics of light mesons with L1 tracking

The addition of charged particle tracking in the hardware trigger opens new opportunities for the L1 trigger, such as the reconstruction of light meson candidates from hadronic decays. In what follows, a case study of the potential performance of a trigger for $B^0_s \rightarrow \phi(K^+K^-)\phi(K^+K^-)$, which is a rare FCNC process forbidden at the tree level in the SM, is described.

Based on this example, further possibilities for other new triggers including $\phi$ or $\rho$ mesons in the flavor and electroweak sectors are discussed.

This section explores whether fully hadronic final states like $B^0_s \rightarrow \phi\phi \rightarrow KK KK$ can be triggered with high efficiency at L1 using only the L1 tracks. The event selection algorithm first reconstructs $\phi$ candidates from pairs of oppositely charged L1 tracks originating from the same vertex and then forms $B^0_s$ candidate(s) from pairs of $\phi$ candidates constrained to come from the same vertex. The $p_T$ of the lowest-$p_T$ kaon lies very close to the lowest possible threshold of the L1 tracking of 2 GeV, as shown in Fig. 4.11 (left), resulting in a major loss of signal efficiency. The same analysis is repeated at the offline level with tracks reconstructed with much higher precision to understand if any further reduction of the efficiency will be incurred at the offline analysis level. Further optimization of the background rejection at the offline level may lead to further improvements with respect to the results shown here.

To optimize signal efficiency and trigger rate, three different baseline selection working points (loose, medium and tight) are defined. The corresponding requirements are listed in Table 4.3.

Table 4.3: Event selection baseline working points. The variable $d_z$ represents the distance between a pair of tracks or trajectories of a pair of reconstructed particles along the beam axis, while $d_{xy}$ represents the distance in the plane perpendicular to the beam axis.

<table>
<thead>
<tr>
<th>Working point</th>
<th>loose</th>
<th>medium</th>
<th>tight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tracks</td>
<td>$p_T \geq 2$ GeV, $</td>
<td>\eta</td>
<td>\leq 2.5$, $\chi^2/ndf \leq 20$, $N_{stub} \geq 4$, $N^{PS}_{stub} \geq 2$</td>
</tr>
<tr>
<td>Track pair $\phi$-pair</td>
<td>$d_{xy} \leq 0.5$ cm, $</td>
<td>d_z</td>
<td>\leq 0.6$ cm</td>
</tr>
<tr>
<td>$\phi$-pair $d_{xy} \leq 0.5$ cm, $</td>
<td>d_z</td>
<td>\leq 0.6$ cm</td>
<td></td>
</tr>
<tr>
<td>$\phi$ mass $0.2 \leq \Delta R(\phi_1, \phi_2) \leq 1$, $\Delta R(K^+, K^-) \leq 0.12$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$B^0_s$ mass $5.0 \leq M_{\phi\phi} \leq 5.8$ GeV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$B^0_s$ mass $5.1 \leq M_{\phi\phi} \leq 5.7$ GeV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$B^0_s p_T \geq 10$ GeV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$B^0_s p_T \geq 12$ GeV</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The distributions of different discriminating variables such as the $\Delta R$ between the $\phi$ candidates, the invariant mass of the $\phi$ candidates, $M_{K^+K^-}$, and the invariant mass of the $B^0_s$ candidates,
$M_{\phi \phi}$ are presented in Fig. 4.11 and Fig. 4.12, at the trigger and offline reconstruction levels for signal events, and at the trigger level for background events. The tails of the signal distribution, for both L1 and offline, look very similar to those of the background, because they are dominated by random combinations of pileup tracks.

Figure 4.12 (right) shows the expected $B_s^0 \rightarrow \phi \phi$ invariant mass distributions from all $\phi$-pairs with separation along the beam axis ($z$) of $|d_z|$ ($\phi$-pair) $\leq 0.6$ cm, distance in the plane perpendicular to the beam axis $d_{xy}$ ($\phi$-pair) $\leq 0.5$ cm, $0.2 \leq \Delta R$ ($\phi$-pair) $\leq 1$, and $\Delta R (K^+, K^-) \leq 0.12$, for an average pileup scenario of 200. Simulations show that an efficiency of $30 - 35\%$ can be achieved at the L1 trigger level, and that the events selected at L1 would be accepted by the subsequent offline analysis with high efficiency. Signal efficiencies of the different trigger working points, the subsequent offline selection efficiency, and the corresponding rates for an average pileup scenario of 200 are shown in Table 4.4. For the “medium” working point, a moderate signal efficiency of around $30\%$ and a reasonable L1 trigger rate of $\sim 15$ kHz are obtained.

The firmware resource usage for such a large number of invariant mass pair calculations must be considered in the context of available resources on the planned FPGAs. We estimate the firmware resource usage required by extrapolating from the current GT, which performs two-track invariant mass calculations. The two-track mass calculation implemented on a VU9P, the target FPGA for the GTT boards, uses two DSPs and is capable of processing a new mass
4.3. New L1 trigger algorithms to extend Phase-1 physics acceptance

Figure 4.12: Left: invariant mass distribution of all track pairs with opposite charge, $|d_z|$ (track-pair) ≤ 0.6 cm, $d_{xy}$ (track-pair) ≤ 0.5 cm, and track $p_T ≥ 2$ GeV, where each track is assumed to be a kaon. Right: Invariant mass distribution of all the $\phi$-pairs with $0.2 ≤ \Delta R$ (φ-pair) ≤ 1, $\Delta R (K^+, K^-) ≤ 0.12$, $|d_z|$ (φ-pair) ≤ 0.6 cm, and $d_{xy}$ (φ-pair) ≤ 0.5 cm. For both the panels, the distributions are normalized to unit area. The signal and background distributions with L1 tracks are shown as red solid lines and green histograms, respectively. The distributions with offline tracks are shown as blue dashed lines. An average pileup scenario of 200 is used.

calculation each clock cycle. The $B_0^0 \rightarrow \phi \phi \rightarrow KKKK$ simulation shows that in more than 99.9% of signal events, less than 135 positively charged tracks, and less than 135 negatively charged tracks with 200 pileup events are expected. A total of 18,225 two-track invariant mass calculations would be needed to fully process each signal event. A time multiplex factor of 6 as planned for the GTT architecture described in Sec. 5.4 is used, and a clock speed of 320 MHz is assumed. This could perform the 18,225 two-track invariant mass calculations in 763 DSPs per FPGA, or in about 11% of the total number of DSPs in a VU9P FPGA. The resource usage is thus considered to be within a reasonable range for the system proposed.

Based on the encouraging performance prospects for $B_0^0 \rightarrow \phi (K^+ K^-) \phi (K^+ K^-)$, other channels that feature light meson candidates are considered. In the flavor sector, $B_0^0 \rightarrow J/\psi (\mu^+ \mu^-) \phi (K^+ K^-)$ is a classic signature for CP-violation studies, which is currently triggered on at CMS only by the di-muon signature. The addition of a $\phi$ candidate and the four track invariant mass would allow the muon momentum thresholds to be reduced and increase acceptance to these decays at L1.

Light meson candidates also offer interesting opportunities to study rare Higgs decays. Decays such as $H \rightarrow \phi \gamma$ or $H \rightarrow \rho \gamma$ have been proposed as probes of light quark Yukawa couplings [73–75], which are otherwise very difficult to measure directly. Without the L1 track finder, these decays can only be triggered on in single photon events, where some acceptance is possible, but the momentum thresholds are generally higher than desirable for Higgs boson decays. A light meson trigger based on a di-track resonance could be combined with a photon to help reduce the $p_T$ threshold on the photon and thus increase the acceptance to these probes of rare Higgs boson decays and potential indirect new physics signatures.

4.3.4 Physics with L1 displaced muons

There are many BSM processes featuring the production of displaced muons in their associated final states. The Phase-1 L1 trigger had a limited potential to efficiently select signatures with
largely displaced muons due to the assumed primary vertex position used in the trigger object reconstruction. Similarly in Phase-2, if the displacement is sufficiently large (more than few centimeters), events from these signal processes would not be selected by the standard trigger algorithms used in the previous Sections targeting prompt objects, such as L1 tracks, standalone muons or track-matched muons. With the improved triggering techniques envisioned in the upgraded system, the relaxation of the primary vertex constraint can be performed, allowing these particular objects to be reconstructed with higher efficiency. The corresponding reconstruction algorithms are referred to as displaced standalone muons and are discussed in Section 3.3.

To illustrate the acceptance gain obtained by using displaced standalone muon triggers, exotic signals that involve the production of dark photons, as described in Ref. [25], are considered. Dark photons here are the mediators of a broken dark U(1) gauge theory that kinematically mix with the Standard Model photons which subsequently decay at a displaced vertex into a pair of fermions. In such models, the dark photon couples to SM charged particles in the same way as a SM photon, except that the couplings are scaled by a parameter $\epsilon$ that indicates the strength of the kinetic mixing. The dark photon lifetime is proportional to $1/\epsilon^2$, and since values of $\epsilon$ can be very small, the dark photon lifetime can be significantly long. If the dark photon has a non-zero transverse momentum, it can have a macroscopic decay length. The most recent CMS result based on LHC Run-2 data [25] uses events collected by HLT trigger algorithms that are efficient for lifetimes of the dark photon only up to 10 cm. The ability to trigger on displaced muons up to larger displacements with high efficiency can significantly improve the sensitivity for these signal models compared to what the Phase-1 triggers could procure. The Phase-2 analysis described in Ref. [76] can be used as reference. In this analysis a very high double-muon trigger efficiency (> 90%) was assumed for displacement up to 300 cm, with offline $p_T$ thresholds of 20 and 15 GeV for each leg respectively. Signal models with dark photon lifetime up to 10 m are explored and the sensitivity projection with 3000 fb$^{-1}$ of data is reported.

As shown in Fig. 3.24 the displaced standalone muon algorithm in the barrel region provides a reconstruction efficiency that ranges from 95% to 50% when reaching 100 cm of displacement in the transverse plane ($d_{xy}$). In the endcap region, the displaced standalone muon algorithm provides a reconstruction efficiency of 50% up to $d_{xy}$ of 100 cm, as show in Fig. 3.35. We have evaluated the trigger rates of dedicated displaced standalone muon seeds: a single muon seed which requires L1 $d_{xy} > 75$ cm ( > 20 cm in the endcap region) and offline $p_T$ thresholds of 22 GeV (matching the same threshold of the single muon seed of the simplified menu), and a double muon seed without L1 $d_{xy}$ requirements and $p_T$ thresholds of 20 and 15 GeV. The efficiency of the L1 $d_{xy}$ cut requirements can be seen, respectively for the barrel and endcap regions, in Fig. 3.24 and Fig. 3.35. The barrel and endcap rates for the single (double) muon seeds are respectively 5.4 (1.6) kHz and 8.7 (0.3) kHz. The total pure rate added to the simplified menu is about 37 kHz.

Displaced muon triggers exploiting a potential dedicated displaced tracking reconstruction algorithm, described in Section 2.1, combined with muon detector information are the subject of future developments. These algorithms are expected to have high efficiency and lower rate for displacements of up to 5–10 cm, allowing a reduction of the trigger thresholds, and are thus useful to target low mass dark photons.
4.3.5 Physics with L1 displaced jets

Triggering on signals that feature fully hadronic final states with soft jets is extremely difficult because of the jet energy scale and energy resolution degradation induced by the high pileup environment. For many processes with low hadronic activity, either multi-jet triggers or hadronic sum triggers such as those based on $H_T$ remain inefficient due to the imposed large thresholds to contain the rate. There are numerous BSM theories considering hidden sectors that predict the decay of the SM Higgs boson into new light scalars. These scalars can be long-lived and decay into SM particles producing hadronic final states ($h \rightarrow \phi \phi \rightarrow bbbb$). The Higgs boson being light, soft hadronic objects are produced when decaying through these hidden states. During Phase-1 the only viable way to study this Higgs boson decay mode was to select events where the Higgs boson is produced in association with a $Z/W$ boson for the price of a reduced cross section. The only handle that can be used to select events, where the Higgs boson is produced via gluon fusion, is to exploit the long lifetime that these hidden states can feature. Therefore, the design of trigger algorithms that target displaced soft jets is a strategic approach to assess a complete new phase space for exotic Higgs decays while keeping the rate to an acceptable level.

The Phase-2 Level-1 trigger system permits the development of two different trigger algorithms to reconstruct efficiently displaced jets, each targeting improved sensitivity for different jet displacements: displaced tracker jets (see Section 4.3.5.1) and time-displaced calorimeter-based jets (see Section 4.3.5.2). The first approach is targeting jet displacement up to 5–10 cm, while the second one provides maximum coverage to larger values between 50 cm and 150 cm.

4.3.5.1 Physics with displaced tracker jets

An extension of the L1 track-finding design described in Section 2.1 provides a new track handle to trigger on BSM physics. We consider the extension of the L1 track finder to off-pointing tracks, and develop a jet lifetime tag for tracks with $|\eta| < 2.0$. Arising from the decay vertex of a long-lived particle, off-pointing tracks do not point back to the primary collision point. This feature is usually quantified in terms of the transverse impact parameter, $d_{xy}$, which gives a measure of the smallest distance between the transverse projection of the track and the primary collision point. A jet with multiple tracks that have large $d_{xy}$ provides a high purity tag for a variety of signals that contain displaced jets. In particular, cases with low hadronic activity ($H_T$), for which the signal acceptance is limited (when using Level-1 $H_T$ triggers) to a point where an offline analysis cannot be performed, are considered. As discussed above, one signal model that results in a “blind spot” at Level-1 is the exotic decay of the Higgs boson into hadronic final states via light scalars ($\phi$), where the signal has very little $H_T$ due to the mass of the Higgs boson. For this scenario, both the decay of the SM Higgs boson with $m_H = 125$ GeV and of a BSM heavy Higgs boson with $m_H = 250$ GeV across a range of $\phi$ lifetimes $\tau = [0, 10]$ cm, are considered.

To quantify the improvement achieved with L1 displaced tracker jets, the signal yields, the signal efficiency and the trigger rate are compared with L1 seeds using nominal track finding reconstruction as described in Section 2.1. Tracker-$H_T$ seeds are considered using the track clustering algorithm described in Section 2.1 for the extended track finding and also for the nominal version, which are referred to here as prompt tracks. The prompt track selection is shown in Table 3.7 and Table 4.5 gives the selection criteria for the extended tracks. The extended track collection with the 5-parameter fit is optimized in two cases for L1 tracks based on the number of track stubs available for the track fit. The case of 4-stubs is further split depending on the $d_{xy}$ of the track. The 5-stub or greater case is kept loose because these are the tracks considered for
displaced jet tagging. The L1 tracker $H_T$ rates become compatible between the extended and prompt track-finding approaches with this track purity selection. The L1 displaced jets require at least 3 tracks with a track fit $d_{xy} > 0.15$ cm for 5-stub tracks and a tighter requirement for 4-stub tracks $d_{xy} > 0.5$ cm.

The improved L1 performance with the displaced jet tag can be seen in the variation of the trigger rate versus the signal efficiency for the SM Higgs boson and heavy Higgs boson model points. Figure 4.13 highlights the improvement for a range of $\phi$ lifetimes and masses. The exotic decay of the SM Higgs boson is a key signature for studying the performance of a displaced tracking approach. The prompt-tracks $H_T$ trigger can only select 5% of the total signal at a large rate of 40 kHz. The extended tracking can improve the signal efficiency on these low $H_T$ signals by almost a factor of three, when considering displaced jets in the final state that allow a large number of signal events offline to be probed with negligible backgrounds.

Considering the large data sample expected at HL-LHC with 3000 fb$^{-1}$, the signal efficiency is evaluated in terms of the number of signal events. The production cross section times the branching fraction to $ff$ for both the SM Higgs boson and the heavy Higgs boson are taken to be the same: $\sigma^{\text{pp}}_{H(250)} B[H(250) \rightarrow \phi\phi \rightarrow 4j] = 10^{-5} \sigma^{\text{pp}}_{H(125)} = 0.55$ fb. The cross section $\sigma^{\text{pp}}_{H(125)}$ is taken from the gluon-fusion production process at center-of-mass energy of 14 TeV for a Higgs boson mass of 125 GeV. The signal yields are shown for two possible L1 tracker $H_T$ rates. Though the optimal $H_T$ trigger rate depends on the available bandwidth in the full L1 menu, an upper bound on the trigger rate at 25 kHz and a lower bound at 5 kHz are chosen to study the improvement as a function of lifetime. Figure 4.14 shows the event yields for exotic scalar decays across a range of lifetimes. We re-weighted the generated samples to interpolate the efficiency across the full range of $crt$ in $[0, 10]$ cm. For all cases the efficiency peaks near $crt \approx 1$ cm. Comparing the yields between prompt and the extended tracking, the largest gain can be seen for the SM Higgs boson signals where the acceptance is increased by almost a factor of 6. This new L1 seed would allow these exotic Higgs boson decays to be triggered at L1 and gives a large enough signal yield across a range of lifetimes to cover this BSM phase-space.

### Table 4.5: Track selection for jet finding with the extended track collection using a 5-parameter track fit.

<table>
<thead>
<tr>
<th>$N_{\text{stubs}}$</th>
<th>$d_{xy}$ (cm)</th>
<th>Track Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\chi^2_{\text{ndof}}$</td>
</tr>
<tr>
<td>4</td>
<td>&lt; 0.5</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>$\geq$ 0.5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>2.75</td>
</tr>
</tbody>
</table>

**4.3.5.2 Physics with displaced jets using timing information**

The use of precision timing information in the L1 trigger significantly enhances our capability to tag long-lived particle (LLP) decays, which exhibit a time delay with respect to the closest bunch crossing. Past studies have suggested that for LLP lifetimes between 10 cm and 100 cm, the delayed timing signature is among the best for triggering and background rejection [77]. In particular, for decays that occur in the vicinity of the calorimeters, between 50 cm and 150 cm away from the primary interaction vertex, timing provides a unique signature to discriminate between signal and background. The barrel ECAL and HCAL subsystems will both provide
4.3. New L1 trigger algorithms to extend Phase-1 physics acceptance

Figure 4.13: Rate of the tracker $H_T$ trigger as a function of signal efficiency for both the prompt and displaced tracking for the SM Higgs boson (left) and the heavy Higgs boson (right).
Figure 4.14: The plots show the number of triggered signal events with $3000 \text{ fb}^{-1}$ for the SM Higgs boson (top) and heavy Higgs boson (bottom). Two L1 rates are considered, a larger rate of 25 kHz (left) and a lower rate of 5 kHz (right). The signal event yields are shown for prompt track finding as dotted lines and for the extended track finding with a displaced jet tag in solid lines.
4.3. New L1 trigger algorithms to extend Phase-1 physics acceptance

timing information to the L1 trigger, while the endcap HGCAL will not have such capability due to bandwidth constraints (see Section 2.3.1).

We study the potential of using the barrel ECAL timing information to trigger on such LLP decays. The signal model considered is the same as the one discussed above: the SM Higgs boson which decays to two LLPs with mass of 50 GeV, each decaying to a pair of b-quarks. As previously discussed, this model produces jets that are relatively low in transverse momentum, peaking at around 20 to 30 GeV, and therefore represents the most difficult signal model to trigger on. We study scenarios with signal proper decay length between 10 cm and 1 m. We reconstruct the timestamp for calorimeter-based jets by averaging the measured timestamp of each ECAL TP comprising the jet, weighted by its measured energy. The time resolution as a function of energy is parametrized based on testbeam measurements of the ECAL module prototype and simulations. The HL-LHC beamspot also has an intrinsic spread in both the longitudinal and time dimensions, which translates into a Gaussian spread of width about 220 ps. A more pessimistic scenario for the time resolution was also studied, where the ECAL timing resolution was assumed to be twice the testbeam measurement and even in the pessimistic case, the calorimeter-based jet timing resolution is dominated by the size of the beamspot for jets above about 15 GeV.

In Fig. 4.15, the single calorimeter-based jet trigger rate is displayed as a function of the signal efficiency when the jet $p_T$ threshold is varied, for four different timing delay requirements. For a given background rate, progressively tighter timing requirements allow higher signal efficiencies to be achieved because of less stringent requirements on $p_T$. With a time delay above 1 ns, a signal efficiency of 20% at a trigger rate ~25 kHz, or 14% at a trigger rate of 10 kHz, can be obtained.

Figure 4.15: The rate of the proposed single time-displaced calorimeter-based jet seed is plotted against the signal efficiency for different requirements on the jet timestamp and $p_T$. A single calorimeter-based jet is required and the different curves represent different jet-time requirements. The points along the curve represent different requirements on the jet $p_T$, varying between 20–100 GeV. The signal model considered is SM Higgs boson production decaying to two scalar long-lived particles of mass 50 GeV, each decaying to a pair of b-quarks. The proper lifetime of the scalar is 1 m.
4.3.6 Dedicated VBF Higgs boson production L1 trigger algorithms based on machine learning techniques

Traditionally, algorithms used to build the trigger menu have employed simple selections on one or more physics objects. With the Phase-1 trigger upgrade, the possibility arose to also build algorithms with requirements on correlations between multiple objects, for example the invariant mass or angular distance between two jets. These types of algorithms can be referred to as “cut-based”, in that they are formed from basic selections on single objects and object correlations. A natural continuation of this evolution towards more complex trigger menu algorithms would be to utilize modern machine learning tools to build more powerful multivariate discriminators. Such techniques have become widespread in the offline analysis of LHC collision events, however they have yet to be used in the trigger menu. The primary reasons for this are that until recently the software tools to synthesize such algorithms into FPGA firmware did not exist, and even if they did the available resources of the FPGAs used in the Phase-1 trigger are tightly constrained. However, as described in detail in Section 3.7, both of these hurdles are overcome with the Phase-2 upgrade.

In this Section, other challenging signal topologies for the HL-LHC are selected: the vector boson fusion (VBF) production of the Higgs boson followed by its subsequent decay into either invisible particles or a pair of bottom quarks. As a baseline, the acceptance to these signals for traditional cut-based algorithms, part of the simplified menu discussed in Section 4.2, was studied. These include the inclusive $E_T^{\text{miss}}$, $H_T$, single-jet or multi-jet triggers, and the “inclusive VBF” trigger that requires two jets with large invariant mass. When combined together and taking into account their degree of overlap, these triggers had a total acceptance of 0.33 (0.36) and a rate of 62.7 (136.5) kHz for the VBF $H\rightarrow$invisible and VBF $H\rightarrow b\bar{b}$ signals, respectively. In addition, in the case of VBF $H\rightarrow$invisible the acceptance of these triggers was checked in a phase space that was optimized for a cut-based analysis based on the expected significance derived using Delphes simulation of the signal and main backgrounds in the offline analysis (predominantly $W$+jets and $Z$+jets), based on the work reported in Ref. [78]. This phase space is defined to be $p_T(\text{jet1})>70$ GeV, $p_T(\text{jet2})>40$ GeV, $\Delta\eta(jj)>4$, $\Delta\phi(jj)<2$, and $m(jj)>1000$ GeV. The selection represents the optimal requirements that would be applied when performing a single bin counting experiment to maximize the expected significance. The acceptance in this phase space for the inclusive cut-based $E_T^{\text{miss}}$ and VBF triggers was found to be 0.76 and 0.70, and the acceptance of their logical OR was found to be 0.82.

Next, it was studied whether a dedicated cut-based trigger for the VBF $H\rightarrow$invisible topology that combines jet and $E_T^{\text{miss}}$ information could improve the acceptance. Several options were considered, and the final configuration that was chosen requires $p_T(\text{jet1})>70$ GeV, $p_T(\text{jet2})>30$ GeV, $m(jj)>200$ GeV, and $E_T^{\text{miss}}>180$ GeV. This dedicated cut-based VBF $H\rightarrow$invisible trigger has an inclusive acceptance of 0.28 and a rate of 57.4 kHz for the VBF $H\rightarrow$invisible signal. In the VBF phase space, the acceptance was found to be 0.81. This trigger performs worse than the logical OR of inclusive $E_T^{\text{miss}}$ and VBF triggers, and it was found difficult to define a cut-based trigger that could exploit the correlation between $E_T^{\text{miss}}$ and many jet observables to improve the signal acceptance in the full phase space without significantly increasing the rate. This trigger would represent a significant fraction of the total bandwidth, and therefore it was investigated whether a machine learning based approach could perform better than the cut-based approach.

A deep neural network (DNN) machine learning approach was adopted, constructed and trained using the TensorFlow software package. The signal was taken as either the VBF $H\rightarrow$invisible or the VBF $H\rightarrow b\bar{b}$ signal MC, and the background was taken to be the same sample used for rate estimations of the menu. Separate trainings were performed for the $H\rightarrow$invisible and
4.3. New L1 trigger algorithms to extend Phase-1 physics acceptance

H → b$b$ signals. This is due to the observation that the DNN can distinguish between the two signals, and therefore a training for an inclusive VBF trigger was not successfully accomplished (although this is the study of ongoing research). The inputs to the DNN include the $p_T$, $\eta$, and $\phi$ of the 3 leading jets; the total event $H_T$; the di-jet $p_T$, invariant mass, and $\Delta R$ of the jet pairs built from these 3 leading jets. Additional input variables related to $E_T^{\text{miss}}$ are also included in the training, to enhance the sensitivity particularly to the H → invisible topology. These include the $E_T^{\text{miss}}$ and $E_T^{\text{miss}}\phi$, as well as the $\Delta \phi$ between the di-jet system and the $E_T^{\text{miss}}$. In total, 24 input variables are used for the training. The network architecture consists of 3 fully connected hidden layers with 72 nodes each, and an output layer with a single node representing the final discriminant score. The activation functions were chosen as ReLU and sigmoid for the hidden layers and output layer, respectively. In between the fully connected nodes, dropout layers are inserted to prevent overtraining. A method known as “pruning” is applied as a stage in the training to remove connections with low weights in order to reduce the total number of multiplications in the final DNN. After pruning, roughly 4300 multiplications are required to use the network to make each prediction. Details on the resource estimates for this network can be found in Section 3.7.

![Graphs showing signal efficiency and rate comparison](Image)

Figure 4.16: Signal efficiency in the full phase space and the rate of different cut-based triggers compared to dedicated DNN triggers trained for VBF H → invisible (left) and VBF H → b$b$ (right).

The performance of the DNN trigger compared to the cut-based algorithms can be seen in Fig. 4.16. The DNN is found to outperform any of the individual cut-based algorithms as well as their logical OR. For the same rate as the logical OR of the inclusive $E_T^{\text{miss}}$ and VBF triggers, the efficiency to the VBF H → invisible signal in the full phase space is found to be 0.38 compared to 0.33. In addition, the DNN trigger was found to have an efficiency in the phase space optimized for a single bin counting experiment of 0.86 compared to 0.81. It should be noted, however, that a full analysis would not utilize just the single most sensitive bin, and instead it could be redesigned to exploit the new phase space accepted by the DNN.

The gain in efficiency in the full phase space at the same rate is even greater for the VBF H → b$b$ signal. In this case, the efficiency of the DNN trigger at the same rate as the logical OR of single-jet, double-jet, $H_T$, and inclusive VBF triggers was 0.49 compared to 0.36 for the logical OR of the cut-based triggers. The efficiency of the DNN trigger at the same rate as the single-jet trigger, which has the highest efficiency for any single cut-based trigger, was 0.38 compared to
0.23 for the single-jet trigger. The larger gains in the VBF $H \to b\bar{b}$ signal efficiency is due to the lack of $E_T^{\text{miss}}$ in the signal process, which is a powerful handle for rejecting the background.

These studies represent a proof of concept for the design of more complicated cross-object triggers that can have improved performance thanks to the high quality L1 PF reconstruction, as well as a proof of concept of a DNN implementation in the Global Trigger that can potentially be used for this or other challenging topologies.

### 4.3.7 Summary of triggers for extended physics reach

Table 4.6 describes the content of the L1 menu which extends the physics reach for CMS Phase-2 beyond that of Phase-1; this is achieved by 7 novel triggers whose selection details are listed, together with the $p_T$ and $E_T$ thresholds, their respective individual trigger rate, and the cumulative rate they add to the simplified menu of the Phase-1 physics acceptance.

A summary diagram displaying the links between the trigger primitives, the trigger objects, the Level-1 algorithms used in the extended menu and the physics channels, is shown in Fig. 4.17.

<table>
<thead>
<tr>
<th>L1 Triggers</th>
<th>Online Threshold(s) (* for Offline)</th>
<th>Rate $\langle PLU \rangle = 200$ [kHz]</th>
<th>Additional Requirement(s) [cm, GeV, ns]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single StaEG ext. eta</td>
<td>36 $\ast$</td>
<td>12</td>
<td>$2.4 &lt; \mid \eta \mid &lt; 3.0$</td>
</tr>
<tr>
<td>Muon-Jet (Cat0+Cat1)</td>
<td>2, 2, 0.5</td>
<td>27</td>
<td>$</td>
</tr>
<tr>
<td>Tracker $B^0_s$</td>
<td>12</td>
<td>15</td>
<td>$</td>
</tr>
<tr>
<td>Displaced Single Muon</td>
<td>22</td>
<td>14</td>
<td>$</td>
</tr>
<tr>
<td>Displaced Double Muon</td>
<td>20, 15</td>
<td>2</td>
<td>$</td>
</tr>
<tr>
<td>Displaced Tracker $H_T$</td>
<td>248(153) $\ast$</td>
<td>20</td>
<td>Jets: $</td>
</tr>
<tr>
<td>Displaced Calo-Jet</td>
<td>40</td>
<td>20</td>
<td>$</td>
</tr>
</tbody>
</table>

| Total rate for above triggers | 110 kHz |

### 4.4 Study of the menu rate evolution with pileup

The HL-LHC might surpass the nominal design parameters and exceed the expected average of 200 pileup events. Moreover, it is reasonable to expect that the average pileup in the HL-
4.4. Study of the menu rate evolution with pileup

LHC collisions will vary on a bunch-to-bunch basis, as observed during the LHC Run-2. The variations in pileup are expected to cause a change in the Level-1 trigger rate, and it is relevant to understand if the targeted trigger bandwidth can withstand these changes. It is important to identify and explain single point failures, caused by the individual triggers or types of triggers, which are particularly sensitive to pileup variations and can compromise the menu rate sustainability.

In this Section, the question of Level-1 trigger rate stability against pileup is addressed, and its implications on the rate of the simplified L1 menu (presented in Section 4.2) for different scenarios of HL-LHC instantaneous luminosity, i.e. different values of average pileup, are discussed. For this study, simulated MC background samples with 140, 200, 250, and 300 average pileup events are used. For each scenario considered, the rate of individual triggers in the menu and the menu total rate are computed. Different models of pileup bunch-to-bunch variations are considered and the effects on the menu rate are evaluated.

4.4.1 Individual trigger rate evolution with pileup

The trigger objects used to implement the L1 trigger seeds were designed and optimized for 200 pileup events, as described in Chapter 3. No modifications to either these trigger objects nor to the online thresholds used in the menu are performed when computing the rates expected for other pileup values. Neither the loss in efficiency, nor the resolution degradation due to such
different pileup conditions are quantified. Non-linear trends of the trigger rates with respect to pileup are being identified. The following results can be reported:

- The rates of Single TkMuon (22 GeV), Single TkElectron (36 GeV), Single PuppiJet (180 GeV), and Puppi$H_T$ (400 GeV) triggers are linear with pileup, as shown in Fig. 4.18 (left).
- The rate of Puppi$E_T^{\text{miss}}$ (200 GeV) and Double PuppiJet (VBF) triggers, which both use energy clusters in HF, are under control for pileup up to 250, but are very high for extreme pileup exceeding 300, as shown in Fig. 4.18 (right).
- The total menu rate increases from $\approx 200$ kHz up to $\approx 500$ kHz, when the average pileup changes from 140 to 250. For extreme pileup of 300, the rate of the menu is no longer under control and triples compared to pileup of 250. This is exclusively due to L1 seeds that use energy clusters in HF.

Overall, the menu robustness against pileup is satisfactory and shows an approximately linear behavior up to average values of 250. Slopes of the linearly interpolated lines for pileup values below 250 are similar, although progressively increasing, as shown in Fig. 4.19 (left). A large rate increase has been identified for the extreme pileup scenarios of 300, reflected in the very large value of the slope for the interpolated line describing rates for average pileup values between 250 and 300.

Figure 4.18: Rates of the triggers used in the menu for four values of average pileup: 140, 200, 250, and 300 (markers), assuming no bunch-to-bunch variation. Intermediate values obtained with linear interpolation (lines). Left: representative triggers with an approximately linear dependence with pileup. Right: representative triggers showing a very non-linear dependence.

Puppi$E_T^{\text{miss}}$ and Double PuppiJet VBF triggers are the largest contributors to the excessive menu rate observed in the very high pileup range. Both of these triggers are constructed from PUPPI objects that also use inputs from the hadronic forward calorimeter (HF). Furthermore, the PUPPI objects used by these triggers algorithms were reconstructed targeting tuned performance for 200 pileup events; improved tuning with higher pileup is possible but was not attempted for this document.
4.4. Study of the menu rate evolution with pileup

4.4.2 Alternative $E_T^{\text{miss}}$ and VBF triggers for extreme pileup

Additional studies were pursued to identify means of mitigating the very non-linear rate trend for trigger physics objects known to be susceptible to pileup, the $E_T^{\text{miss}}$ and the very forward jets for VBF signatures. As described in Section 3.5.3.2, the PUPPI algorithm has been designed to take advantage of the event pileup estimate, which can be derived from the GTT track quantities or from the calorimetry deposit information as was done in Phase-1. Since the estimate is used via a logarithmic term of its ratio to the nominal pileup, having a high precision for the estimate is not critical for the performance of this algorithm. The knowledge of the event pileup has not been exploited in the PUPPI reconstruction for studies discussed up to this point. For pileup significantly larger than 200, injecting the knowledge of the event pileup in the PUPPI reconstruction is an efficient means for reducing the rate for well-above nominal pileup scenarios. If this additional information is used at pileup 300, the rates of Puppi$E_T^{\text{miss}}$ and VBF Double PuppiJet triggers are reduced by 41% and 25% respectively. This translates into rate values of 486 kHz and 427 kHz instead of the values shown Fig. 4.18 (right). As expected, for a pileup value of 250, the size of this effect is smaller, and the rates of the two triggers discussed here are reduced by 18% and 8%, corresponding to values of 55 kHz and 116 kHz, respectively.

One way to further control the rate of the $E_T^{\text{miss}}$ trigger is to restrict the reconstruction of PUPPI objects to a region less susceptible to pileup. This technique has been successfully used for commissioning and controlling the $E_T^{\text{miss}}$ trigger rates induced by large pileup during Run 2. For Phase-2, we construct two types of alternative $E_T^{\text{miss}}$ triggers that can be used for this purpose:

- **PuppiNoHF$E_T^{\text{miss}}$** - with $E_T^{\text{miss}}$ defined in region $|\eta| < 3$, which excludes calorimetry information from the HF.
- **PuppiL1Tracker$E_T^{\text{miss}}$** - with $E_T^{\text{miss}}$ defined in region $|\eta| < 2.4$, which excludes calorimetry information beyond the fiducial region of the L1 tracker.

By construction, the two $E_T^{\text{miss}}$ objects defined in the restricted $\eta$ region have an inferior resolution and overall lower signal efficiency to that of the $E_T^{\text{miss}}$ object defined in the full accessible region if used at the same online thresholds. However, at extreme pileup the two alternative PUPPI-based $E_T^{\text{miss}}$ triggers are measured to have substantially lower rates compared to that of the standard Puppi$E_T^{\text{miss}}$. At the offline threshold of 128 GeV, used for Puppi$E_T^{\text{miss}}$ in the simplified menu (see Table 4.1) and the 50% efficiency online-to-offline scaling, the rates of PuppiNoHF$E_T^{\text{miss}}$ and PuppiL1Tracker$E_T^{\text{miss}}$ triggers are 30 (153) kHz and 21 (40) kHz for pileup 250 (300), respectively. For offline thresholds of 100 GeV and the same scaling the rates of the two triggers are 52 (120) kHz and 48 (76) kHz for pileup 200 (250), respectively.

The reasonable rates of the alternative $E_T^{\text{miss}}$ triggers at the nominal 128 GeV or lower offline thresholds make them viable alternative triggers for rate control in case of an above-nominal pileup in HL-LHC. Similarly, the triggers can be used for commissioning of the system in case of high rates due to potential mis-calibration of energy deposits in the HF or in the more forward parts of the HGCAL.

4.4.3 Menu rate and bunch-to-bunch pileup variations

The L1 menu rates shown previously assume that the number of pileup events follows a Poisson distribution of given mean $\mu$ (with $\mu = 140, 200, 250$ or 300), and that $\mu$ is identical for all the bunches in the beam. However, several effects may lead to a dispersion of $\mu$ across the bunches. Such bunch-to-bunch pileup variations cause an increase of the trigger rates, which have a non-linear behavior with respect to pileup. To evaluate this increase on the total menu rate, four different scenarios of the pileup variations across the bunches in the HL-LHC are
considered and modeled with the following distribution functions:

- uniform
- Gaussian truncated at 2σ
- Gaussian truncated at 3σ
- Gaussian truncated at 4σ

The uniform model is very conservative. It assumes a HL-LHC bunch to have an average pileup of any value in a given range with equal probability. It represents a “worst-case” scenario of bunch-to-bunch variation. The models with Gaussian distributions are motivated by observations of LHC bunches in Run-2 and are used as proxy to describe potential bunch-to-bunch variation scenarios for the HL-LHC.

The average pileup in each bunch is modeled to be distributed between $200 \times (1 - \Delta_{\text{max}})$ and $200 \times (1 + \Delta_{\text{max}})$, where $\Delta_{\text{max}}$ is increased from zero (no bunch-to-bunch variations) to 50%. The total rate of the simplified menu at a given fixed value of pileup is determined from the linear interpolations shown in Fig. 4.19 (left).

The simplified total menu rate assuming the nominal average pileup of 200 for given bunch-to-bunch variation model as a function of $\Delta_{\text{max}}$, is depicted in Fig. 4.19 (right).

For relative bunch-to-bunch variations as large as ±50%, and for the case of the conservative uniform model, the menu rate remains below 490 kHz. Going from larger to smaller bunch-to-bunch pileup spread, the three Gaussian models truncated at 2σ, 3σ and 4σ at the maximal value of variation predict the menu rate to remain below 420 kHz, 380 kHz, and 360 kHz, respectively.

Efficiency and resolution degradation inflicted by an above-nominal average pileup will be evaluated in the future. A more precise estimation of the impact of the increased pileup on the physics acceptance, as achieved by the L1 menu proposed here, would be required to derive...
the offline-to-online scaling for each pileup scenario separately. Moreover, should HL-LHC operate at an average pileup much larger than 200, the algorithms would be re-optimized for such conditions, improving performance further.
Chapter 5

Conceptual design of the Phase-2 L1 Trigger

5.1 Introduction

The architecture of the Phase-2 L1 trigger system described in this Chapter is the result of studies conducted to optimize processing resources and interconnections in order to reach a design with high flexibility and robustness. These studies were conducted considering the minimum requirements of this design in terms of algorithm features and associated performance in order to comply with the physics program described in Chapter 3. In addition, the full physics menu exercise performed in Chapter 4 defines the necessary physics object types and their complementarity. The system organization proposed enables a suitable flow of data and an interconnection scheme conforming with these choices and compliant with the latency budget.

This Chapter presents separately the different components of this trigger architecture as illustrated in the functional diagram shown in Fig. 1.3. A description of functionality, interface specifications, input/output bandwidth, as well as object size and format is provided for each component. In the case of the Global Trigger, this extends to interfaces with HLT, DAQ, TCDS and external triggers. The assessment of the algorithm firmware resources and latency, presented in Chapter 3, defines the provision of resources for each component with a sufficient number of boards and links. Where possible, the hardware requirements specified refer to the current R&D described in Chapter 6. For all components, descriptions of how contingency is realized and how a potential expansion of the system could be achieved are included. The supply of functional triggers for commissioning purposes is also considered. A proposal for a 40 MHz scouting system and its interfaces to the L1 trigger are specified. An overview of the complete architecture concludes this Chapter.

