



The CDF Data Acquisition System for Tevatron Run II

Arnd Meyer (Fermi National Accelerator Laboratory, Batavia, IL 60510, U.S.A)
for the CDF DAQ group

Abstract

The CDF experiment at the Fermilab Tevatron has been significantly upgraded for the collider Run II, which started in March 2001 and is scheduled to last until 2006. Instantaneous luminosities of $10^{32} \text{ cm}^{-2}\text{s}^{-1}$ and above are expected. A data acquisition system capable of efficiently recording the data has been one of the most critical elements of the upgrade. Key figures are the ability to deal with the short bunch spacing of 132 ns, event sizes of the order of 250 kB, and permanent logging of 20 MB/s. The design of the system and experience from the first months of data-taking operation are discussed.

Keywords: Data Acquisition, Colliders

1 Introduction

A functional block diagram of the new CDF readout scheme along with a schematic of the data acquisition system is shown in fig. 1. With minor exceptions, all electronics systems had to be replaced to accommodate the larger instantaneous luminosity in the Tevatron Run II, with a similar increase in data transfer rates, and more importantly, to allow for a bunch separation of as low as 132 ns.

The Level 1 trigger system is synchronized with the 132 ns clock, and forms a decision without incurring deadtime at the end of a 42-crossing deep pipeline ($5.5 \mu\text{s}$). Upon receiving a Level 1 trigger, the data on each front-end board are transferred to one of four Level 2 buffers. The Level 2 trigger operates asynchronously and has an average decision time of $20 \mu\text{s}$. Upon a positive Level 2 decision, the event is read out into DAQ buffers and then transferred via a network switch to the Level 3 filter farm, where the complete event is assembled, analyzed, and, if accepted, sent to the data logger.

The data acquisition is designed to be *partitionable*: parts of the system can be read out independently through the central DAQ, by assigning them to one of eight partitions. In general this is possible down to the level of single front-end crates, with the exception of the physics trigger and the Silicon systems. Partitioning is extremely useful for the efficient use of accelerator downtime to take various types of calibrations and for detector commissioning.

2 Front-End Electronics

All front-end and trigger electronics are packaged as standard VME modules and housed in 21-slot commercial VIPA (VME International Physics Association) crates; there are a total of about 120 crates in the system. In addition to the front-end modules, each crate contains at least one Motorola PPC processor board (MVME 2301 or better) for hardware initialization, event readout and/or monitoring, running under the VxWorks operating system.

All front-end and trigger crates except those serving the Silicon system¹ contain a standardized "Trigger and Clock + Event Readout" (TRACER) module. The TRACER provides the interface between the front-end modules and the Trigger System Interface (TSI) through fast serial control messages. Timing signals from the MasterClock are received by the TRACER and, as well as the trigger signals, distributed via the J2 backplane. It also provides the data interface between the front-end modules and the event building network, and sends information back to the TSI about when a Level 2 buffer can be reused, when error conditions occur, or when a backup from higher levels in the DAQ prevents a new trigger from being accepted.

¹The Silicon Readout Controller (SRC) performs similar functions in this case.

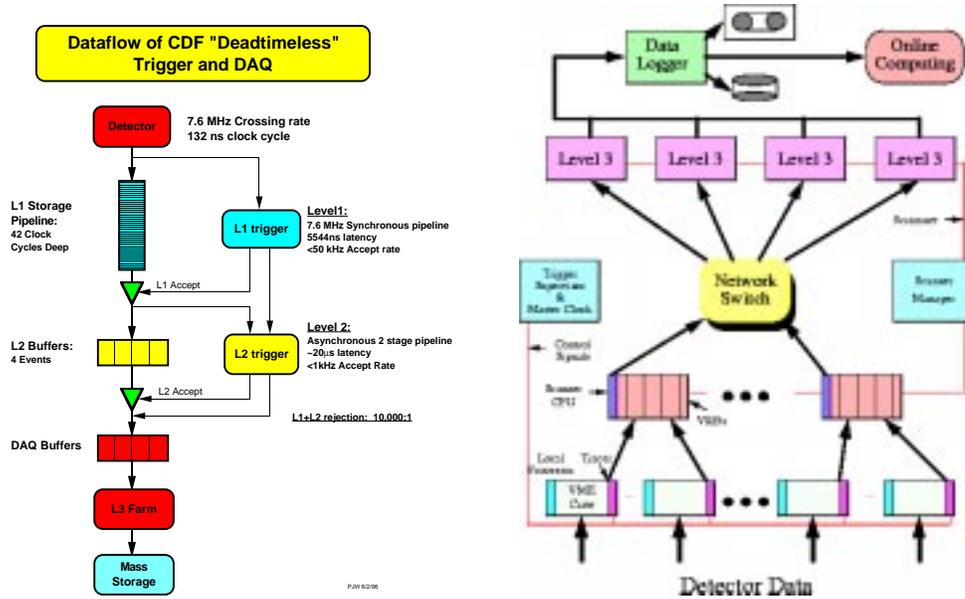


Figure 1: Schematic views of the CDF data acquisition and trigger system.

3 Trigger

The rates at the three trigger stages are about an order of magnitude larger than in Run I. The Level 1 accept rate will be about 40 kHz. Its decision is formed with the help of dedicated hardware, and uses information of the central drift chamber, the calorimeters, the muon systems, and the luminosity system, as well as combinations thereof including matching of central tracks to muon and electron signatures. A total of 64 different trigger conditions are presently possible.

Using refined and additional information, Level 2 reduces the rate to about 300 Hz. At this stage, information from the Silicon detectors is available in the form of a displaced vertex trigger; also, calorimeter clustering and improved matching of central drift chamber tracks to hits in the muon systems are beneficial. The Level 2 decisions are formed in a set of custom CPU boards based on DEC Alpha processors. It is worth noting that the bulk of the Level 1 trigger bandwidth (more than 50%) will be used by the hadronic B decay triggers, which can easily be reduced by a displaced vertex condition at the second trigger level.

4 Event Builder and Level 3 Trigger

The event builder is implemented around a commercial ATM switch running with 16 input ports delivering event fragments to 16 output ports. The I/O ports are connected via OC-3 optical fibers with a bandwidth of 16.2 MB/s. On the input side, MVME2603 processors scan data from the VME readout boards (VRB) which function as FIFO buffers into the switch. On the output side, 16 Intel processor based PCs running Linux, so-called converter nodes, receive and assemble the events for shipment to the Level 3 processing nodes (fig. 2) via fast Ethernet. There are 128 processing nodes in the current system, organized in 16 subfarms, all of which are Pentium III based dual-processor systems. The design input rate into Level 3 is about 75 MB/s (300 Hz), which is reduced by filter algorithms using standard CDF offline C++ code to ~ 20 MB/s (75 Hz). One or more subfarms are connected to an output node which receives accepted events and ships them to the data logger.

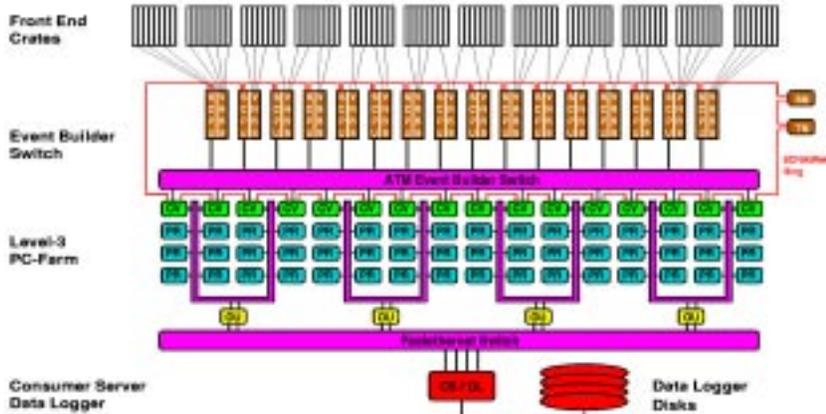


Figure 2: The Level 3 trigger architecture. SCPU: scanner CPU; CV: converter node; PR: processing node (presently 8 processing nodes per converter node); OU: output node.

An alternative data path exists for calibration runs and debugging purposes: event fragments can directly be sent over Ethernet from the VME readout controller in each front-end crate to a multi-threaded application running on a Linux PC. This *software event builder* assembles the events, albeit at a much lower rate than the main event builder, and sends them to the data logger.

5 Data Logger and Event Monitoring

The events that pass the Level 3 trigger are collected by a logger process [1] running on an SGI 2200 server. Data are logged in several streams (of the order 4 to 10) to dual-port fiber channel disk arrays, read from the second port in the computer center, and written to Sony AIT tapes. In addition to data logging at a sustained rate of 20 MB/s, some of the events are sent to remote “consumer” processes for online monitoring, at a total rate of up to 10 MB/s. Sending events to consumers must not impact the logging rate to permanent storage.