5.2 Calorimeter Trigger system

The CMS calorimeter is comprised of four sub-detector elements to provide pseudorapidity coverage up to $|\eta| = 5$. The central region is instrumented with the barrel electromagnetic (ECAL) and hadronic (HCAL) calorimeters, while the forward region is covered by the endcap high-granularity calorimeter (HGCAL) and the hadron foward calorimeter (HF). To optimally utilize these detectors for selecting events, the information from all of them should be combined to produce similar trigger performance over the whole $\eta$ range. The main calorimeter trigger objects are: electrons, photons, jets, hadronically decaying taus, as well as various energy sums. The calorimeter trigger system receives trigger primitives from the backend electronics of the four calorimeter sub-detectors and is presented schematically in Fig. 5.1. The calorimeter trigger must process the data from the backend electronics and send it to the correlator trigger, as well as calculate the calorimeter trigger quantities for the whole calorimeter sending that output to the global trigger (GT). The trigger system includes a regional calorimeter trigger (RCT)...
Chapter 5. Conceptual design of the Phase-2 L1 Trigger

that combines information from ECAL and HCAL. The output of the RCT is sent to the global calorimeter trigger (GCT) that serves two tasks:

- prepare and send the HF and RCT information to the correlator trigger (CT);
- combine the ECAL, HCAL and HF with the HGCAL for calorimeter-only-based triggers to be sent to GT.

After the HF and RCT information is sent to the Correlator Trigger, the GCT combines all calorimeter information, calculates the calorimeter objects and sends them to GT. This must be done efficiently in order to stay within the allocated latency budget, which corresponds to $5 \mu s$ from the bunch crossing for providing outputs to CT, and $9 \mu s$ for providing outputs to GT.

<table>
<thead>
<tr>
<th>System</th>
<th>Fibers</th>
<th>Content</th>
<th>Bits/TMUX</th>
<th>Total number of bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endcap(HGCAL)</td>
<td>216×2</td>
<td>Clusters/Towers</td>
<td>352/18</td>
<td>352×18×432</td>
</tr>
<tr>
<td>HF</td>
<td>18×2</td>
<td>Towers</td>
<td>352/1</td>
<td>352×1×36</td>
</tr>
<tr>
<td>ECAL</td>
<td>3060</td>
<td>Crystals</td>
<td>352/1</td>
<td>352×1×3060</td>
</tr>
<tr>
<td>HCAL</td>
<td>144</td>
<td>Towers</td>
<td>352/1</td>
<td>352×1×144</td>
</tr>
<tr>
<td>RCT</td>
<td>144</td>
<td>Clusters/Towers</td>
<td>352/1</td>
<td>352×1×144</td>
</tr>
<tr>
<td>GCT to Correlator</td>
<td>180</td>
<td>Clusters/Towers</td>
<td>352/6</td>
<td>352×6×(180/6)</td>
</tr>
<tr>
<td>GCT to GT</td>
<td>6</td>
<td>Calorimeter objects</td>
<td>520/1</td>
<td>520×1×6</td>
</tr>
</tbody>
</table>

The data content, the number of fibers, and the amount of information transferred from each element of the system are summarized in Table 5.1. Each calorimeter system sends their trigger primitives via 16 Gb/s links to the trigger system (upper rows in Fig. 5.1). At 16 Gb/s, a single fiber has the capacity to transmit 352 bits/BX of information. In the case of the barrel and HF, the cards can also operate at 25 Gb/s, thus increasing the bandwidth for each single fiber to 520 bits/BX. The endcap information provided by HGCAL are transmitted within $5 \mu s$ of latency, and contains clusters and towers as summarized in Table 5.1. HGCAL sends its information time-multiplexed with TMUX=18 simultaneously to the Correlator Trigger and to the GCT. The HF information contains only towers and is sent to the GCT at TMUX=1, where it is calibrated, time-multiplexed to TMUX=6, and sent to the correlator. It is also used in the GCT to compute calorimeter objects. The ECAL information is sent for each single crystal to the RCT at TMUX=1, where it is combined with the HCAL tower information. The RCT performs clustering of $e/\gamma$ objects and an initial calibration. The output of the RCT is sent to the GCT at TMUX=1 where it is time-multiplexed and sent to the Correlator Trigger at TMUX=6.

The details of the barrel part of the system are shown in Fig. 5.2. The small squares represent the single HCAL towers, the corresponding ECAL regions contain $5 \times 5$ ECAL crystals. The geometrical coverage of the ECAL and HCAL backend cards is also indicated. The information from these cards is sent to the RCT. One RCT card, shown in the same figure, covers $17 \times 4$ towers in $\eta \times \phi$ and receives 85 ECAL links with crystal level information and 4 HCAL links with towers, yielding 89 input links for a single RCT card. This number of links is provided by the APx card (see Section 6.3.1) that has up to 100 inputs and outputs. There are 36 RCT cards needed to cover the entire barrel calorimeter surface. In the RCT cards, the ECAL crystals are combined into $e/\gamma$ clusters. The remaining unclustered ECAL crystals are combined with the HCAL to build trigger towers. The RCT card sends the 12 highest $p_T$ $e/\gamma$ clusters and unclustered ECAL+HCAL energy for each tower to the barrel part of the GCT via four links as summarized in Table 5.1. The information in one link contains 17 towers $\times$ 16 bits (10 bits $E_T$, 3 bits H/E, 3 spare) and 3 clusters $\times$ 24 bits (10 bits $E_T$, 12 bits position, 1 bit shape, 1 bit spare). The tower position is encoded systematically by the link and bit position within the
5.2. Calorimeter Trigger system

Figure 5.1: Calorimeter trigger architecture. The upper part of the plot presents the calorimeter systems. Barrel ECAL and HCAL are combined in the regional calorimeter trigger (RCT). The information from RCT and HF is processed in the global calorimeter trigger (GCT) and is sent to the Correlator Trigger. The endcap (HGCAL) information is sent directly to the correlator and is duplicated to GCT. After the calorimeter information is sent to the Correlator Trigger, GCT combines all the calorimeter information, produces calorimeter-based trigger quantities and sends them to the global trigger (GT). The left side of the plot shows the approximate time required for all processes. Information should be delivered to the Correlator Trigger within $5\mu s$ and to GT within $9\mu s$ after BX.
Figure 5.2: Diagram of the barrel calorimeter trigger geometry; Small squares represent single HCAL towers and correspond with $5 \times 5$ ECAL crystals. There are 72 towers in $\phi$ and 34 towers in $\eta$. The ECAL and HCAL backend card geometrical regions are shown in dark lines. Each ECAL card covers $3 \times 4$ towers, 25 crystals each, with some cards covering only $2 \times 4$ towers. HCAL cards cover $16 \times 4$ towers each. The sizes of the RCT and GCT cards are also shown. The vertical regions of different colors represent sizes of calorimeter supermodules (see Section 2.2.1).
data packet that is transferred to the GCT.

There are three barrel GCT cards, each receiving information from 12 unique RCT cards, thus subdividing the whole barrel region into three regions in \( \phi \). A small overlap in information from RCT is required to provide independent calculations on each GCT card; it corresponds to the coverage of one RCT card in \( \phi \), as shown in Fig. 5.2. The total inputs to one GCT card originate from the output of \( 2 \times 6 \) unique RCT cards, and 2 more cards on each \( \phi \) side to account for overlap. Each GCT card receives 12 clusters from each RCT card and, when needed, combines them at the borders of the RCT cards. Of the 12 clusters, only the 8 highest-\( p_T \) clusters within the geometrical acceptance of one RCT card are considered. Clusters that are ignored are added to the tower information. The clusters and towers are then sent via 48 links to the Correlator Trigger system.

![Diagram of GCT architecture](image)

Figure 5.3: Global calorimeter trigger architecture. Each circle represents three GCT cards that cover a certain \( \eta \) area for the whole \( \phi \) region. The input and output information for each card is specified at the top part of the diagram. The colored arrows (red and yellow) demonstrate additional data flow between GCT cards of different \( \eta \) regions and the same \( \phi \), while open arrows demonstrate the data flow between cards of different \( \phi \) regions for the same \( \eta \). The numbers indicate the number of links needed to transfer the information. The bottom part shows the GCT SUM card, where information from single GCT cards is combined and sent to GT.

The full GCT architecture is shown in Fig. 5.3. It consists of three groups of cards, one for the positive \( \eta \) side of the HGCAL, one for barrel, one for the negative \( \eta \) side of the HGCAL, and a sum card. The barrel GCT cards process RCT information and send data to the correlator within 4\( \mu s \) after bunch crossing. At the same time each HGCAL GCT card receives information from HF and sends it to the correlator after processing and within 5\( \mu s \) after the bunch crossing. At this point, both the barrel and HF information is stored on the GCT cards and can be combined with the HGCAL information to build the calorimeter-based trigger objects. The GCT will also run the PF and PUPPI forward algorithms (see Section 5.5). The HGCAL backend electronics start to deliver information to the GCT HGCAL cards at 4\( \mu s \); the complete information is sent within 1\( \mu s \) at TMUX=18. The input information is demultiplexed upon reception. To build calorimeter objects, one needs to overlap the information in both \( \eta \) and \( \phi \) directions between different calorimeter systems. Therefore, part of the information corresponding to the four
trigger towers in $\eta$ for the whole region in $\phi$ at the border between barrel and HGCAL, is shared between cards. This requires six links as shown in Fig. 5.3. The $\phi$ overlap is taken into account by the RCT for barrel, but for the HGCAL an additional two links are required to share four towers in $\phi$. After all information is shared between the HGCAL and barrel cards, each card processes data independently and sends the 12 highest-$p_T$ objects for each of $e/\gamma$, jets, hadronic taus, and 9 partial sums to the GCT sum card. The GCT sum card selects 6 highest-$p_T$ objects in each category for positive and negative $\eta$, calculates total sums, $H_T$, $E^{\text{miss}}_T$, etc., and sends them via 6 links at 25 Gb/s to GT. The maximum number of links for a single card in the GCT system is 90.

The system can be easily scaled in case it is needed for stable operation. If the amount of information to be exchanged between the GCT cards needs to be increased, it can be done by increasing the link speed from 16 to 25 Gb/s without changing the configuration. The same is true for the HF and RCT parts of the system. Thus the data volume for these parts can be increased by 50%. All GCT algorithms run in a single GCT card. Thus, if the system needs to be extended, it can be done by running GCT at TMUX=2. In this case, a single GCT card will work with every second event. This will imply a duplication of the HGCAL and barrel GCT cards and require 9 additional cards, and optical links between GCT and the Correlator Trigger, and changing the RCT and HF outputs to TMUX=2, while keeping the number of optical links the same.

### 5.3 Muon Trigger system

The function of the muon trigger is to identify muon tracks in the experiment and measure their momenta and other parameters for use in the global trigger menu. In addition to the muon detector upgrades that include improved electronics and new sub-detectors, the presence of a L1 track finder in CMS will bring some of the offline muon reconstruction capability to the L1 trigger, delivering unprecedented reconstruction and identification performance.

The muon trigger will provide redundancy by identifying and measuring both standalone muons, reconstructed solely within the muon detector systems as with the previous Phase-1 trigger system, and combined muons by matching muon information to tracks from the L1 track finder. While the combination with the tracker improves performance, the standalone reconstruction provides new capabilities, such as triggering on muon-like particles from long-lived particle decays displaced from the collision vertex, as well as on heavy stable charged particles (HSCP) that reach the muon system with velocity $\beta < 1$.

The inputs to the muon trigger system arrive in the form of muon stubs (32-64 bits each) containing position information from the electronics associated with all muon detector technologies: drift tubes (DT) in the barrel, resistive plate chambers (RPC) in the barrel and endcaps including the very forward extension (iRPC), cathode strip chambers (CSC) in the endcaps, and gaseous electron multiplier (GEM) detectors in the endcaps. As with the Phase-1 system, these inputs will be combined prior to the track-building step in order to increase redundancy and improve efficiency. Because of the large number of tracks from the L1 track finder and the large size of the track object (96 bits), L1 tracks arrive time-multiplexed with a period of 18. In contrast, because of the low data volume from the muon detectors, muon stub data arrives without time-multiplexing. Detectors with upgraded backend electronics (e.g., the barrel muon systems) can transmit information either way based on the requirements of the muon trigger.

Because of the different technologies and detector conditions in each sub-detector region, the
reconstruction of standalone muons will be segmented into three regions in $\eta$ as done in the Phase-1 system: a barrel muon track finder (BMTF), an overlap muon track finder (OMTF), and an endcap muon track finder (EMTF). The output of these systems will be sent to a time-multiplexed Global Muon Trigger (GMT) system, which will also receive tracks from the L1 track finder system. The BMTF results of Section 6.4.2.2 demonstrate that the algorithm firmware is small and can be consolidated onto the same boards as the GMT. It does not need to be implemented as a separate physical system like OMTF and EMTF, thus reducing the latency and hardware required.

As with offline muon reconstruction, the GMT will reconstruct two types of muons: global muons matched from standalone muons to tracks from the tracker and tracker muons matched from individual muon stubs to the propagated track parameters. Tracker muon reconstruction has demonstrated improved efficiency (especially in the gaps between muon detectors), therefore, the OMTF and EMTF will transmit trigger primitives along with standalone tracks to the GMT. A diagram of the muon trigger architecture is shown in Fig. 5.4.

Figure 5.4: Diagram illustrating the components comprising the muon trigger architecture and their connections to the detector inputs and subsequent L1 trigger processing layers. Arrows with solid lines illustrate the data flow, while arrows with dashed lines show optional data flow connections under consideration. Approximate latency is shown along the vertical axis. The GTT is the Global Track Trigger discussed in Section 5.4.

### 5.3.1 Endcap Muon Track Finder

The EMTF receives trigger primitive data from the endcap muon systems (without any time-multiplexing), in the region $1.2 < |\eta| < 2.4$. Its function is to find standalone muons (including displaced, slow, and other signatures) in the endcaps and assign the track parameters. The processing is partitioned into six regions in azimuth ($\phi$) per endcap denoted “sectors” (12 proces-
sors in total). To efficiently cover the boundaries between sectors, the input optical fibers from the muon systems are split from a 10–20° detector region near each boundary and connected also to the neighboring processor on the other side of the boundary (about 20% of fibers).\(^1\) The fiber input to each processor is summarized in Table 5.2, where the total fiber input per board is expected to be 75 links. In an alternate configuration, under study, the fiber inputs for the inner ring of the first two stations of CSCs (ME1/1 and ME2/1 chambers) would come directly from optical Trigger Mother Boards, bypassing the Muon Port Cards. This has two advantages: it makes the transmission of CSC data over the 8 MPC fibers more robust during operations (data from one of 9 CSCs does not need to be shared across the 8 available MPC fibers), and more than 2 cathode strip patterns could be transmitted per CSC for the inner ring chambers, making the system more robust against pileup at high \(\eta\). The fiber count increases by 3 in this case, to a total of 78 fibers.

For the RPC and iRPC links, some concentration of fibers upstream of the EMTF is assumed either in the backend electronics directly or in a small set of dedicated boards. Furthermore, while board designs exist that are able to accommodate up to 100 optical link inputs as required by the EMTF (e.g. Section 6.3.1), we are also considering the concentration of the lowest bandwidth links from the CSC system into a reduced number of higher bandwidth links. A dedicated system using inexpensive FPGAs (each taking approximately 40 low-speed input links) could provide this concentration, as well as data reformatting into universal tracking coordinates with position alignment corrections, for a modest latency increase (≈0.25 \(\mu\)s). This could be realized in an architecture with 6 (12) boards with 2 (1) FPGAs each. The possibility of extending this layer to receive also hits from the GEM muon system, such that they can be re-formatted and combined with CSC segments into super-primitives, remains an option as well and would offload some of the preprocessing work done by the EMTF. Additionally, the precisely measured angle between the GEM and CSC hits would improve the measurement of the standalone muon \(p_T\). Finally, the number of optical links per EMTF processor could be reduced by approximately 30–40 links from that shown in Table 5.2 for concentrated link bandwidths of 16–25 Gb/s.

The output of each processor, including the identified muons and all non-zero trigger primitives, will be time-multiplexed to the GMT with an anticipated multiplex factor of 18 using a 25 Gb/s fiber bandwidth. The data sent will cover a several BX wide time interval around the central time period in order to facilitate commissioning, operations, and the search for slow HSCPs. Given the effective bandwidth available by time-multiplexing the event data, all output data can be concentrated onto one link per target destination even for time-multiplexing periods as short as 3.

A notable feature of each EMTF board is that it will contain an external bank of memory (≈128 GB) for use as a large look-up table for a flexible and low-latency momentum assignment calculation (see Section 6.3.1); this approach has successfully been used for the complex momentum assignment in the endcap for the Phase-1 and earlier endcap muon trigger systems. Moreover, large neural network inferences also can be implemented into the FPGA for complementary capability along these lines (see Section 3.3.4).

### 5.3.2 Overlap Muon Track Finder

The OMTF receives trigger primitive data from regions of both the barrel and the endcap muon systems (without any time-multiplexing) covering \(0.8 < |\eta| < 1.2\). Its function is to find stan-

\(^1\)A given processor only needs the neighbor inputs from one of its sector boundaries for the entire system to seamlessly cover tracking over \(2\pi\) radians in \(\phi\).
Table 5.2: Summary of the optical link inputs from each muon system to each EMTF sector processor. The link count after the “+” denotes the split fibers from the neighboring sector. The links from the inner CSCs (ME1/1 and ME2/1) could come directly from the frontend, leading to the numbers in parentheses.

<table>
<thead>
<tr>
<th>System</th>
<th>Link Bandwidth (Gb/s)</th>
<th>Link Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSC</td>
<td>3.2</td>
<td>40+9 (43+9)</td>
</tr>
<tr>
<td>RPC</td>
<td>25</td>
<td>4+1</td>
</tr>
<tr>
<td>tRPC</td>
<td>16</td>
<td>2+1</td>
</tr>
<tr>
<td>GEM ME0</td>
<td>25</td>
<td>2+1</td>
</tr>
<tr>
<td>GEM GE1/1</td>
<td>10</td>
<td>8+2</td>
</tr>
<tr>
<td>GEM GE2/1</td>
<td>25</td>
<td>4+1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>75 (78)</td>
</tr>
</tbody>
</table>

dalone muons (including displaced, slow, and other signatures) in this region and assign the track parameters. Processing is partitioned into three sectors in azimuth on each side of the barrel (6 processors in total). To efficiently cover the boundaries between sectors and between the BMTF and EMTF track finders, the input optical fibers from the muon systems will be either passively or actively split. The fiber input to each processor is summarized in Table 5.3, where the total fiber input per board is expected to be 88 links when the RPC endcap concentration is assumed. As with the EMTF, the number of input links can be further reduced with a dedicated CSC data concentration system. The output of each processor, including the identified muons and all non-zero trigger primitives, will be sent time-multiplexed to the GMT with an anticipated period of 18 at a 25 Gb/s fiber bandwidth.

Table 5.3: Summary of the optical link inputs from each muon system to each OMTF sector processor. The link count after the “+” denotes the split fibers from the neighboring sector.

<table>
<thead>
<tr>
<th>System</th>
<th>Link Bandwidth (Gb/s)</th>
<th>Link Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barrel (DT+RPC)</td>
<td>25</td>
<td>12+3</td>
</tr>
<tr>
<td>CSC</td>
<td>3.2</td>
<td>60+5</td>
</tr>
<tr>
<td>Endcap RPC</td>
<td>25</td>
<td>7+1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>88</td>
</tr>
</tbody>
</table>

5.3.3 Barrel Muon Track Finder

The BMTF receives and processes trigger primitive data from the barrel muon systems in the region $|\eta| < 0.8$. The barrel DT and RPC hits will already have been combined by the last layer of the barrel muon electronics into “super-primitives”. The BMTF logic will reside in the GMT boards described in the next section. The input to BMTF comes from Layer-1 of the barrel muon systems, which is segmented into 60 sectors. For the configuration option where these 60 sectors are mapped onto 60 boards with a single output fiber to each target, the number of links to each GMT board will need to be reduced. Several reduction options are considered including using 30 boards with two FPGAs (with a fiber output from one FPGA) or data concentration using a dedicated system of 3–4 boards. If the 60 barrel sectors are mapped onto 30 output fibers to the BMTF target, then this concentration layer is not strictly necessary. The output link bandwidth is 16 Gb/s or possibly 25 Gb/s in these schemes.

5.3.4 Global Muon Trigger

The GMT receives trigger primitive data from the barrel, standalone muons and stubs from EMTF and OMTF, and tracking data from the L1 track finder all with a time-multiplexing pe-
iod of 18 that is aligned with the track finder period. The L1 track finder data are sent from 9 separate azimuthal sector processors for each time sample. In one possible consolidation option, the L1 tracks would be sent to the GMT after first being consolidated by the Global Track Trigger (GTT) system (see Section 5.4), which would reduce the number of optical links by a factor of 2 for a modest increase in latency ($\approx 0.25$–$0.5 \mu s$).

In order to maintain high efficiency, facilitate commissioning and operations where detector timing could be off, and maintain acceptance for a slow HSCP, the muons and stubs will be transmitted within a few BX time window around the central time period. One of the functions of the GMT is to find standalone muon tracks (including displaced, slow, and other signatures) in the barrel and assign track parameters using the Kalman Filter algorithm. The GMT then must cancel duplicate muons found by multiple track finder processors at their boundaries. These operations do not affect the overall latency budget since they are performed in the shadow of the L1 track finder processing.

The GMT then receives tracks streamed through 18 optical links and propagates them to the muon system to reconstruct global or tracker muons. The compatibility of the propagated track parameters with nearby muon stubs or tracks is tested to tag global muons. For both cases, the muon momentum is assigned using the tracker measurement, which has better resolution. For displaced muon candidates, a track veto can be applied to reject candidates coming from the primary vertex. Additional functionality considered for the GMT include identifying isolated muons using the tracks in a cone around each muon, and finding unique topologies such as the $\tau \rightarrow 3\mu$ lepton-flavor violating process. The output of the GMT is sent to the Correlator Trigger system.

Conceptually the GMT will process each time-multiplexing period independently of the others, which typically corresponds to 18 processors. The number of optical link inputs coming from the BMT Layer-1, OMTF, EMTF, and the L1 track finder system are summarized in Table 5.4. The L1 track finder data arrive on 18 links, 2 from each of 9 processors for a given time slice. This would be 9 links if the data came via the GTT system. The number of the input links for a time-multiplexed processor is 66, when the BMT Layer-1 sends its data on 30 links. Based on the firmware implementation results of the candidate algorithms, a cost effective FPGA (Xilinx KU15P or equivalent) may have enough resources and bandwidth to satisfy the requirements of the system in this scheme. It might be possible to pack more than one time slice into a board using a larger FPGA or multiple FPGAs, assuming the number of input links also fits the board, further reducing the system size and cost.

Table 5.4: Summary of the optical link inputs from each muon track finder system, the barrel Layer-1 system, and the L1 track finder to each GMT processor for each time sample (of 18). The range for the barrel muon system depends on whether the links come from an intermediate preconcentrator or directly from 30 boards. The link count would be reduced by 9 if tracking data were sent via the GTT.

<table>
<thead>
<tr>
<th>System</th>
<th>Link Bandwidth (Gb/s)</th>
<th>Link Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMTF</td>
<td>25</td>
<td>12</td>
</tr>
<tr>
<td>OMTF</td>
<td>25</td>
<td>6</td>
</tr>
<tr>
<td>Barrel</td>
<td>16–25</td>
<td>30–18</td>
</tr>
<tr>
<td>TT (GTT)</td>
<td>25</td>
<td>18 (9)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>66–54 (57–45)</strong></td>
</tr>
</tbody>
</table>

The output data of each GMT processor are sent to two destinations. One copy goes to the Correlator Trigger described in Section 5.5. Another copy is sent to a demultiplexing card,
5.4 Global Track Trigger system

Among the novel features of the Phase-2 upgrade of the CMS Level-1 trigger system is the ability to reconstruct charged particle tracks within the full outer silicon tracker volume at the 40 MHz collision rate. The role of the Global Track Trigger (GTT) is to build high-level objects out of the tracks. As discussed in Section 3.6, hadronic triggers such as $H_T$ and $E_{\text{miss}}$ benefit significantly from the rejection of pileup tracks possible with the precision of the silicon tracker. As discussed in Section 3.4, a precise primary vertex position can be found from the tracks to aid in the mitigation of the pileup effects in the particle-flow reconstruction of the event as well. This section describes an architecture for the GTT that achieves those objectives.

The baseline architecture for the GTT consists of 12 cards, each hosting a single Xilinx VU9P FPGA. The cards are arranged as a time-multiplexed system with a period of 6, with two cards processing each event. Each card receives the full set of tracks from the track finder and sends identified stand-alone track objects to the Global Trigger (GT). One card for each event will forward the identified primary vertex (or potentially vertices) to the Layer-1 Correlator Trigger. A schematic drawing of this architecture is shown in Fig. 5.5. An alternative arrangement of the 12 cards is possible, when one of the two cards processing an event receives the tracks and forwards them to the second card, which in turn passes the output objects to the GT. This option minimizes the interfaces with the track finder and GT, but adds inter-dependencies among the 12 GTT cards. An option also exists for one card for each event to aggregate the tracks and pass them to the Global Muon Trigger (GMT).

Figure 5.5: Architecture of the Global Track Trigger system. Each board receives a copy of all of the tracks from three time slices from the track finding system and delivers the primary vertex and track stand-alone object information to the correlator and global trigger systems, respectively. The latency estimated for this architecture is discussed in Section 6.4.3.

which collects the data from each time sample for consolidation and transfer to the Global Trigger, as described in Section 5.6. This demultiplexing card could also be used to transmit muon data to any detector sub-system (triggering or not) as may be needed for calibration, etc.
As described in Section 2.1, the L1 track finder is divided into 9 sectors in $\phi$ and 18 time slices. Each of the sectors can deliver all of the tracks for an event on two 25 Gb/s fibers over the 18-bx time period. Therefore, all of the tracks in an event can be delivered to a GTT board in 18 fibers. The GTT operates at TMUX = 6, while the track finder operates at TMUX = 18, and thus each GTT board must handle three time slices from the track finder. The result is a total of 54 fibers from the track finder system into each GTT card. This is possible with 25 Gb/s links.

The processing resources on the GTT cards must be sufficient to identify the primary vertices and to reconstruct the track stand-alone (SA) objects, including track jets, track $E_T^{miss}$, and single isolated tracks. The hardware demonstration studies described in Section 6.4.3 show that these algorithms are an appropriate fit on a single VU9P FPGA occupying approximately 70% of the chip with reasonable latency and timing.

Beyond these baseline requirements, many other possibilities exist for the GTT algorithms. For example, Section 4.3.3 describes the potential for using tracks to reconstruct light resonances such as $\phi$ mesons; light resonance reconstruction was never previously possible in the CMS Level-1 trigger. Potential applications include rare Higgs decays and $CP$-violation studies in the flavor sector. Additionally, the availability of tracks displaced from the primary vertex is under study as an extension to the track finding system. Such tracks may be indicative of new long-lived physics signatures and would open new trigger opportunities, such as displaced jets and displaced isolated tracks. We extrapolate the performance of the demonstrated algorithms in Section 6.4.3 to conclude that a second VU9P FPGA per event would allow for the inclusion of these exciting physics opportunities in the GTT to be pursued, thus motivating the choice of a 12 card system.

The design with two boards per event allows for the separation of the logic for the various algorithms to be divided among the two boards. The final splitting of responsibility between the two boards can be decided based on how to most efficiently use the logic resources. Copies of the same firmware for the algorithm reconstruction can be used for each time slice.

Each of the 12 GTT boards has one 25 Gb/s link to the GT. This link has sufficient bandwidth to carry multiple collections of objects from the GTT to the GT, as detailed in Section 5.6. One set of 6 GTT boards will also be responsible for delivering the primary vertex information to the Correlator Layer-1 system (particle-flow). Since the vertex object is small and only a few (at most) vertices will be passed, the bandwidth on one link is more than sufficient to transmit the vertex information. The Correlator Layer-1 system is divided into 6 slices in $\eta$ and 6 slices in time, as described in Section 3.5. Each Correlator Layer-1 board needs the primary vertex, and thus each GTT board will deliver the vertex to all 6 slices in $\eta$. As is typical of a trigger system, the bandwidth into the GTT (54 links) is much larger than the bandwidth out (up to 7 links), and thus there is significant spare output capacity available in the design for potential data scouting at 40 MHz as described in Section 5.7.

The Global Track Trigger is thus able to receive all of the reconstructed tracks from each event and compute track-based stand-alone triggerable objects, such as $H_T$, $E_T^{miss}$, and the primary vertex. The results are passed downstream to the Correlator Layer-1 and to the global trigger.

### 5.5 Correlator Trigger system

In this Section, we describe the system architecture of the Correlator Trigger (CT). The primary function of the CT is to aggregate inputs from all upstream systems and optimally combine the information from the various sub-systems to achieve the best possible trigger performance.
for the most challenging physics topologies. The CT is at the nexus of the trigger system and both its upstream inputs and downstream systems are depicted in Fig. 5.6. The CT is shown in red boxes and the primary path through the CT system is shown with blue arrows while other paths to the Global Trigger are shown in gray dotted lines.

Figure 5.6: Correlator Trigger architecture: the Correlator Trigger is shown here in red and split into its two functional layers. The first layer focuses on building particle-flow candidates and passes those to the second layer which builds physics objects. All the upstream input systems and their TMUX factors are shown, as well as the downstream Global Trigger system.

The architecture of the CT consists of two layers. Section 5.5.1 describes the Correlator Trigger Layer-1 (CTL1), which creates particle-flow candidates which can be filtered based on a primary vertex ansatz (PUPPI candidates). Those candidates are passed to a second stage of FPGAs, Correlator Trigger Layer-2 (CTL2), which reconstructs physics objects, e.g. electrons, jets, hadronic taus, and $E_T^{miss}$. The CTL2 is described in Section 5.5.2.

**5.5.1 Correlator Trigger Layer-1: particle-flow candidates**

The particle-flow algorithm optimally combines the tracker, calorimeter, and muon system information into particle-flow candidates. The PUPPI algorithm gives a probability of those particle-flow candidates to originate from a given primary vertex. The primary vertex is provided to the CT by the GTT system. As indicated in Fig. 5.6, the tracker, muon, and calorimeter sub-systems are inputs to the CT. We summarize the inputs to the CT in Table 5.5. In this table, we show the upstream sub-system inputs in terms of their TMUX factor and the regionalization ($\eta \times \phi$) in which that sub-system processes their data on an FPGA. This defines the origin of the detector data and the time interval in bunch crossings (BX) for which the data arrive. We also present the payload in bits per event which approximately defines how much data are being sent and thus how many optical links are needed in the CT to accept that payload. We assume for the architecture that each optical link sends data at a rate of 16 or 25 Gb/s. Given these input interface definitions, a CTL1 architecture can be defined. The full description of the particle-flow trigger algorithm performance and implementation in
Table 5.5: Summary of the data flow in the Correlator Trigger system

<table>
<thead>
<tr>
<th>System</th>
<th>TMUX (in BX)</th>
<th>bits/event</th>
<th>$\eta \times \phi$ regions</th>
<th>Links/FPGA (Tx)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tracker</td>
<td>18</td>
<td>81k</td>
<td>$1 \times 9$</td>
<td>2</td>
</tr>
<tr>
<td>HGCAL</td>
<td>18</td>
<td>120k</td>
<td>$2 \times 3$</td>
<td>4</td>
</tr>
<tr>
<td>HF</td>
<td>6</td>
<td>14k</td>
<td>$2 \times 1$</td>
<td>1</td>
</tr>
<tr>
<td>Barrel Calorimeter</td>
<td>6</td>
<td>60k</td>
<td>$1 \times 3$</td>
<td>6</td>
</tr>
<tr>
<td>Muon</td>
<td>18</td>
<td>6k</td>
<td>$1 \times 1$</td>
<td>1</td>
</tr>
<tr>
<td>GTT (vertices)</td>
<td>6</td>
<td>1k</td>
<td>$1 \times 1$</td>
<td>1</td>
</tr>
</tbody>
</table>

firmware is given in Section 3.5. The algorithm will output particle-flow candidates of the types: electron, photon, muon, charged hadron, neutral hadron. Each candidate will also have a probability of the particle being from pileup based on the PUPPI algorithm. To recall, the particle-flow algorithm is an inherently local association algorithm ($\Delta R$ comparisons) and processes its input data in $\eta \times \phi$ regions that are roughly $0.5 \times 0.7$. Each particle-flow region includes an overlap buffer region of $0.25 \eta$ and $\phi$ units to find particle candidates on the boundary. The particle-flow region size is optimized to not be too small, such that the region overlaps dominate the reconstruction area, and to not be too large, such that we waste FPGA resources doing associations of objects that are far apart from each other. This is described in more detail in Section 3.5.5, where we show the resource usage for a range of different region sizes and the number of inputs into that region. In Fig. 5.7, we show the geometrical regions corresponding to the input regions and the particle-flow processing regions. One important feature of this design is that the particle-flow regions in the Forward Calorimeter ($|\eta| = 3–5$) will be processed within the Global Calorimeter Trigger FPGAs and will be passed through the CTL1 FPGAs directly to CTL2. Because there is no track linking in the forward region, the particle-flow algorithm is a null algorithm and the PUPPI forward algorithm will be performed in the GCT.

Figure 5.7: Correlator Trigger input regions are shown for the calorimeters and trackers in red, blue, green, and magenta. The particle-flow processing regions are shown in black including the overlap areas ($0.5$ in $\eta$).
5.5. Correlator Trigger system

Based on the resources needed to process each of the particle-flow regions, the CTL1 trigger is designed for 36 FPGAs. We use the highest performing FPGAs (VU9P) envisioned for the system due to both the high bandwidth and processing requirements of the CT system. This optimization is based on system requirements limiting the total algorithm firmware to below 50% the total allowed FPGA resources. From Section 3.5.5, we find that the particle-flow + PUPPI algorithm will be approximately 35–40% of the FPGA resources including the regionizer (data flow organization) firmware. This provisions some extra resources for additional algorithm firmware for related algorithms that may require some more processing resources, such as an optimized particle-flow electron algorithm or dedicated track + e/\gamma taus.

To fully define the CTL1 architecture, we define the TMUX factor and the CTL1 input slices. This will define how many regions and in what spatial layout each FPGA in CTL1 will process the particle-flow regions, as well as its TMUX factor. Based on the input subsystem TMUX factors, we choose a TMUX of 6 bunch crossings as a trade-off of data organization and latency. This then divides the detector into 6 data organization slices which receive the data from the upstream systems. Within the data organization slices, we further route the data into the particle-flow processing regions. The 6 data organization regions are shown in Fig. 5.8.

![Figure 5.8: Definition of the particle-flow Layer-1 regions.](image)

The data organization regions are split geometrically into 6 slices in $\eta$. We choose $\eta$ regions because $\phi$ regionalization requires an accounting for overlaps in upstream systems that are $\phi$ regionalized and, thus, demands more requirements on the processing resources of the upstream systems. The 6 $\eta$ regions in Fig. 5.8 are chosen to align with different particle-flow processing requirements based on detector boundaries. Regions 2, 3, and 4 cover the central detector regions, which require association with the barrel calorimeter and the tracker. Regions 1 and 5 cover the overlap between HGCAL and tracker. Region 6 does not include tracker.
information anymore and works only on the HGCAL. Reconstruction of the forward region beyond the HGCAL, HF, is actually performed upstream in the GCT as it does not depend on any other sub-system and the latency of the HF trigger is quite low.

Given this alignment, the number of input links needed are:

- Region 3: 54 (TF) + 18 (GCT, barrel) + 3 (GMT) + 1 (GTT) = 76 links
- Region 2, 4: 27 (TF) + 18 (GCT, barrel) + 3 (GMT) + 1 (GTT) = 49 links
- Region 1, 5: 27 (TF) + 36 (GCT, endcap) + 3 (GMT) + 1 (GTT) = 67 links
- Region 6: 36 (GCT, endcap) x 2 + 3 (GMT) + 1 (GTT) = 76 links

This number of input links fits comfortably within the 96 allowed links available in the VU9P FPGA that will be used by the CT.

The number of output links per FPGA are determined by the total output payload of the CTL1. Each particle-flow candidate is defined by 64 bits and the total payload is 60k bits (20k for PUPPI candidates and 40k for PF Candidates). To be delivered over 6 BX, this requires 3 Tx links out of the CTL1 to the CTL2 trigger per FPGA. These outputs can be cloned so as to send them to several different CTL2 boards performing different algorithms. With a total of 6 different regions going into the CTL2, CTL2 receives a total of 18 links corresponding to the 3 outputs for each CTL1 board.

**5.5.2 Correlator Trigger Layer-2: object reconstruction**

The Correlator Layer-2 uses the particle-flow candidates identified at Layer-1 to reconstruct physics objects: jets, hadronic taus, electrons, photons, and global energy sums. These objects are then sent to the Global Trigger for use in the menu.

The Correlator Layer-2 comprises 30 cards, each housing a single Xilinx VU9P FPGA. The default TMUX factor is chosen to be 6, the same as Layer-1. Hence, a single event is processed by 5 cards. As detailed below, the full list of PF candidates from the entire detector can be received by a single card at TMUX 6 and no division of processing by region required. This is clearly important in the case of global energy sums and seamless jet finding across the full detector. The processing work is therefore divided between cards by algorithm:

1. Jets
2. Energy sums
3. Taus
4. Electrons & photons (w/ and w/o isolation)
5. PF muons (w/ and w/o isolation)

Data are transferred from Correlator Layer-1 to Layer-2 on 25 Gb/s links. Sufficient bandwidth is allocated to ensure the probability of event truncation is less than $10^{-4}$. Each Layer-1 board sends up to 162 PF or PUPPI candidates, on 3 links, to each Layer-2 board. Each Layer-2 board therefore receives 18 links from Layer-1. Up to 3 links are provisioned for reception of HF PF candidates from the GCT, giving a total of 21 input links per Layer-2 board. To receive a whole event in the current TMUX architecture, at most 3 local regions per clock are required. This means that as many as 32 candidates per region can be sent into a single board. We expect roughly 20 PUPPI candidates per region, well within the tolerance of the system. Significant
headroom is therefore available for expansion, which may for example include transmission of tracks above some $p_T$ threshold for use in lepton isolation or tau algorithms.

The CT output data are sent to the GT on thirty 25 Gb/s links; one fiber per Correlator Layer-2 board. This provides capacity for up to 27 objects per board, at 128 bits per object. The baseline GT menus require between 8 and 24 objects from each of the 5 processing board types, giving some room for future expansion. A copy of the data sent to GT is also sent to the scouting Global System (sGS), as described in Section 5.7.

The firmware requirements in Correlator Layer-2 are expected to be driven by jet finding, energy sum and isolation algorithms, which must intrinsically compare input from large regions of the detector. As a first estimate of the Correlator Layer-2 firmware requirements, a prototype sliding window jet algorithm was implemented, as described in Section 6.4. Extrapolating the FPGA resource usage to the final system (including known optimizations) indicates that a jet finder can be implemented in a VU9P FPGA using 11% of flip-flops, 47% LUTs, and 4% DSPs. This jet algorithm is indicative of histogramming of detector regions and represents a worst-case scenario for resource use in histogram-based algorithms. The latency of the algorithm is 20 clock cycles at 360 MHz, or approximately 55 ns.

We anticipate that firmware resource use can be optimized using more sophisticated algorithm implementations that are currently under study. In particular, we expect more widespread use of algorithms that directly perform object reconstruction on the PF or PUPPI particle inputs, without histogramming. A first implementation of such a particle-based algorithm has been implemented on a VU9P development board. The algorithm is a neural network based hadronic tau identification algorithm, and is discussed in more detail in Section 6.4. The algorithm assumes up to 2592 PUPPI input candidates arrive on up to 72 input streams, which are divided by particle type, as described in Section 5.5.1, and sorted by $p_T$. Resources for the algorithm, including input and output buffering, were found to be 27% of the Block RAM, 22% of the DSPs, 16% of the LUTs, and 8% of flip-flops. The latency is 255 clocks at 240 MHz, or 1.065 $\mu$s. While this latency is relatively large it also includes both input and output buffering, and we anticipate further reductions in future. We expect that further development of particle-based algorithms for Correlator Layer-2, will result in optimal physics performance, at minimal cost in terms of FPGA resources requirements and latency.

Finally, we note that the design of the Correlator Layer-2 system is quite flexible. Future object reconstruction algorithms may not require the full set of PF candidates and/or may not use large amounts of FPGA resources. In this case, some flexibility is retained to use lower TMUX factors for particular algorithm sets. The lower TMUX factors would allow for a larger set of individual FPGA based algorithms to be applied.

### 5.6 Global Trigger system

The Global Trigger system is responsible for implementing the trigger menu (described in Section 4) - i.e. for the parallel evaluation of $O(1000)$ trigger algorithms that each select a specific event signature. It needs to receive inputs from all upstream trigger systems, buffer data to compensate for different arrival times, de-multiplex data and provide them to the global trigger algorithms discussed in Section 3.7, which are all evaluated in parallel. The GT system further provides functionality to prescale algorithms and to monitor their rate, both before and after taking into account dead time. Finally, the GT system computes the trigger type based on the trigger algorithms that fired and sends it to the Phase-2 Timing and Control Distribution System (TCDS-2) [59]. As in Phase-1, the trigger menu is expected to evolve continuously to
adapt to the operating conditions provided by the LHC and to new insights gained. The experience from Run-1 and Run-2 shows that more and more complex algorithms tend to be added over time, which are tailored to very specific physics channels. As these algorithms tend to use a lot of resources, the Global Trigger system needs to be scalable. In general, it is desirable that in order to update the trigger menu, only the Global Trigger system needs to be updated, i.e., that all cuts defined in the menu are applied in the Global Trigger. It is likely that the firmware of the Global Trigger will need to be regenerated when the menu changes - at least for major changes of the menu (as is the case for the Phase-1 Global Trigger). The software infrastructure to automatically create new firmware based on a description of the menu will be an important deliverable of the Global Trigger project.

5.6.1 Inputs to the Global Trigger

The Global Trigger system receives collections of physics objects from the Global Calorimeter Trigger, the Global Muon Trigger, the Global Track Trigger and the Correlator Trigger Layer-2 as detailed in Table 5.6. All input links are expected to operate at 25 Gb/s, using the same protocol. In general, di-object correlations, such as $\Delta\eta$, $\Delta\phi$, $\Delta R$, invariant mass and others are calculated in the Global Trigger from the single objects (see Section 3.7). The Global Trigger will be able to perform the necessary de-multiplexing if upstream systems are time-multiplexed, except in cases where receiving one link per time-slice would lead to low link utilization and unnecessarily large number of input links (e.g. GMT at TMUX=18). In the latter case, upstream systems are expected to de-multiplex their outputs using a dedicated de-multiplexer card. For objects with a width other than 64 bits (the width of data words on the link), the format of data on input links will have to be arranged to facilitate un-packing of the data frames. Object collections on time-multiplexed links are expected to be ordered to minimize latency in the GT processing whenever possible. Moreover, common scales should be used across objects wherever possible. Objects within a collection are expected to be ordered as required by the algorithms run in the GT (e.g. by decreasing transverse momentum). As can be seen from Table 5.6, the total expected bandwidth into the GT system corresponds to 41 links at 25 Gb/s. After taking into account the overhead due to having to receive at least one link per TMUX period and some contingency, about 54–58 links with lower occupancy will be required. Should the MIP Timing Detector system be up-scope to send information to the GT, 2 additional links would be needed. This amount of links can be fed into a single FPGA, such as the Virtex Ultrascale VU9P, on one of the processing boards described in Section 6.3 with adequate contingency, especially when considering boards with 96 inputs to a single FPGA.

5.6.1.1 External trigger inputs

In addition to collections of physics objects, the Global Trigger will also be able to receive external trigger conditions (as described in Section 2.6). The external trigger conditions can be used as part of any trigger algorithm in the algorithm step. Some of the external signals may be used in the default trigger menu while others may be needed only for special runs and/or for a limited time. In general, it should be possible to integrate new external trigger conditions on a short timescale.

Since the BPTX sub-system will only send the BEAM1 and BEAM2 signals, when a bunch is present in beam 1 or 2 respectively, the GT will include logic to compute any derived signals, such as BPTX-AND and BPTX-OR (the logical AND and OR of these signals, see Section 2.6), which in turn may be used in trigger algorithms. Additionally, the GT may provide functionality to histogram the BPTX signals and verify the filling scheme published by the LHC.

For external systems that are based on the same or similar electronics as the L1 trigger system,
### 5.6. Global Trigger system

<table>
<thead>
<tr>
<th>System</th>
<th># bits / obj</th>
<th># obj / coll</th>
<th># collections</th>
<th># bits</th>
<th>min. @25Gb/s</th>
<th># links planned</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMT</td>
<td>128</td>
<td>8</td>
<td>2: standalone, track+muon (track+stub, global)</td>
<td>2624</td>
<td>5</td>
<td>5-9</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>2</td>
<td>2: B_{st}, \tau</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>64</td>
<td>1: TkMuStub/MuStub trigger conditions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCT</td>
<td>64</td>
<td>12</td>
<td>3: jet, e/\gamma, \tau</td>
<td>3072</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>GTT</td>
<td>64</td>
<td>12</td>
<td>5: prompt jets, displaced jets, \tau, \phi, isolated (prompt/displaced) track</td>
<td>4352</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>12</td>
<td>1: energy sums</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>2</td>
<td>1: B_s</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlator L2</td>
<td>128</td>
<td>12</td>
<td>8: PUPPI jet/\tau, sums, tk+e/\gamma x3, tk e/\gamma \tau \times 2, muons</td>
<td>12288</td>
<td>22</td>
<td>30</td>
</tr>
<tr>
<td>Ext. Cond</td>
<td>1</td>
<td>64</td>
<td></td>
<td>64</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>41</td>
<td>54-58</td>
</tr>
</tbody>
</table>

Table 5.6: Object collections received by the Global Trigger system. The given number of bits per object are an upper bound. The definitions of the collections are expected to evolve as upstream algorithms evolve. The number of links from the GMT depends on whether data are received directly from TMUXed GMT boards or from a de-multiplexer card.

the baseline is to transfer external triggers using a 25 Gb/s link with the same protocol as used elsewhere in the L1 trigger system. While such a link is poorly utilized on external trigger signals, using the same protocol simplifies the integration and allows for the use of existing infrastructure in the monitoring of the link to be performed. Moreover the task of synchronizing an external signal to the correct bunch crossing is in this case a clear responsibility of the external system.

Should an external system not be able to provide a 25 Gb/s optical link, a backup solution using LVDS links over copper, as has been available in Phase-1, will be maintained. In this case, the LVDS signal will be sampled at 160 MHz by the GT system and the phase and bunch crossing alignment of the signal will be done by the GT.

In order to keep the GT system as homogeneous as possible, the same type of board as for the GT algorithm boards is planned to be used to receive external conditions. This type of board needs to provide a number of LVDS input pairs. The physical interface for LVDS signal can be RJ45 connectors as in Phase-1.

In order to reduce the number of cards in the GT system, the reception of external conditions and the Final-OR function (Section 5.6.2.2) can be implemented on the same FPGA in a single board.

#### 5.6.2 Global Trigger architecture

The logic of the Global Trigger is foreseen to run at a multiple of the LHC bunch crossing frequency. Frequencies of 240 MHz and 480 MHz are currently under study. The typical size of input object collections will be 12 objects for jets/taus/e/\gamma, and 8 objects for muons (as in Phase-1). Assuming that an operation on an object can be applied in 1 clock cycle, processing
Chapter 5. Conceptual design of the Phase-2 L1 Trigger

12 objects would take 1 or 2 BX. An appropriate time-multiplexing factor for the Global Trigger Logic thus would be 1 or 2. In order to evaluate inter-BX conditions (discussed in Section 3.7.1), data from consecutive bunch crossings need to be available, which is only possible at a TMUX factor of 1. The final trigger decision needs to be communicated to TCDS at a TMUX factor of 1, meaning that at least the final stage of the GT would need to run at TMUX=1. Moreover, neural net based algorithms (described in Section 3.7.2) show a large penalty in latency when run at TMUX factors larger than one. This is because the calculation of one layer needs to be completed before the following layer can start computing.

Higher time-multiplexing factors (such as 6 or 18) have been studied for the earlier stages of the GT. However, they imply that operations need to be scheduled in time to preserve an optimal use of resources. For example, the computation of the invariant mass of all combinations of 12 $\times$ 12 objects over a specific class of objects would need to be performed during the first 1 or 2 BX of the time-multiplexing period, then for the next class of objects we would need to perform the same mass calculations and so on; otherwise, the computation unit would be idling for most of the time. This scheduling would add an additional complexity to the GT logic which may increase the risk of problems surfacing when a new menu is deployed. The preferred TMUX factor for the entire GT system is therefore 1, which does not preclude that part of the logic internally runs at a TMUX factor of 2. A TMUX factor of 1 further has the advantage that the GT system size can be scaled easily with a granularity of 1 board at a time.

5.6.2.1 GT algorithm boards

The GT architecture with the preferred TMUX factor of 1 thus consists of multiple algorithm boards that each receive a copy of all inputs and each will execute a subset of the trigger algorithms as illustrated in Fig. 5.9. The maximum number of GT boards in this architecture is limited by the available outputs of the upstream systems. In general, enough output links are available for time-multiplexed upstream systems that send at most 2 links per board to each GT algorithm board. Non-time-multiplexed upstream systems (GCT, potential GMT demux board) that send 6 links per board to each GT algorithm board, are not on the latency critical path. A number of solutions are available to avoid limiting the number of GT algorithm boards, such as fan-out boards, sharing data over an inter-FPGA bus in a two-FPGA system, or passive optical splitting. As discussed in Section 3.7.3, the baseline GT system contains 12 algorithm boards.

5.6.2.2 GT Final-OR board

A GT Final-OR board will receive the outputs of the algorithm boards and perform the final OR (per trigger type). It may also perform functions, such as prescaling and monitoring in order to offload this functionality from the algorithm boards. The GT Final-OR board may also include the functionality to receive external trigger conditions and to distribute them to the algorithm boards.

If overall latency proves to be critical for the Phase-2 Level-1 Trigger, a solution absorbing the GT Final-OR board into the TCDS - albeit without any monitoring - may be envisaged in order to save the latency of one optical link.

5.6.2.3 Latency

The total latency of the global trigger system including three optical links (up to TCDS) and deserialization is illustrated in Table 5.7. For input links from upstream systems and the output link to TCDS, in each instance half of the latency of the link is attributed to the GT. The full
5.6. Global Trigger system

Figure 5.9: Architecture of the Global Trigger system. Each algorithm board receives a copy of the same inputs. The number on the top right of the upstream systems indicates the assumed time multiplexing factor of the upstream system. However, other time multiplexing factors can be employed in these systems, as long as the total number of links does not change. Other inputs may include MTD if up-scope in the future. Both time axes show the worst-case latency of the GT system, that would be needed, if the upstream system(s) on the critical path had a time-multiplexing factor of 18.

Latency of de-multiplexing data is attributed to the GT. This brings the total latency of the GT to 1 μs, if the upstream system on the critical path used the highest supported TMUX factor of 18.

<table>
<thead>
<tr>
<th>Item</th>
<th>Latency [BX]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input link</td>
<td>2.5</td>
</tr>
<tr>
<td>Deserializer</td>
<td>nTMUX</td>
</tr>
<tr>
<td>Algo</td>
<td>10 (see algorithm section)</td>
</tr>
<tr>
<td>Link to Final-OR card</td>
<td>5</td>
</tr>
<tr>
<td>Processing on Final-OR card</td>
<td>2</td>
</tr>
<tr>
<td>Link to TCDS</td>
<td>2.5</td>
</tr>
<tr>
<td>Total Latency</td>
<td>22 + nTMUX</td>
</tr>
</tbody>
</table>

Table 5.7: Total latency of the GT system. The latency of the deserializer may be less than the time-multiplexing period (nTMUX), if links are not fully utilized. nTMUX only has to be counted fully if the upstream system with the highest nTMUX is on the critical path.

5.6.3 Interfaces to TCDS, DAQ and HLT

5.6.3.1 Interface to TCDS

The baseline for the output to TCDS is a single 10 Gb/s optical fiber transmitting the trigger decision for each bunch crossing. The trigger decision will be encoded as a field of bits, each bit denoting a trigger type with a list of example trigger types given in Table 5.8. TCDS passes the field of trigger bits on to the sub-detectors. The actual list of trigger types and their meaning for each detector remains to be defined.

The interface to TCDS also foresees a reverse link that is reserved for future use, such as communicating the dead time or the warning state back, to be used in the trigger throttling.
Chapter 5. Conceptual design of the Phase-2 L1 Trigger

<table>
<thead>
<tr>
<th>trigger type</th>
<th>action in TCDS / sub-detectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard physics</td>
<td></td>
</tr>
<tr>
<td>zero bias</td>
<td></td>
</tr>
<tr>
<td>min bias</td>
<td></td>
</tr>
<tr>
<td>high priority</td>
<td></td>
</tr>
<tr>
<td>Long-lived particle (LLP)</td>
<td>detectors (if supported) to read out wider window of bunch crossings</td>
</tr>
<tr>
<td>First event of a 2-event LLP</td>
<td>must not be suppressed if first event wasn’t suppressed</td>
</tr>
<tr>
<td>Second event of a 2-event LLP</td>
<td>read out additional event information for validation</td>
</tr>
<tr>
<td>Periodic validation event</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.8: Examples of trigger types that may be generated by the GT.

5.6.3.2 Interface to DAQ and High Level Trigger

Each of the GT boards will send a readout record to the DAQ. The readout record will contain the GT inputs and several bits per trigger algorithm, indicating whether the algorithm fired, passed the prescale, and passed the bunch mask\(^2\). The readout record further contains information needed by the High Level Trigger: the HLT needs to receive the union of all objects that possibly contributed to firing any given trigger algorithm. The information is used to seed the HLT reconstruction with L1 objects, thus improving its performance - and to reject possible volunteer objects (objects found by the HLT but not by the L1). The latter is necessary to properly measure trigger efficiencies. This information can be represented by a bitmap for each trigger algorithm in which each bit represents an input object of the GT. Assuming 1024 algorithms and 200 input objects, such an uncompressed bitmap would add some 25 kB to the readout record. Various compression schemes may be used to reduce the size.

For validation and debugging, a full copy of the inputs and outputs of each board can optionally be added to the readout record for a reduced set of events (validation events). All information can optionally be read out from multiple bunch crossings in order to facilitate timing studies or similar.

5.6.4 Using the 40 MHz Scouting System for trigger monitoring

As explained earlier the GT is reprogrammed several times per year. It is thus imperative to detect any data inconsistencies as fast as possible.

By receiving both the inputs and outputs of the GT, the 40 MHz Scouting System (described in Section 5.7) can verify correct system operation. Furthermore, new algorithm developments can be prototyped and tested rapidly in the scouting system using the GT inputs. To do so, the GT will send to the scouting system two bits for each of up to 1024 algorithms indicating whether the algorithm fired and whether it passed the prescale, as well as a few bits per bunch crossing to indicate the trigger type(s) sent to TCDS-2 and whether the event accept has passed the TCDS-2\(^3\). The inputs to the GT will be sent to the scouting boards by the upstream systems themselves, except for the external inputs which will be forwarded by the GT.

The algorithm and the decision information, as well as external triggers, can be sent on four 25 Gb/s links from a combined Final-OR / external condition board (or on 4+1 individual links if the boards are not combined). Sufficient additional output links are available in the GT Final-OR Board and GT algorithm boards to send more information to the scouting system, if this

---

\(^2\)A bunch mask defines for each BX in the LHC orbit, whether a certain trigger is prevented from firing in this BX.

\(^3\)Event accepts may be blocked in TCDS-2 due to trigger rules, trigger throttling or other dead times.
5.7 Scouting system

Besides the study of specific physics channels, as discussed in Section 1.4.2, the Scouting System will enable real-time trigger component diagnostics, providing invaluable monitoring functionality for the Level-1 trigger, in particular for the GT, as outlined in Section 5.6.4. Diagnostics for the lower-level systems will be greatly facilitated by capturing their inputs and outputs. As an example, high statistics heat maps from the regional triggers may provide instantaneous indications of transient detector problems affecting them; re-running specific algorithms in software from the captured input can help identifying corner cases not covered by offline tests. Selecting and reconstructing specific physics objects or processes without the rate limitations inherent to the full DAQ or the bias of a Level-1 algorithm will allow cross-checks of the luminosity measurements to be performed and simplify the comparison with other experiments.

5.7.1 General architecture

Intermediate data from the different trigger processors will be read out and processed at the Level-1 accept rate by the data acquisition system for triggered events, in much the same way as detector data. The Scouting system represents a second, parallel readout chain running as an "opportunistic experiment", processing copies of the output streams feeding the different components of the Level-1 at 40 MHz (see Fig. 5.10). Trigger data will be extracted directly from the Level-1 using spare outputs of the processing boards, over the same type of high-speed serial optical links used for the Level-1 interconnects and using the same protocols. For this to be possible, provision must be made for sufficient spare output capability at each and every layer of the Level-1 system. In general, this fits well with the Level-1 architecture. The scouting system must be instrumented with dedicated input boards receiving data streams from multiple serial links, decoding the protocol, collecting and preprocessing relevant data that is subsequently fed over PCIe into standard processors (I/O nodes) for short-term storage and subsequent further analysis in a distributed system.

No handshake with the Level-1 control system is required, but the scouting system will receive Run Control commands and will be integrated in the operation sequences of the experiment. Note that since no back-pressure modulates the input traffic into the scouting system, if one or more of the downstream components cannot keep up with the input throughput, data will simply be lost.

The output of the GT Final-OR, information on prescaling and other TCDS functions, as well as external triggers status at the GT input, are captured in the scouting Decision System (sDS). The output of the four global systems, GCT, GMT, GTT, and the Correlator Trigger (PFT), is captured by the scouting Global System (sGS). The sDS and sGS constitute a first and independent stage of the scouting (referred to as stage1 in the following), with relatively modest throughput requirements, providing vital trigger diagnostic functionality for the GT and interesting physics functionality. The sLS (scouting Local System) captures the output of the local muon and regional barrel calorimeter triggers and the endcap calorimeter primitives. The regional barrel and endcap calorimeter triggers, the barrel, overlap and endcap muon track finder systems are all distinct and independent. As such, each of them can be included in the scouting system as needed. The capture of cluster and trigger tower data from the endcap calorimeter has throughput requirements similar to those of the scouting Track System (sTS), which captures the output of the Level-1 track finder (TF). Collectively the sLS and sTS form the second stage of the scouting system (referred to as stage2 in the following), with steps in...
Figure 5.10: Principle architecture of the Phase-2 Level-1 including the 40 MHz scouting system.

throughput and compute requirements of roughly a factor 3–4 between the sGS and sLS and between the latter and the sTS. Capturing primitives from the endcap and barrel calorimeter back-ends (sPS) involves throughput figures which are one order of magnitude larger than the ones discussed so far. Making the case for such a system to be built is beyond the scope of the Level-1 TDR and its discussion is therefore postponed for possible further evaluation at a later time.

The capture of Level-1 intermediate data at the bunch-crossing rate has been demonstrated on a small scale in Run 2 and is discussed in Section 6.4.

Even though hard scheduling is not required, the input throughput and average latency requirements of the different scouting systems are very similar to the corresponding components of the Level-1 trigger. The logic resources available on the input boards will be used to perform zero-suppression and local preprocessing; this preprocessing can include the application of selection criteria, corrections, and calibration constants, to reduce input data and overhead of the following processing stages to the maximum extent possible. Clearly, even after fine-grained zero-suppression, the long-term storage of the huge amount of raw data produced by the trigger processors, in view of a subsequent “classic” multi-tiered offline analysis and reduction, does not represent a viable long-term approach. Fast, local short-term storage⁴ in the I/O node will collect preprocessed input data. Asynchronous distributed algorithms running on the I/O nodes and using a low-latency interconnect will combine multiple inputs and identify relevant physics objects or features. The short term storage will be sized to provide enough buffer space to match the overall latency of the distributed processes.

The distributed processing of data from multiple sources will be demonstrated in Run-3 as outlined in Section 6.4.

---

⁴Realized with a suitable technology guaranteeing large volumes and low latency.
Features deemed interesting will be handed, over the interconnect, to a distributed stream-processing system, to be further selected and combined into higher-level data, organized and stored in medium-term storage. The intermediate storage decouples high-level analysis from the feature search in much the same way as reconstruction and analysis phases in classic multi-tiered offline processing, while allowing convenient reorganization of the data to be achieved. The distributed algorithms will be easily re-configurable without affecting the performance of the experiment, allowing new ideas to be tested, which can then possibly be imported in the Level-1.

Multiple parallel analyses/searches, possibly query-based, will be run off the medium-term storage content to produce a synthesis of relevant information, in the form of e.g. high level distributions, “tuples” or collections of candidate “events”, which will be used for analysis, diagnostics and monitoring, and stored long-term.

There is great flexibility in implementing the architecture described above. As an example, the stage1 scouting, which is de facto part of the GT system, will work only on the output of the GCT, GTT, Correlator Trigger, and GMT. The GTT will produce primary vertex and global track quantities estimates, along with a small number of track-only physics objects. The limited amount of logic available on the GTT processors may require a minimum $p_T$ of the tracks combined to form two-track candidates, say e.g. $\phi$ candidates. Capturing the TF output stream, part of the stage2 scouting, will allow that limitation to be circumvented. In other cases, the limitation may come from the way the GCT handles primitives from the HGCAL and require capturing the HGCAL primitives at the output of the HGCAL TP boards. Given the significant complexity and cost of hardware for the complete stage2 system, it is envisaged to initially implement the stage2 in such a way as to enable one or the other functionality using the same input hardware. The expansion of stage2 to support both at the same time could be later justified by a compelling physics case. Connectivity, compute, and storage resources can be expanded separately as the system increases in complexity or when new use cases arise.

The following section discusses the requirements of the different stages in some detail.

### 5.7.2 Inputs, memory, and connectivity requirements

It is assumed that the inputs to the scouting system are in all cases 16 or 25 Gb/s serial optical links, in some case time-multiplexed. It is also assumed that the protocol, not yet completely defined, will be similar to the one used in the Phase-1 system. There is no need for prior demultiplexing of time-multiplexed inputs, as the scouting system works asynchronously and can distribute work to additional compute resources over the interconnect. For time-multiplexed systems, however, BX assignment will have to be done link-by-link, depending on how its connections are configured, and will require special care. Where regional processing is used, as is more likely to be the case in lower level layers, there is no need to combine inputs from different regions, since the global view is accessible to the distributed algorithms via the interconnect.

Input devices with a significant amount of optical input bandwidth are readily available from the market or on parallel developments. The baseline used is an eight-optical-input half-size PCIe board with a modest FPGA. A board similar to the one used for the GMT scouting demonstrator, but providing 8 PCIe Gen4 lanes and supporting four optical links with 25 Gb/s transceivers exists, and similar devices with 8 or more optical inputs can be expected from

---

5For efficient operation of the subsequent analysis steps, the medium-term storage needs to be organized around some form of database technology allowing random access to, and reorganization of, features using dynamic criteria for analysis. For example, search-engine technologies could be adapted to work on mainly numerical data, organized and indexed using columnar formats or a key-value store.
Xilinx or other FPGA vendors in the coming two-to-three years. One advantage of such a form factor is that it can be fitted in an inexpensive 1U rack-mount server. The nominal input bandwidth of an 8-link board, i.e. 200 Gb/s, matches that of a 16-lane PCIe Gen4 slot. In the following, this form factor for the input board is used as a convenient unit and it is assumed, unless otherwise stated, that one or two input boards will be installed in each I/O node.

A special scouting I/O node will be dedicated to receiving the TCDS-2 stream. This will enable the system to detect a global resync and deal with it at every individual component, avoiding situations where the input firmware may otherwise get into an unrecoverable error state. It will also be possible to redistribute specific TCDS information to the individual I/O nodes, translating the TCDS stream to an asynchronous stream compatible with the specific link speed (16 or 25 Gb/s). This will use one additional optical input and is only necessary if the corresponding input streams do not carry enough information to properly tag inputs with the correct orbit/bunch crossing. In what follows, link counts do not take into account this additional link.

In general, it is expected that an input frame from a trigger processor will correspond to one or multiple orbit(s) and carry orbit number, with every filled bunch crossing accounted for. It must be noted that the fixed-size format lists of candidates used, for example, for the GT input may have a variable but significant level of sparsity. In the GMT demonstrator, suppressing the empty BX frames reduces the effective throughput by a factor of 20. Further reductions can come from finer zero suppression into a variable-size format. A global factor four reduction is therefore a very conservative assumption and will be used in the following to estimate the requirements in short-term storage, assuming a buffering time of 120 s, similar to the one currently used for the HLT, and the number of uplinks to the HPC interconnect, assuming 200 Gb/s links as a baseline. A fully non-blocking switch or switch complex at 200 Gb/s will be sufficient to even transport the entire raw candidate information from every node to at least another node. Performing a classic event building and storing this information could be achieved with such an interconnect, for example using InfiniBand. Since anyway the proposed architecture offers many more opportunities for data reduction, discussing a classic approach is besides the point.

The requirements for the different scouting systems are summarized in Table 5.9.

Since all input links to the sGS are 25 Gb/s, the total minimum number of boards required is 8. One dedicated I/O processor per input board is preferable for these stage1 processors, since the nominal maximum throughput is 200 Gb/s, in line with the throughput capabilities of a 16-lane PCIe Gen4 board. The stage1 scouting system can be realized with today’s technology at a modest cost, since all the components are readily available.

The stage2 sLS input capture is over 16 Gb/s links. Capturing only eight links per I/O node is clearly sub-optimal. For the baseline, two input boards per I/O node are envisaged, assuming the minimal bandwidth “overbooking” can be absorbed by zero suppression and preprocessing. A compression of the input streams by a factor of three to six will be necessary to enable their operation with inexpensive interconnects. Notice that direct capture of output from the muon track finders requires a large number of links, but links are under-utilized. In this case, not shown in Table 5.9, intermediate boards with higher optical input multiplicity could operate as concentrators with a similar final result. Clearly, a denser solution for the input board would be desirable. This is in particular the case for endcap calorimeter primitives (in Table 5.9, two boards per I/O node are assumed). For example, a FELIX prototype board for ATLAS with 24 optical inputs and using PCIe Gen4 is currently in preparation. A similar form factor can, if necessary, be developed in house. FPGAs supporting 16-lane PCIe Gen4 at an accessible price
5.8. Overall trigger system architecture

In order to provide the required scientific performance, the Phase-2 L1 trigger system relies on the delivery of highest precision inputs representing the largest bandwidth ever handled at Level-1 (63 Tbit/s of information, see Table 2.11). The design of the architecture is guided by the

<table>
<thead>
<tr>
<th>System</th>
<th>Links (speed Gb/s)</th>
<th>I/O boards (I/O nodes)</th>
<th>Total (120 s) S/T Storage (TB)</th>
<th>200 Gb/s links to I/C</th>
</tr>
</thead>
<tbody>
<tr>
<td>sDS (GT+TCDS)</td>
<td>3+1 (25/10)</td>
<td>1</td>
<td>small</td>
<td>1</td>
</tr>
<tr>
<td>sGS (GMT)</td>
<td>8 (25)</td>
<td>1</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>sGS (GCT)</td>
<td>8 (25)</td>
<td>1</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>sGS (GTT)</td>
<td>16 (25)</td>
<td>2</td>
<td>1.5</td>
<td>2</td>
</tr>
<tr>
<td>sGS (PF)</td>
<td>32 (25)</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Total stage1</td>
<td>63+5</td>
<td>9 (9 nodes)</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>sLS (RCT+HF)</td>
<td>144+36 (16)</td>
<td>22</td>
<td>43</td>
<td>15</td>
</tr>
<tr>
<td>sLS (xMTF)</td>
<td>18+6+12 (16)</td>
<td>6</td>
<td>2.5</td>
<td>3</td>
</tr>
<tr>
<td>sLS (HGCAL)</td>
<td>432 (16)</td>
<td>54</td>
<td>50</td>
<td>34</td>
</tr>
<tr>
<td>sTS</td>
<td>324 (25)</td>
<td>41</td>
<td>120</td>
<td>21</td>
</tr>
<tr>
<td>Total stage2</td>
<td>648(16)+324(25)</td>
<td>123 (62 nodes)</td>
<td>230</td>
<td>73</td>
</tr>
<tr>
<td>sPS</td>
<td>thousands</td>
<td>hundreds</td>
<td>2-5k</td>
<td>hundreds</td>
</tr>
</tbody>
</table>

Table 5.9: Summary of the input, connectivity, and short-term storage requirements for the different scouting systems. For HGCAL and TF only a factor two reduction from zero-suppression and low-level processing is assumed.

seem within reach on the timescale of Run-4. Commercial I/O boards with similar capabilities are far from being excluded on the timescale of the coming five years, as well as Gen5 PCIe and 400 Gb/s interconnects. Commercial motherboards with multi-Tbps I/O capabilities will definitely follow.

With two links for each of the 18 time slices in one \( \phi \) sector, using two 8-link boards per node, the sTS would require four and a half input boards for each of the 9 \( \phi \) sectors. In the minimal configuration with 21 I/O nodes, the latter would have to deal with 400 Gb/s nominal maximum input. A modest factor two reduction could be sufficient to enable short-term storage and full non-blocking connectivity over a standard 200 Gb/s switch with 36 ports.

A single fully non-blocking interconnect realized with a monolithic switch or a switch complex is the most flexible solution to support distributed processing of low-level data as well as feeding intermediate streams to the global processing system. For each stage, a conservative assumption is that as many switch ports will be required to connect the global processing as those required for the I/O nodes. Clearly, stream traffic from the I/O nodes can be routed to multiple heterogeneous compute resources from each of those ports, making the system expandable at will. Likewise, the amount of medium- and long-term storage deployed can be tailored to match applications as they come online.

It must be noted that only the I/O nodes are required to be installed underground, due to the limited range of the optical links from the Level-1 boards. Assuming 1U processors are used, the stage1 and stage2 systems, excluding the sLS(HGCAL) and the sTS, can comfortably fit in a single standard PC rack, while one additional rack is required to fit the latter two. The rest of the system, including the switch complex, can be installed in the surface data center, and use the same type of equipment as - and even be physically part of - the standard HLT farm.
need to efficiently organize the distribution and processing of primitive data throughout the system, the need to provision the appropriate computing resources and interconnections, and the capability to retain headroom for future flexibility and robustness. The key design feature is the implementation of a Correlator Trigger that harvests the regionally processed information from the sub-detectors (clusters of energy and charged-particle tracks) to feed sophisticated algorithms to produce higher-level objects. The full architecture exercise considers the following constraints: the number of input links per board should not exceed 100 and the occupancy of the FPGA resources should remain under 50%. The limit on FPGA resources used for a given processor board is motivated by the need to allow the algorithms to be improved with time\textsuperscript{6}, and the limit on the number of input links is driven by physical constraints of the electronics boards presented in Chapter 6. The main constraint on the design is the latency budget, which cannot exceed 9.5 $\mu$s, including the arrival latency of tracks and HGCAL clusters, which are not expected before 5 $\mu$s (See Table 2.11 in Chapter 2). Prior to performing a full architecture exercise, the design of each component of the Phase-2 L1 trigger has been presented with a clear specification of their function in the previous sections.

An initial approach studied the implementation of a first layer of Correlator Trigger that received information from the muon track finders, the barrel calorimeter trigger, the endcap calorimeter backend, and the track finder. This resulted in a design with the number of input links nearing the limit and estimated FPGA resources nearing 50%. The redistribution of processing into four lanes was then studied through the addition of a Global Calorimeter Trigger, a Global Muon Trigger and a Global Track Trigger. This redistribution allowed for a reasonably sized Correlator Trigger while requiring only a lightweight exchange of data between the global lanes. Moreover, this partitioning offers the advantage of providing independently functioning subsystems for making L1 trigger decisions, which is expected to be convenient during commissioning periods, where not all detectors or electronic systems may be available. This choice leads to the high-level organization of the Phase-2 trigger architecture visible in Fig. 5.11; the latency spent on each processing step is also illustrated. The first 5 $\mu$s is spent on regional processing, track finding, and HGCAL clustering, while the next 4.5 $\mu$s is consumed by the triggerable object matching and reconstruction, followed by the global trigger decision.

For each trigger subsystem, various architecture options have been explored. The division of the overall processing requirements among processor boards may follow a regional approach or a time-multiplexed one. The latter was originally proposed for the Phase-1 upgrade of the calorimeter trigger [5]. These different approaches and their potential combinations have been discussed in the interim TDR [3]. This document presents the baseline approach for the different layers and portions of the system based on realistic simulations of data-flow and prototype algorithm implementations. Further tests performed with hardware demonstrators allow for architecture options to be tested. The results of these tests are described in Section 6.4.

Focusing first on the Correlator Trigger architecture, the baseline approach considers a combination of data processing subdivided by timing, region, and task. A study of input object multiplicity and the required associated FPGA resources to process these inputs led to a regionalized approach to the Correlator Layer-1 system. Natural subdivisions of input data are obtained by following the sub-detector organization in $\eta$ and their associated backend TMUX period. To retain maximum flexibility in the algorithms studied, a separation in processing task and time is also proposed for Layer-2. A common intermediate TMUX period of 6 is proposed for the two layers. All particle-flow and PUPPI candidate particles are distributed across Cor-

\textsuperscript{6}Another important consideration is the difficulty encountered by the firmware routing software in meeting the timing requirements. The complex computing sequence envisioned requires a high clock speed to keep the latency to a minimum
5.8. Overall trigger system architecture

Figure 5.11: Diagram of the CMS L1 Phase-2 trigger design. The Phase-2 L1 trigger receives inputs from the calorimeters, the muon spectrometers, and the track finder. The calorimeter trigger inputs include trigger primitives from the barrel calorimeter (BCAL), the hadronic forward calorimeter (HF), and the high-granularity calorimeter (HGCAL). The calorimeter trigger includes a regional barrel calorimeter trigger (BCT) and a global calorimeter trigger (GCT). The muon trigger receives input from barrel and endcap muon stations. The muon trigger is composed of a barrel Layer-1 processor (BMT L1) and muon track finders (BMTF, OMTF, and EMTF) transmitting their muon candidates to the global muon trigger (GMT) where association with tracking information is implemented. The track finder (TF) provides tracks to the GMT, the global track trigger (GTT), and the correlator trigger (CT). All objects are sent to the global trigger (GT) issuing the final L1 trigger decision. The arrow on the left represents the latency spent for each of the processing steps.

Concerning the choice of architecture for the GCT, GMT, and the GTT, the division of data among processing boards is generally dictated by the organization of the sub-detectors back-end electronics providing the inputs. TMUX depth can be chosen conveniently as divisors of the input system, so that FPGAs on the receiving end can be simply partitioned to handle the data from different time slices. For example, the GTT has a TMUX depth of 6 while the track finder backend has a depth of 18. Given the resources needed for vertex reconstruction and tracker-only algorithms, a minimum of 6 processing boards is required (12 are provisioned, for extension to additional algorithms and contingency). Although the same TMUX period across the global lanes may simplify the overall architecture design, an optimal use of resources may not always be achieved by doing so. For this reason, the structure of the global trigger systems are designed to be easily convertible between different TMUX depths for both input and output. For example, the global calorimeter trigger receives input data within the same time slice from the barrel while the endcap data are time-multiplexed with a period of 18. The output from the Regional Calorimeter Trigger (See Section 5.2) may be time-multiplexed to any period before entering the GCT, making this a flexible choice. This argument is also true for the multiplexing of the barrel ECAL, HCAL, and HF data although this falls under sub-detector projects with additional constraints. The choice of the GCT structure proposed here is to fully demultiplex the input endcap data and proceed with regional processing, followed by
a global processing step which merges or sums the regional outputs. Given the rather simple calorimeter-only object reconstruction algorithms and the available processing power to perform them, the performance achieved is not directly impacted by this choice. For example, the GCT design remains completely convertible to a fully time-multiplexed approach where all the data from barrel and endcap can be processed by the same board while offering a more adaptive interface to the track finder, should future requirement changes result in preferring it. In the case of the GMT, the choice to align the TMUX period with that of the track finder is motivated by the main processing task of this system: correlate tracks and muon information. The firmware resource estimations indicate that lighter hardware is required (See Section 5.3).

Figure 5.12: Diagram of the CMS L1 Phase-2 trigger design. The calorimeter trigger is represented on the left and composed of a barrel calorimeter trigger (BCT) and a global calorimeter trigger (GCT). The track finder in the center transmits tracking information to the correlator trigger (CT), the global track trigger (GTT), and the global muon trigger (GMT). The muon trigger architecture is represented on the right and composed of three muon track finders: EMTF, OMTF, and BMTF. The CT in the center is composed of 2 layers for particle-flow processing. The global trigger (GT) receives all trigger information for the final decision. For each architecture component, the information about the time-multiplexing period (TMUX), the regional segmentation (RS) in η or ϕ, the functional segmentation (FS), and the number of FPGAs are specified.

Figure 5.12 displays the baseline architecture chosen for the Phase-2 Level-1 trigger system. This diagram represents all the components of the foreseen system and their interconnections. The number of processing boards, ϕ or η segmentation (x axis), and TMUX period (y axis) are represented. The architecture modeled relies on the use of generic processing boards to equip each of the subsystems. The trigger components directly interfacing with sub-detectors are subject to constraints on the number of links and assignment of data fibers. At the time of this writing, most of the sub-detector backend electronics designs have been finalized and the trigger primitive formats specified. In some cases, the format was directly optimized to achieve the best algorithm performance or to optimize the resources on the receiving end. For some sub-detector interfaces, a baseline format was assumed and it was verified that reasonable...
changes have negligible consequences on the overall design. Later stages of the architecture displayed in Fig. 5.12 show more flexibility in their design. This allows contingency for future improvements and additions.
Chapter 6

Instrumentation of the Phase-2 trigger system

6.1 Introduction

This Chapter describes the current hardware R&D being carried out within the Phase-2 L1 trigger project. These R&D activities cover a large range of subjects directly connected to the system requirements, such as the choice of infrastructure and form factor, the use of high-speed optical link technologies, the design of generic processing engines equipped with large FPGAs, and the development of advanced software interfaces. There are currently four ongoing R&D lines aimed at an electronics board implementation. After a short presentation of the design philosophy of each of these projects, details of optical link tests and protocols, as well as board controls and results from thermal tests are provided. Descriptions of software progress and examples of firmware implementation are also included. The following section is dedicated to current demonstrator tests organized within the project. Test stands have been set up to demonstrate the function of each of the components of the trigger design specified in Chapter 5. The test setups presented implement single-board or multi-board configurations using target hardware when available. After describing the test strategy and the experimental protocols, results obtained with pattern tests and comparisons with software emulators are described. In the case of multi-board tests, board communication and synchronization are included. Experience from these demonstrator tests provides feedback on both the architecture and the structure of the algorithm firmware. Plans for slice tests including sub-detectors, and other potential tests performed during LHC Run-3 operation, are also discussed.

Later in this Chapter, system-wide technical issues are discussed. The challenges of integrating, installing and commissioning this system are identified. The experience acquired through the operation of the Phase-1 L1 trigger drives the strategy discussed. Amongst various topics, the hardware platforms, crate organization and firmware management are discussed. Trigger link alignment and synchronization procedures are presented, as well as a strategy for latency measurements. A key feature of the Phase-1 L1 trigger operation was the development of a reliable monitoring system and data validation procedure. This point is developed along with routine system health checks. A prospective commissioning plan concludes this Chapter.

6.2 System infrastructure and services

For the Phase-2 CMS upgrade the ATCA [79] form factor has been chosen as the crate system for all electronics modules located outside the CMS experimental cavern, thus in the service cavern. These crates will replace the VME [80, 81] and µTCA [82] systems from respectively the original construction [4] and the CMS Phase-1 upgrade [5].

ATCA offers a flexible, high density, and high performance backplane specified to support
Chapter 6. Instrumentation of the Phase-2 trigger system

Figure 6.1: CMS dual star backplane signal conventions. Each connection shown represents a ‘star’ connection from each of the two hub slots to each of the node slots.

several of the most common serial standards in use today (e.g., 1/10/100 GbE, PCIe, etc.), with a preference towards the Ethernet standards. Different backplane configurations allow optimization of the inter-board connectivity to specific use cases.

Each ATCA crate (called a ‘shelf’ in the PICMG documentation) contains one or two hub slots, identified by logical slot numbers 1 and 2. In a dual star configuration the fabric interface made up of four bidirectional high quality signal pairs connects each hub slot with each of the node slots. In addition the base interface consists of a 1000base-T Ethernet connection, three multi-drop clock lines, and a 100base-T Ethernet connection to the shelf manager.

The ATCA configuration adopted for the CMS Phase-2 upgrades is based on a 14-slot shelf endorsed by the CERN ATCA evaluation project [83], equipped with a CERN-endorsed shelf manager. The baseline foresees a dual star, or optionally a full mesh, backplane. Figure 6.1 shows the baseline backplane connectivity scheme between the hub and node slots. For overall system reasons (e.g., power and cooling load, availability of spare slots) the CMS recommendation is to reserve two of the twelve node slots as spares or for future needs.