The three main components of the consumer system are online monitor analysis programs, a display GUI to visualize results of the monitoring programs, and display servers to establish communications between the monitoring programs and the display GUI through sockets. The analysis (monitor) programs are written by subsystem experts in a common framework. The system is coded in C++ using ROOT packages for physics analysis, network communications, and graphical manipulations. The display GUI is either run locally in the control room or remotely. The display and the display server are coded independently of any specific monitors. A total of about 10 core consumer programs are run routinely during data taking; typical examples are the online event display or consumers monitoring trigger performance.

6 Run Control

The CDF Run Control [2] is the top level application that controls the data acquisition activities across front-end and trigger VME crates and related service processes (fig. 3). Run Control is a real-time multi-threaded application implemented in Java 1.3 with several flexible state machines allowing sequencing and sophisticated error handling and recovery, and controlling the synchronization across the system. Run Control communicates to its distributed clients via connections to a commercial publish/subscribe message passing system, SmartSockets by Talarian. SmartSockets is also used by various Run Control clients to publish monitoring

information that is available both in the control room and on the web. Run Control clients are written in Java, C, or C++, and a custom high level API on top of SmartSockets is used to facilitate message passing. Some clients (e.g. Level 3 trigger) implement a proxy for communication with Run Control.

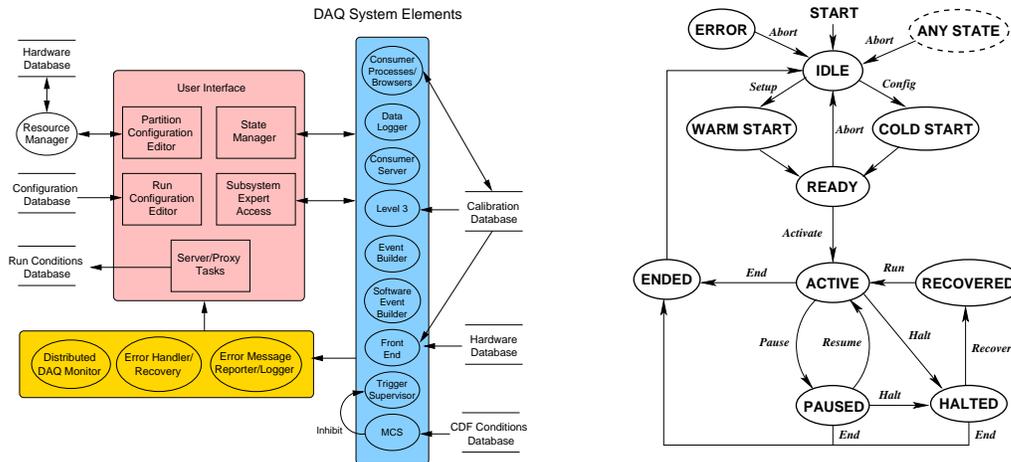


Figure 3: Components of Run Control, and the state machine used for physics data taking.

Run Control configures its clients using an Oracle database containing hardware, trigger, calibration and run configuration information; the custom database API uses JDBC and implements realtime and synchronized updating. Run Control also provides the graphical user interface for data taking and calibrations, as well as scripting capabilities using JPython. It can be run on any machine in the online cluster, consisting of about 40 control room PCs running Linux and several Linux and IRIX based servers.

7 Current Performance and Outlook

Most of the performance figures mentioned in previous sections have been reached or surpassed at the time of writing. Notable exceptions are the Level 2 trigger, which is still being commissioned, and about 50% of the Silicon system, which is currently being integrated. From observations with parts of these systems and test runs it can be concluded that no unsurmountable problems are to be expected that would prevent reaching the design specifications.

After collecting an integrated luminosity of about 2 fb^{-1} until the end of 2003 (Run IIa) and a shutdown of about 8 months, it is expected that about 15 fb^{-1} will be delivered by the Tevatron to CDF until 2006 (Run IIb). After the shutdown the accelerator will likely switch to 132 ns bunch spacing, in contrast to the 36 bunch operation (396 ns) that is currently being used. Studies are underway to identify parts of the DAQ that will require upgrades or replacement for the much larger trigger and data rates; in particular Level 3 input rates of 1 kHz may require replacement of parts of the event builder system (ATM switch). Upgrades to the Level 1 and Level 2 trigger system and incremental upgrades of the Level 3 computing power are also foreseen.

References

- [1] B.J. Kilminster *et al.*, “The CDF Consumer-Server/Logger system for Run II at the Tevatron”, and T. Arisawa *et al.*, “Online Monitor System of CDF Run II”, these proceedings
- [2] W. Badgett *et al.*, “CDF Run Control for Tevatron Run II”, these proceedings