The ATCA crates will be powered by CERN-endorsed external mains-to-48 V power converters. For installation in the existing electronics rack infrastructure in the CMS service cavern (USC), racks require vertical airflow for cooling. The USC rack cooling system is rated for a maximum heat load of 10 kW per rack, with an average of 6 kW. In addition, the HVAC system of the USC is designed to cope with a maximum heat leakage of 50 W per rack. Based on this it is recommended that node boards be designed for a maximum power dissipation of 300 W.

The interface between a Phase-2 CMS ATCA crate and the surrounding systems is provided through hub boards. The baseline design foresees a DAQ and TCDS Hub (DTH) [84] with integrated Ethernet switch in the first hub slot, with optionally a high speed Ethernet switch in
the second hub slot. If needed, the DAQ throughput for a shelf can be increased by adding additional DTHs, either in the second hub slot or in node slots.

In the baseline architecture proposed for the Phase-2 CMS DAQ, sub-system data (i.e., detectors and Level-1 trigger systems) are aggregated in sub-system-dependent backend boards. The data are then sent using mid-board optics over short, optical, point-to-point links using a custom protocol via the front-panel to the DTH. The DTH in turn aggregates data from multiple backend boards in a single backend crate and combines the individual streams for transmission on the data-to-surface network for processing and storage. In this position, the DTH provides the decoupling between the detector and networking time domains: fluctuations in event size and or trigger rate are buffered in the backend boards, and fluctuations in network throughput are buffered in the DTH.

The DTH is also responsible for distributing the LHC bunch clock, Level-1 accepts, and fast commands, as well as trigger control instructions for calibration, synchronization, etc., to the backend boards, all on dedicated point-to-point backplane links. In the reverse direction, backend boards report their data-taking readiness to the DTH, where these signals are aggregated for trigger throttling and data-taking recovery procedures.

The CMS trigger and timing control and distribution system (TCDS) revolves around two information streams exchanged between the TCDS master and the sub-system electronics. A Trigger Timing and Control (TTC) stream from the master to the sub-systems distributes the LHC bunch clock, beam-synchronous timing commands, and the L1 triggers. A Trigger Throttling Stream (TTS) in reverse direction notifies the master of changes in data-taking readiness on the sub-system side. For Phase-2 both the timing (TTC2) and the throttling (TTS2) streams are based on high-speed serial links on the backplane, running at a line rate of approximately 10 Gb/s, and synchronous to the LHC bunch-crossing clock. The TTC2 stream is an evolution of the Phase-1 TCDS TTC stream. Apart from the link speed and technology, the most notable new features are:

- The distribution of Level-1 event/trigger types, including a multi-bit physics trigger type from the Global Trigger. (See also Section 5.6.3.1.)
- Simultaneous (i.e., in the same bunch crossing) distribution of multiple synchronization commands, removing most of the scheduling constraints present in the current TTC system.
- Expansion of the BRILDAQ functionality to all receivers. This allows opportunistic use of any sub-system in measurements related to luminosity, and/or beam-induced background. This functionality will also enable sub-systems to (re)synchronize to ongoing data-taking runs without waiting for a start-of-run synchronization moment, hence easing commissioning and trouble-shooting.
- The distribution of dedicated ‘BRIL’ triggers for luminosity, and/or background measurements. For more detail, please refer to [85].

The throttling stream is an evolution of the TCDS TTS. The main difference is the increased bandwidth and the large number of backend links aggregated per TTS link.

The DAQ and TCDS Hub is being designed by the CMS central DAQ group in collaboration with the CERN electronics group. Figure 6.2 shows a photo of a prototype DTH in a test chassis.
Figure 6.2: A first DTH prototype in a test chassis. The front panel is oriented to the left. The two large, black heatsinks hide the DAQ (top) and TCDS FPGAs (bottom). On the top right side of the board the on-board COMExpress controller is visible. Connectivity to central services is provided via front-panel SFPs (for TCDS2) and QSFPs (for the data-to-surface network). Connections to sub-system node boards takes place through backplane connectors (on the right, for TCDS2) and via Firefly mid-board optics (of which two are mounted on this board) providing front-panel optical inputs for sub-system data.
6.3 Hardware research and development

6.3.1 ATCA processor: APx consortium

6.3.1.1 APx philosophy

The Advanced Processor (AP) Consortium includes several CMS institutes which have pooled efforts in ATCA FPGA-based processing hardware, firmware and software. The consortium is designed to leverage engineering resources, to provide a platform for communication and sharing of designs, hardware, software and firmware, and to minimize duplication of effort. Building from the successful CTP7 and MTF7 µTCA architectures in the CMS Phase-1 upgrade, which introduced to CMS the embedded Linux and high bandwidth memory lookup technology in FPGA processing boards, the consortium is developing multiple ATCA board and mezzanine types utilizing a modular design approach, with an emphasis on reusable circuit, firmware and software elements. The first generation of these boards has passed successful testing as described below.

The main processing board family (APx) is designed to provide for either one or two large processing Ultrascale+ FPGAs. The APx model uses an integrated CPU/FPGA System-on-Chip device, the Xilinx ZYNQ, for the primary embedded Linux control point. This control point uses integrated FPGA logic to allow the Linux system to effectively manage the FPGAs, clock synthesizers, optical modules and other components present in the platform. For thermal management of large FPGAs, the APx approach favors direct soldering of the FPGA on the main ATCA card so as to make maximum use of the standard ATCA slot width of 6HP (1.2 inch) for cooling. Direct soldering provides better electrical connections and optimal heat conduction.

The main APx boards host 5 subsidiary boards, including: Embedded Linux Mezzanine (ELM), Large Look-up Memory Table Mezzanine (LLUT), ATCA Rear Transition Module (RTM), Intelligent Platform Manager Control ATCA control point (IPMC), and Ethernet Switch Mezzanine (ESM). After successful testing, these 5 subsidiary boards along with their designs, firmware and software have been presented to and made available to CMS and ATLAS groups.

6.3.1.2 Hardware description

APd1 ATCA card  The first APx demonstrator board (APd1, Fig. 6.3) is built for trigger system usage in the ATCA form factor, based on a single Xilinx XCVU9P FPGA in a C2104 package.

The APd1 has a total of 76 optical links at up to 28 Gb/s (depending on Xilinx FPGA speed grade), which are supported at 19 Samtec Firefly optical transceiver positions on the main board, and 24 additional MGT (Multi Gigabit Transceiver) channels, which are routed to the RTM connector. These 24 links to the RTM bring the total links to 100 at 28 Gb/s. The APd1 also contains a 10GbE Ethernet PHY (physical transceiver), to support a 10GbE connection from ATCA Hub Slot 2 to the ELM board. It also supports a PCIe Solid State Drive (SSD) connection to the embedded Linux system. The APd1 provides multiple options for a CMS Timing and Control Distribution System (TCDS)-style timebase connection through a fabric of crosspoint switches, including a front panel SFP, or the standard backplane interconnection as specified in CMS documentation [84]. The first APd1s have passed full board and system tests, including all optical links running at 28 Gb/s.

Hardware control summary  Controls in the APx are provided by the IPMC and the ELM. The IPMC board is responsible for crate power on/off control and sensor readouts, and com-
municates with the crate IPM controller, known as the Shelf Manager. When the ATCA card is inserted (or crate is powered up), only the IPMC and ESM cards initially have power (the latter to provide the IPMC with network connectivity). Upon successful negotiation for crate resources with the Shelf Manager, the IPMC enables power to the ELM and main board components (FPGA, optics, etc.), and the LUT and RTM if present. Once booted, the ELM Linux system provides configuration and operational support for the platform. This includes initialization of FPGAs (bitfile loading and register/memory initialization) and configuration of support devices such as Firefly optical modules, clock switches and jitter-cleaner/synthesizers. Communication between APx FPGA(s) and the ELM board occurs over an AXI (Advanced eXtensible Interface) bridge which brings the FPGA registers and memories into the address space of the ZYNQ CPU at the hardware level. AXI is an interface standard supported by industry and well integrated into the FPGA core catalog and process flows of Xilinx Vivado.

**ZYNQ-IPMC** The ZYNQ-IPMC (Fig. 6.4) is an IPM control point for ATCA blades. It is based on a Xilinx XC7Z020 or XCZ7Z014S ZYNQ device in a DIMM-244 form factor, and is powered from a single +3.3V supply. The board contains 256MB of DDR3 memory and 64MB of QSPI flash memory. The QSPI is divided into 4 partitions: a recovery partition, two operational partitions (for fallback to the last installed image), plus a development partition. Flash
upgrades are supported via a custom FTP server running on the card. Additionally, the card contains 16 ADC channels on two SPI interfaces, capable of monitoring of main board signals at 3kSamples/sec. A 32 Kbyte EEPROM provides non-volatile storage of operational parameters. ZYNQ-IPMC software runs in the FreeRTOS environment providing core IPM controller functions that are mandated by the ATCA specifications. It implements JTAG-over-LAN service called Xilinx Virtual Cable, including JTAG access into the ELM ZYNQ or FPGA devices on the platform. Both versions of the ZYNQ-IPMC have been fully tested successfully, including programming ability and communications with main board and crate devices. The entire project (hardware, firmware, software) is in the process of being published as an open-source project for wider use in non-APx applications [86].

![Image of ELM1 and ELM2 mezzanine boards.](image)

**Figure 6.5:** Left: ELM1 and, Right: ELM2 mezzanine boards. The ZYNQ SoC device provides a Linux endpoint with IO customized to the specific needs of the host blade. The ELM1 is based on the ZYNQ-7000 family while the pin-compatible ELM2 is based on the ZYNQ-Ultrascale+ family.

**ELM**  
The ELM (Fig. 6.5) is a 75mm x 84mm custom form factor mezzanine presently containing a ZYNQ device plus supporting peripherals. Two variants of the ELM are built use the Xilinx devices XC7Z045 in ELM1 and XCZU04CG in ELM2. The ELM is designed to allow the use of newer or more economical ZYNQ devices in the future. The ELM form factor has two main board connectors: a 220-pin connector and a Samtec QSH/QTH header, that can support up to 20 bidirectional MGT channels at up to 25 Gb/s. In both the ELM1 and ELM2 8 of those 20 pairs are used, with the ELM1 supporting 8 GTX channels, and the ELM2 supporting 4 GTR channels and 4 GTH channels. The ELM has a dual synthesizer programmable reference clock (refclk) architecture, permitting timebase sharing with the main board clock system. The ELM-based logic can also operate on the LHC clock as distributed through the TCDS link. The ELM contains two on-card boot media sources: a QSPI flash for the backup/recovery image, and a microSD card for the operational boot image. It also supports two USB 2.0 ports through an on-card hub and ULPI (low-pin, low-cost, small form-factor interface) USB transceiver.

**ESM**  
The ESM (Fig. 6.6) is a 6-port GbE Ethernet switch based on the Microsemi VSC7512 Integrated Circuit. It comes in a 35mm x 40mm form factor, and is powered from a single +3.3V supply, such as the Management Power supply for the ATCA card. On the ESM four of the six ports are 1000BASE-T, with integral magnetics on the ESM, and the remaining two are SGMII-type. The board interface includes a UART console and SPI interfaces for monitoring board operation and updating switch software. In the APd1 application, 1000BASE-T connections are made to the backplane hub slots, the ELM and the IPMC cards. The ESM has successfully passed testing.
Chapter 6. Instrumentation of the Phase-2 trigger system

Figure 6.6: The 6-port Gigabit Ethernet Switch board.

Large LUT  The Large Lookup Table (LLUT) module (Fig. 6.7) is a mezzanine card 146 mm x 107 mm in size, designed to be installed on the APd1 ATCA card. The LLUT was developed to support trigger applications, like the Endcap Muon Track Finder, that use deep, high bandwidth memory to support FPGA-based processing. It contains 128 GB of DDR4 memory which can be used as a very large Lookup Table or as a general purpose very large high-speed data storage buffer. The module uses an XCKU035-1FFVA1156 FPGA as a control device. Two DDR4 Load-Reduced Dual Inline Memory Modules (LRDIMM), p/n MTA72ASS8G72LZ, are available on the LLUT module. Each of the LRDIMMs is connected to its own FPGA interface, so they can be used completely independently of each other, or together as one memory array. The LLUT is also compatible with LRDIMMs of a smaller capacity that can help to reduce cost if a smaller memory array is needed. The connection to the host APd1 board is implemented as a parallel bus with two differential clock lines and 196 single-ended bidirectional signals. The first LLUT prototype only uses 46 single-ended signals. Additionally, a Samtec Firefly link (4 RX + 4 TX channels) is provided that can be used as a serial interface to LLUT. The LLUT prototype has successfully passed testing.

RTM  The Rear Transition Module (RTM) (Fig. 6.8) provides flexible optical link connectivity for 24 MGT links of the APd1 ATCA card, approximately 25% of the total, using the standard ATCA RTM form factor. It contains 6 connectors for Firefly links (4 RX + 4 TX variety). It is compatible with links running up to 28 Gb/s and specifically with both 25.78125 Gb/s and 14 Gb/s Firefly modules. The connection with the APd1 card is done via a high-speed connector from the ExaMAX family. The RTM also includes two sockets for legacy optical receivers (Avago AFBR-820, up to 10 Gb/s), 16 general-purpose LVDS Input-Output lines (8 inputs and 8 outputs), and I2C logic and ADC for programming and monitoring optical links and reading values of power voltages. The RTM has passed testing.

APx firmware shell  The APx Firmware Shell (APx-FS) refers to a service environment within the FPGA that provides MGT link instantiation, TCDS connectivity, DAQ support, and an AXI interface to the controlling system. The shell is configurable to accommodate a range of FPGA devices and applications. At the core of the shell is the MGT link instantiation block. It contains the port interface logic, including the 64b66b interface, the protocol engine. The clock-domain-crossing FIFO connecting the link and algorithm interface clock domains, data integrity checking logic (CRC or checksum) are included. The playback and capture memories interfaces that support a finite amount of link data to be transmitted from and/or captured by RAM are also included. Figure 6.9 shows the occupancy of the APx-FS, with the core logic clustered in the clock regions adjacent to the dedicated MGT ports in the FPGA. The interface
6.3. Hardware research and development

Figure 6.7: The Large Lookup Table board with 128GB of DDR4 Memory.

Figure 6.8: The APd1 RTM with 24 MGT Links.
between APx-FS and the algorithm domain is an AXI stream interface of 64 or 128 bits in width. A sideband channel of 8 or 16 bits carries metadata for the stream, including word-valid, frame start/end, orbit start, and frame error flags. TCDS and DAQ connectivity are similar to the approaches used in the Phase-1 upgrade, with the exception that the TCDS connection is now MGT based. The AXI interface is implemented using the concept of sectors. A single sector represents an AXI interconnection point in the shell, servicing a group of MGT ports. The sector concept is useful for managing routing and timing in FPGAs that utilize the Xilinx stacked-silicon interconnect technology, where multiple dies are combined in a single package.

**Simplified framing trigger protocol**  
Simplified Framing (SF) is the name given to a protocol developed by the AP Consortium for CMS trigger applications that provides all three levels of link alignment (word, frame and orbit) with a minimum of overhead, thereby maximizing the bandwidth available for physics data. It is a synchronous protocol that operates at asynchronous line rates, such that the line rate slightly exceeds the underlying synchronous data rate. That small amount of excess bandwidth is used for alignment and transmitter identification purposes.

The SF payload structure is agnostic with respect to data rate or frame size but for the constraint that a frame must contain an integer number of words (64-bit words for 64b66b encoding) in a time interval corresponding to an integer number of bunch crossings. Figure 6.10 lists line/data rates and frame sizes for a 25.78125 Gb/s link operating at different time-multiplexed lengths. For a given line rate, there is an underlying relationship between frame size and the clock speed for the algorithm interface.
rates and frame sizes for different combinations of time-multiplexing interval and interface clock speeds. An added benefit of encompassing all alignment levels within the protocol is that SF-based links are automatically self-aligning, provided that the transmitter and receiver have properly configured TCDS connections. This includes automatic link lock error detection and recovery at the individual channel level.

Figure 6.11: A mosaic of eyescans with an optical loopback on the 76 Firefly channels on the main APd1 ATCA card at a line rate of 25.78125 Gb/s, with retimers turned on.

Figure 6.12: A sample of eyescans for Firefly modules mounted on the RTM, operating without errors at 26.0 Gb/s, with retiming enabled.

**MGT links testing methodology and results** MGT links are tested using the Xilinx Integrated Bit Error Rate Tester (IBERT) core. PRBS-31 patterns are used for all tests, as this type of pattern most closely matches the bit patterns associated with 64b66b link encoding. Optical links that operate at bit rates of 25.78125 or 28.0 Gb/s make use of retimers in the optical modules. The optical modules are therefore capable of recovering a clock from the input bit stream, and using it to time the output bits. Ideally “In-path” retiming restores eye integrity before bit errors occur. Link testing was done with retimers both enabled and disabled. Figure 6.11 shows a mosaic of eyescans with an optical loopback on the 76 Firefly channels on the main APd1 ATCA card at a line rate of 25.78125 Gb/s, with retimers turned on. These links ran for
Figure 6.13: A single 28.0 Gb/s eye on the main board in optical loopback, with retiming enabled.

Figure 6.14: A single 25.78125 Gb/s eye for a main board firefly in optical loopback, with retiming disabled.
48-hours with zero bit errors with minor transmit-side tuning on a small number of channels, after which the test was stopped.

Figure 6.12 shows a sample of eyescans for Firefly modules mounted on the RTM, operating without errors at 26.0 Gb/s, with retiming enabled.

Figure 6.13 shows a single 28.0 Gb/s eye on the main board in optical loopback, with retiming enabled.

Figure 6.14 shows a single 25.78125 Gb/s eye for a main board firefly in optical loopback, with retiming disabled. Disabling the retimers in the transmit and receive fireflies does degrade link performance and can produce errors, but as a test mode it permits transmit-side parameter changes to be visible in the eye scan.

**Power and cooling**  All APd1 power supply rails have test points with both DC and integral AC coupling, to facilitate precision DC level and AC ripple measurements on the supplies under realistic load conditions with all optical modules installed and the FPGA programmed to draw operational levels of MGT and core logic power. Cooling performance is evaluated by loading bitfiles into the FPGA, and measuring the die temperature of the FPGA as a function of the FPGA load power, for a given crate environmental condition. This approach has the advantage of capturing all of the performance elements, including the thermal resistance of the FPGA interfaces, to the heat sink and the PCB, under conditions that test up to the full FPGA capability, beyond that expected in regular use. The test bitfile used for °C/W measurements uses a large amount of internal logic (BRAMs and DSPs) tied to a runtime-configurable internal clock synthesizer. This ensures that the incremental changes in FPGA power are proportional to changes in clock frequency, in a range from 10 MHz to 500 MHz. The delivered VCCINT (0.85V) current to the FPGA core reaches about 105A at the top setting. A target thermal performance of 0.5 °C/W was established. The FPGA heat load of 120W corresponds to a +60 °C temperature rise, for a die temperature of 85 °C at an ambient temperature of 25 °C. Against this criteria, multiple heat sinks were evaluated, including copper heat sinks with long (26 mm) and short (10 mm) fins. At moderate crate fan speeds (5/10 setting), the long-finned heat sink, which is suitable for double-width slot applications, met the 0.5 °C/W target. Whereas, at the 5/10 setting, the short finned heat sink gives a thermal performance of 0.52 °C/W. At higher fan settings, the short finned heat sink performs better than the target.

**Mezzanines (ELM, LLUT, IPMC, ESM)**  The ELM, IPMC and ESM cards all received initial testing in dedicated test fixtures, in which power supply operation and basic connectivity were verified. The dedicated ESM test fixture was used to test and program the ESMs to act as fully functional Ethernet switches on power-up. The IPMC and the ELM boards contain ZYNQ devices are programmed to match the specific ATCA main board. The initial testing in the fixture utilizes generic programming. Those preliminarily qualified devices are installed on specific ATCA cards, like the APd, where the testing of the unit is completed.

**RTM testing**  The RTM has its own power connector for initial standalone tests. The tests completed include power up sequence and voltage measurements, first without and then with all optical modules installed. Once the power has been checked, the RTM is connected to an APd1 board for the tests of main functionality. These tests include Firefly optical link tests at 26 Gb/s and 16 Gb/s, and legacy Avago link tests at 10 Gb/s, running via optical loopback. Serial link tests are performed using Xilinx IBERT core, with PRBS-31 pattern. The LVDS general-purpose I/O interface was tested by using an electrical loopback cable from outputs to inputs, with a special version of test firmware.
LLUT testing The LLUT has its own power connector for initial standalone tests. The tests completed include power up sequence and voltage measurements, first without and then with the DDR4 memory and the Firefly modules installed. A test firmware is then downloaded via the JTAG cable into the FPGA. That firmware is designed to exercise and test the memory modules. The Firefly module is used for the connection to the control computer. Once the memory modules have been tested, the LLUT is connected to the APd1 for the parallel interface tests. These tests are performed using special test firmware.

6.3.2 ATCA processor: Serenity consortium

The development of ATCA blades requires substantial resources to develop not just the hardware, but more importantly the large amount of firmware and software. Given the flexible nature of the cards, where one card can serve many applications, it makes sense for institutions to create common solutions. The Serenity Collaboration was established to realize this objective, initially for applications in the Trigger, Tracker and High Granularity Calorimeter, but increasingly elsewhere in CMS. Serenity (shown in Fig. 6.15) is an ATCA data-processing platform consisting of three elements:

1. An ATCA blade which provides the common board services, specifically: power, clocking, optical interfaces, electrical interconnections between daughter-cards, the intelligent platform management controller (IPMC) functionality and an on-board CPU for higher-level control and monitoring functions.

2. Daughter-cards which host the data-processing elements themselves. These can be changed to suit the application, providing substantial flexibility and the possibility to mitigate risk by splitting the manufacturing of the low cost base board from the high cost FPGAs. They are mounted onto the base board with a new, high density, compression connector from Samtec [87] that was designed to operate at line speeds up to 28 Gb/s. The connector comes in the form of a factory-customized 1 mm thick interposer that is inserted between the PCBs. Miniature springs on each side of the interposer connect an array of 1990 pads on each PCB, which are arranged on a 1 mm pitch.

3. A framework of generic, flexible firmware and software for both the board management (e.g. monitoring of power, temperature, etc) and the infrastructure that surrounds the application specific processing of the daughter cards (e.g. high speed SerDes (Serializer/Deserializer) links to other processing cards and DAQ, fast control & feedback, etc).

Fifteen copies of the Serenity v1.1 board have been produced to date. They have been used to evaluate the design of the hardware itself, develop the framework firmware and software, support the development of application-specific firmware, and carry out system tests, as described in the following sections. Notably, recent tests have focused on the robustness of board to board links at 25 Gb/s (e.g. transfer of 200 Pb over 48 × 25.6 Gb/s optical links with zero errors, corresponding to a BER of less than 1.5 × 10^{-17} at 95 % CL), the full path jitter performance when communicating to frontend electronics (i.e. DTH-Serenity-lpGBT) and board management capability (e.g. graphical tools for real-time monitoring).

6.3.2.1 Hardware

The aim of Serenity is to be as simple and flexible as possible, and to that end, uses Commercial Off-The-Shelf (COTS) components wherever possible. Specifically, the IPMC functionality required by ATCA blades is provided by the CERN IPMC module [88], which itself runs the
Figure 6.15: Left: Serenity without optical heatsinks. Right: Serenity with optical heatsinks. On 2 of the 4 optical banks only 3 out of the 6 possible optical sites are populated. The 2 FPGA processing sites are both covered with large heatsinks to dissipate in excess of 100 W. The CPU is located in the top-right, while the IPMC (DIMM form factor) is in the center-right. Up to 12 MTP12 or MTP24 optical connectors (center-left) provide up to 288 links, each operating at up to 28 Gb/s.

commercial PigeonPoint software [89]. Likewise, the board-level control is provided by a COM Express type-10 computer-on-module which, being an industry standardized form-factor [90], does not restrict us to a single vendor. A single PCI Express lane (Gen-2 or Gen-3) connects the CPU to each daughter card site, and also to a Xilinx Artix-7 FPGA which provides the protocol and voltage conversions necessary to interface to the JTAG and I2C chains which control the board. This arrangement provides a very clean separation of hardware, firmware and software, simplifying the parallel development of the three. Since the COM Express modules used for many of the prototypes produced to date contain an Intel CPU, they run the same x86 build of the Linux operating system as the rack servers of the experiment’s online cluster (currently CentOS7), which simplifies long-term maintenance.

Serenity has been developed as a generic, open processing platform like its predecessor, the MP7 [16]. It has been designed for use in multiple CMS Phase-2 upgrade projects: the Level-1 Trigger system, and the trigger primitive and readout systems for the high-granularity endcap calorimeters [10] and the outer-tracker readout system [8]. These systems all have very different requirements in terms of logic resources, cost considerations and connectivity. To meet such disparate requirements, the chosen solution was to provide a standardized interface to a very large number of connectivity options, which could be used if required and simply ignored if not. This was done by defining a footprint for a standardized daughter-card onto which a FPGA in any package, from any family or generation, from potentially any vendor, could be mounted. The daughter-card design selects which carrier resources are connected to the the FPGA, thus allowing the blade to be tailored to suit the application. Locating the FPGAs on daughter-cards has a second advantage in that, by placing the FPGA on the simple daughter-card, financial risk is reduced by isolating the FPGAs, which constitute the bulk of the cost, from the carrier, which carries the bulk of the potential failure modes.

The connectivity options for the daughter-cards are 144 channels, nominally 72 Tx + 72 Rx, routed to Samtec Firefly connectors [91], 64 differential pairs routed between the two daughter-card sites (the so-called inter-interposer bus) and 2 Tx + 2 Rx channels to a Quad Small Form-factor Pluggable (QSFP) optical module. Although the Firefly signals were given a nominal
Chapter 6. Instrumentation of the Phase-2 trigger system

Figure 6.16: Left: The layout of the top surface of the Serenity ATCA carrier highlighting the two daughter-card sites (labelled North and South although they are functionally identical), the 64 inter-interposer pairs, the positions of the Firefly optical channels (including the optional “spare” channels), the QSFP for DAQ applications, and the on-board CPU. Right: The Serenity ATCA carrier hosting two different designs of daughter cards each utilizing Xilinx KU115 FP-GAs, the “North” receiving data optically and transmitting across the inter-interposer bus and the “South” receiving data on the inter-interposer bus and transmitting optically.

direction assuming the use of 12-channel unidirectional optical Firefly modules, bidirectional and passive electrical (non-directional) modules are also available and compatible with the Serenity design (e.g. the nominal balance of optical modules is 6 Tx and 6 Rx, but this could be changed to 4 Tx and 8 Rx if more Rx links were needed). The inter-interposer bus is passive and makes no assumption on directionality. The ATCA carrier includes provision for an additional 48 channels of Firefly connectivity (nominally 24 Tx + 24 Rx) for each daughter-card, accessed via a minor revision of the carrier that reroutes part of the inter-interposer bus to eight “spare” Firefly connectors. This increases the optical bandwidth by up to 33% at the cost of reducing the inter-site connectivity by up to 75%. These features are shown in Figs. 6.15 and 6.16.

A wide variety of daughter cards have been designed and manufactured, hosting Xilinx Ultra-scale (KU115) and Ultrascale+ (KU15P, VU7P and VU9P) parts with different connectivity (i.e. some have more optical links, whereas others have more inter-FPGA connectivity). A daughter card with a recently released Xilinx Versal part is under consideration. A small production of 15 base boards is underway to accompany the daughter cards, with 9 already delivered to several different locations and institutes. This has enabled rapid progress in the development of the board management control software and firmware (Section 6.3.2.2), which unlike the infrastructural logic cannot easily be developed using commercial development cards. Discussions are currently underway on how Serenity should evolve in a next iteration to ensure it meets the needs of all users, now that the requirements in terms of logic and connectivity are starting to crystallize.

Performance  Given that the primary feature of Serenity is its large amount of serial connectivity, testing the signal integrity at 25 Gb/s was a priority. At these speeds CDR (Clock & Data Recovery) circuits are frequently employed to restore the high speed SerDes signal. This can make analysing the performance of the board more difficult because not all CDR circuits are accompanied by built-in SerDes analysis tools to measure the quality of the received signal
Figure 6.17: A 28 Gb/s eye diagram measured in two different locations along the transmission path. Left: The optical signal measured with a real-time scope. Right: The electrical signal measured at the FPGA-Rx with the Xilinx internal IBERT analysis suite. CDR was used at both the Tx and Rx side of the optical module.

(e.g. by measuring the eye diagram). For the optical links there is a CDR circuit in both the Tx and Rx stages without the debugging capabilities described above. In the absence of these tools the eye diagram was not only measured at the end of the SerDes link (FPGA), but the optical signal was also probed with a large bandwidth optical oscilloscope (33 GHz); examples of these measurements are shown in Fig. 6.17. Given that the optical signal and the electrical signal (Opto-Rx to FPGA-Rx) are clean and that the FPGA-Tx to Opto-Tx path is very similar to the measured Opto-Rx to FPGA-Rx we have a good degree of confidence in the signal integrity.

For some applications outside of the trigger system, Serenity is intended as the clock-distribution system requiring a high-performance (320.624 MHz) LHC clock. A further advantage of having the FPGAs mounted on daughter-cards is that the performance of the carrier and interposers can be measured independently of the performance of the FPGA and firmware. A daughter-card was produced to expose the clock-signals on SMA connectors. A low-jitter (1.3 ps RMS) clock was injected through the ATCA zone-2 connector and the signals measured on the daughter-card using a 6 GHz, 20 GS/s oscilloscope configured to acquire 1.3 million clock-cycles with a large acquisition window (20 million samples) corresponding to a 1 ms continuous acquisition, thereby scanning jitter frequencies above 1 kHz. Across all 18 available clock lines (9 on each interposer site), the channel-to-channel RMS jitter was measured to be 2.8 ps, demonstrating that Serenity can be considered an ideal clock-distribution node.

**Thermal studies**

A test-stand was constructed at CERN to understand the power, cooling, thermal and acoustic limitations of Serenity, and ATCA systems in general. Custom kapton heaters were designed and mounted on aluminium blocks to represent up to 6 optical modules (nominal power of 10 W for 36 links at 14-16 Gb/s and 20 W for 36 links at 25-28 Gb/s) and FPGAs (nominal 90 W, max 120 W). These assemblies were mounted on blank Serenity and daughter-card PCBs, along with off-the-shelf heatsinks and twelve Platinum RTDs to probe the air temperature at various locations (Fig. 6.18 - left). Identical cards, albeit without temperature sensors, were placed on either side of the card under test to mimic the heat load from neighbouring cards. The remainder of the slots were populated with filler cards, which contained a baffle to mimic the air impedance of a real card. The tests were carried out in two shelves, from different manufacturers: one with a nominal rating of 300 W per slot (shelf A - standard product with front to back airflow); and the other 450 W per slot (shelf B - customized product with vertical airflow, which has been used for the ATLAS Phase-1 upgrade). Measurements were made at various fan-speeds and thermal-loads, and the results compared against thermal simulations made with Ansys Icepak [92].
Figure 6.18: Left: Thermal mock-up to validate thermal simulation. Center-left: The maximum temperature measured on the two 90 W FPGAs and four 10 W optical banks on a mock-up of the card. Center-right: The same configuration but from a thermal simulation of the card. Right: The optimized design with pins swapped for fins on the heatsinks, the optical banks enlarged and the power increased to 20 W per optical bank to allow for 25G optics. All measurements and simulations were performed with a fan speed of 10 out of 15 on a standard CERN LHC ATCA Phase-1 shelf.

The first tests, using shelf A, focused on shelf configuration. Specifically, they evaluated the impact of having powered neighbouring cards and whether baffles were needed to balance the air impedance across the shelf. In both instances the impact was negligible with differences of $\pm 1^\circ C$ and $1-3^\circ C$ respectively.

Ansys Icepak simulations were found to match well the observed behavior for both shelves at several different fan settings. For brevity, only the results for shelf B are shown in Fig. 6.18.

Measurements were made with a fan speed of 10 because operating the fans at the maximum setting of 15 consumes considerable power (2 kW) compared to the shelf electronics (~4 kW) and generates significant noise (> 90 dBA) while providing only a modest improvement in cooling performance (15% reduction in the FPGA temperature above ambient). At a fan speed setting of 10, the fans consume 0.5 kW and the noise is ~75 dBA (Fig. 6.19). Designing the blades for operation in these conditions provides a saving of 3 kW for a rack fitted with 2 shelves. This allows 9 kW, rather than just 6 kW, of the 10 kW allocated per rack to be used for electronics. If the fan power was not restricted, then the average blade power in a fully populated rack would be limited to 210 W rather than 320 W.

The measurements indicate that, even with the original naive heatsink designs (Fig. 6.18 - left, center-left and center-right), the FPGAs are below their operating limit of 100°C and the optics can be kept below the 50°C threshold required for them to survive for the 10 to 15 year lifetime of the experiment (see Fig. 6.18 (centre)). However, there is insufficient margin for 25 Gb/s optical modules (i.e. if the 14-16 Gb/s optical modules reach 42°C then 25 Gb/s modules, which need approximately twice the power, are unlikely to stay below the limit) and there is little thermal margin for the FPGA (i.e. it reaches 91°C, which is close to the limit). Using heatsinks based on fins rather than pins and increasing the area of the optical heatsink allows for 25 Gb/s parts and provides significantly more headroom for FPGA power dissipation (Fig. 6.18 (right)).

6.3.2.2 Firmware and software

The firmware and software developed by the Serenity Collaboration has been designed to match the modularity of the hardware platform, and to meet the requirements of a range of applications, from the testing and commissioning phases to long-term operations, while keeping maintenance and support at manageable levels. Modern code development practices, such
6.3. Hardware research and development

Figure 6.19: Left: The fan noise as a function of shelf fan setting for a closed shelf, but not fully closed rack. The increase is \( \sim 3.1 \) dB per fan setting. Right: Fan power consumption as a function of fan setting. In both cases, measurements were made with Shelf B (i.e. shelf used for the ATLAS Phase-1 upgrades).

as build and test automation integrated into the source code version control system, are extensively used to ensure an effective and efficient quality assurance process.

The firmware and software can be split into two functional domains: an infrastructural framework that implements the logic required for synchronized high-speed data transfer, processing and readout on the daughter cards; and I2C, SPI and JTAG interfaces and TCDS multiplexing logic in the Artix-7 Service FPGA. In both cases, the control, configuration and monitoring software runs on the COM Express module, which acts as the gateway for control and monitoring by remote (off-board) control applications over the experiment/laboratory network. Reads and writes are issued to both the service and data-processing FPGAs using the uHAL library and IPbus firmware [93], that have been extended to support PCIe as a transport layer.

**Board management**

The Artix-7 FPGA on the base board is a master of the JTAG, I2C and SPI chains, which are used to program the daughter cards, and configure all non-FPGA on-board components such as the power regulators and optical Firefly modules. The Artix-7 also receives clock and TTC signals from the backplane, and transmits them to the daughter cards; this enables it to act as a board-local TTC and clock source, allowing for synchronized data transfer between the two daughter card sites without a DTH.

Across the projects using the Serenity, the board will have a wide range of physical configurations. The multiplicities and types of optical modules used, as well as the chosen daughter card design, will depend on the use case. The board control and monitoring software must support all possible physical configurations without adding significantly to the complexity of the implementation. This has been achieved by developing a new framework, SMASH (the Serenity MAnagement SHEll), in which all the on-board components, their physical connections, and their bus interfaces are described by configuration files, such that software applications can control and monitor on-board components without the need to hard-code representations of the physical connections and bus topologies in any library. Access to all on-board components via the Artix-7 I2C, SPI and JTAG interfaces using SMASH has been validated. The I2C and SPI functionality was tested by verifying that expected values were read from all registers with fixed values and by confirming that the response of components to writes matches expectations. The JTAG functionality was validated by repeatedly using the SMASH Xilinx Virtual Cable (XVC) server to program firmware images to Serenity daughter cards. The receipt of clock and legacy TTC signals from the backplane, and transmission to daughter-card sites has been tested, and used for synchronized data transfer between daughter cards on different boards via optical links. Firmware that distributes a board-local clock and generates user-specified run-
time defined sequences of TTC commands in the Artix was developed for standalone operation and debugging purposes.

Testing was initially conducted in the lab, but now also takes place in remote ATCA test stands with multiple cards and users. Modern industry-standard, open-source frameworks for collecting and displaying monitoring data (see Fig. 6.20) have been used to simplify the process of validating the stability of these systems.

**EMP framework** Application firmware on the daughter card FPGAs will need to send and receive data to/from other FPGAs in a synchronized manner, and send input/output data for a subset of events to a DTH module. The EMP (Extensible, Modular data Processor) framework consists of highly-configurable firmware designs and software libraries, designed to provide infrastructure for use cases from multiple CMS upgrade projects. The use of this common firmware and software implementation across multiple applications reduces the total effort expended developing and maintaining infrastructural components, and allows application developers to focus solely on application-specific functionality. The design and implementation of the EMP framework are based on the MP7 framework, which is successfully used across several different sub-systems of the Phase-1 Level-1 trigger upgrade. The top-level EMP firmware designs instantiate the EMP infrastructural components, and connect them to a user-created payload firmware block – e.g. a particle/event reconstruction algorithm – that conforms to a standard payload interface. The multiplicity and static configuration of the infrastructural firmware components (e.g. clock frequencies and the multiplicity of I/O buffers) are controlled by build-time constants whose values can be changed for each use case.

The major components from the EMP firmware are shown in Fig. 6.21. External clock and TTC signals are received by the TTC block, which has the ability to capture the TTC command history and inject commands for debugging purposes, as well as to locally generate the periodic TTC commands that are required for standalone testing of data-processing algorithms. The payload’s principal data ports are connected to the FPGA pins via the datapath logic that includes buffer memories and the link firmware blocks. Data can be transferred between all input/output buffers and the on-board computer via the control bus. The datapath logic can be configured at runtime to synchronously play arbitrary data from the buffers into the payload or transceivers, and the buffers can be configured to capture the output of the payload or transceivers. The link firmware blocks use the FPGA transceivers to transmit and receive data.
with fixed, low latency over 16 Gb/s and 25 Gb/s links according to the asynchronous protocol described in Section 6.3.4. Source code or IP cores that implement an application-specific payload (typically particle reconstruction algorithms in the trigger) are integrated into generic application-agnostic top-level designs by referencing the application-specific code in a build tool configuration file along with a file that specifies the top-level design and the board-specific configuration and constraints.

The framework is designed such that it can be easily ported between different devices, FPGA families and boards with minimal code duplication, which notably allowed algorithm firmware to be tested on development boards before several prototypes of the Serenity were available. The framework has been used on three different FPGAs (Xilinx KU115, KU15P and VU9P) situated on two development boards (the HiTech Global K800 and the Xilinx VCU118) and the MPUltra, in addition to the Serenity daughter cards. Support for other boards and FPGAs, including the Serenity VU7P daughter card, is also planned.

The EMP firmware is controlled and monitored using an object-oriented software library, which includes a command-line interface that is used for testing both the infrastructure and application-specific algorithms. The core functionality of the framework has been validated, including that required to test payload firmware, receive clock and legacy TTC signals from other boards, and transmit and receive data over electrical/optical links at 16 Gb/s and 25 Gb/s using the protocol specified in Section 6.3.4. The EMP framework has been successfully used to test and validate several algorithms from different stages of the trigger, including: the TMTT track finder algorithm, the ‘FastHisto’ vertexing algorithm (Section 6.4.3), and a particle-flow jet clustering algorithm. In order to test the algorithms, data from offline simulations are written to the input buffers, and the datapath logic is configured to play these data into the payload. Data from the payload are then captured in output buffers, the contents of which are read and compared with the output values predicted by the offline simulation. The framework also includes a testbench that allows one to simulate the algorithm logic with inputs specified and outputs recorded using the same file format for I/O data as is used with hardware.

The EMP framework is validated using a Python test suite that uses the EMP software library to repeatedly reset and configure the main firmware components in each of the operational modes required for the supported high-level use cases. The test suite reads status registers to check for erroneous values and inconsistencies, and whenever feasible, reads datapath buffers, TTC command history memories and other block memories to check that their contents are consistent with the relevant patterns loaded into ‘upstream’ firmware components. For example, the link firmware has been validated by repeatedly resetting and reconfiguring it, playing known
patterns into the links, recording the data received at the other end of the links and comparing them with expectations, whilst monitoring for erroneous values of status registers. At present, these tests involve a relatively small number of links (e.g. two cards with 48 links), but are run for several hours. As more boards are produced, these tests must be repeated in larger systems in order to identify and fix any rarer problems that may not have been uncovered to date.

CMS will have \(~1000\) cards and many more optical links. There will be occasional bit errors. As long as these do not have a significant impact on the trigger rate (either missing or additional triggers) then they are acceptable. Assuming we want these unwanted triggers to be less than 0.01%, then a bit flip rate of 100 Hz for the whole of CMS would be acceptable, however it could probably be much larger. In many instances a single bit flip is unlikely to have a large impact (e.g. if the least significant bit of a calorimeter energy deposit is changed it is highly unlikely to change whether a L1A is issued). There will also be occasions when links or entire cards malfunction. In these cases CMS should be capable of running in a “degraded” mode, which may still be acceptable for physics analysis. While at the same time trying to automatically fix the part that has failed.

\subsection*{6.3.3 ATCA processor: Ocean}

Ocean is a general-purpose ATCA processor featuring a large System On Chip (SoC) that contains programmable FPGA logic for trigger algorithms together with a multi-core processing system that can be used for configuration, monitoring, and real-time data analysis (40 MHz scouting) at HL-LHC. The SoC is connected to on-board optical transceivers providing a total I/O bandwidth of 3 Tb/s. Trigger processors of typical design feature large FPGAs and a management device (SoC, CPU or small FPGA) that is used to control the peripherals, load bitstreams, and monitor the system. The philosophy of the Ocean design to merge management and processing devices into a single chip reduces the cost and the complexity of the board. The Ocean design is also inline with upcoming trends in FPGA technology as shown by Xilinx and Intel’s recent announcements of their next generation chips that feature high speed links up to 112 Gb/s, large FPGA logic, and faster multi-core processors. The purpose of the Ocean R&D program was to study the single chip approach with the technology available today. Although the board is designed as a general purpose processor for the CMS L1 Trigger, the presence of many different processing elements in a common package with a dense interconnect extends its purpose to other applications such as heterogeneous computing and real time data analysis (scouting).

\subsubsection*{6.3.3.1 Main features}

The processing power of the Ocean blade is provided by a Xilinx ZYNQ Ultrascale+ SoC (ZU19EG-2) that provides similar programmable logic (PL) and bandwidth as two Phase-1 trigger boards at the cost of one Phase-1 trigger board. The chip also features a quad core ARM Cortex-A53 processor that runs at 1.3 GHz, a dual core real time processor, and a small GPU. The processing system (PS) is connected to a SoDIMM memory module supporting DDR4 memory up to 16 GB. The different elements of the SoC (Processor, programmable logic, memory) are connected through a high speed interconnect featuring several AXI4 lanes. Those AXI lanes support bus widths up to 128 bits and can run at a frequency up to 333 MHz, providing capabilities for configuration, monitoring and real-time scouting. Some of the available AXI lanes provide low-latency high-throughput paths from the programmable logic to the processor cache making Ocean a platform for possible future heterogeneous computing applications. The board features several peripherals for debug and control such as SD card, Serial ATA SSD drive, USB, Display Port, and serial UART.
Figure 6.22: Ocean prototype. The ZYNQ Ultrascale+ SoC is located in the middle of the board while the service mezzanine with the Artix 7 is located at the top right. The GTY transceivers are connected to the Firefly modules on the left and the GTH transceivers are connected to the modules on the right. The MTP connectors are located in the top and the bottom of the front panel.

The I/O bandwidth of 3 Tbit/s is realized by 28 GTY transceivers that can reach speeds up to 28 Gbit/s and 44 GTH transceivers that can run up to 16 Gbit/s. The full bandwidth of the SoC is exploited by connecting all GTY and GTH transceivers to Firefly on-board optics provided by Samtec. The GTY transceivers are instrumented with one optical module per quad (4 transmitters and 4 receivers) that supports rates up to 28 Gbit/s. The 28 Gbit/s optical module features clock and data recovery (CDR) circuits that improve signal integrity in a small rate window that is configurable to be around 25 Gbit/s or 28 Gbit/s. Each pair of 28 Gbit/s modules is connected to a single MTP connector (Y module) delivering 16 fibers to the front panel. The 44 GTH transceivers are connected to four Firefly modules that support 12 transmitter lanes and four firefly modules that support 12 receiver lanes resulting in eight modules that reach rates up to 16 Gbit/s. The output optical cables of a transmitter-receiver pair are combined to a common MTP connector of 24 fibers forming a Y module. Using Y modules in GTH and GTY requires only 8 MTP connectors in the front panel.

The backplane interface to the DTH requires three lanes operating at a maximum data rate of 10 Gbit/s. While those lanes could be connected to high speed GTH or GTY transceivers, the design choice made was to implement a wide LVDS bus from the ZYNQ to a small FPGA that will serve as SerDes and TCDS simulator. This way, a high speed transceiver quad is not sacrificed for services and can be used for physics. For the first version of the Ocean blade, a
mezzanine featuring a small Artix 7 FPGA is used (Figure 6.22). The mezzanine is connected to the main board through two Samtec Razor Beam connectors. Three 16 bit LVDS buses connect the Razor beam connectors to the ZYNQ serial I/O banks achieving the required rate for the backplane signals. The Artix 7 GTP transceivers can run at rates up to 5 Gb/s.

The clock network receives four clocks: the LHC and high-quality LHC clock from the backplane, a 40 MHz on-board oscillator, and an external clock input in the front panel. All the clocks are buffered and supplied to both the ZYNQ and the Artix 7. A pair of clock synthesizers is used to create the frequency required by the GTH, GTY, and Artix GTP transceivers. The connectivity of the transceiver clock signals was chosen so that it is possible to implement both synchronous and asynchronous links with respect to the LHC clock. Both clock synthesizers and optical modules support control through an I2C bus. The processing system supports two I2C buses that are used to connect and manage clocks and optical modules through software running in the processing system.

The management of the blade is done through standard IPMI using the CERN IPMC mezzanine. The IPMC mezzanine reads all sensors on the blade and performs power operations. The sensors include temperature, humidity, voltage, current, and power. Two I2C buses on the IPMC are connected to the sensors and the power regulators. The first bus is used for monitoring and the second one is used for power sequencing. Both the ZYNQ and the IPMC support Ethernet connections that are routed through a 5-port ethernet switch on the board. One port of the switch is connected to the ZYNQ, one to the IPMC, two ports to the backplane, and one port to an external RJ45 connector.

The board is powered by a voltage of -48V through the ATCA chassis and two on-board modules provide power for the IPMC and 12V power that is distributed to the components via on-board switching regulators. The core FPGA voltage of 0.85 V is supplied through a single micro-module by Analog Devices (LTM 4700) that supports a maximum current of 100 A which is more than sufficient for the ZYNQ Ultrascale+ SoC. While we do not expect challenges in cooling the SoC, a custom heatsink with a height of 17.3 mm was designed to exploit all the available space in the z direction while complying with the ATCA specifications. The heatsink is designed to keep the temperature of the SoC well under 100°C with an airflow as low as 1 m/s (20% fan setting) when the power is the maximum obtained from the Xilinx Power Estimator (50 W).

### 6.3.3.2 PCB layout

The board was designed with a 16 layer stack-up resulting in a board thickness of 2.4 mm. The high speed signals have been routed using both microstrip and stripline configurations and power planes were combined with signal traces in several inner layers. Panasonic Megtron 6 dielectric material was chosen for its low dielectric losses and a plane rotation of 90° was applied to compensate for the effects of the fiberglass weave. The priority while placing the components was to achieve the shortest possible length for the 28 Gb/s traces and to allocate enough space to efficiently route the DDR4 memory connections.

### 6.3.3.3 Performance

The design was completed in March 2019 and the first assembled prototype was received in June. Testing started by verifying the voltages of all power rails in a board with no SoC using the software implemented to read on-board sensors via the IPMC mezzanine. The ethernet switch was also tested by verifying the communication of both ethernet backplane ports with the IPMC mezzanine and the ZYNQ. Subsequently, the board was powered and programmed
through JTAG. The next step was to verify the quality of the layout by studying the signal integrity of the board. Tests on both GTY and GTH transceivers were performed using the integrated Bit Error Rate Test (IBERT) framework by Xilinx. The GTY links were tested both 25 and 28 Gb/s while the GTH links were tested at 15.7 Gb/s. PRBS 31 bit sequences were used for both transceivers. Each test lasted for 20 hours and resulted in no errors for any of the

![GTY QUAD 130, CH0, 28Gbps, CDR](image1)

![GTH QUAD 228, CH3, 15.7Gbps](image2)

Figure 6.23: Left: Ocean eye diagrams for GTY transceiver at 28 Gb/s. Right: GTH transceiver at 15.7 Gb/s. Derived using the IBERT core. The GTY transceiver has CDR enabled. In both cases, the transmitters are connected to the receivers of the same quad through optical loopback.

72 links for the default configuration. Figure 6.23 shows the IBERT results for GTY and GTH transceivers. The GTY results have been obtained for a rate of 28 Gb/s with the CDR enabled and the GTH results have been obtained for a rate of 15.7 Gb/s. The quad transmitters were connected to the quad receivers with an optical loopback module.

While the nominal operation of the links will exploit the presence of the CDR, tests were repeated for all GTY links without it to stress further the signal integrity of the board. Figure 6.24 shows the results of the tests with CDR disabled. No errors were observed for all links up to rates of 25 Gb/s even without CDR demonstrating excellent link performance. At the maximum rate of 28 Gb/s supported by the Firefly modules, six out of the 28 links had errors. Since it is not recommended to operate the optical modules at 28 Gb/s without the retimer, it is not clear if the degradation of performance can be attributed to signal integrity or to limitations of performance of the optical modules.

![GTY QUAD 129, CH2, 25Gbps, CDR disabled](image3)

![GTY QUAD 129, CH3, 28Gbps, CDR disabled](image4)

Figure 6.24: Left: Ocean eye diagrams for GTY transceivers at 25 Gb/s and, Right: for 28 Gb/s with the retimers disabled.
6.3.3.4 Management and configuration infrastructure

The main philosophy behind the Ocean design was to perform all management, configuration, and monitoring using the processing system and use the programmable logic explicitly for trigger algorithms and high speed optical I/O. The board is managed by a software component residing in the processing system and a firmware IP core residing in the programmable logic. Those two components communicate with each other using the AXI interconnect by mapping an area in the processor memory to registers in the programmable logic. Both physical on-board peripherals (such as clock synthesizers and optical modules) and firmware components are managed by a common software application that supports communication through I2C, SPI, and AXI direct memory access.

Figure 6.25: Snapshot of the Ocean framework IP core that controls clocks, transceivers, I/O buffers and algorithm configuration

The IP core running in the programmable logic (Figure 6.25) is an AXI slave that communicates with the firmware modules (algorithms) and other FPGA components such as high speed transceivers. The IP core uses the Hermes link protocol (see Section 6.3.4.4) for both GTH and GTY transceivers delivering a range of different speeds and settings. For each Rx-Tx pair, the IP core adds buffers to support data playback and capture for debugging the firmware during runtime. The Artix 7 FPGA on the service mezzanine is managed by the same tools by implementing a physical AXI chip-to-chip bridge between the ZYNQ and the service mezzanine. The device constraints including pin connections and floorplanning of firmware components are applied automatically during implementation based on the specific configuration of the IP core.

6.3.3.5 Future developments

After the successful tests of the Ocean prototype, a second version is implemented, upgrading the Artix 7 mezzanine with a Kintex 7 FPGA directly mounted on the board. The Kintex 7 FPGA supports transceivers that run at rates up to 10 Gb/s satisfying the specifications of the backplane links for TCDS2 and DAQ. In addition, an SFP optical channel is added to support legacy TCDS for integration with the Phase-1 running trigger system. The second version of the board will be installed in CMS during LS2 and Run-3 implementing a Phase-2 muon trigger chain in parallel with the standard operations.

6.3.4 Specific processing board: Barrel Muon Trigger (BMT) demonstrator

In addition to the multi-purpose/generic ATCA bladed discussed previously, and within the scope of Phase-2 R&D, a new Level-1 Trigger development processor board was designed, by the University of Ioannina/Athens CMS Trigger team, to provide a hardware environment for
developing and evaluating new Level-1 muon trigger designs and technologies. The board comes with state-of-the-art fiber optics technologies, using micro footprint optical interconnects. Control and monitoring of the board will be carried out by a new instance of a Cell, similar to those that have been used during Phase-1, making use of the IPBus interface. A new instantiation of a Cell has been implemented using packages from the L1 repositories (such as uHAL and the SWATCH core libs). In addition, the related library which holds the controller for the EMP firmware is integrated too. This board is described in detail in the following sections.

For testing purposes, a new firmware is also integrated, implementing asynchronous 16 Gb/s GTH links. The links use a 64b/66b encoding scheme with an overhead of 2 coding bits per 64 bits that is considerably more efficient than the previously-used 8b/10b encoding scheme. The link’s speed can increase to 25 Gb/s, if the firmware is implemented in an FPGA which integrates GTY transceivers that support that speed.

Furthermore, after the successful design and testing of the development board, the group has embarked on the design of a prototype processor board for the Layer-1 barrel muon trigger. The new board under design will be able to handle at least 125 inputs and will host all the stub-finding algorithms. Based on these requirements, we foresee a Layer-1 system which has 30 ATCA boards. In addition these boards will serve as the readout of the muon detector to the DAQ.

6.3.4.1 Hardware

The board is powered by the XCKU040 Kintex UltraScale FPGA, providing 20 GTH transceivers, that reach speeds up to 16.3 Gb/s. The high performance interconnect system uses active optical engines over 12 full-duplex channels, at data rates up to 16 Gb/s. Furthermore, four FPGA transceivers are routed to a QSFP28 connector, allowing data rates of up to 16 Gb/s per channel over 4 channels to be reached. In total, the board’s 16×16 Gb/s links add up to a total optical bandwidth of approximately 256 Gb/s in each direction, making it a high-performance all-optical data-stream processor (see Fig. 6.26). A Xilinx ZYNQ SoC device is used as the control interface for the Kintex UltraScale FPGA. The system controller configures and queries on-board resources, such as the power controllers and programmable clocks.

6.3.4.2 PCB

The board provides 32 high speed differential pairs running at 16 Gb/s. The Firefly transceivers were placed very close to the right side of the FPGA, achieving closer proximity to the Banks where all the MGTs reside, to simplify board layout and enhance signal integrity. For the substrate, the Panasonic Megtron-6 was chosen, due to the excellent high-frequency performance and impedance properties. A ground plane has been placed between each pair of layers containing high-speed traces, resulting in a 16-layer stack-up. All high-speed differential pair route lengths were matched by using serpentine routing as seen in Fig. 6.27. Finally, to avoid additional signal distortion caused by the plates through-hole (PTH) vias, the excess stub material of the vias was removed, using a technique known as back-drilling (Drill from the rear through the plated through hole with a drill large enough to remove the copper from the via).

6.3.4.3 Clocking

The board includes four low-jitter programmable clock sources as seen in Fig 6.28. The GTH transceivers connected to the high-speed Firefly modules are clocked by a dedicated, low jitter, quad clock generator (Si5338). A low-jitter frequency generator (Si570) is connected to the
Chapter 6. Instrumentation of the Phase-2 trigger system

Figure 6.26: The Layer 1 Barrel Muon Trigger demonstrator board hosting a Xilinx KU040 FPGA located at the center of the board. Twelve transceivers are connected to the two FireFly modules located at the bottom center of the board. Four more GTH transceivers are connected to the QSFP28 module located at the bottom left corner. The four DDR4 RAM modules are placed near the right side of the FPGA and the ZYNQ XC7Z010 SoC controller is placed on the left side. The right half of the board is populated mainly by the power modules and the programmable clock generators.

Figure 6.27: PCB details. Left: the TOP layer of the PCB with all the major components. Right: Detail of the PCB where the serpentine routing is visible.

QSFP28 transceivers and can also be used as a secondary clock source to the FireFly transceivers. A jitter attenuator (Si5328B) is used to reduce the jitter of a received recovered clock. A fixed frequency clock source (Si5335A) provides four frequencies (33.3 MHz, 90 MHz, 125 MHz, 300 MHz) that can be used as a free running clock for reset and initialization FSMs. Finally, an SMA external clock input is also available. All programmable clocks are accessed through a dedicated I²C bus. The Si5328B and the Si570 can be programmed using the ZYNQ system controller and the Si5338 from the Kintex FPGA.

6.3.4.4 High-Speed optical links

The data-interface consists of 16 optical links, operating in 16 Gb/s, making use of four of the MGTs available on the Kintex Ultrascale FPGA. Twelve (12) of the optical links are routed to the FireFly optical flyover assembly next to the FPGA. The FireFly configuration consists of 12 separate transmitter (TX) and receiver (RX) optical modules, joined in a “Y” configuration and terminated onto a single 24 fiber MPO connector. The connectors are placed mid-board and the data “fly” over the PCB, allowing easier routing to be performed. Four more MGTs are routed
6.3. Hardware research and development

Figure 6.28: The clock distribution tree of the L1 BMT demonstrator board.

to a QSFP28 (Finisar FTLC9551REPM) transceiver module. This module has a hot-pluggable QSFP28 form factor and supports 103.1 Gb/s of aggregate bit rate. However, the rate is limited by the MGT's maximum speed to 64 Gb/s.

The FireFly optical links were tested using the integrated Vivado serial I/O analyzer (IBERT). Eyescans taken with the IBERT are shown in Figs. 6.29 and 6.30, for line-rates of 10 Gb/s and 16 Gb/s respectively. For the scans we used a 10 m MTP cable to create a 20 m-long optical loopback path.

Figure 6.29: Eyescans for all twelve FireFly optical links at 10 Gb/s.

6.3.4.5 Firmware implementation and tests

The CMS Trigger groups identified 16 Gb/s and 25.78125 Gb/s line rates with 64b/66b encoding for wide use in Phase-2 trigger applications. The trigger protocol and the amount of physics payload carried over the links have important implications for algorithm logic and algorithm developments. For these reasons, a new payload framing link protocol was developed along with the firmware infrastructure. The L1 BMT board was extensively used for the development and testing of this protocol. The protocol uses a 64b/66b encoding scheme with an overhead of 2 coding bits per 64 bits. CRC (Cyclic Redundancy Check) error detection logic was also included. The design was named after the Greek mythological god Hermes, who was the emissary and messenger of the gods.

Hermes is a lightweight, link-layer protocol that can be used to move data point-to-point across
one or more high-speed serial lanes. It supports simplex operation with continuous data transfer. The links are asynchronous, meaning that the main algorithmic logic is clocked with a lower frequency than the link clock, allowing more flexibility to be achieved when choosing the logic clock. This is achieved by using asynchronous FIFOs in the receiving and transmitting sides. To compensate for the difference in frequencies, padding words are injected on the transmitting side and are stripped away on the receiving side. The link initialization, synchronization and error handling are based on the checking of those padding words and the 2-bit header of the 64b/66b encoding.

For testing purposes the local clock is running at 240 MHz and the link clock at 250 MHz. The link encoding transforms 64-bit data to 66-bit line code to provide enough state changes to allow reasonable clock recovery and alignment of the data stream at the receiver to be achieved. The protocol overhead of 64b/66b encoding is 2 coding bits for every 64 payload bits or 3.125%.
This makes the encoding considerably more efficient than the 20% overhead of the previously-used 8b/10b encoding scheme, which added 2 coding bits to every 8 payload bits. In Fig. 6.32, the Hermes firmware implementation is shown in the KU040 FPGA die. All 5 quads of the FPGA are instantiated.

Figure 6.32: Hermes firmware implementation in the FPGA die with all 5 quads of the XCKU040 FPGA instantiated. In orange the Hermes links and in green the service firmware (ipbb).

The Hermes Framing protocol is a general purpose protocol offering a simple way to send data grouped in packets (frames). However, an alternative protocol was designed to meet the needs of sub-systems that stream data without the requirement of framing. The simple One-Word protocol uses just one extra coding word or SoF (Start of Frame) containing the CRC code. In addition, to avoid adding extra words as a payload, the CRC/SoF word is periodically injected as a padding word, as required to cross between processing and link clock domains. Special care must be taken to inject the CRC/SoF word at a frequency lesser than the maximum padding rate. For this simple protocol, the CRC checksum must be generated and checked at the link domain, where the padding words are injected at the transmitter. At the receiving end, the CRC/SoF padding words are first checked for errors and then stripped from the data stream, before crossing to the processing clock domain.

6.4 Trigger demonstrators and slice tests

The following sections describe tests of the algorithms and architecture choices discussed in previous chapters. In many cases, algorithms were initially tested using the available Phase-1 upgrade hardware, generally based on Xilinx Virtex-7 FPGAs, and later tested on the prototype Phase-2 hardware, described earlier in this chapter. In some limited cases, tests were possible during Run-2 CMS data-taking as described. In most cases, algorithms were tested in single cards, simulating the data-flow from upstream and downstream systems by filling input and output data buffers with simulated data. A few preliminary multi-board tests with prototype hardware are described, and for each subsystem, a program of future tests is provided. As during Run-2, it will also be possible to test certain part of the upgrade system during CMS Run-3 data-taking. This possibility is also discussed in the following.
Figure 6.33: The simplified scheme of the connection of the CTP7 cards in the demonstrator.

Figure 6.34: The test stand with 3 CTP7 cards.
6.4. Trigger demonstrators and slice tests

6.4.1 Calorimeter trigger demonstrators

The calorimeter trigger is responsible for collecting information from the barrel (ECAL, HCAL), Endcap and HF calorimeters. It is designed as a tiled structure with two layers, the regional calorimeter trigger (RCT) and the global calorimeter trigger (GCT) as described in Section 5.2. The RCT receives inputs from the barrel calorimeter: the crystal-level information from the ECAL and the tower-level information from the HCAL, and partitions it into 36 cards, each covering \(17\eta \times 4\phi\) towers region. Each RCT card receives \(17\eta \times 4\phi \times 25\) ECAL crystals and \(16\eta \times 4\phi\) HCAL towers, and creates Egamma clusters and towers with unclustered energy as described in Section 3.2. The data from \(17 \times 4\) towers and 12 highest-\(p_T\) clusters are sent to the GCT via four links as described in Section 5.2. The GCT collects data from the barrel calorimeters via RCT as well as data from the endcap and HF calorimeters. A total of 10 GCT cards are needed to cover the entire calorimeter.

For demonstration purposes the RCT and GCT algorithms were implemented in firmware that runs on Phase-1 Virtex7-based CTP7 cards. The first version of the firmware for VU9P and APd1 card, see Section 6.3.1, is currently under development. The demonstrator described in this section was restricted to the barrel part. The firmware was implemented using C High Level Synthesis tools. The clock rate was optimized to achieve minimum latency. Each CTP7 card has 48 I/O links running at 10 Gb/s with the 8b10b encoding scheme, 3.6k DSPs, and \(\sim 433k\) LUTs available. Because of the limited number of input links available, a scaled down version of the RCT algorithm was tested in the demonstrator setup. The geometry of the RCT was scaled down from \(17\eta \times 4\phi\) to \(5\eta \times 4\phi\). A version of the GCT algorithm was also deployed. The RCT and GCT algorithms run at a clock rate of 160 MHz and achieve a total latency of 10 and 18 LHC bunch crossings respectively. Table 6.1 shows a comparison of the resource estimates of the Virtex 7 FPGA for the RCT and GCT algorithms. A three card setup was successfully tested where the output of two RCT cards was forwarded to a GCT card as shown in Fig. 6.33. The test stand is shown in Fig. 6.34.

Table 6.1: Resource Utilization for RCT and GCT in a Virtex7/CTP7 and VU9P/APd1 card.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>LUT</th>
<th>FF</th>
<th>BRAM</th>
<th>DSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCT (CTP7)</td>
<td>22%</td>
<td>17%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>RCT (APd1)</td>
<td>4%</td>
<td>3%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>GCT (CTP7)</td>
<td>48%</td>
<td>30%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

The Phase-2 calorimeter trigger demonstrator will be using APdx cards (see Section 6.3.1) with VU9P FPGA running at 240 or 360 MHz. Similar test of the resources requirements as described above was performed on the RCT algorithm for APd1 card, the results are presented in Table 6.1. A latency of 66 (40) clock cycles was measured for 360 (240) MHz clock respectively.

A more realistic setup using APdx boards is currently under test. The links are operated at 25 Gb/s and with 64b/66b encoding as expected in the full design. The firmware shell is described in Section 6.3.1. The demonstrator hardware and firmware shell allow both RCT and GCT to be implemented. Trigger primitive generation can be simulated as well. In this test stand also multiple-card configuration can be tested using inputs based on simulated \(tf\) events with 200 pileup.

6.4.2 Muon trigger demonstrators: muon track finder and global muon trigger

6.4.2.1 Demonstration of the DT local trigger primitives

During the second LHC long shutdown, four DT chambers (MB1 to MB4 of a single DT sector) have been instrumented with Phase-2 on-board DT electronics (OBDT) [11] to setup a demon-
Chapter 6. Instrumentation of the Phase-2 trigger system

strator of the upgraded system (DT slice test). In the two innermost chambers of the DT slice test, the legacy on-board DT electronics have been fully replaced by Phase-2 prototypes, which were installed with a setup that is as close as possible to the final one. This is done to maximize the acquisition of expertise that will be needed to integrate the new components in the full upgrade. In the two outermost chambers of the sector, the signals coming from a fraction of the chamber frontends were split and sent in parallel to the legacy and to the Phase-2 electronics. This setup allows for an event-by-event comparison of the response of the two readout and trigger chains.

By means of optical link connections, the OBDT streams TDC hits with 25/30 ns time bins from the detector directly to the backend electronics in the service cavern. At present, the backend electronics of the DT slice test consists of boards hosting Virtex 7 XC7VX330T-3FFG1761E FPGAs. These boards, called AB7, are identical to the ones used for the DT Phase-1 upgrade of the readout and local trigger (TM7), but run a dedicated firmware version. Each AB7 board runs both event building and a complete version of the DT Analytical Method (AM) algorithm for trigger primitive generation previously described in Section 2.4.1. The logic resources required for event building and trigger primitive generation within a chamber are presented in Table 6.2. A diagram that describes the architecture of the DT slice test is shown in Fig. 6.35 (right). Within the current setup, one AB7 board processes information from up to three OBDTs, therefore a single AB7 covers each of the three innermost chambers (MB1, MB3 and MB3). Two AB7 boards are required for the larger outermost chamber (MB4), which is instrumented with four OBDTs, labelled in Fig. 6.35 top (T) and bottom (B), depending on which AB7 board they are connected to.

Table 6.2: Resource utilization for the full readout and trigger logic based on the Analytical Method algorithm in a Virtex 7 AB7 board.

<table>
<thead>
<tr>
<th>AM algorithm Virtex 7</th>
<th>LUT</th>
<th>BRAM</th>
<th>DSP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>61056 (30%)</td>
<td>71 (10%)</td>
<td>164 (15%)</td>
</tr>
</tbody>
</table>

The performance of the trigger primitive generation by the AB7 firmware is measured comparing to: (i) the response of the legacy trigger, (ii) the offline DT segment reconstruction, and (iii) the emulator for the AM algorithm developed in C++ and integrated in the CMS offline Software Framework (CMSSW). As an example performance figure, Fig. 6.36 shows the difference between the crossing time of a cosmic muon computed by the DT offline reconstruction and by the trigger. Results from both the AB7 firmware (blue) and the legacy trigger (red) are presented. For the legacy trigger, which is only capable of BX identification, the corresponding BX assignment is multiplied by 25 to obtain a time measurement in units of ns. The plot shows a clear improvement for Phase-2, due to the sub-BX precision achieved at trigger level. For the Phase-2 trigger, the time resolution can be further improved by a channel-by-channel inter-calibration of the response of each DT cell, that has not been deployed yet into the AB7 firmware.

The AB7 firmware has also been validated in the laboratory using injected data based on a CMS 2016 proton-proton collision sample. This data, originally recorded with a fixed latency with respect to the L1 trigger accept signal, was modified to simulate the continuous streaming of TDC hits produced by the Phase-2 on-detector electronics.

The fitting logic of the AM method was tested by comparing the trigger primitives from the AB7 firmware and the emulator code that are built by the same set of TDC hits. The results, presented in Fig. 6.37, show exact match for all probed trigger primitive quantities.

The bunch crossing identification block of the DT HB algorithm (described in Section 2.4.1)
6.4. Trigger demonstrators and slice tests

Figure 6.35: Diagram comparing the DT readout (RO) and trigger chains as they are in the Phase-1 system and in the DT slice test. The central part of the figure highlights how many Phase-2 on-board DT electronics (OBDT) boards are installed in each DT chamber. The left part of the diagram describes the Phase-1 system: readout and trigger primitive generation are performed by the legacy on-board DT electronics (Minicrates). Information, transmitted by optical fibers from the detector balconies to the counting room, is further processed independently for readout (µROS) and trigger (TwinMux). The right part of the diagram describes the slice test: TDC data is streamed by each OBDT to AB7 boards, hosted in the counting room, which are in charge both of event building and trigger primitive generation.

Figure 6.36: Difference between the crossing time of incoming cosmic muons computed by the DT offline reconstruction and the DT local trigger. Results, coming from the DT Slice Test setup, are provided for both the Phase-2 (blue) and the legacy (red) trigger. A minimal quality requirement, asking for a trigger primitive to be reconstructed using at least all four DT layers of a single superlayer, out of the eight layers available from the two DT $r - \phi$ SLs, is applied.
Figure 6.37: Left: BX as obtained by the firmware (red squares) or the emulator (blue lines), computed relative to the event BX-ID of proton-proton collision data injected on an AB7 board. Right: DT trigger primitive position versus the position obtained by the emulator of the AM method computed using the same data as in the left plot. The insets in the plots show the difference between the time information computed with sub-BX precision (left) and the position (right) obtained with firmware and emulation. The bin width of the insets, corresponds to the trigger resolution. Exact match is found for both histograms.

was implemented on a Kintex UltraScale XCKU115-2 FPGA and tested with cosmic muons. This FPGA equipped the backend of a setup operating in the INFN laboratory in Legnaro and consisting of four reduced-area DT super-layers, called mini-chambers, designed around the same technology as the CMS Drift Tubes. Two Xilinx VC707 boards, running a firmware similar to the one of the OBDT, are used to emulate the new DT Phase-2 TDCs, providing unfiltered streaming on fiber using the GBT protocol. The trigger was adapted to read and process the GBT-encoded TDC. In terms of FPGA occupancy, 16K Look-Up-Tables are needed to process a group of 18 wires, called a macro-cell. A dynamic allocation of downsized macro-cells, currently under design, will further reduce this area. The setup collected data without interruption for 25 hours and 45 minutes with no errors. The agreement between emulated triggers and the actual hardware was 98.3% for the output crossing time and 99.4% for the trigger quality.

Figure 6.38 shows the parent muon BX assignment performance of this latter setup. Figure 6.38 (left) shows the distribution of hardware triggers with respect to the time of the high quality coincidences in external mini-chambers. Figure 6.38 (right) shows the drift time within a DT mini-chamber, computed with respect to the BX assigned by emulated triggers and hardware triggers.

6.4.2.2 Demonstration of the kalman filter BMTF in CMS data-taking

The Barrel Kalman Muon Track Finder algorithm discussed in Section 3.3.2 was implemented in firmware and demonstrated in hardware during Run-2 operations in 2018.

The algorithm implements two major operations in firmware: track propagation and parameter update. Track propagation requires about 800 multiplications that are relatively expensive in firmware resources. The resource utilization for those operations was substantially reduced by mapping them to the DSP slices present in the FPGA. Each XC7VX690T FPGA contains 3.6k DSP slices, where each slice can perform the operation $x + y \times z$ with no additional resource
usage. Since there are four stations with two stubs each and a well-defined track requires at least two stubs, there exist 22 possible track combinations with at least two stubs. All those combinations were implemented in parallel track reconstruction chains. Each track parameter update in any of those track chains corresponds to a different Kalman gain, which was precalculated and mapped into one block RAM. Tracks from different chains were then compared for overlapping stubs and lower quality tracks were rejected. Final muon candidates were sorted using a bitonic sorting algorithm. The firmware was implemented with C High Level Synthesis tools. The clock rate was optimized to achieve minimum latency. The version of the algorithm that was deployed runs at a clock rate of 160 MHz and achieves a total latency of 6.5 LHC bunch crossings.

The KBMTF algorithm was implemented in the Phase 1 BMTF system that runs in 12 MP7 boards featuring Virtex 7 690T FPGAs. Figure 6.39 shows how the two algorithms were implemented in the same chip. Muon stub data arrive over the 10 Gb/s optical links from the
Twinmux processor and are formatted and deserialized inside the chip. The incoming muon stubs are split and sent to both algorithms that process the event with different latencies. The output of the default BMTF algorithm is forwarded to the next layer of the L1 Trigger chain, and the output of KBMTF is stored in the DAQ.

Table 6.3: Resource Utilization for both track finders in a XC7VX690T FPGA.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>LUT</th>
<th>FF</th>
<th>BRAM</th>
<th>DSP</th>
<th>Latency (bx)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legacy BMTF</td>
<td>43%</td>
<td>23%</td>
<td>35%</td>
<td>0%</td>
<td>6.5</td>
</tr>
<tr>
<td>Kalman BMTF</td>
<td>16%</td>
<td>11%</td>
<td>15%</td>
<td>25%</td>
<td>9.0</td>
</tr>
</tbody>
</table>

Table 6.3 shows a comparison of the resource utilization of the Virtex 7 690T FPGA. The usage of 25% of the DSP cores results in a smaller utilization for the Kalman filter even if it is a more complicated algorithm. Both track finders occupy about half of the FPGA, leaving a significant margin for improvements. The total latency including serialization and deserialization was 9.6 bunch crossings.

Figure 6.40: Left: Agreement between data and emulator for the muon transverse momentum $p_T$. Right: Comparison of the trigger efficiency as a function of the transverse impact parameter in 2018 data for the vertex unconstrained KBMTF trigger vs the vertex constrained KBMTF and BMTF triggers (right).

The KBMTF physics and firmware performance is studied using CMS data by analyzing the KBMTF candidates that were collected with orthogonal triggers in CMS. Figure 6.40 (left) shows the agreement between data and the C++ software emulator that is measured to be better than 99%, and Fig. 6.40 (right) shows the displaced trigger performance of the KBMTF algorithm as measured in cosmic muon data in 2018. The presence of a vertex unconstrained measurement provides improved muon trigger efficiency for large displacement validating the expected results presented in Section 3.3.2 with data.

In the upcoming Run-3 of the LHC, one sector of the detector will be instrumented with the Phase-2 Drift Tube electronics. The new trigger primitives will be readout in parallel with the nominal trigger primitives. The plan is to continue this work by interfacing the KBMTF algorithm with the new trigger primitives during Run-3, to study the impact of the new trigger primitives and possible algorithm adaptions towards an HL-LHC CMS barrel muon trigger.

### 6.4.2.3 Demonstration of the global muon trigger concept

This section describes an implementation of a Global Muon Trigger in the Ocean prototype described in Section 6.3.3. The algorithm implemented for track-muon matching is the TPS
algorithm described in Section 3.3. The system is implemented assuming a time-multiplexing period of 18, as described in Chapter 5.

Figure 6.41: Left: Data flow and main payload components of the GMT Demonstrator firmware. Right: configuration of the loopback in a single demonstration board.

Figure 6.41 (left) shows the main components implemented in the demonstrator. The system receives tracks from the global track trigger layer over 9 links at a rate of 25.6 Gb/s and muon trigger data over 24 links running at a rate of 16 Gb/s. The track finder tracks have a data width of 96 bits. Each 64 bit word arrives from the link at a clock rate of 360 MHz. The track data are unpacked delivering two tracks every three words, resulting in an algorithm clock of 240 MHz and a total number of 972 tracks to be processed by the GMT for each bunch crossing.

The muon data in this demonstrator are received before the track finder tracks in the form of Regions-Of-Interest (ROI) that result from clustering the muon hits in different detector layers. Each ROI is represented by a set of global coordinates \((\eta, \phi)\) for the inner stub and position and quality information for muon stubs up to 11 detector layers that cover muon tracks crossing the whole detector in every possible region in pseudorapidity. Each ROI is represented by 384 bits arranged as six words of 64 bits. The 24 links at 16 Gb/s deliver 72 ROIs in 18 clock cycles at 240 MHz. The TMT period of 18 corresponds to 108 clocks at 240 MHz, therefore there is sufficient time to serialize all ROIs to memory before the ROIs of the next event appear at the same board.

The TPS algorithm starts by propagating the coordinates and the resolution of each track to all 11 layers using the DSP cores in the FPGA. The two closest ROIs are found by accessing the register file of the coordinates. Consequently, the ROI stubs are fetched from the memory.Stub matching to track finder tracks is implemented and muon candidates are formed based on the number of matched muon stubs. Since most tracker tracks are not matched to any muon stubs, a very large suppression of the output rate is observed at the end of the track matching step, therefore a zero suppression layer is implemented. The muon candidates that pass basic identification requirements are collected in a set of 32 registers in the demonstrator. After the event has been processed up to 32 muons are propagated to the next steps of overlap cleaning and sorting. The zero suppression layer outputs data once per bunch crossing or once per 108 clock cycles at 240 MHz.

Exploiting the long TMUX period of 18 allows us to reduce the resources required for overlap cleaning and sorting, reusing the logic up to 108 times before producing the output. Muons with overlapping stubs are compared based on their qualities and stubs are removed from the lower quality muons. Each muon is compared with all remaining 31 candidates. The same comparison logic block is reused 32 times to perform overlap cleaning for all 32 muon candidates.
Finally, sorting of muons is implemented with an odd-even sort algorithm. One odd-even sort step is implemented and reused 16 times providing a sorted muon collection. Output muon candidate data are serialized through a single output link at 25.6 Gb/s. The total latency of the firmware components is 10 bunch crossings. This latency is added to the latency required to stream all 972 tracks resulting in a cumulative latency of 28 bunch crossings (or 0.7 $\mu$s) for the whole demonstrator, which is within the latency budget for the Global Muon Trigger.

Table 6.4: Resource utilization of the GMT demonstrator in Ocean prototype (ZU19EG-2 device) including management framework, high speed links, loopback infrastructure and TPS algorithm firmware.

<table>
<thead>
<tr>
<th>Name</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUT</td>
<td>21%</td>
</tr>
<tr>
<td>Flip-flop</td>
<td>17%</td>
</tr>
<tr>
<td>Block RAM</td>
<td>43%</td>
</tr>
<tr>
<td>DSP</td>
<td>13%</td>
</tr>
</tbody>
</table>

The demonstrator was implemented in a single Ocean board with loopback through optical fibers as shown in Fig. 6.41 (right). The input data are first entered in buffers through the playback and capture interface of the Ocean framework. The buffers are connected to the high-speed transmitters sending the data out. The data return to the board through the high speed receivers and are fed to the firmware block. The output data are then stored in capture buffers that are readout through the Ocean framework.

Table 6.4 shows the resource utilization for all components including high speed links, framework infrastructure, and algorithm firmware. Figure 6.42 shows the mapping of the individual components of the firmware on the ZYNQ Ultrascale+ FPGA. The total resource utilization with the assumptions considered for this demonstration is very modest, leaving enough space in the FPGA for implementing additional components.
6.4.2.4 Demonstration of the endcap muon track finder

The Phase-2 algorithm for muon reconstruction in the CMS endcaps (EMTF++), as described in Section 3.3.4, improves the Run-2 Endcap Muon Track Finder by introducing more granular patterns for track identification and reconstruction and by using a neural network (NN) for the estimation of the muon momentum. The EMTF++ algorithm is being rewritten in Vivado HLS to assure maintainability and flexibility of development throughout the HL-LHC operations, but the corresponding firmware is not yet complete. Therefore, the hardware demonstrator tests use the Run-2 EMTF firmware in combination with the NN $p_T$ assignment. The first demonstration setup used the current MTF7 [94] micro-TCA hardware, and the second demonstration used the ATCA prototype described in Section 6.3.1.

An experimental setup based on the MTF7 board [94] was used first to test the NN performance on a Virtex-7 device. The MTF7 was programmed with a control firmware that is interfaced with the algorithm firmware. The control firmware reads test data in the form of 2048-bit words from a buffer memory implemented in the FPGA, forwards them to the implemented algorithm, and stores the result back in the buffer memory. Data can be written and read to/from the MTF7 board via a PCIe interface, effectively allowing extensive testing of the algorithm performance with a large number of events to be performed. The inherent latency of 2 clock cycles from the control firmware is subtracted when reporting the results. A picture of the setup installed at the University of Florida is shown in Fig. 6.43.

![Figure 6.43: Picture of the endcap muon track finder demonstrator based on the MTF7 hardware at University of Florida.](image)

The NN model is converted into firmware using the HLS4ML software [44] and the resulting HLS is then translated to Verilog and synthesized for the MTF7 together with the demonstrator control firmware. The resource usage is reported in Table 6.5 for both the standalone NN and the combined synthesis of NN and EMTF firmware for the Virtex 7 FPGA and the Virtex 9 Ultra-scale+ FPGAs. The NN mostly uses digital signal processing (DSP) resources, complementing the usage of resources of the EMTF track reconstruction firmware. The latency achieved by the NN evaluation corresponds to 48 clock cycles when running the device at a clock frequency of 200 MHz.

The output of the NN implementation in the MTF7 board was compared to the prediction from the HLS simulation using a set of $2 \times 10^4$ muon patterns. A perfect agreement was found between the simulation and the implementation for the predicted $p_T$ value.

The combined NN and EMTF firmware was also synthesized for a target VU9P device as re-
Table 6.5: Summary of the resource usage for the standalone NN and NN + EMTF firmware for the synthesis in a Virtex 7 and Virtex 9 Ultrascale+ devices. For comparison, the resource usage for the EMTF without NN in the Virtex 9 Ultrascale+ device is shown in the last column.

<table>
<thead>
<tr>
<th>Algorithm (target FPGA)</th>
<th>LUT</th>
<th>Flip-flop</th>
<th>Block RAM</th>
<th>DSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN (V7)</td>
<td>11%</td>
<td>5%</td>
<td>7%</td>
<td>54%</td>
</tr>
<tr>
<td>NN + EMTF (V7)</td>
<td>71%</td>
<td>26%</td>
<td>62%</td>
<td>58%</td>
</tr>
<tr>
<td>NN + EMTF (VU9P)</td>
<td>28%</td>
<td>8%</td>
<td>30%</td>
<td>30%</td>
</tr>
<tr>
<td>EMTF (VU9P)</td>
<td>18%</td>
<td>6%</td>
<td>13%</td>
<td>1%</td>
</tr>
</tbody>
</table>

Reported in Table 6.5. A single NN was synthesized, so the usage of DSP resources should be increased by a factor of two if NNs are to be used for the $p_T$ assignment of both prompt and displaced muons. This estimation does not include any optimization of the resource usage; for example, the NN can be pruned to remove connections with low importance and reduce the resource usage, or one of the two NNs can be implemented with precomputed values in the large DDR external memories of the Phase-2 endcap muon hardware.

The EMTF firmware (without the additional NN) and the ancillary test setup firmware required to load and readout test data were ported from the MTF7 setup to the ATCA APx prototype described in Section 6.3.1 for demonstration purposes. The optical link protocols were kept the same as for the MTF7, namely that the CSC muon inputs (49 links) were kept at 3.2 Gb/s synchronous to the 40 MHz clock and the muon output link to the GMT was kept at 10 Gb/s asynchronous. For initial demonstration purposes, the 10 Gb/s RPC data links from the concentrator preprocessor and fan-out (CPPF) system were not enabled. Data were written to and read from buffers via the ELM2 control interface described in Section 6.3. The CSC data were sent out from the playback buffer via the Firefly optical links and looped back as inputs to the EMTF algorithm. Likewise, the GMT output data link was looped back to the FPGA to be recorded. The set up is shown in Fig. 6.44. A sample of 600 events were sent in this test, and the hardware output matched the Verilog simulation exactly for all events. The sample size was chosen to verify the correctness of the implementation, while further tests with more events are planned in a subsequent stage with an updated firmware to study in detail the performance of the implemented algorithm.

The resource usage of the EMTF firmware in the VU9P FPGA is shown in Table 6.5, and a visual map of the FPGA usage is shown in Fig. 6.45. Concerning the complete EMTF++ firmware, an estimation of its resource usage can be obtained by considering that most of the resources are used for the comparators that implement the pattern finding algorithm. Consequently, the amount of LUT and BRAM resources used is expected to scale linearly with the number of these comparators, which is proportional to the number of input values to be compared. Considering the new muon detector inputs, we estimate the usage to be about 50-60% for LUT and BRAM resources.

### 6.4.2.5 Demonstration of the muon and track correlator

The correlation of tracks reconstructed from muon chambers and silicon tracker sub-detectors is demonstrated in different setups. One demonstrator implements the track and muon correlation in a MTF7 board, and another implements the correlation with a CTP7 board. In both cases, the correlation is implemented in a Virtex-7 device.

The MTF7 demonstrator uses the same setup as the one described in Section 6.4.2.4 and shown in Fig. 6.43 for the standalone endcap muon track reconstruction. The implementation is designed to accept tracks from each of the nine tracker sectors, correlating the input tracks in
Figure 6.44: Test setup for the EMTF demonstrator using a prototype APx ATCA card.

Figure 6.45: Image of the EMTF firmware usage in the Virtex 9 Ultrascale+ FPGA. Shown are the EMTF logic core (yellow), input links (red), input link data deframers (light pink), control link to the ELM2 (magenta), and other logic (green).
parallel to the standalone muons. Each correlation requires the looking-up of precomputed values that describe the region of a valid match, effectively encoding the propagation of the track spatial coordinates and the expected muon scattering. As such, the same implementation is easily extensible to the implementation of the track and muon stub correlator algorithm.

A summary of the resources used by the implementation is reported in Table 6.6. The implementation considers 12 standalone muons matched in parallel to the tracks coming from 9 tracker sectors. Each sector is assumed to transmit serially 18 track candidates.

The implementation is validated using a set of muon patterns generated from a MC simulation of muons in a collision environment with 200 pileup interactions on average. The efficiency achieved in the demonstrator is reported in Fig. 6.46.

Table 6.6: Summary of the resource usage for the muon and track correlator demonstrator based on the MTF7 hardware. The resource usage corresponds to the firmware for a Virtex-7 device.

<table>
<thead>
<tr>
<th></th>
<th>LUT</th>
<th>Flip-flop</th>
<th>Block RAM</th>
<th>DSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage</td>
<td>13%</td>
<td>2%</td>
<td>3%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Figure 6.46: Efficiency of the muon correlator in the MTF7 demonstrator. The efficiency is shown for a set of muon patterns from a MC simulation with 200 PU interactions on average, and is shown for reconstructed objects of $p_T > 20 \text{ GeV}$ as a function of the generated muon transverse momentum.

A second demonstrator was realized using the CTP7 board. Simulated data were encoded in binary files and injected into the board with dedicated software. A loopback injection scheme was implemented with the test data being sent out on the optical links and read back by the CTP7, thus verifying the synchronisation of input and output links. The correlator firmware runs on the injected data and the output is saved in buffer memories for subsequent analysis. A picture of the CTP7 setup and a scheme of the data flow are shown in Fig. 6.47.

A summary of the resource usage of the CTP7 correlator demonstrator is reported in Table 6.7. A total latency of the algorithm of about 7 bunch crossings for a clock speed of 240 MHz was achieved.
6.4. Trigger demonstrators and slice tests

(a) Picture of the setup  
(b) Data flow scheme

Figure 6.47: Left: Picture of the CTP7 setup at Texas A&M. Right: schematic representation of the data flow.

Table 6.7: Left: Summary of the resource usage for the CTP7 correlator demonstrator for the correlation of $6 \mu$ candidates with 24 tracker tracks. Right: 36 $\mu$ candidates with 60 tracker tracks.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>LUT</th>
<th>Flip-flop</th>
<th>Block RAM</th>
<th>DSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$6\mu + 24$ TT</td>
<td>4%</td>
<td>2%</td>
<td>8%</td>
<td>4%</td>
</tr>
<tr>
<td>$36\mu + 60$ TT</td>
<td>40%</td>
<td>16%</td>
<td>19%</td>
<td>10%</td>
</tr>
</tbody>
</table>

6.4.3 Global track trigger demonstrator

The Global Track Trigger (GTT) is responsible for collecting all of the charged particle tracks from an event and performing the reconstruction of track stand alone (SA) objects, such as the primary vertex for the event, track-based jets, and track-based $E_T^{miss}$, as described in Sections 3.4 and 3.6. In this section, we describe the performance of hardware demonstration tests to reconstruct these objects. First, demonstrations of vertex finding algorithms in standalone tests are described. Second, we show results for track jets, track $E_T^{miss}$, and vertexing algorithms integrated into a common framework.

6.4.3.1 Vertex finding demonstrator

The implementation of the ‘Fast Histo’ algorithm described in Section 3.4 was tested on different boards supported by the EMP Framework, which is described in Section 6.3.2.2. These include the Xilinx VCU118 development kit (with VU9P FPGA), the MPUltra, and Serenity (both with KU115 FPGAs). In all cases the Fast Histo firmware was inserted as the ‘payload’ into the package, with no changes to the algorithm firmware.

For testing, tracks were digitized and written into pattern files. Several consecutive events were written back-to-back in each file, to study any effects which might overload the firmware. These were written to the input link buffers in the FPGA using the EMP Framework software. After propagating the data through the algorithm firmware, the data written to the output buffers were captured using the same software.

A simple Python implementation of the Fast Histo algorithm was used to compare performance with the firmware. This created a numpy histogram of the $z_0$ of all tracks in each event, weighted by the track $p_T$. In the software, the $z_0$ and $p_T$ were digitized identically to the firmware. This comparison was designed to use the same quantized mathematical operations
as the firmware, but with no timing information such as the ordering of tracks. Any effects due to truncation or overloading of the firmware would not be emulated by the software, and would cause a discrepancy in the comparison.

![Graph](image)

Figure 6.48: Left: Distributions of primary vertex $p_T$ and, Right: $z_0$, from the Fast Histo firmware and the software reference. Out of 1500 events, two events are observed with a mismatch in the found vertex $z_0$.

For testing, 1500 $t\bar{t}$ events with 200 average pileup interactions were injected into the algorithm according to the procedure described. Distributions of the $p_T$ and $z_0$ of the primary vertex, as found by the Fast Histo algorithm in these events, are shown in Fig. 6.48. The firmware and software implementations of the algorithm produce excellent bit-wise agreement, with the $z_0$ of the found primary vertex differing in only two events.

### 6.4.3.2 Integrated track standalone demonstration

The GTT is designed to implement multiple reconstruction algorithms on the same board. We demonstrate this functionality with the simultaneous reconstruction of the primary vertex, track jets, and track $E^\text{miss}_T$. The integrated track standalone demonstration is based on the APx prototype hardware described in Section 6.3.1. It includes an infrastructure firmware shell, which takes inputs on the optical links and passes them between the link clock domain and the algorithm clock domain using FIFOs controlled by an AMC13 interface. This allows the link clock and algorithm clock to be operated at different clock frequencies, each optimized separately. The link clock was set to 320 MHz reading in packets with 64b/66b encoding, assuming links are operating at 25 Gb/s. The L1 track objects are 96 bits arriving at 213 MHz, which is used for the algorithm clock after the clock domain boundary for the demonstration described in this Section.

In the proposed architecture described in Section 5.4, tracks arrive at the GTT on 18 links. We simulate this configuration by storing the track patterns for simulated collision events, as described in Section 2.1, in 18 streams and load them into 18 FIFOs on the demonstration board. The tracks are then read into separate firmware blocks for each algorithm on the board. In the demonstration described here, separate blocks of firmware are tested for primary vertex finding, track jet reconstruction, and track $E^\text{miss}_T$ reconstruction as shown in the schematic in Fig. 6.49.
This structure allows for the independent development of the algorithm for each block to be performed without the need to redesign the others and gives the flexibility to place different algorithms on each of the two GTT boards that serve a single event in the proposed architecture. In the example here, the algorithm for the track $E_T^{\text{miss}}$ has been written in Vivado high level synthesis (HLS), the algorithm for the track jet finding has been written in traditional HDL (verilog), and the algorithm for the primary vertex finding has been written in a combination of HDL and HLS. We describe the design and performance of the three algorithm blocks below.

Figure 6.49: Block diagram of Global Track Trigger demonstrator board. Tracks are read in from the track-finding system and used in separate blocks of algorithm firmware to find the primary vertex, track jets, and track $E_T^{\text{miss}}$. The products of the reconstructed algorithms are passed out to the Global Trigger.

The demonstration of track $E_T^{\text{miss}}$ follows the algorithm described in Section 3.6.3.1. Since the primary vertex is required to select tracks consistent with it, the tracks are stored in the 18 FIFOs while the vertex reconstruction is performed. When the vertex is available, track quality selection and primary vertex consistency are applied. The selected tracks are passed to the $E_T^{\text{miss}}$ algorithm, which performs the vector sum of their $p_T$ simultaneously for the 18 streams of tracks. The resulting central value and direction are computed to determine the track $E_T^{\text{miss}}$ for the event. The performance of the firmware algorithm is compared to the expectation from the bit-wise emulation of the algorithm. An exact match is obtained for the track $E_T^{\text{miss}}$ for a sample of 50k $t\bar{t}$ + 200 PU simulated events, as shown in Fig. 6.50.

The FPGA resource usage for the track $E_T^{\text{miss}}$ algorithm is shown in Table 6.8, and is within 3% or less for all categories on the VU9P. The algorithm has been run at clock speeds up to 360 MHz, and completes in a latency of 146 clock cycles after the arrival of the primary vertex. This corresponds to a total of 171 (267) clock cycles from the arrival of the last (first) track for the event, which is within the requirement to arrive at the Global Trigger in time for the event decision.

The demonstration of primary vertex finding utilizes the implementation described in Section 3.4.4.1. The tracks are streamed into the algorithm from 18 FIFOs and processed imme-
The FPGA resource usage for the vertex algorithm is shown in Table 6.8, and is within 7% or less for all categories on the VU9P. The algorithm has been run at clock speeds up to 360 MHz, and completes in a latency of 25 (121) clock cycles after the arrival of the last (first) track for the event, which is within the requirement to be passed to the Correlator Trigger Layer-1 in time to perform the PUPPI pileup rejection.

The demonstration of reconstructing track jets follows the algorithm described in Section 3.6.1.2. The tracks are streamed into the algorithm from 18 FIFOs containing tracks from the 18 fibers per event from the track finding boards. The tracks are processed by the jet finding algorithm which performs a two-step jet clustering technique performed in bins along the $z$ axis. The incoming tracks are assigned to the overlapping $z$ bins. For each $z$ bin, the algorithm first searches for clusters in the $\phi$ dimension in slices of $\eta$, followed by clustering in $\eta$ as well. The clusters are built into a list of candidate track jets. The performance of the firmware algorithm is compared to the expectation from the bit-wise emulation of the algorithm. An exact match is obtained for a sample of 200 $t\bar{t}$ simulated events at 200 PU. For this initial demonstration, no track quality selection is applied in the algorithm, though a process similar to that described for the track $E_T^{\text{miss}}$ will be implemented for the final design.

The FPGA resource usage for the track jet algorithm is shown in Table 6.8, and is within 15% or less for all categories on the VU9P. The algorithm has been run at clock speeds up to 360 MHz. It completes in a latency of 175 (271) clock cycles after the arrival of the last (first) track for the event, and is capable of beginning the processing of a new event every 36 clock cycles, which is within the requirement to arrive at the Global Trigger in time for the event decision.
Table 6.8: Resource utilization of the VU9P FPGA for the GTT demonstrator system.

<table>
<thead>
<tr>
<th>VU9P</th>
<th>FF</th>
<th>LUT</th>
<th>BRAM</th>
<th>DSP</th>
<th>f_{clk,max} [MHz]</th>
<th>Latency (clocks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Available</td>
<td>2364480</td>
<td>1182240</td>
<td>2160</td>
<td>6840</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>FW shell</td>
<td>165548 (7%)</td>
<td>128899 (11%)</td>
<td>392 (18%)</td>
<td>0 (0%)</td>
<td>320</td>
<td>–</td>
</tr>
<tr>
<td>MET algo.</td>
<td>31921 (1%)</td>
<td>35752 (3%)</td>
<td>2 (0%)</td>
<td>74 (1%)</td>
<td>360</td>
<td>146</td>
</tr>
<tr>
<td>Jets algo.</td>
<td>153890 (7%)</td>
<td>135117 (11%)</td>
<td>325 (15%)</td>
<td>0 (0.0%)</td>
<td>360</td>
<td>175</td>
</tr>
<tr>
<td>Vertex algo.</td>
<td>29137 (1%)</td>
<td>83741 (7%)</td>
<td>1 (0%)</td>
<td>1 (0%)</td>
<td>360</td>
<td>121</td>
</tr>
<tr>
<td>MET+jets+PV (1 copy)</td>
<td>380496 (16%)</td>
<td>383509 (32%)</td>
<td>860 (33%)</td>
<td>75 (1%)</td>
<td>213</td>
<td>175</td>
</tr>
<tr>
<td>MET+jets+PV (3 copies)</td>
<td>810392 (34%)</td>
<td>892729 (76%)</td>
<td>1376 (64%)</td>
<td>225 (3%)</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

The total resource usage for the integrated GTT demonstration is shown in Table 6.8 in the row “MET+jets+PV (1 copy),” which includes one copy of each of the three algorithms, plus the associated infrastructure firmware shell and track parameter conversion modules needed to convert from the format arriving from the track finding boards into that needed for the GTT algorithms. The resource placement footprint for this demonstration is shown in Fig. 6.51. The design architecture requires each GTT board to handle three time slices from the track finding system. This can be accomplished by including three copies of each GTT algorithm. Scaling from the one copy demonstration, the resource usage for that configuration is also shown in Table 6.8, though this configuration has not been demonstrated in the hardware.

![Resource placement on VU9P FPGA for the GTT demonstration. Red shows the firmware shell plus track conversion modules, green shows the primary vertex finding algorithm, blue is the track jet finding algorithm, and yellow is the track $E_T^\text{miss}$ algorithm.](image)

Figure 6.51: Resource placement on VU9P FPGA for the GTT demonstration. Red shows the firmware shell plus track conversion modules, green shows the primary vertex finding algorithm, blue is the track jet finding algorithm, and yellow is the track $E_T^\text{miss}$ algorithm.

All algorithms on the GTT board require up to 450 ns (18 bx) to read in all tracks for the event, plus the algorithm processing time. After 450 ns, the tracks from the next event will begin to arrive at the GTT board. The system must process multiple events at a time, as depicted in the latency diagram in Fig. 6.52. The FPGA resources are reused for subsequent events as the needed information is passed to later steps in the algorithms.

6.4.3.3 Plans for multi-board demonstrators

The demonstrators described above for the GTT are based on single boards with simulated inputs. To test the full chain of the GTT system, we will perform multi-board demonstrations. A first board-to-board demonstration was performed using two Serenity boards with KU15P FPGAs. One board was used to perform vertex reconstruction, as described in Section. 6.4.3.1, and the other ran the particle-flow and PUPPI algorithms, described in Section. 3.5.3.1. The
Vertexing and MET Timing Example

GTT Firmware Concept

Figure 6.52: Latency diagram for the Global Track Trigger demonstrator board. Tracks stream in from the track finding system over a period of 18 bunch crossings (bx) for each event. They are processed to identify the primary vertex for the event, as well as a collection of track jets and missing energy. The system begins processing the next event after 18 bx and will simultaneously work on multiple events streaming through the board.

Figure 6.53: A diagram of the multi-board Vertex Finder to particle-flow and PUPPI set-up.

PUPPI algorithm uses the particle-flow candidates and the position of the Primary Vertex (PV) to scale the $p_T$ of the neutral candidates, so correct receipt and synchronization of the PV with the PF + PUPPI algorithm is required to obtain the correct PUPPI particles at the output. A single 25 Gb/s link was used to transmit the $z_0$ of the Primary Vertex. A schematic diagram of the setup is shown in Fig. 6.53.

Several changes were required compared to the single board tests of the same algorithm components. For the Vertex Finder, the output PV $z_0$ was linearly transformed into the units expected by the PF and PUPPI block using a single DSP and small amount of logic. Additionally, the PV was sent repeatedly over several frames to match the sequential computation over regions of the PF and PUPPI algorithm. In a final design, this PV duplication could be shifted to the board hosting the PF and PUPPI algorithm. The PF and PUPPI algorithm deployed for this test was a smaller version of the intended algorithm, accepting 14 tracks; 10 electromagnetic calorimeter clusters; 10 combined calorimeter clusters; and 2 muons. Since these objects are sourced from a pattern file, whereas for this test the PV was sent over a link, a synchronization step was required. A BRAM buffer was added for each ‘link’ sourced from a pattern file. This RAM was written while the data arriving from its source was ‘valid’, and read while the data arriving from the link on which the vertex was sent was ‘valid’. The synchronization block
pipelined the vertex link by the 2 cycle latency of these buffers, such that the track, calorimeter and muon objects were aligned with the vertex at the output.

For the test procedure, this buffering layer meant that the tracks, calorimeter objects and muons were first written to the link buffers on the PF and PUPPI board. They are read into the algorithm, which holds them in the additional buffers awaiting the vertex. After this, the tracks are written into the link buffers on the vertexing board, read into, and processed by, the vertexing algorithm and output onto a 25 Gb/s optical fiber. When the vertex is received in the receiving board, the vertex, tracks, calorimeter objects and muons then propagate through the PF and PUPPI algorithm, with the PUPPI candidates written into buffers at the output links. Events of \( \text{t}\) with 200 average pileup interactions were used.

Figure 6.54 compares the PUPPI \( p_T \) captured at the output buffers of the PF + PUPPI board - after receiving the Primary Vertex position from the first board running the Vertex Finder - with the PUPPI \( p_T \) expected given the same inputs processed by the HLS simulation of the PF + PUPPI algorithm. In the simulation, the Primary Vertex position is computed by a software reference of the Fast Histo algorithm. Exact agreement is observed between the two collections of particles. While it is generally to be expected that the behavior of a Xilinx HLS design should be identical between a CPU simulation and in the FPGA, the exact matching here additionally shows that the Primary Vertex was computed, transmitted, and synchronized correctly across the two boards.

This multi-board demonstration will continue to be made more realistic in the future. In the final system, the Global Track Trigger and Correlator Layer-1 would be expected to receive track inputs approximately simultaneously. The PF algorithm and Vertex Finder can process simultaneously, because the PV is needed only for the PUPPI algorithm, which also uses the PF candidates. It is expected that the Vertex Finder latency will be smaller than the PF latency, so the synchronization step used for this multi-board demonstration can be reversed. The Primary Vertex can be buffered, and read out when valid PF candidates are produced. This synchronization implementation will additionally benefit from the splitting of the PF and PUPPI
algorithms into separate HLS IP cores, such that the synchronization will take place outside of the algorithm IPs themselves.

## 6.4.4 Correlator trigger demonstrators

### 6.4.4.1 Demonstration with legacy hardware

Multiple components need to be tested in the context of a demonstrator for the Correlator Trigger system. The most crucial for Layer-1 are the particle-flow and PUPPI firmwares, which have been successfully tested using a demonstrator setup comprised of 3 CTP7 boards. For this test the CTP7's were arranged as shown in Fig. 6.55, with two boards each providing inputs to the third board. The inputs from each board were sent over 24 links operating at 10 Gb/s with 8b/10b encoding. This demonstrator was used to test versions of the particle-flow and PUPPI firmware designed for a smaller number of expected inputs; specifically 15 tracks, 15 electromagnetic calorimeter clusters, and 15 combined calorimeter clusters. In the test, the first two boards were designed to simply pass inputs over links to the third board running particle-flow and PUPPI. This algorithm was run at a clock frequency of 120 MHz with an initiation interval of three clock cycles. Simulated tt events at an average pileup of 140 events per bunch crossing were used to generate inputs, and 100% bit-wise agreement between emulated firmware and the results from running on hardware was achieved.

To conduct this test the firmware shell shown in Fig. 6.56 was used. This firmware shell takes inputs from the optical links and passes them to both the “link clock domain” and the “algorithm clock domain” using FIFOs. This allows the link clock and algorithm clock to be operated at different clock frequencies, each optimized separately. In the tests conducted for particle-flow and PUPPI, the link clock was set to 250 MHz, and the algorithm clock was set slower, at 120 MHz. In the firmware shell, the link alignment and the clock domain crossing are controlled by an AMC13 interface. The firmware shell allows data to be used as input to the algorithm to come either from the links or from dedicated BRAM memory buffers. The input data can also be stored in buffers before being sent to the algorithm. Similarly, the output of the algorithm block can be stored in buffers before being sent along to another board in the system. The output sent by a board can also come directly from a different set of dedicated memory buffers. The actual algorithm core code used can be in the form of HDL or HLS. For tests of particle-flow and PUPPI the algorithms used were written in HLS. The physical setup of CTP7's used for this test is the same one shown in Fig. 6.34.
A ZYNQ on the CTP7 is used for the slow controls needed for this test. In particular, it is used to specify the desired operating mode and the inputs and outputs to the algorithm block. Specifically, after the desired bitfile for a particular algorithm is programmed, the ZYNQ is able to configure the location of the inputs and destination of the outputs (optical links or BRAMs). The ZYNQ is also used to place the desired inputs into the input playback buffers in the case of their use, and to read the output buffers in the analogous configuration. It is also capable of further specific memory access to the Virtex-7 FPGA on the CTP7, which allows for detailed debugging to be conducted.

The same configuration has also been used to test a basic implementation of a tau lepton identification algorithm. In this test, the first two boards were used to run particle-flow and PUPPI and then pass the PUPPI particles over links to the third board running a neural network designed for tau lepton identification. In addition to the actual neural network, the third board was also implemented with a realistic sorting and preparation of the input PUPPI particles as would be necessary in a Correlator Layer-2 board. Again, the link clock was set to 250 MHz and the algorithm clock was set at 120 MHz. These tests represent a simplified but effective test of all basic components of the Correlator design.

6.4.4.2 Demonstration with expected hardware

While the CTP7 setup described above enables the testing of small versions of the particle-flow and PUPPI firmware, it is not capable of also testing the regionizer in a meaningful way due to the limited resources. To establish the viability of certain configurations for the regionizer, particle-flow, and the PUPPI firmware, a simulation of the complete setup with VU9Ps and a placeholder for the firmware shell was created. While this was useful for the purposes of prototyping, a full design with firmware to operate the links and manage other board signals was necessary.

To accurately simulate the complete Correlator Layer-1 design, a more realistic board setup using an early revision of the APd1 board was used. Here the links were operated at 25 Gb/s, with 64b/66b encoding as expected in the full design. The firmware shell itself is very similar to the one shown in Fig. 6.56, with the major exception being additional control over the link clock and algorithm clock frequencies. The additional updates to both the demonstrator hardware and firmware shell allow all components of the Layer-1 firmware to be implemented: the board
Table 6.9: Resource utilization of the VU9P FPGA for the Correlator Layer-1 demonstrator.

<table>
<thead>
<tr>
<th>VU9P</th>
<th>$n_{FF}$</th>
<th>$n_{LUT}$</th>
<th>$n_{LUTRAM}$</th>
<th>$n_{BRAM}$</th>
<th>$n_{DSP}$</th>
<th>$n_{UltraRAM}$</th>
<th>Latency (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used</td>
<td>2.36M</td>
<td>1.18M</td>
<td>592k</td>
<td>2160</td>
<td>6840</td>
<td>960</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>785k (33%)</td>
<td>528k (45%)</td>
<td>47k (8%)</td>
<td>871 (40%)</td>
<td>1021 (15%)</td>
<td>243 (25%)</td>
<td>693</td>
</tr>
</tbody>
</table>

firmware shell, the regionizer, the particle-flow, and the PUPPI firmware. A schematic of this setup is shown in Fig. 6.57.

Although initial iterations of this design were able to be placed successfully on a VU9P, they were unable to meet timing requirements. This issue was due in large part to the Super Logic Regions (SLRs) present on VU9P chips. While the resources necessary for the PF and PUPPI firmware are less than 50% of those available on the VU9P, the implementation of the combined algorithm is very close to the limits of one SLR. This implies high levels of congestion on the utilized SLRs and therefore presents serious difficulties for meeting timing constraints.

The solution to this problem was to systematically modify the most problematic paths in the design. Problematic paths in the design were identified by analyzing the timing reports available from implementation runs, and then flip-flop stages were inserted in the HLS as a buffer in those paths. By inserting buffers in the particle-flow firmware between the EM calorimeter and calorimeter cluster subtraction and the track and calorimeter cluster linking steps, and inserting buffers between the particle-flow and PUPPI algorithms, the new design was able to meet timing. This design requires the resources shown in Table 6.9. One caveat is that this design lacked a dual-clock FIFO to allow the outputs and particle-flow and PUPPI to properly write into the output links of the APx board. The need for such a FIFO arises because the particle-flow and PUPPI firmware runs at a clock frequency of 240 MHz, whereas the links operate at 320 MHz. However, the ability of this design to meet timing with minor modifications suggests that the techniques employed here are fully capable of successfully relaxing timing constraints. We expect that the design discussed here is capable of meeting timing with minimal additional adjustments of the firmware. In this test, all components are implemented as they would be in a full system, and simulated $\tau$ events generated with 200 pileup are used to produce inputs.

![Figure 6.57: Demonstrator schematic for testing a complete Layer 1 board configuration implemented on a VU9P.](image)

For the Correlator Layer-2, another demonstrator based on the Serenity board was used. Here the set of algorithms that are tested is numerous, but each is implemented in a similar way within the EMP framework. For each algorithm, the algorithm firmware is separated from the infrastructure firmware for clocks, links and all board management. The framework itself is
6.4. Trigger demonstrators and slice tests

Figure 6.58: One of the Serenity demonstrator setups at Imperial College London. This setup, with a KU115 daughter card, was used to test the histogrammed PF jet algorithm.

Described in Section 6.3.2.2.

For the Correlator Layer-2 demonstrators, Serenity boards with KU115 FPGAs were used. The COM Express on-board CPU was used to write files of simulated Correlator Layer-1 outputs into the input buffers on one, or both, of the FPGAs. Algorithm outputs are written into the output buffers, which are stored on the COM Express. In the demonstrations performed so far, the algorithms were contained within one of the two possible FPGA sites on the Serenity card.

The histogrammed PF jet algorithm presented in Section 3.6.1.3 was implemented in firmware and loaded onto a Serenity demonstrator card. 8000 ± 200 PU events, with PF candidates in at least one of the two regions implemented, were played through the firmware algorithm. The histogramming stage of the firmware bins the incoming PF candidates into a fixed array and was designed to accept candidates according to the architecture and regions described for the Layer-1 particle-flow. The demonstrator setup is shown in Fig. 6.58.

Figure 6.59 shows the comparison of hardware output with software emulation of the algorithm for jet $p_T$, $\eta$, and $\phi$, with 98% of events having perfect agreement. The most significant cause of remaining discrepancies is the difference in numeric precision between firmware and emulator.

To expand the histogrammed PF jet demonstrator and begin testing the integration of the hardware implementation of the algorithm within the trigger architecture, a multi-board demonstrator was set up. One Serenity card was programmed with the PF algorithm, another with the histogrammed PF jet algorithm, and the cards were connected optically via $24 \times 25$ Gb/s optical links. A pattern containing inputs to the PF algorithm for 50 ± 100 MC events was generated and played through the PF card, and the reconstructed PF Candidates for two regions were transmitted optically to the histogrammed PF jet algorithm on the receiving card, which reconstructed jets from the PF Candidates. The histogrammed PF jets produced by the multi-board demonstrator were then compared with those produced by the HLS emulation of the two algorithms in sequence, running over the same inputs, for the two regions.

Firmware for $\tau$ lepton identification using a neural network has been successfully implemented.
and operated at 240 MHz. Firmware for jet reconstruction and for computation of lepton isolation have also been implemented. In each case, the inputs are generated using simulated $\tau\bar{\tau}$ events as in the demonstrator for Layer-1. In both Layer-1 and Layer-2 demonstrators, the firmware shell used to operate links and control the board is designed to be fully realistic.

For the $\tau$ lepton demonstration, the full Serenity infrastructure was used and, although a preliminary version was synthesized for the Serenity board with a Xilinx Kintex Ultrascale 115, the full implementation was done on a Virtex 9 Ultrascale Plus. To perform the demonstration, a Vivado HLS IP project was prepared that performed both the input routing and organization of the PF candidates, and then applied the tau lepton identification. This IP was synthesized in two scenarios. The first implementation had 36 inputs that contained 15 tracks, 10 photons, 10 neutral hadrons, and one muon. This corresponds roughly to a single large region that is processed per time multiplex period in Correlator Layer-1. A second implementation was also synthesized, implemented and tested on the demonstrator board with 72 inputs (doubling the number of particle inputs in each case). Within the demonstrator board, an additional VHDL wrapper with a 15 clock FIFO-based input buffer was added to buffer in the inputs before they are sent to the HLS IP. This buffer is needed to account for timing restrictions found on the Virtex 9 Ultrascale when going between dies of the chip, referred to as SLRs. When the input pins in the algorithm were placed on the same SLR no FIFO buffer was needed. Additionally, on the output of the HLS IP, a 5 clock FIFO-based buffer was needed before transmission of the signal to the output regions. For the 72 input version of the tau algorithm, the output buffer length needed to be increased from 5 clocks to 15 clocks.

Within the HLS IP performing the $\tau_{h}$ identification, two components were implemented. First, a component that deals generically with particle inputs. This component assumes a single
6.4. Trigger demonstrators and slice tests

A 64-bit PF candidate comes on each input at each clock and that the board is running with a time multiplex factor of six. The algorithm is clocked at 240 MHz, so this corresponds to 36 clocks before a new event comes in. That equates to 36 regions or a maximum 1292 particles for the 36 input version, and 2592 for the 72 input versions, both cover more than the expected multiplicity of all PUPPI particles in each event. However, currently implementations assume either 72 or 54 regions for tracker volume of particle-flow. In addition to separating the inputs by particle type, we also assume that within a single region the inputs are sorted by $p_T$. From the inputs, the particles are sent into dedicated BRAMs. Each BRAM stores either the leading-$p_T$ charged particle for all regions, the sub-leading charged particle, the leading neutral, and so on until we cover all 36 inputs. The total number of BRAMs correspond to the number of inputs. Following the storage of the inputs into BRAM, seeds are found whereby the leading $p_T$ charged particle in all the regions are sorted and the 10 highest-$p_T$ particles are selected. These 10 are then filtered to remove any overlapping seeds where the distance between the particles $\Delta R < 0.4$. Next for each seed, we employ a large look-up table binning the particle $\eta$ and $\phi$ position. This LUT returns the memory locations of the four neighboring regions within the BRAM. Finally, all the particles from these 4 neighboring regions are pulled out of the BRAM and the particles within $\Delta R < 0.4$ with respect to the seed particle $\eta$ and $\phi$ are used for the $\tau_0$ algorithm. This whole procedure takes roughly 60 clocks at 240 MHz when designed for a time multiplex of 36 clocks (every sixth event) and could be quickly adapted for the application of particle isolation, jet reconstruction, or other algorithms. A full VHDL-based version of this seeding algorithm is being developed. This algorithm will have the added feature of being able to take multiple seeds from a single region.

Once all particles within a cone of $\Delta R$ with respect to the charged particles seed are obtained, the algorithm sorts the particles by $p_T$ and then takes the top 10 as input to a deep neural network. The deep neural network has roughly 2000 weights and is capable of taking in a single tau every other clock cycle. While the neural network inference is running, the $p_T$ sum of the top 10 candidate particles within a cone of $\Delta R < 0.1$ with respect to the seed is computed and selected to be the seed $p_T$ of the final $\tau_0$ candidate. The neural network output discriminant is then associated to the particle-flow object and the top 6 candidates are transferred to the output buffer. The whole algorithm, including the input and seeding, takes up roughly 60% of a single SLR of a VU9P (there are three SLRs on an VU9P), consequently several additional algorithms could be run at the same time. In fact, based on the seeding algorithm embedded in this design, it is foreseen to add additional lepton isolation, or jet-style algorithms with minimal additional resources.

Currently, the whole $\tau_0$ implementation has been tested on a single board using BRAMs as inputs and outputs. The BRAMs are conveniently placed along the edges of the chip to mimic the actual BRAMs that would be used with the fiber inputs. In this scenario, several hundred events have been passed through and no deviation has been observed when comparing the output with the HLS emulation of the algorithm. This HLS-based emulation has further been checked against the CMSSW implementation. For the $\tau_0$ implementation to be completely interfaced with Layer-1, an additional AXI-Stream based HLS IP would be needed to deal with the inputs. This IP has already been implemented but it has not yet been tested on the demonstrator.

The tests detailed above provide a high degree of confidence in the feasibility of the Correlator Trigger design. All tests show good agreement between emulation and the hardware results, and have been designed to mimic the full design as closely as possible. In some cases the hardware used for the demonstrator is less powerful than the hardware in the full design, giving additional confidence in the viability of the design.
6.4.4.3 Future demonstrator tests

Although the demonstrator tests that have already been performed are extensive, they represent only a subset of possible demonstrator tests. The following section describes future demonstrators that will provide further validation of the full trigger design and the role of the correlator trigger in the CMS trigger.

The first would be a test of both the GTT system and the Correlator Layer-1. This would require at least two boards, one each to run the GTT and Layer-1. Although these components have been successfully tested separately, they must be synchronized in the full design and this needs to be tested. Specifically, the GTT system must compute the $z$ position of the primary vertex, and send this information to the Correlator Layer-1 system in time to be used as an input for the PUPPI algorithm. A test of this coordination cannot be done effectively without both systems operating in parallel. The success of the test requires proper handling of the vertex information, which is transmitted asynchronously with respect to the other inputs to the Correlator Layer-1.

Another planned test would be of both the calorimeter GCT (either the barrel or forward) and the Correlator Layer-1. This also requires a minimum of two boards and would help to ensure synchronization across different trigger sub-systems. In this case the arrival of the inputs would be synchronous with the start of the layer-1 algorithm processing, but there would be many more inputs and links being utilized. This would therefore represent a more complete test of the two systems’ abilities to operate synchronously over multiple links and with large data volumes.

These two tests could be done with existing hardware and test stands, and would represent the first multi-board demonstrators involving the Correlator Layer-1 system. In addition to these two tests, further multi-board tests are also planned that will require additional preparation and assembly of test stands. A test of note would be similar to the planned test of the GTT and Correlator Layer-1, but would utilize the track-trigger boards as well to provide inputs to both systems. This would require multiple tiers of coordination, and test multiple layers of the trigger design simultaneously. A successful test would indicate that the systems are fully capable of functioning asynchronously and maintaining performance operating at full design specifications. A final planned test would be of the full Correlator system. This would require coordination across multiple board species and multiple types of inputs. Both the Correlator Layer-1 and Layer-2 systems have been tested separately, but a test of them together would give assurance that they are capable of operating together in the final design.

6.4.5 Demonstration of the 40 MHz scouting system

6.4.5.1 40 MHz Scouting demonstration during 2018 data-taking

At the end of the 2018 data-taking year, scouting at 40 MHz was demonstrated using a Xilinx KCU1500 FPGA Acceleration Development Kit as a scouting board that captured muon candidates from the Phase-1 Global Muon Trigger (uGMT) as seen in Fig. 6.60.

The uGMT ranks 108 input muons and sorts them in two stages before sending the eight highest ranked muon candidates to the Global Trigger. The outputs of the first sorter stages – so-called intermediate muon candidates – are buffered to be included in the uGMT readout record for diagnostic purposes. For the demonstrator system the uGMT was modified in order to send up to eight final muon candidates as well as up to eight intermediate muons from the barrel sorter to the scouting board. The muon candidates were then transmitted to the scouting board via eight 10 Gb/s optical links.
Because the KCU1500 does not allow interfacing to the TCDS system, it cannot be operated synchronously to the LHC clock, which operates the Level-1 trigger. We therefore exploited the asynchronous 10 Gb/s link protocol used by the trigger boards. Data at 240 MHz are transmitted over a 250 MHz link with padding words inserted and removed at the sender and receiver respectively. Individual links were aligned based on comma words. For data synchronisation at the analysis stage bunch crossing and orbit counters as well as other auxiliary information (e.g. link IDs) were added to the data package in the uGMT. The KCU1500 board provided two PCIe Gen3 x8 interfaces bifurcated to one x16 edge connector that were used to transfer the incoming data over DMA to the host machine. To reduce the load on the host machine the incoming muon data were zero suppressed, removing all bunch crossings that did not contain any muon candidates.

Data acquisition and system control for the scouting system was achieved using three computers: the host for the scouting board, a storage node, and a control machine. The scouting board host applied a second, more granular zero suppression algorithm to the transferred muon data, before it was transferred to the storage node via 10 GbE. The storage node compressed the data using the BurrowsWheeler algorithm (bzip2) before storing the data in a 8 TB RAID system. Optionally, the data could also be transferred to the CMS Lustre storage system via Infiniband, however this was not used in the 2018 run.

The control machine interfaced with the scouting board via JTAG over USB. Rudimentary control and monitoring software was written to achieve (near) autonomous operation, automatically starting data-taking at run start and initiating recovery procedures when system failure was detected. CMS global data-taking was left fully unaffected by any issues affecting the scouting system.

The scouting at 40 MHz demonstrator system was operated for the last week of Run-2 proton-proton data-taking as well as the entire 2018 HI run. For the pp run up to 1.1 TB of data per 24 hour beam day could be recorded after compression. In general about 2.1 GB per 1 pb\(^{-1}\) were stored. The data-taking efficiency was measured to be about 50%. The relatively low efficiency was due to the rudimentary control and to the fact that the link alignment and the DMA were debugged in operation.

In the HI run 26 MB per 1 \(\mu\)b\(^{-1}\) were recorded, but as expected a large contribution of cosmic...
muons was seen.

The use of the GMT scouting to track luminosity is illustrated in Fig. 6.61. The analysis of the emittance scan at the end of the LHC fill 7333 using data from the scouting system gives results consistent with those of other luminometers.

![Figure 6.61: The number of GMT muons per $2^{12}$ orbits follows the emittance scan evolution over time. A large statistics can be achieved for relatively short intervals even for a single bunch.](image)

6.4.5.2 Scouting at 40 MHz during Run-3

After demonstrating the capability to record data from a single Level-1 trigger system, the aim for Run-3 is to demonstrate the ability to operate on distributed data. To this end, the scouting system will be extended to include the Calorimeter Layer-2 objects (jets, $e/\gamma$, $\tau$, energy sums), thereby capturing all Global Trigger inputs (except for external signals). As an extension the inclusion of the BMTF output muon candidates is currently under consideration. A possible system design for Run-3 is shown in Fig. 6.60. The Run-3 scouting will therefore be a scaled-down version of the Phase-2 system, capturing all inputs to the GT.

6.5 System-wide technical issues

The CMS Phase-2 Level-1 Trigger represents a huge leap forward in performance and complexity. Compared to the Phase-1 trigger system there will be a significant increase in the number of boards, computing power in individual chips, number of interconnects, as well as link speed itself. In addition, the widespread use of embedded CPUs and the adoption of High-Level Synthesis will bring significant advantages in terms of board control and monitoring as well as algorithm development. However, these gains in system capabilities come at the cost of a larger, more intricate system to design, commission, and operate. Experience with the Phase-1 trigger system have impressed upon us how important it is to devise strategies for global commissioning, operation, and monitoring at the design stage of the system.
This section will address the specific issues related to ensuring this key upgrade system can be commissioned and validated for data-taking, along with the strategies to minimize the complexity of operation throughout HL-LHC running.

6.5.1 Hardware issues

6.5.1.1 Hardware platforms

As described in Section 6.3 multiple ATCA processing boards are currently in development for the Phase-2 upgrade. Experience from the Phase-1 trigger upgrade has shown that a standardized board interface to software and infrastructure firmware with common functionality accessible in the same way between hardware platforms is mandatory when operating a mixture of distinct processing board types.

Processing boards need to be provided with common infrastructure firmware and software libraries for control and monitoring access. The minimum set of firmware available must include the DTH interface, link firmware with spy buffer\(^1\) functionality, test pattern injection at both inputs and outputs, as well as standard monitoring facilities.

A common view of all processing boards could have profound implications on on-call training and organization. While the trigger has so far been operated with a general on-call (the L1 Detector-On-Call) with minimal training for hardware interventions backed by at least one on-call expert per sub-system, exposing a more homogeneous view of the system would allow the training of the primary L1 on-call expert for more advanced diagnostics and interventions. Every sub-system expert would be able to react to many problems in any of the sub-systems, as the infrastructure layer has been homogenized, and could therefore act as a primary on-call expert. Dedicated sub-system domain experts would only need to be called in case of more subtle issues. For hardware issues each hardware producing group must provide their own dedicated experts.

6.5.1.2 Embedded CPUs

Embedded CPUs will feature prominently on all proposed Phase-2 trigger boards. They have already been shown to be potentially highly useful tools in the Phase-1 Layer-1 calorimeter trigger where they have been used for advanced diagnostics such as in situ eye diagrams. For Phase-2 embedded CPUs could also be exploited to realize powerful system tests (see Section 6.5.8). Furthermore these devices can be used to implement a largely board-independent interface, as advocated for in Section 6.5.1.1.

The large number and high versatility of these devices makes a central management strategy necessary. Experiment-wide discussions have begun on how to perform tasks such as monitoring, as well as software upgrades and deployments. Good practices, such as standardized software for common tasks as well as minimal local data storage should be followed.

6.5.1.3 Crate organization

The crate organization in the Phase-1 trigger follows traditional segmentation by sub-system. Crates are under the control of a single group of people and are usually populated with a single processing board type. Such a scheme has clear advantages, among them the well defined line of responsibility for maintenance as well as the on-call escalation chain. Furthermore a given crate is a fairly simple system as it is composed of a moderate number of identical components. In particular, this point makes early commissioning of the crate itself a comparably easy task.

---

\(^1\)Facilities to allow the non-intrusive inspection of the data stream into or out of the processing board.
However, it also has some drawbacks, in particular concerning flexibility during periods of heavy testing; if a given system needs to test even one of their boards in local running this would effectively make that entire system unavailable for global runs or integration testing, as a crate is connected to the central DAQ and TCDS systems via a single interface. Similarly, tests of multiple systems require experts from each of these systems due to the high level of inter-crate heterogeneity. More importantly, crate failures would make entire sub-systems unavailable for extended periods, potentially precluding CMS from taking data.

As such, an alternative architecture has been proposed in which crates form a fundamental slice (or two) of a fully time-multiplexed trigger. As has been described in Section 5 the Phase-2 trigger will largely be implemented in a time-multiplexed way, each processing node operating on a subset of all bunch crossings. In such an environment one can view the system as composed of functionally independent units, so-called time-slices, each of which operates on a subset of the bunch crossings. In the proposed architecture, processing boards for a single time-slice are grouped in the same crate, see Fig. 6.62, allowing simple integration tests to be performed without an external timing system, but purely with clock and synchronization signals supplied by the DTH. Furthermore, such a scheme could mitigate otherwise catastrophic events, such as crate failures because in these cases just one time-slice (and thus only a subset of the data) would be lost and data-taking could potentially even continue, whereas a crate failure would render an entire sub-system non-functional in the traditional setup. Replacing a full crate would also require a smaller number of spare processing boards per sub-system, reducing the load on a single sub-system in the case of water damage or similar issues.

To enable such a scheme to work efficiently the trigger would need to mandate a common board interface to the online control and monitoring software; this was provided for in Phase-1 through the sub-systems adopting SWATCH, and a similar project would be necessary for Phase-2. As described in Section 6.5.1.1 this would have additional benefits during standard data-taking, allowing fewer on call experts to solve problems more efficiently.

As demonstrator systems are developed, and experience with ATCA infrastructure increases, a dedicated technical review of the choices above will be held to decide on which option to pursue.

While the above reasons make the heterogeneous crate organization very attractive there are other potential issues surrounding commissioning that should also be considered. The initial integration could prove to be more difficult if sub-systems require significant time for individual tests using central services, however later integration tests may prove to become easier due to the removal of external dependencies (TCDS and central DAQ). Furthermore it may become more difficult to allow individual systems to perform tests without affecting the global run. Sufficient organizational oversight and scheduling (i.e. assigning different time slices to different sub-systems) will be a strong requirement to ensure this period is successful.

### 6.5.2 Firmware and software deployment

#### 6.5.2.1 Firmware management

All firmware used for production data-taking must be built from code that is accessible and versioned using standard tools. Bitfiles destined for use at Point-5 should be required to be packaged together with the associated address tables, as well as the following auxiliary information:

- a list of the source files and their last changed date
- the version control branch/tag/revision
6.5. System-wide technical issues

Figure 6.62: An example of a single ATCA crate containing all processors necessary for two BXs of processing. The GMT, GCT, GTT, refer to Global Muon Trigger, Global Calorimeter Trigger, and Global Track Trigger, respectively. These boards can process stand-alone quantities, e.g. Missing energy using Tracker information only. The PF+PUPPI and Obj Reco boards refer to the two stages of combining calorimeter and muon quantities with tracking information. The first stage would be the particle-flow and pileup per particle identification, followed by full object reconstruction.

- the timestamp of the build
- the output of the version control system confirming that the working directory is free from local changes

Providing this information makes a potential later post mortem analysis much easier and greatly reduces the risk of unintended features being introduced to the firmware by ad-hoc changes.

The trigger project will provide facilities to make central firmware builds easy and preferable to local builds. Providing a central build system allows to keep an audit trail for builds that will later be deployed at Point 5, however care must be taken to provide firmware specialists the required tooling as well as flexibility to configure the central build of their sub-system to their specifications.

Production firmware must only be deployed from a central firmware repository. This repository should allow sub-system experts to upload archives containing firmware bitfiles, as well as auxiliary information to a bucket for the system under their responsibility and then provide an audit trail that includes the information collected in the build step described above. From the sub-system control PC the bitfile archive should be easily accessible in order to deploy it to a given board. An audit trail should also exist for the deployment to any production board and the global run control system or at the least the Level-1 trigger online software should be aware when firmware has been deployed that is not regarded as production-ready. Such an audit trail could be implemented by performing production bitfile loading exclusively via the system configuration keys.
6.5.2.2 Online software deployments

Learning from Phase-1, a change to the method of production online software deployment should be enforced. Deployment of production online software should only be allowed via central means, complete with an audit trail and the possibility for fast rollbacks. Such a system would dramatically reduce the risk of human error and provide a clear mechanism for knowing the exact state of each sub-system’s online software at any point in time. Any advance here must be made in conjunction with the CMS system administrators in order to profit from central support while also improving the CMS-wide infrastructure.

6.5.3 Trigger links and system synchronization

The Phase-1 trigger is operated with a common link protocol for communication between trigger processing boards. This allowed for faster commissioning of systems due to being able to profit from other groups’ experience as well as shared benefits when bugs were found and fixed. Similarly the Phase-2 trigger should at a minimum be operated with a common physical layer. Furthermore, standardization of the packet format is highly desirable. The introduction of subtle bugs at this layer of the system has the potential to highly disrupt the commissioning effort. Any deviation from the standardized format should be well motivated.

The above arguments extend to the link interface with the TPGs. Links entering the trigger should therefore be required to conform to a common protocol unless precluded by fundamental issues such as incompatible hardware.

A database mapping connections between input and output ports becomes mandatory at the scale of the proposed Phase-2 trigger system. The link protocol should provide fields to store package origin and destination in order to allow each system to automatically verify correct wiring by comparing this information with the data obtained from the link mapping database.

6.5.3.1 System synchronization

Within the Phase-1 trigger, the 10 Gb/s asynchronous link protocol is exploited for system synchronization, aligning the links on commas sent at a specified point in the orbit (the so-called TCDS “dark gap” during which no triggers are distributed). Within the current time-multiplexed system a once-per-orbit marker is implemented as part of the link protocol which is used to align to. At the TPG level a mix of synchronization strategies is used. For Phase-2 a common scheme should be implemented. If synchronization markers from TPGs cannot be carried over into the trigger system, they should be generated at the edge of the system after which they can be carried along exploiting the common link protocol.

6.5.4 Software-firmware interface

6.5.4.1 High level synthesis and trigger emulation

Since the inception of the CMS Level-1 trigger emulation software has existed that exactly reproduces the behavior of the system implemented in firmware. Such emulation software is used both offline and online. In the offline domain emulators are used to estimate the effect of changes, calculate trigger rates while developing a new trigger menu, and during Monte Carlo production campaigns as part of the detector simulation. In the online domain the emulators process a subset of the recorded data and any mismatches with the hardware are reported.

Vivado High Level Synthesis allows the development of firmware in a high level language such as C++ before it is synthesized to the register-transfer level (RTL) by a software package. It can be a useful tool to bring powerful algorithms to the Level-1 trigger, however as the
complexity of the algorithms increases care must be taken to maintain the current high level of firmware validation. This means that independently developed emulators (distinct from the HLS code) must continue to be provided for each trigger sub-system at commissioning time in order to validate the synthesized firmware. An independently written emulator increases the possibility of discovering logic bugs that would otherwise only be discovered significantly later.

### 6.5.4.2 Online software

In the Phase-1 trigger upgrade, one of the main design goals for the online software was to significantly increase the fraction of software that was common across different subsystems. In order to achieve this goal we developed a new control and monitoring framework (SWATCH) [95] that represents the commonalities of the different trigger subsystems, and implements all common functionality, whilst providing the flexibility that is required for different subsystems to be able to control their electronics through subsystem-specific libraries, and monitor different quantities. One of the key features in this framework is the fine granularity of the subsystem-independent hardware description, control primitives and monitoring primitives, based on detailed common models of the data processing boards and the subsystems. Building on the framework’s feature-rich subsystem-independent interfaces, we implemented common GUIs that could run all configuration procedures and access all monitoring information, developed common libraries for retrieving configuration data, and defined a single database schema that was used for all subsystems. The resulting reduction in the fraction of subsystem-specific code significantly reduced the effort required for both development and long-term maintenance. Furthermore, having an identical user interface for all trigger boards has simplified training of operational personnel, and has also proven immensely valuable during high stress operational periods, by allowing central trigger experts to diagnose problems in trigger subsystems as central monitoring metrics are found across subsystems.

In preparation for Run-3, we are currently improving several components of the online software. Notably, we are updating deployment procedures to increase the amount of automation, using open-source industry-standard solutions including GitLab CI and Kubernetes. We are investigating the use of industry-standard monitoring and alerting tools such as Prometheus, in order to further reduce the amount of time required to detect and resolve problems, and to simplify the post-mortem analysis of such problems. Based on the operational experience from Run-2, we are also redesigning the tools used to develop and validate trigger menus and to edit the configuration parameters, in order to increase the efficiency of menu developers and operational personnel.

In the Phase-2 trigger system, electronics boards will be housed in ATCA shelves rather than MicroTCA and each will host an on-board processor (i.e. a Zynq device, or Intel CPU). However the level of homogeneity in hardware and major firmware components across different subsystems will remain the same as in the Phase-1 trigger. As such, all major design features of the Phase-1 trigger’s online software that were essential for increasing the fraction of common software and ensuring uniformity in user interfaces will be equally possible in the Phase-2 trigger. We plan to control and monitor the Phase-2 trigger using a future iteration of the SWATCH framework used in Run-3, updated to take advantage of new open-source industry-standard tools and libraries, and to adapt to changes in the CMS-wide online software ecosystem. Additionally, since the trigger system will contain hundreds of on-board processors, on which the low-level control and monitoring software will run, in designing the online software we must ensure that these on-board processors will be managed in a robust, systematic and efficient manner.
6.5.5 Commissioning and parallel running

Given the critical position of the Level-1 trigger for the ability of CMS to take data a minimal working trigger needs to be provided early during commissioning. At a minimum this could consist of one of the Global Trigger processing boards with a simple muon trigger for commissioning with cosmics. If the muon trigger chain has not been fully commissioned the external inputs could be used to receive signals directly from any of the muon systems. These data could for instance indicate the number of muon stubs detected for a given bunch crossing.

Experience during the Phase-1 trigger commissioning (as well as from GEM commissioning in 2018) has shown that it is highly useful for a new system to profit from inclusion in a global run. Conversely the introduction of initially unstable systems into the global run has been seen to be highly disruptive. It would therefore be beneficial to introduce the concept of “test FEDs” into the CMS run control software and central DAQ. FEDs marked in such a way would be included in the run and expected to send valid data to the central DAQ, however if they at some point begin sending invalid data, these would be ignored instead of leading to a run termination. Data from these test FEDs could be sent to a special HLT stream to ensure that production data can never be contaminated by faulty hardware.

As both central DAQ and TCDS, as well as all TPGs, will undergo a complete upgrade in parallel to the trigger, it is mandatory for the system to be able to do rudimentary slice tests without the support of (full) central services or TPG input. As was described in Section 6.5.1.1 a heterogeneous crate solution would allow significant commissioning even in the absence of a fully functioning TCDS system, however even if the trigger is implemented with homogeneous crates it is vital to be able to commission the system without inputs from the TPGs and readout from central DAQ. To ensure this it must be possible to inject and read out patterns at defined positions in the LHC orbit both at the inputs and outputs of each Level-1 trigger system. Such facilities are described in Section 6.5.8, where we further elaborate on their use during standard operations.

Later stages of trigger commissioning may proceed using both cosmic muons and beam splashes. Cosmic muons will allow the trigger input links to be validated, the performance of track finders, as well as the correct timing of the read out logic. Using beam splashes will enable us to globally time in all inputs to the system and synchronize them with the beam presence signals. Data is taken using special calorimeter triggers for these events, making the commissioning of this system with the help of patterns a critical task.

6.5.6 Monitoring and data validation

6.5.6.1 Monitoring data

Monitoring data shall be recorded and evaluated for each triggered bunch crossing. It is required to detect and diagnose potential problems during the operation of the trigger. Examples are the detection of hot calorimeter towers, prefiring, or failed links. Such data can be restricted to a subset of either the outputs or inputs of each system, but ideally should be taken for each triggered bunch crossing.

6.5.6.2 Validation data

These data are required to validate the firmware against the emulator as well as ensure that the output of a given system corresponds to the input of the downstream system. They can be significantly larger in size (in particular for systems closer to the TPGs) and therefore it is acceptable to read these data at a significantly reduced rate during standard operations. The
readout should include both the inputs and outputs to the system as this would allow the source of errors to be identified more easily. In particular during commissioning such data will be invaluable to find problems on the link level or in a system’s readout logic. Synchronized read out of validation data between sub-systems as well as TPGs can be achieved using special trigger types issued by TCDS or by counting triggered events within the subsystem logic and agreement to read out validation data for every Nth event. The latter method will be used in the event that not all TPGs can support validation data read out using trigger types and therefore all trigger subsystems are required to support both methods.

6.5.6.3 Realtime medium-granularity monitoring

It has been proven highly beneficial to provide monitoring data without relying on the readout system. (For instance to debug problems that block data-taking, but also because it tends to be faster.) The increase in system size will likely lead to a corresponding increase in monitoring data. We are currently in the process of evaluating industry standards for monitoring large distributed systems (among others moving towards principles derived from the SRE and Devops movements) for Run-3 and expect to profit from this experience for the Phase-2 system monitoring.

6.5.7 Latency measurements and checks

Measuring the absolute latency of the trigger accurately is a key requirement of the Phase-2 system. Such functionality is not currently available and both during commissioning and operations significant time was lost due to the inability to quickly and reliably establish the system latency. The increased complexity of the Phase-2 trigger further enhances our need for a mechanism to not only measure the whole system latency but also the latency between two trigger sub-systems. While the system is being installed and firmware updates are regular, a quick latency check in-situ will become an essential part of the validation program. The simplest method to have a quick measurement of latency is to have an absolute timing reference. Discussions are on-going with the TCDS group to determine the feasibility of including this in the TCDS2 design. Other options are also being explored which would involve common firmware blocks in all trigger sub-systems time-stamping a data signal as it moves through the chain. This method would require the fiber lengths to be well-known. Additionally, an independent partial verification of the latency measurement can be achieved by connecting two systems directly to an oscilloscope, though this relies on appropriate output connectors on all the trigger sub-systems. It should be noted that this is not feasible for a full system latency verification. A technical review of the options available and how to proceed will occur when the above options have been sufficiently examined. This should form part of the overall trigger Electronics Systems Review.

6.5.8 System health check

The Phase-2 trigger should be equipped with a system to inject patterns at the inputs and outputs of every stage of the system as well as read out these test data. Such a system could be implemented internally to the trigger by configuring the system locally to:

- Block the outputs at the test system boundaries. (Care must be taken not to send triggers to the TCDS, potentially interfering with other CMS sub-systems.)
- Inject patterns at a known point in the LHC orbit.
- Read out the data at the above mentioned known point via the online software.
And then verifying the recorded data versus the expected outputs generated with the emu-
lation software. Such a scheme has the benefit of being independent of most external infras-
tructure: Even for integration tests only basic TCDS functionality (clock and orbit markers) are
needed. Tests could be performed from the inputs of the first trigger layers up to the Global
Trigger, in particular singling out specific parts of it, e.g. the links between two systems.

Alternatively the CMS run control software could be augmented with a test mode. This scheme
could work similar to the above described one with the following additional features:

- Additional read out via the central DAQ triggered by the TCDS system
  - as a consequence the validation facilities provided by central DQM can
    be used.
- Can use TCDS BGos to trigger pattern injection and read out.
- Also non-trigger systems can be specially configured when test mode is entered.
  - In particular TPGs can be centrally configured to enter test mode.

Both schemes should ideally support parallel tests if resources used do not overlap. For the
second scheme, each test slice could be configured to react to a different TCDS trigger type
and BGo for the various test commands. As also described in Section 6.5.1.1 the functionality
to inject and read out test patterns must be uniform across all trigger boards in order to allow
meaningful integration tests to be performed.

While invaluable during commissioning of the system, such an integration test system would
also be beneficial during normal operations; especially following routine deployments of new
firmware to the Level-1 trigger.

### 6.5.9 Demonstrator systems in Run-3

For Run-3 we expect several sub-systems to be able to demonstrate key aspects of their Phase-2
upgrades:

- Scouting at 40 MHz
  While receiving data from one trigger system has been demonstrated at the end of
  the 2018 run, it is foreseen to show the feasibility of distributed scouting by receiving
data from multiple trigger sub-systems in Run-3 (see Section 6.4.5).

- Novel algorithms in the Global Trigger
  Work is ongoing to provide the full set of Global Trigger inputs to both an indepen-
dent test crate as well as the 40 MHz scouting system. It is envisaged to use these
two facilities to test novel algorithms for use in the Phase-2 Global Trigger within a
non-latency constrained environment.

- Triggering on displaced muons in the barrel region
  In Run-3 the Barrel Muon Track Finder will utilize the Kalman Filter algorithm tested
  at the end of 2018. Upgrades of the Global Trigger and Global Muon Trigger are
  currently ongoing in order to be able to use the displaced muon tracks in trigger
  algorithms (see Section 6.4.2.2).

- Neural network-based transverse momentum assignment
  The Endcap Muon Track Finder group is planning to demonstrate the use of neural
  networks for transverse momentum assignment in Run-3.

Furthermore it may be possible to install a prototype Phase-2 ATCA-based processing board in
the service cavern, using inputs from the Drift Tube Phase-2 test system.
Chapter 7

Organization, schedule, and costs

7.1 Project organization and work breakdown structure

The organization of the CMS L1 Trigger (L1T) system is shown in Fig 7.1. The L1 trigger upgrade project has two Upgrade Coordinators who report to the L1 trigger project manager. They are appointed by the L1 trigger project manager for two year terms subject to approval by the Trigger and data acquisition system (TriDAS) Institution Board (IB). The Institution Board consists of one member per institution working on the TriDAS project. Upgrade coordinators report regularly in L1 Trigger Project meetings, and consult regularly with Technical, DPG, and Software Coordinators to ensure the relevant deliverables meet requirements. Technical Coordinators oversee integration testing, installation, and commissioning activities at P5 and CERN, and Upgrade Coordinators define and execute those aspects together.

![Figure 7.1: Organization of the L1 trigger project.](image)

The work breakdown structure (WBS) for the project is shown in Fig. 7.2. In the CMS Upgrade project, L1 Trigger is a Level-2 CMS management WBS area consisting of seven Level-3 WBS sub-areas:

1. Calorimeter Trigger
2. Muon Trigger
3. Global Track Trigger
4. Particle Flow Trigger
5. Global Trigger
6. Data Scouting System

7. Integration and Commissioning

For the first six L3 WBS areas, the deliverables are of two types:

- ATCA boards for FPGA-based data processing, along with their accompanying firmware and software and
- electronics infrastructure (ATCA crates, optical fibers connecting them, DTH cards) supporting a trigger subsystem.

The different varieties of ATCA boards for each WBS L3 have a WBS L4 area (such as the regional and global calorimeter trigger boards). For each WBS L3, there is also a WBS L4 area specified for Infrastructure and Integration, encompassing the delivery of electronics infrastructure as well as system integration activities for that WBS L3 prior to delivery (such as slice tests or commissioning tests). The last area encompasses integration of delivered production boards, installation at P5, and subsequent commissioning of the installed system. Each WBS L3 area has subsystem coordinator(s) reporting to the L1 trigger upgrade coordinators.

![Work breakdown structure of the L1 Trigger upgrade project](image)

**Figure 7.2: Work breakdown structure of the L1 Trigger upgrade project.**

### 7.2 Project schedule and milestones

#### 7.2.1 Schedule

The project schedule for each WBS L3 area is structured similarly, with division into seven phases: prototyping, preproduction, pilot production, production, integration testing, installation, and commissioning. A schedule illustrating these seven phases is shown in Fig. 7.3.

In the prototype phase, initial prototype ATCA boards were designed and tested for each production line. Prototype firmware was developed and deployed on the hardware and compared with simulated expectations. The results of this development program, which began in 2016 af-
7.2. Project schedule and milestones

Figure 7.3: High-level schedule of the L1 trigger upgrade project.

The commissioning of the Phase-1 trigger upgrade and concluded in March 2020, informed the baseline design and are documented in Section 6.

During the preproduction phase, prototype hardware and associated firmware will be utilized and further advanced to fulfill the requirements for the production system. This comprises the use of the selected FPGAs, optical link protocols, infrastructure firmware, and algorithms. A minimal number of trigger boards and hardware infrastructure will be produced and tested to demonstrate each function of each subsystem and their interfaces. Components will initially be tested individually in labs before being assembled into time slices at CERN and Fermilab to verify subsystem compatibility. Demonstrator systems for GMT, GTT and the Correlator will furthermore be set up at the Tracker Integration Facility (TIF). This stage will also be used to demonstrate the heterogeneous crate concept described in Section 6.5.9. This phase starts at the beginning of 2020 and finishes in December of 2021. The resulting hardware and firmware will be the basis for an Electronics System Review (ESR) in June 2022, preceding the production phases.

In the pilot production phase, a full slice of the proposed system will be constructed from production-grade hardware based on the final system design approved in the ESR. After testing components individually, partial subsystems will be assembled into vertical slices and have all interface functionality tested. The CMS integration facility will be used as it will provide TCDS and DAQ facilities. Crates involved in the tests will be equipped with DTH boards. The pilot production phase ensures that all design specifications are met and system integration is proven to be reliable prior to committing the bulk of production procurement funds. This phase will take place from March 2021 to January 2024.

In the production phase, procurement, assembly, and testing of the remaining production components are conducted, concluding in the shipment of a full complement of successfully tested components to CERN. This phase will conclude by December 2024. Upon delivery to CERN, the subsystems will be integrated and subsystem testing will be repeated on the delivered components. This integration phase will last twelve months, starting with receiving the first production boards at CERN in December 2023 and concluding upon receiving the last production boards in December 2024. The purpose of the integration phase is to conduct a system-wide test including the interfaces with the sub-detector backend electronics, the clock distribution system (TCDS), and DAQ. CMS has been conducting a series of integration tests in a dedicated integration facility (IF) at CERN. The facility has been maintained throughout LS1 and LS2 and will be used during LS3. The integration facility has been used to integrate upgraded back-end systems prior to installation in the Underground Service Cavern (USC). The L1T project will install crates hosting the various components of the system and exercise the interface with
Chapter 7. Organization, schedule, and costs

TCDS. Verifying the communication and protocols among these components constitutes the first step of integration; establishing the link communication and alignment with sub-detectors backend electronics will be performed next. Depending on the results previously obtained on the first tests of the heterogeneous crate approach, the integration testing will proceed accordingly. Furthermore, the integration phase will help validate the boards with the final system software infrastructure, including the interface software for board control and firmware management. CMS USC is available for installation in January 2026; this allows twelve months of float to complete production and integration phases prior to installation.

The infrastructure installation, including the electronics crates and DAQ interface can start as early as January 2026. Boards installation and initial tests will be conducted along with the crate installation. Both DTH and TCDS boards will be available and installed with them. Crates and optical patch panels dedicated to the scouting system will be installed as well. The steps performed to establish communication among the various L1T components, at the integration center, will be repeated in the USC. The commissioning of the L1T is estimated to last six months, at which time the system is available for triggering at CMS. Commissioning with the sub-detectors including timing alignment of links and initial assessment of latency will be initiated. Integration to central CMS DAQ system will start taking place in July 2026, where CMS-wide global runs will be organized along with the first cosmic ray data-taking. The L1T system will be used to deliver triggers for commissioning and integration of sub-detectors, with completion expected by the end of 2026. The Level-1 trigger will be ready for operations no later than the start of HL-LHC beam commissioning in July 2027. This schedule allows for six months of float to meet that need-by date.

7.2.2 Milestone list

A list of high-level milestones is shown in Table 7.1. The six remaining phases of the project are demarcated by horizontal lines. Each WBS L3 area will have milestones for hardware and firmware development for each of the project phases (L1T.X.nn), along with milestones for the completion of global project goals (L1T.nn). The L3 milestones (L1T.X.nn) are optimized for each subsystem and for those this table reflects an average over L3 areas.

7.3 Institution interests and construction responsibilities

Table 7.2 lists all of the institutions within CMS TriDAS contributing to the L1 trigger upgrade project. Table 7.3 shows the areas of interest for each institution, broken down by WBS L2 area.

7.4 Project costs

The cost of the L1 trigger upgrade is estimated from a cost breakdown structure (CBS) which mirrors the project WBS down to L4. The summary of the costs, as defined by the CORE (LHCC Cost Review Committee), is shown in Table 7.4. CORE costs represent the value of the equipment installed in the experiment, without contingency. They include costs for fabrication, construction, installation, and integration. They do not include R&D and prototype costs, costs for infrastructure and facilities at CMS institutions, nor institution personnel costs. The CORE cost estimates include spare parts to cover production losses (sufficient to cover no less than 5% per production line of each WBS L4 area related to board production), while spares to support long term maintenance and operation are paid separately from the CMS M&O budget. Costs are reported here in 2019 CHF, with no correction for inflation to future years. Exchange rates
Table 7.1: Milestones for the Phase-2 upgrade of the CMS L1 trigger. L1T milestones denote WBS L2, and L1T.X milestones denote each of the WBS L3 production areas: (1) Calorimeter Trigger, (2) Muon Trigger, (3) Global Track Trigger, (4) Particle Flow Trigger, (5) Global Trigger, and (6) Data Scouting System. Integration, installation and commissioning activities belong to WBS area (7) Integration & Commissioning.

<table>
<thead>
<tr>
<th>Milestone</th>
<th>Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1T.X.Y.1</td>
<td>12-Nov-2020</td>
<td>X.Y preproduction board and firmware ready for testing</td>
</tr>
<tr>
<td>L1T.X.1</td>
<td>31-Dec-2021</td>
<td>X preproduction complete</td>
</tr>
<tr>
<td>L1T.1</td>
<td>30-Jun-2022</td>
<td>L1T Electronic Systems Review</td>
</tr>
<tr>
<td>L1T.6.1.1</td>
<td>31-Mar-2022</td>
<td>DS prototype specification complete</td>
</tr>
<tr>
<td>L1T.X.Y.2</td>
<td>31-May-2022</td>
<td>X.Y production board design and firmware specification complete</td>
</tr>
<tr>
<td>L1T.6.1.2</td>
<td>29-Jul-2022</td>
<td>DS prototype ready for testing</td>
</tr>
<tr>
<td>L1T.6.1.3</td>
<td>25-Nov-2022</td>
<td>DS prototype testing complete</td>
</tr>
<tr>
<td>L1T.X.Y.3</td>
<td>28-Feb-2023</td>
<td>X.Y production pilot board and firmware ready for testing</td>
</tr>
<tr>
<td>L1T.X.2</td>
<td>11-Jan-2024</td>
<td>X production pilot system integration testing complete</td>
</tr>
<tr>
<td>L1T.2</td>
<td>11-Jan-2024</td>
<td>L1T ready for final production procurement</td>
</tr>
<tr>
<td>L1T.6.1.4</td>
<td>24-Jan-2024</td>
<td>DS Stage 1 production specification complete</td>
</tr>
<tr>
<td>L1T.6.1.5</td>
<td>26-May-2024</td>
<td>DS Stage 1 production ready for testing</td>
</tr>
<tr>
<td>L1T.X.Y.4</td>
<td>19-Jun-2024</td>
<td>X.Y production final board and firmware ready for testing</td>
</tr>
<tr>
<td>L1T.6.1</td>
<td>23-Sep-2024</td>
<td>DS Stage 1 production complete</td>
</tr>
<tr>
<td>L1T.X.Y.5</td>
<td>31-Dec-2024</td>
<td>X.Y production complete</td>
</tr>
<tr>
<td>L1T.3</td>
<td>31-Dec-2024</td>
<td>L1T production complete</td>
</tr>
<tr>
<td>L1T.7.1.X</td>
<td>31-Dec-2024</td>
<td>X integration testing complete</td>
</tr>
<tr>
<td>L1T.4</td>
<td>31-Dec-2024</td>
<td>L1T integration testing complete</td>
</tr>
<tr>
<td>L1T.6.1.6</td>
<td>1-Apr-2025</td>
<td>DS Stage 2 production specification complete</td>
</tr>
<tr>
<td>L1T.6.1.7</td>
<td>2-Aug-2025</td>
<td>DS Stage 2 production ready for testing</td>
</tr>
<tr>
<td>L1T.6.2</td>
<td>29-Nov-2025</td>
<td>DS production complete</td>
</tr>
<tr>
<td>L1T.7.2.X</td>
<td>30-Jun-2026</td>
<td>X installation complete</td>
</tr>
<tr>
<td>L1T.5</td>
<td>30-Jun-2026</td>
<td>L1T installation complete</td>
</tr>
<tr>
<td>L1T.6</td>
<td>31-Dec-2026</td>
<td>L1T ready for physics</td>
</tr>
</tbody>
</table>

based on the average value in 2019 have been adopted to convert quotes collected in currencies different from CHF.

CBS L4 areas are composed of either procurement of ATCA boards required for the associated WBS L4 area, or procurement of electronics infrastructure needed for that L3 subsystem. ATCA board costs are estimated from the procurement cost of required components, primarily the FPGAs for data processing, the optical transceivers driving the optical fibers, the base ATCA circuit board, and any mezzanine boards accompanying it. FPGA and optical transceiver costs are estimated from recent vendor quotes, and the base board and mezzanine board costs are estimated from recent prototype board procurement. Subsystem infrastructure costs are derived from vendor quotes for the required optical fibers, patch panels, ATCA crates, and DTH cards needed.
Table 7.2: List of institutions expressing interest in participation in the Phase-2 upgrade of the L1 trigger.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Institution name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vienna</td>
<td>Institut fur Hochenergiephysik, Wien, Austria</td>
</tr>
<tr>
<td>Beijing</td>
<td>Institute of High Energy Physics, Beijing, China</td>
</tr>
<tr>
<td>Cyprus</td>
<td>University of Cyprus, Nicosia, Cyprus</td>
</tr>
<tr>
<td>Tallinn</td>
<td>National Institute of Chemical Physics and Biophysics, Tallinn, Estonia</td>
</tr>
<tr>
<td>LLR</td>
<td>Laboratoire Leprince-Ringuet, CNRS/IN2P3, Ecole Polytechnique, Institut Polytechnique de Paris, Palaiseau, France</td>
</tr>
<tr>
<td>NKUA</td>
<td>National and Kapodistrian University of Athens, Athens, Greece</td>
</tr>
<tr>
<td>NTUA</td>
<td>National Technical University of Athens, Athens, Greece</td>
</tr>
<tr>
<td>Ioannina</td>
<td>University of Ioannina, Ioannina, Greece</td>
</tr>
<tr>
<td>SINP</td>
<td>Saha Institute of Nuclear Physics, HBNI, Kolkata, India</td>
</tr>
<tr>
<td>TIFR</td>
<td>Tata Institute of Fundamental Research-B, Mumbai, India</td>
</tr>
<tr>
<td>Bologna</td>
<td>INFN Sezione di Bologna, Universita di Bologna, Bologna, Italy</td>
</tr>
<tr>
<td>Padova</td>
<td>INFN Sezione di Padova, Universita di Padova, Padova, Italy</td>
</tr>
<tr>
<td>Warsaw-IEP</td>
<td>Institute of Experimental Physics, Faculty of Physics, University of Warsaw, Warsaw, Poland</td>
</tr>
<tr>
<td>Warsaw-INS</td>
<td>National Centre for Nuclear Research, Warsaw, Poland</td>
</tr>
<tr>
<td>Belgrade Vinca</td>
<td>Vinca Institute of Nuclear Sciences, University of Belgrade, Belgrade, Serbia</td>
</tr>
<tr>
<td>CIEMAT</td>
<td>Centro de Investigaciones Energeticas Medioambientales y Tecnologicas, Madrid, Spain</td>
</tr>
<tr>
<td>Oviedo</td>
<td>Universidad de Oviedo, Oviedo, Spain</td>
</tr>
<tr>
<td>CERN</td>
<td>CERN, European Organization for Nuclear Research, Geneva, Switzerland</td>
</tr>
<tr>
<td>Zurich</td>
<td>Universitat Zurich, Zurich, Switzerland</td>
</tr>
<tr>
<td>Bristol</td>
<td>University of Bristol, Bristol, United Kingdom</td>
</tr>
<tr>
<td>RAL</td>
<td>Rutherford Appleton Laboratory, Didcot, United Kingdom</td>
</tr>
<tr>
<td>Imperial</td>
<td>College Imperial College, London, United Kingdom</td>
</tr>
<tr>
<td>FNAL</td>
<td>Fermi National Accelerator Laboratory, Batavia, USA</td>
</tr>
<tr>
<td>Colorado</td>
<td>University of Colorado at Boulder, Boulder, USA</td>
</tr>
<tr>
<td>MIT</td>
<td>Massachusetts Institute of Technology, Cambridge, USA</td>
</tr>
<tr>
<td>UIC</td>
<td>University of Illinois at Chicago, Chicago, USA</td>
</tr>
<tr>
<td>Davis</td>
<td>University of California, Davis, Davis, USA</td>
</tr>
<tr>
<td>Florida</td>
<td>University of Florida, Gainesville, USA</td>
</tr>
<tr>
<td>Rice</td>
<td>Rice University, Houston, USA</td>
</tr>
<tr>
<td>UCLA</td>
<td>University of California, Los Angeles, USA</td>
</tr>
<tr>
<td>UCSD</td>
<td>University of California, San Diego, USA</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>University of Wisconsin - Madison, Madison, USA</td>
</tr>
<tr>
<td>Princeton</td>
<td>Princeton University, Princeton, USA</td>
</tr>
</tbody>
</table>
Table 7.3: Areas of interest for institutions in the L1 trigger upgrade project. Areas correspond to the WBS areas described in Fig. 7.2.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Vienna</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beijing</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyprus</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tallinn</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLR</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NKUA</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTUA</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ioannina</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SINP</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIFR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bologna</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padova</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Warsaw-IEP</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Warsaw-INS</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belgrade Vinca</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CIEMAT</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oviedo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CERN</td>
<td></td>
<td>o</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zurich</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bristol</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAL</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imperial</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FNAL</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colorado</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIT</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UIC</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Davis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Florida</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rice</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCLA</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCSD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wisconsin</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Princeton</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 7.4: CORE cost estimates for the L1 trigger upgrade project. Areas correspond to the WBS areas described in Fig. 7.2.

<table>
<thead>
<tr>
<th>WBS</th>
<th>Cost (kCHF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Barrel Regional Calorimeter Trigger</td>
<td>494</td>
</tr>
<tr>
<td>1.2 Global Calorimeter Trigger</td>
<td>306</td>
</tr>
<tr>
<td>1.3 Infrastructure and Integration</td>
<td>237</td>
</tr>
<tr>
<td>1. Calorimeter Trigger</td>
<td>1036</td>
</tr>
<tr>
<td>2.1 Endcap Muon Trigger</td>
<td>686</td>
</tr>
<tr>
<td>2.2 Barrel Muon Trigger</td>
<td>579</td>
</tr>
<tr>
<td>2.3 Overlap Muon Trigger</td>
<td>231</td>
</tr>
<tr>
<td>2.4 Global Muon Trigger</td>
<td>500</td>
</tr>
<tr>
<td>2.5 Infrastructure and Integration</td>
<td>565</td>
</tr>
<tr>
<td>2. Muon Trigger</td>
<td>2560</td>
</tr>
<tr>
<td>3.1 GTT boards</td>
<td>443</td>
</tr>
<tr>
<td>3.2 Infrastructure and Integration</td>
<td>75</td>
</tr>
<tr>
<td>3. Global Track Trigger</td>
<td>518</td>
</tr>
<tr>
<td>4.1 Layer 1</td>
<td>940</td>
</tr>
<tr>
<td>4.2 Layer 2</td>
<td>876</td>
</tr>
<tr>
<td>4.3 Infrastructure and Integration</td>
<td>274</td>
</tr>
<tr>
<td>4. Particle Flow Trigger</td>
<td>2090</td>
</tr>
<tr>
<td>5.1 GT boards</td>
<td>414</td>
</tr>
<tr>
<td>5.2 Infrastructure and Integration</td>
<td>66</td>
</tr>
<tr>
<td>5. Global Trigger</td>
<td>480</td>
</tr>
<tr>
<td>6.1 DS boards</td>
<td>296</td>
</tr>
<tr>
<td>6.2 Infrastructure and Integration</td>
<td>78</td>
</tr>
<tr>
<td>6. Data Scouting System</td>
<td>374</td>
</tr>
<tr>
<td>L1 trigger upgrade total</td>
<td>7058</td>
</tr>
</tbody>
</table>
Appendix: Glossary of Special Terms and Acronyms

- $\mu$ROS = $\mu$TCA based Read-Out System, for DT readout.
- AB7 = Alpha Backend 7 board, equipping the DT slice test and featuring the Xilinx FPGA Virtex-7 chip.
- ADC = Analog-to-digital converter.
- ALCT = Anode Local Charged Track segment, theta view, part of the CSC system.
- AM = Analytical Method algorithm for DT trigger primitive generation.
- AMC13 = A $\mu$TCA data concentration and clock distribution card specifically designed for the CMS experiment.
- AMC = Advanced Mezzanine Card (from the ATCA specification).
- AP = Advanced Processor.
- APd = Advanced Processor demonstrator.
- APx = Advanced Processor board type.
- APD = Avalanche Photo Diode.
- ASIC = Application Specific Integrated Circuit.
- ATCA = Advanced TeleCommunications Architecture.
- AXI = Advanced eXtensible Interface.
- BCT = Barrel Calorimeter Trigger.
- BASE-T = A standard for Gigabit Ethernet over copper wiring.
- BCP = Barrel Calorimeter Processor.
- BDT = Boosted Decision Tree, is a multivariate analysis method.
- BMTF = Barrel Muon Track Finder.
- BMTL1 = Barrel Muon Trigger Layer-1.
- BRAM = Block RAM (FPGA resource).
- BRIL = Luminosity and beam monitoring detectors.
- BSM = Beyond the Standard Model.
- BW = Bandwidth.
- BX = Bunch crossing.
- CAD = Computer Aided Design.
- CDB = Controller Development Board.
- CDR = Clock and Data Recovery.
- CE = Calorimeter Endcap.
- CE-E = Calorimeter Endcap Electromagnetic.
- CE-H = Calorimeter Endcap Hadronic.
- CE-TPG = Calorimeter Endcap Trigger Primitive Generator.
- CERN = European Laboratory for Particle Physics.
- CFEB = Cathode Front-End Board, part of the CSC system.
- CHT = Compact Hough Transform.
- CL = Confidence Level.
• CLCT = Cathode Local Charged Track (cathode view muon stub), part of the CSC system.
• CMC = Common Mezzanine Card.
• CMS = Compact Muon Solenoid experiment.
• CMSSW = Compact Muon Solenoid Software, is the CMS experiment software package.
• COM-Express = Computer On Module form factor, a compact compute core used as part of System on Chip devices.
• CPU = Central Processing Unit.
• CRC = Cyclical-redundancy check, a widely-used family of algorithms for identifying data corruption.
• CSC = Cathode Strip Chamber.
• CT = Correlator Trigger.
• CTL1 = Correlator Trigger Layer-1.
• CTL2 = Correlator Trigger Layer-2.
• CTP7 = Calorimeter Trigger Processor 7 card, featuring the Xilinx FPGA Virtex-7 chip.
• CT-PPS = TOTEM precision proton spectrometer trigger.
• DAQ = Data AcQuisition.
• DDR = Double Data Rate, the transfer of data on a synchronous digital link at both the rising and falling edges of a clock.
• DDR4 SO-DIMM = Fourth-generation small outline dual in-line memory module, a type of dynamic random-access memory with a high-speed, synchronous data transfer interface
• DIMM = Dual in-line memory module.
• DMA = Direct Memory Access.
• DSP = Digital Signal processing.
• DT = Drift Tubes.
• DTH = DAQ and TCDS Hub
• EB = Barrel portion of ECAL covering pseudorapidity below 1.5.
• EB TPG = ECAL Barrel Trigger Primitive Generator.
• ECAL = Electromagnetic CALorimeter.
• ECON-T = Endcap CONcentrator Trigger.
• EDR = Engineering Design Report.
• EEPROM = Electrically-Erasable Programmable Read-Only Memory.
• EG = Electron-Gamma, normally used in the context of a trigger algorithm that will select events with electron or gamma candidates.
• ELM = Embedded Linux Mezzanine.
• EM = Electro-Magnetic.
• EMP = Extensible Modular data Processor
• EMTF = Endcap Muon Track Finder.
• EoF = End of file
• ESM = Ethernet Switch Mezzanine
• ESR = Electronic Systems Review.
• ETmiss = 2-vector sum of transverse energy over the calorimeter systems.
• ET = Scalar sum of transverse energy components over the calorimeter systems.
• FEB = Front End Board.
• FEE = Front-End Electronics.
• FE = Front-End.
• FF = Flip-flop.
• FIFO = First-In First-Out logic device that can be used to store and retrieve data.
• FPGA = Field-Programmable Gate Array.
• FSM = Finite State Machine.
• FTP = File transfer protocol
• GbE = Gigabit Ethernet, is a term describing various technologies for transmitting Ethernet frames at a rate of a gigabit per second.
• GBT = Gigabit Transceiver Project developed at CERN, source of the GBTX and associated chips.
• GCT = Global Calorimeter Trigger.
• GEM = Gas Electron Multiplier, is a type of gaseous ionization detector used in nuclear and particle physics.
• GMT = Global Muon Trigger.
• GPIO = General Purpose Input Output.
• GT = Level-1 Global Trigger.
• GTT = Global Track Trigger.
• GTx = Gigabit transceiver types.
• HB = Barrel portion of HCAL covering pseudorapidity less than 1.2, also: Histogram Based algorithm for DT trigger primitive generation.
• HCAL = Hadronic CALorimeter.
• HDL = Hardware Description Language.
• H/E = Ratio of energy deposits between HCAL and ECAL trigger towers.
• HF = Very forward portion of HCAL covering pseudorapidity between 3 and 5.
• HGCROC = High-Granularity Calorimeter ReadOut Chip.
• HI = Heavy Ions, at the LHC refers to collisions between lead ions.
• HL-LHC = High Luminosity Large Hadron Collider, is an upgrade planned for the current LHC machine.
• HLS = High-level Synthesis, a compiler that translates code written in high-level programming language to a register transfer level hardware description language.
• HLS4ML = HLS for Machine Learning
• HLT = High Level Trigger, a collection of software trigger algorithms.
• HP = Horizontal pitch unit (5.08 mm)
• HPS = Hadron Plus Strips, a tau lepton reconstruction and identification algorithm.
• HSCP = Heavy Stable Charged Particles.
• HT = Magnitude of the vectorial sum of
• HVAC = Heating, ventilation, and air conditioning
• I2C = Inter-Integrated Circuit chip-to-chip communications protocol.
• IBERT = Integrated Bit Error Rate Tester.
• ID = Identification, algorithm/method to identify particle candidate objects.
• II = Initialization interval, number of clock cycles after which a firmware algorithm can accept a new set of inputs.
• I/O = Input/Output.
• IPBus = is a protocol to control and communicate with Ethernet-attached xTCA hardware.
• IP = Impact Parameter, minimum distance between a trajectory and a selected vertex. Alternatively, Intellectual Property, an FPGA firmware algorithm that may be protected by industrial patents.
• IPMC = Intelligent Platform Management Controller.
• IPMI = Intelligent Platform Management Interface, is a standardised computer system interface used by system administrators.
• iRPC = Improved Resistive Plate Chambers, additional Resistive Plate Chambers that extend the coverage of an upgraded Muon System.
• JTAG = Joint Test Action Group; test and diagnostic bus standard by IEEE1149.1.
• KBMTF = Kalman filter Barrel Muon Track Finder
• KU115 = an FPGA from the Xilinx Kintex Ultrascale series
• L1A = Level-1 Accept.
• L1 = Level-1.
• L1T = Level-1 Trigger.
• L1HPS = L1 Hadron Plus Strips, a tau lepton reconstruction and identification algorithm designed for the Level-1 Trigger and based on the offline HPS algorithm.
• LB = Link Board.
• LCT = Local Charged Track, or muon stub, part of the CSC system.
• LHCC = Large Hadron Collider Experiments Committee.
• LHC = Large Hadron Collider.
• lpGBT = low power GigaBit Transceiver.
• LS2 = Long Shutdown 2, second LHC long shutdown scheduled for around the end of 2018.
• LS3 = Long Shutdown 3, third LHC long shutdown scheduled for around the end of 2023.
• LSB = Least Significant Bit.
• LSP = Lightest Supersymmetric Particle, is the generic name given to the lightest of the additional hypothetical particles found in supersymmetric models.
• LLUT = Large Look-up Memory Table Mezzanine
• LUT = Look-Up Table (memory, or FPGA resource).
• LVDS = Low Voltage Differential Signaling, a specification for differential digital
logic.

- MB = Muon Barrel Stations.
- ME = Muon Endcap Stations.
- MET = Missing Transverse Energy.
- MGT = Multi-Gigabit Transceiver.
- minbias = Soft proton-proton collision.
- MiniDIMM = Mini Dual In-line Memory Module.
- MIP = Minimum Ionizing Particle (muon).
- ML = Machine Learning
- MLB = Master Link Board.
- MMT = Majority Mean Timer.
- MP7 = Master Processor 7 card, featuring the Xilinx FPGA Virtex-7 chip.
- MPC = Muon Port Card, part of the CSC system.
- MPO = Optical connector standard.
- MTF7 = Muon Track Finder 7 card, featuring the Xilinx FPGA Virtex-7 chip.
- MTP = Optical fiber connector standard allowing up to 12 fibers in a single connector.
- MVA = Multivariate analysis
- µTCA = Micro Telecommunications Computing Architecture.
- NN = Neural Network.
- OBDT = On-Board Electronics for DT system.
- OCEAN =
- OH = OptoHybrid board
- OMTF = Overlap Muon Track Finder.
- OTMB = Optical Trigger motherboard
- PCB = Printed Circuit Board, is used to mechanically support and electrically connect electronic components.
- PCIe = Peripheral Component Interconnect Express, is a high-speed serial computer expansion bus standard.
- PCI = Personal Computer Interface Bus.
- PC = Personal Computer.
- PetiROC = extended analogue silicon photomultiplier Read-Out Chip.
- PF = Particle-flow reconstruction.
- pileup = overlapping of multiple soft interactions during a single LHC beam crossing.
- PICMG = PCI Industrial Computer Manufacturers Group
- PHY = Physical transceiver.
- PRBS31 = PseudoRandom Binary Sequence involving a generating polynomial of the form $x^{31} + x^{28} + 1$.
- pseudorapidity ($\eta$) = $-\ln \tan \frac{\theta}{2}$, where $\theta$ is the angle of the particle momentum with respect to the anti-clockwise beam direction.
• pT = Transverse momentum of a physics object.
• PU = Pileup, average quantity of particle collisions per bunch crossing.
• PUPPI = Pileup Per Particle Identification.
• PV = Primary Vertex.
• PYTHIA = A program to simulate high energy particle interactions.
• QCD = Quantum Chromodynamics - a theory that describes strong interactions.
• QSFP = Quad Small Form-factor Pluggable
• QSH =
• QSPI = Queued Serial Peripheral Interface
• RAID = Redundant Array of Independent Disks.
• RAM = Random Access Memory.
• RAW = Raw Data, Data tier format that contains the information read directly from the detector.
• RB = Readout Board, also: RPC Barrel.
• RCT = Regional Calorimeter Trigger.
• R&D = Research and Development.
• RE = RPC Endcap.
• RECO = Reconstructed Data, Data tier format that contains the reconstructed physics objects information from the CMS recorded data.
• RO = Readout.
• RPC = Resistive Plate Chamber.
• RTD = Resistance Temperature Detector.
• RTL = Register Transfer Level
• RTM = Rear Transition Module.
• RTOS = Real Time Operating System.
• RUI = Readout Unit Interface.
• Rx = Optical Receiver.
• SF = Simplified framing.
• SGMII = serial gigabit media-independent interface
• SLB = Slave Link Board.
• SL = SuperLayer.
• SLR = Super Logic Region
• SoC = System on Chip.
• SoF = Start of file
• SM = Standard Model of Particle Physics, is a theory concerning the electromagnetic, weak, and strong nuclear interactions.
• SMA = SubMiniature version A connector
• SMASH = Serenity management shell
• SP = Combined DT + RPC Super Primitive.
• SPI = Serial Peripheral Interface.
- SSD = Solid State-Drive.
- SUSY = SUper SYmmetry - an as yet unobserved symmetry between fermions and bosons.
- SW = Software.
- SWATCH = SoftWare for Automating the conTrol of Common Hardware
- TIA = Trans-Impedance Amplifier
- TCDS = Trigger Control and Distribution System.
- TDC = Trigger Data Concentrator card, also: Time-to-digital converter.
- TF = Track Finder corresponding to the Track Trigger Primitive Generator.
- TP = Trigger Primitive.
- TMUX = Time-Multiplex.
- TMB = Trigger motherboard
- TM7 = TwinMux 7 board, equipping the Phase-1 barrel muon trigger and featuring the Xilinx FPGA Virtex-7 chip.
- TMT = Time-Multiplexed Trigger, trigger design that processes events in parallel rather than sequentially.
- TPG = Trigger Primitive Generator.
- TriDAS = Trigger and DAQ.
- TRK = Outer Tracking Detector.
- TTC = Trigger Timing and Control, a system for distribution of clocking and control.
- TTS = Trigger Throttle System.
- TwinMux = Twin Multiplexer, data concentrator for muon barrel track finder.
- Tx = Optical Transmitter.
- UART = Universal asynchronous receiver-transmitter.
- uHAL = µTCA Hardware Access Libraries
- USC = Underground Service Cavern, where the CMS counting room is located.
- VBF = Vector Boson Fusion.
- VHDL = Very-high-speed integrated circuits Hardware Description Language.
- VME = Electronics and mechanics standard for crates, buses and application boards.
- VU9P = an FPGA based on the Xilinx Virtex Ultrascale+ series.
- XVC = Xilinx Virtual Cable.
- ZYNQ = Xilinx embedded compute core, used in System on Chip devices.
References


CMS Collaboration

Yerevan Physics Institute, Yerevan, Armenia
A.M. Sirunyan, A. Tumasyan

Institut für Hochenergiephysik, Wien, Austria

Institute for Nuclear Problems, Minsk, Belarus
V. Chekhovsky, A. Litomin, V. Makarenko, J. Suarez Gonzalez

Universiteit Antwerpen, Antwerpen, Belgium

Vrije Universiteit Brussel, Brussel, Belgium

Université Libre de Bruxelles, Bruxelles, Belgium

Ghent University, Ghent, Belgium

Université Catholique de Louvain, Louvain-la-Neuve, Belgium

Centro Brasileiro de Pesquisas Fisicas, Rio de Janeiro, Brazil
G.A. Alves, G. Correia Silva, C. Hensel, A. Moraes

Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil

Universidade Estadual Paulista, Universidade Federal do ABC, São Paulo, Brazil
C.A. Bernardes, L. Calligaris, T.R. Fernandez Perez Tomei, E.M. Gregores, D.S. Lemos, P.G. Mercadante, S.F. Novaes, Sandra S. Padula

Institute for Nuclear Research and Nuclear Energy, Bulgarian Academy of Sciences, Sofia, Bulgaria
A. Aleksandrov, G. Antchev, I. Atanasov, R. Hadjiiiska, P. Iaydijiev, M. Misheva, M. Rodozov, M. Shopova, G. Sultanov
University of Sofia, Sofia, Bulgaria
M. Bonchev, A. Dimitrov, T. Ivanov, L. Litov, B. Pavlov, P. Petkov, A. Petrov

Beihang University, Beijing, China
W. Fang², X. Gao², Q. Guo, H. Wang, L. Yuan

Department of Physics, Tsinghua University, Beijing, China
M. Ahmad, Z. Hu, Y. Wang

Institute of High Energy Physics, Beijing, China

State Key Laboratory of Nuclear Physics and Technology, Peking University, Beijing, China
A. Agapitos, Y. Ban, C. Chen, G. Chen, A. Levin, J. Li, L. Li, Q. Li, X. Lyu, Y. Mao, S.J. Qian, D. Wang, Q. Wang, J. Xiao

Sun Yat-Sen University, Guangzhou, China
Z. You

Zhejiang University, Hangzhou, China
M. Xiao

Universidad de Los Andes, Bogota, Colombia
C. Avila, A. Cabrera, C. Florez, J. Fraga, A. Sarkar, M.A. Segura Delgado

Universidad de Antioquia, Medellin, Colombia
J. Mejia Guisao, J.D. Ruiz Alvarez, C.A. Salazar Gonzalez, N. Vanegas Arbelaez

University of Split, Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, Split, Croatia
D. Giljanovic, N. Godinovic, D. Lelas, I. Puljak, T. Sculac

University of Split, Faculty of Science, Split, Croatia
Z. Antunovic, M. Kovac

Institute Rudjer Boskovic, Zagreb, Croatia
V. Brigljevic, D. Ferencek, D. Majumder, B. Mesic, M. Roguljic, A. Starodumov⁹, T. Susa

University of Cyprus, Nicosia, Cyprus

Charles University, Prague, Czech Republic

Escuela Politecnica Nacional, Quito, Ecuador
E. Ayala

Universidad San Francisco de Quito, Quito, Ecuador
E. Carrera Jarrin

Academy of Scientific Research and Technology of the Arab Republic of Egypt, Egyptian Network of High Energy Physics, Cairo, Egypt
H. Abdalla¹¹, Y. Assran¹²,¹³, E. Salama¹³,¹⁴
Center for High Energy Physics (CHEP-FU), Fayoum University, El-Fayoum, Egypt
A. Lotfy\textsuperscript{15}, M.A. Mahmoud

National Institute of Chemical Physics and Biophysics, Tallinn, Estonia
S. Bhowmik, A. Carvalho Antunes De Oliveira, R.K. Dewanjee, K. Ehataht, M. Kadastik, M. Raidal, C. Veelken

Department of Physics, University of Helsinki, Helsinki, Finland
P. Eerola, L. Forthomme, H. Kirschenmann, K. Osterberg, M. Voutilainen

Helsinki Institute of Physics, Helsinki, Finland
E. Brücken, F. Garcia, J. Havukainen, V. Karimäki, M.S. Kim, R. Kinnunen, T. Lampén, K. Lassila-Perini, S. Laurila, S. Lehti, T. Lindén, H. Siikonen, E. Tuominen, J. Tuominiemi

Lappeenranta University of Technology, Lappeenranta, Finland
P. Luukka, T. Tuuva

IRFU, CEA, Université Paris-Saclay, Gif-sur-Yvette, France

Laboratoire Leprince-Ringuet, CNRS/IN2P3, Ecole Polytechnique, Institut Polytechnique de Paris, Paris, France

Université de Strasbourg, CNRS, IPHC UMR 7178, Strasbourg, France

Université de Lyon, Université Claude Bernard Lyon 1, CNRS-IN2P3, Institut de Physique Nucléaire de Lyon, Villeurbanne, France

Georgian Technical University, Tbilisi, Georgia
A. Khvedelidze\textsuperscript{10}

Tbilisi State University, Tbilisi, Georgia
Z. Tsamalaidze\textsuperscript{10}

RWTH Aachen University, I. Physikalisches Institut, Aachen, Germany

RWTH Aachen University, III. Physikalisches Institut A, Aachen, Germany
Budapest, Hungary

Wigner Research Centre for Physics, Budapest, Hungary
G. Bencze, C. Hajdu, D. Horváth, F. Sikler, V. Veszpremi, G. Vesztergombi

Institute of Nuclear Research ATOMKI, Debrecen, Hungary
N. Beni, S. Czellar, J. Karancsi, J. Molnar, Z. Szillasi, D. Teyssier

Institute of Physics, University of Debrecen, Debrecen, Hungary
P. Raics, Z.L. Trocsányi, B. Ujvari

Eszterházy Karoly University, Karoly Robert Campus, Gyongyos, Hungary
T. Csurgo, S. Lőkös, F. Nemes, T. Novák

Indian Institute of Science (IISc), Bangalore, India
S. Choudhury, J.R. Komaragiri, D. Kumar, L. Panwar, P.C. Tiwari

National Institute of Science Education and Research, HBNI, Bhubaneswar, India

Panjab University, Chandigarh, India

University of Delhi, Delhi, India
A. Ahmed, A. Bhardwaj, B.C. Choudhary, R.B. Garg, M. Gola, S. Keshri, A. Kumar, M. Naimuddin, P. Priyanka, K. Ranjan, A. Shah, R. Sharma

Saha Institute of Nuclear Physics, HBNI, Kolkata, India

Indian Institute of Technology Madras, Madras, India

Bhabha Atomic Research Centre, Mumbai, India
D. Dutta, V. Jha, D.K. Mishra, K. Naskar, P.K. Netrakanti, L.M. Pant, P. Shukla

Tata Institute of Fundamental Research-A, Mumbai, India
T. Aziz, M.A. Bhat, S. Dugad, R. Kumar Verma, U. Sarkar

Tata Institute of Fundamental Research-B, Mumbai, India

Indian Institute of Science Education and Research (IISER), Pune, India
S. Dube, B. Kansal, A. Kapoor, K. Kothekar, S. Pandey, A. Rane, A. Rastogi, S. Sharma

Department of Physics, Isfahan University of Technology, Isfahan, Iran
H. Bakhshiansohi
Seoul National University, Seoul, Korea

University of Seoul, Seoul, Korea

Yonsei University, Department of Physics, Seoul, Korea
H.D. Yoo

Sungkyunkwan University, Suwon, Korea

Riga Technical University, Riga, Latvia
V. Veckalns

Vilnius University, Vilnius, Lithuania
A. Juodagalvis, A. Rinkevicius, G. Tamulaitis

National Centre for Particle Physics, Universiti Malaya, Kuala Lumpur, Malaysia
W.A.T. Wan Abdullah, M.N. Yusif, Z. Zolkapli

Universidad de Sonora (UNISON), Hermosillo, Mexico
J.F. Benitez, A. Castaneda Hernandez, J.A. Murillo Quijada, L. Valencia Palomo

Centro de Investigacion y de Estudios Avanzados del IPN, Mexico City, Mexico
H. Castilla-Valdez, E. De La Cruz-Burelo, I. Heredia-De La Cruz, R. Lopez-Fernandez, A. Sanchez-Hernandez

Universidad Iberoamericana, Mexico City, Mexico
S. Carrillo Moreno, C. Orobeza Barrera, M. Ramirez-Garcia, F. Vazquez Valencia

Benemerita Universidad Autonoma de Puebla, Puebla, Mexico
J. Eysermans, I. Pedraza, H.A. Salazar Ibarguen, C. Uribe Estrada

Universidad Autónoma de San Luis Potosí, San Luis Potosí, Mexico
A. Morelos Pineda

University of Montenegro, Podgorica, Montenegro
J. Mijuskovic, N. Raicevic

University of Auckland, Auckland, New Zealand
D. Krofcheck

University of Canterbury, Christchurch, New Zealand
S. Bheesette, P.H. Butler

National Centre for Physics, Quaid-I-Azam University, Islamabad, Pakistan

AGH University of Science and Technology Faculty of Computer Science, Electronics and Telecommunications, Krakow, Poland
V. Avati, L. Grzanka, M. Malawski

National Centre for Nuclear Research, Swierk, Poland
Institute of Experimental Physics, Faculty of Physics, University of Warsaw, Warsaw, Poland

Laboratório de Instrumentação e Física Experimental de Partículas, Lisboa, Portugal

Joint Institute for Nuclear Research, Dubna, Russia

Petersburg Nuclear Physics Institute, Gatchina (St. Petersburg), Russia
G. Gavrilov, V. Golovtcov, Y. Ivanov, V. Kim, E. Kuznetsova, V. Murzin, V. Oreshkin, I. Smirnov, D. Sokn, V. Sulimov, L. Uvarov, S. Volkov, A. Vorobyev

Institute for Nuclear Research, Moscow, Russia

Institute for Theoretical and Experimental Physics named by A.I. Alikhanov of NRC ‘Kurchatov Institute’, Moscow, Russia
V. Epshteyn, V. Gavrilo, N. Lychkovskaya, A. Nikitenko, V. Popov, I. Pozdnyakov, G. Safronov, A. Spiridonov, A. Stepenov, M. Toms, E. Vlasov, A. Zhokin

Moscow Institute of Physics and Technology, Moscow, Russia
T. Aushev

National Research Nuclear University 'Moscow Engineering Physics Institute' (MEPhI), Moscow, Russia
M. Danilov, P. Parygin, D. Philippov, S. Polikarpov, E. Popova

P.N. Lebedev Physical Institute, Moscow, Russia
V. Andreev, M. Azarkin, I. Dremin, M. Kirakosyan, A. Terkulov

Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University, Moscow, Russia

Novosibirsk State University (NSU), Novosibirsk, Russia
V. Blinov, T. Dimova, L. Kardapoltsev, I. Ovtin, Y. Skovpen

Institute for High Energy Physics of National Research Centre ‘Kurchatov Institute’, Protvino, Russia

National Research Tomsk Polytechnic University, Tomsk, Russia
A. Babaev, A. Iuzhakov, V. Okhotnikov

Tomsk State University, Tomsk, Russia
V. Borchsh, V. Ivanchenko, E. Tcherniaev
University of Belgrade: Faculty of Physics and VINCA Institute of Nuclear Sciences, Belgrade, Serbia
P. Adzic, P. Cirkovic, M. Dordevic, P. Milenovic, J. Milosevic, M. Stojanovic

Centro de Investigaciones Energéticas Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain

Universidad Autónoma de Madrid, Madrid, Spain
C. Albajar, J.F. de Trocóniz, R. Reyes-Almanza

Universidad de Oviedo, Instituto Universitario de Ciencias y Tecnologías Espaciales de Asturias (ICTEA), Oviedo, Spain

Instituto de Física de Cantabria (IFCA), CSIC-Universidad de Cantabria, Santander, Spain

University of Colombo, Colombo, Sri Lanka
MK Jayananda, B. Kailasapathy, D.U.J. Sonnadara, DDC Wickramarathna

University of Ruhuna, Department of Physics, Matara, Sri Lanka
W.G.D. Dharmaratna, K. Liyanage, N. Perera, N. Wickramage

CERN, European Organization for Nuclear Research, Geneva, Switzerland

Paul Scherrer Institut, Villigen, Switzerland
ETH Zurich - Institute for Particle Physics and Astrophysics (IPA), Zurich, Switzerland

Universität Zürich, Zurich, Switzerland

National Central University, Chung-Li, Taiwan
C. Adloff, C.M. Kuo, W. Lin, A. Roy, T. Sarkar, S.S. Yu

National Taiwan University (NTU), Taipei, Taiwan

Chulalongkorn University, Faculty of Science, Department of Physics, Bangkok, Thailand
B. Asavapibhop, C. Asawatangtrakuldee, N. Srimanobhas

Çukurova University, Physics Department, Science and Art Faculty, Adana, Turkey

Middle East Technical University, Physics Department, Ankara, Turkey
B. Isildak, G. Karapinar, K. Ocalan, M. Yalvac

Bogazici University, Istanbul, Turkey
I.O. Atakisi, E. Gülmez, M. Kaya, O. Kayas, Ö. Özcelyk, S. Tekten, E.A. Yetkin

Istanbul Technical University, Istanbul, Turkey
A. Cakir, K. Cankocak, Y. Komurcu, S. Sen

Istanbul University, Istanbul, Turkey
F. Aydogmus Sen, S. Cerci, B. Kaynak, S. Ozkorucuklu, D. Sunar Cerci

Institute for Scintillation Materials of National Academy of Science of Ukraine, Kharkov, Ukraine
B. Grynyov

National Scientific Center, Kharkov Institute of Physics and Technology, Kharkov, Ukraine
L. Levchuk

University of Bristol, Bristol, United Kingdom

Rutherford Appleton Laboratory, Didcot, United Kingdom
K.W. Bell, A. Belyaev, C. Brew, R.M. Brown, D.J.A. Cockerill, K.V. Ellis, K. Harder, S. Harper,
California Institute of Technology, Pasadena, USA

Carnegie Mellon University, Pittsburgh, USA

University of Colorado Boulder, Boulder, USA

Cornell University, Ithaca, USA

Fermi National Accelerator Laboratory, Batavia, USA

University of Florida, Gainesville, USA

Florida International University, Miami, USA
Y.R. Joshi

Florida State University, Tallahassee, USA

Florida Institute of Technology, Melbourne, USA

University of Illinois at Chicago (UIC), Chicago, USA

The University of Iowa, Iowa City, USA
Johns Hopkins University, Baltimore, USA
O. Amram, B. Blumenfeld, L. Corcodilos, M. Eminizer, A.V. Gritsan, S. Kyriacou,
P. Maksimovic, C. Mantilla, J. Roskes, M. Swartz, T.Á. Vámi

The University of Kansas, Lawrence, USA
C. Baldenegro Barrera, P. Baringer, A. Bean, A. Bylinkin, T. Isidori, S. Khalil, J. King, G. Krintiras,
A. Kropivnitskaya, C. Lindsey, W. Mcbrayer, N. Minafra, M. Murray, C. Rogan, C. Royon,
S. Sanders, E. Schmitz, J.D. Tapia Takaki, Q. Wang, J. Williams, G. Wilson

Kansas State University, Manhattan, USA

Lawrence Livermore National Laboratory, Livermore, USA
F. Rebassoo, D. Wright

University of Maryland, College Park, USA
E. Adams, A. Baden, O. Baron, A. Belloni, S.C. Eno, Y. Feng, N.J. Hadley, S. Jabeen, G.Y. Jeng,
K. Wong

Massachusetts Institute of Technology, Cambridge, USA
D. Abercrombie, B. Allen, R. Bi, S. Brandt, W. Busza, I.A. Cali, Y. Chen, M. D'Alfonso,
G. Gomez Ceballos, M. Goncharov, P. Harris, D. Hsu, M. Hu, M. Klute, D. Kovalskyi, J. Krupa,
Y.-J. Lee, P.D. Luckey, B. Maier, A.C. Marini, C. McGinn, C. Mironov, S. Narayanan, X. Niu,
J. Wang, T.W. Wang, B. Wyslouch

University of Minnesota, Minneapolis, USA
R.M. Chatterjee, A. Evans, S. Guts†, P. Hansen, J. Hiltbrand, Sh. Jain, M. Krohn, Y. Kubota,
Z. Lesko, J. Mans, M. Revering, R. Rusack, R. Saradhy, N. Schroeder, N. Strobbe, M.A. Wadud

University of Mississippi, Oxford, USA
J.G. Acosta, S. Oliveros

University of Nebraska-Lincoln, Lincoln, USA
K. Bloom, S. Chauhan, D.R. Claes, C. Fangmeier, L. Finco, F. Golf, J.R. González Fernández,
I. Kravchenko, J.E. Siado, G.R. Snow†, B. Stieger, W. Tabb

State University of New York at Buffalo, Buffalo, USA
G. Agarwal, C. Harrington, I. Iashvili, A. Kharchilava, C. McLean, D. Nguyen, A. Parker,
J. Pekkanen, S. Rappoccio, B. Roozbahani

Northeastern University, Boston, USA
G. Alverson, E. Barberis, C. Freer, Y. Haddad, A. Hortiangtham, J. Li, G. Madigan, B. Marzocchi,
D.M. Morse, V. Nguyen, T. Orimoto, L. Skinnari, A. Tishelman-Charny, T. Wamorkar, B. Wang,
A. Wisecarver, D. Wood

Northwestern University, Evanston, USA
S. Bhattacharya, J. Bueghly, Z. Chen, A. Gilbert, T. Gunter, K.A. Hahn, N. Odell, M.H. Schmitt,
K. Sung, M. Velasco

University of Notre Dame, Notre Dame, USA
R. Bucci, N. Dev, R. Goldouzian, M. Hildreth, K. Hurtado Anampa, C. Jessop, D.J. Karmgard,

The Ohio State University, Columbus, USA
J. Alimena, B. Bylsma, B. Cardwell, L.S. Durkin, B. Francis, C. Hill, W. Ji, A. Lefeld, B.L. Winer, B.R. Yates

Princeton University, Princeton, USA

University of Puerto Rico, Mayaguez, USA
S. Malik, S. Norberg

Purdue University, West Lafayette, USA

Wayne State University, Detroit, USA
P.E. Karchin, N. Poudyal, J. Sturdy, P. Thapa
University of Wisconsin - Madison, Madison, WI, USA

‡: Deceased
1: Also at Vienna University of Technology, Vienna, Austria
2: Also at Université Libre de Bruxelles, Bruxelles, Belgium
3: Also at IRFU, CEA, Université Paris-Saclay, Gif-sur-Yvette, France
4: Also at Universidade Estadual de Campinas, Campinas, Brazil
5: Also at Federal University of Rio Grande do Sul, Porto Alegre, Brazil
6: Also at UFMS, Nova Andradina, Brazil
7: Also at Universidade Federal de Pelotas, Pelotas, Brazil
8: Also at University of Chinese Academy of Sciences, Beijing, China
9: Also at Institute for Theoretical and Experimental Physics named by A.I. Alikhanov of NRC ‘Kurchatov Institute’, Moscow, Russia
10: Also at Joint Institute for Nuclear Research, Dubna, Russia
11: Also at Cairo University, Cairo, Egypt
12: Also at Suez University, Suez, Egypt
13: Now at British University in Egypt, Cairo, Egypt
14: Now at Ain Shams University, Cairo, Egypt
15: Now at Fayoum University, El-Fayoum, Egypt
16: Also at Purdue University, West Lafayette, USA
17: Also at Université de Haute Alsace, Mulhouse, France
18: Also at Erzincan Binali Yildirim University, Erzincan, Turkey
19: Also at CERN, European Organization for Nuclear Research, Geneva, Switzerland
20: Also at RWTH Aachen University, III. Physikalisches Institut A, Aachen, Germany
21: Also at University of Hamburg, Hamburg, Germany
22: Also at Department of Physics, Isfahan University of Technology, Isfahan, Iran, Isfahan, Iran
23: Also at Brandenburg University of Technology, Cottbus, Germany
24: Also at Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University, Moscow, Russia
25: Also at Institute of Physics, University of Debrecen, Debrecen, Hungary, Debrecen, Hungary
26: Also at Physics Department, Faculty of Science, Assiut University, Assiut, Egypt
27: Also at Institute of Nuclear Research ATOMKI, Debrecen, Hungary
28: Also at MTA-ELTE Lendület CMS Particle and Nuclear Physics Group, Eötvös Loránd University, Budapest, Hungary
29: Also at IIT Bhubaneswar, Bhubaneswar, India, Bhubaneswar, India
30: Also at Institute of Physics, Bhubaneswar, India
31: Also at G.H.G. Khalsa College, Punjab, India
32: Also at Shoolini University, Solan, India
33: Also at University of Hyderabad, Hyderabad, India
34: Also at University of Visva-Bharati, Santiniketan, India
35: Also at Indian Institute of Technology (IIT), Mumbai, India
36: Now at Princeton University, Princeton, USA
37: Also at Deutsches Elektronen-Synchrotron, Hamburg, Germany
38: Also at Department of Physics, University of Science and Technology of Mazandaran, Behshahr, Iran
39: Now at INFN Sezione di Bari a, Università di Bari b, Politecnico di Bari c, Bari, Italy
40: Also at Italian National Agency for New Technologies, Energy and Sustainable Economic Development, Bologna, Italy
41: Also at Centro Siciliano di Fisica Nucleare e di Struttura Della Materia, Catania, Italy
42: Also at Riga Technical University, Riga, Latvia, Riga, Latvia
43: Also at Consejo Nacional de Ciencia y Tecnología, Mexico City, Mexico
44: Also at Warsaw University of Technology, Institute of Electronic Systems, Warsaw, Poland
45: Also at Institute for Nuclear Research, Moscow, Russia
46: Now at National Research Nuclear University 'Moscow Engineering Physics Institute' (MEPhI), Moscow, Russia
47: Also at St. Petersburg State Polytechnical University, St. Petersburg, Russia
48: Also at University of Florida, Gainesville, USA
49: Also at Imperial College, London, United Kingdom
50: Also at P.N. Lebedev Physical Institute, Moscow, Russia
51: Also at INFN Sezione di Padova a, Università di Padova b, Padova, Italy, Università di Trento c, Trento, Italy, Padova, Italy
52: Also at Budker Institute of Nuclear Physics, Novosibirsk, Russia
53: Also at Faculty of Physics, University of Belgrade, Belgrade, Serbia
54: Also at Università degli Studi di Siena, Siena, Italy
55: Also at Trincomalee Campus, Eastern University, Sri Lanka, Nilaveli, Sri Lanka
56: Also at INFN Sezione di Pavia a, Università di Pavia b, Pavia, Italy, Pavia, Italy
57: Also at National and Kapodistrian University of Athens, Athens, Greece
58: Also at Universität Zürich, Zurich, Switzerland
59: Also at Stefan Meyer Institute for Subatomic Physics, Vienna, Austria, Vienna, Austria
60: Also at Laboratoire d’Annecy-le-Vieux de Physique des Particules, IN2P3-CNRS, Annecy-le-Vieux, France
61: Also at Şırnak University, Sirnak, Turkey
62: Also at Department of Physics, Tsinghua University, Beijing, China, Beijing, China
63: Also at Near East University, Research Center of Experimental Health Science, Nicosia, Turkey
64: Also at Beykent University, Istanbul, Turkey, Istanbul, Turkey
65: Also at Istanbul Aydin University, Application and Research Center for Advanced Studies (App. & Res. Cent. for Advanced Studies), Istanbul, Turkey
66: Also at Mersin University, Mersin, Turkey
67: Also at Piri Reis University, Istanbul, Turkey
68: Also at Adiyaman University, Adiyaman, Turkey
69: Also at Ozyegin University, Istanbul, Turkey
70: Also at Izmir Institute of Technology, Izmir, Turkey
71: Also at Necmettin Erbakan University, Konya, Turkey
72: Also at Bozok Üniversitesi Rektörlüğü, Yozgat, Turkey
73: Also at Marmara University, Istanbul, Turkey
74: Also at Milli Savunma University, Istanbul, Turkey
75: Also at Kafkas University, Kars, Turkey
76: Also at Istanbul Bilgi University, Istanbul, Turkey
77: Also at Hacettepe University, Ankara, Turkey
78: Also at Vrije Universiteit Brussel, Brussel, Belgium
79: Also at School of Physics and Astronomy, University of Southampton, Southampton, United
Kingdom
80: Also at IPPP Durham University, Durham, United Kingdom
81: Also at Monash University, Faculty of Science, Clayton, Australia
82: Also at Bethel University, St. Paul, Minneapolis, USA, St. Paul, USA
83: Also at Karamanoğlu Mehmetbey University, Karaman, Turkey
84: Also at California Institute of Technology, Pasadena, USA
85: Also at Bingol University, Bingol, Turkey
86: Also at Georgian Technical University, Tbilisi, Georgia
87: Also at Sinop University, Sinop, Turkey
88: Also at Mimar Sinan University, Istanbul, Istanbul, Turkey
89: Also at Nanjing Normal University Department of Physics, Nanjing, China
90: Also at Texas A&M University at Qatar, Doha, Qatar
91: Also at Kyungpook National University, Daegu, Korea, Daegu, Korea
92: Also at Laboratório de Instrumentação e Física Experimental de Partículas, Lisboa, Portugal