FERMILAB-CONF-24-0233-ETD

Smart Pixels: In-pixel AI for on-sensor data filtering

Benjamin Parpillon[1,2], Chinar Syal[1], Jieun Yoo[2], Jennet Dickinson[1], Morris Swartz[3], Giuseppe Di Guglielmo[1,4], Alice Bean[5], Douglas Berry[1], Manuel Blanco Valentin[4], Karri DiPetrillo[6], Anthony Badea[6], Lindsey Gray[2], Petar Maksimovic[3], Corrinne Mills[2], Mark S. Neubauer[8], Gauri Pradhan[1], Nhan Tran[1,4], Dahai Wen[3], Farah Fahim[1]

[1]Fermi National Accelerator Laboratory (FNAL), [2]University of Illinois Chicago, [3] Johns Hopkins University, [4] Northwestern University, [5] University of Kansas, [6] The University of Chicago, [8] University of Illinois Urbana-Champaign, Champaign

Abstract*—We present a smart pixel prototype readout integrated circuit (ROIC) designed in CMOS 28 nm bulk process, with in-pixel implementation of an artificial intelligence (AI) / machine learning (ML) based data filtering algorithm designed as proof-of-principle for a Phase III upgrade at the Large Hadron Collider (LHC) pixel detector. The first version of the ROIC consists of two matrices of 256 smart pixels, each 25×25 μm^2 in size. Each pixel consists of a charge-sensitive preamplifier with leakage current compensation and three autozero comparators for a 2-bit flash-type ADC. The frontend is capable of synchronously digitizing the sensor charge within 25 ns. Measurement results show an equivalent noise charge (ENC) of $\sim$30e$^-$ and a total dispersion of $\sim$100e$^-$ The second version of the ROIC uses a fully connected two-layer neural network (NN) to process information from a cluster of 256 pixels to determine if the pattern corresponds to highly desirable high-momentum particle tracks for selection and readout. The digital NN is embedded in-between analog signal processing regions of the 256 pixels without increasing the pixel size and is implemented as fully combinatorial digital logic to minimize power consumption and eliminate clock distribution, and is active only in the presence of an input signal. The total power consumption of the neural network is $\sim$ 300 μW. The NN performs momentum classification based on the generated cluster patterns and even with a modest momentum threshold, it is capable of 54.4% – 75.4% total data rejection, opening the possibility of using the pixel information at 40MHz for the trigger. The total power consumption of analog and digital functions per pixel is $\sim$ 6 μW per pixel, which corresponds to $\sim$ 1 W/cm^2 staying within the experimental constraints.

I. Introduction

Detectors at forthcoming high-energy colliders will confront substantial technical hurdles. Managing the unprecedented volume of particles anticipated in each event demands highly detailed silicon pixel detectors with billions of readout channels. With event frequencies reaching 40 MHz, these detectors will produce immense amounts of data every second. To facilitate discoveries while adhering to stringent bandwidth and latency constraints, future trackers must possess the ability to swiftly, efficiently, and robustly reduce data at the source while operating in high radiation environments. This prototype development targets Phase III pixel detector upgrades projected for 2034, as well as other future colliders.

II. Analog Frontend for Hit Detection and Digitization with Test results

Reducing the pixel sizes by a factor of four while maintaining the same sensor thickness leads to charge sharing which can be positively exploited to improve the effective position resolution [1]. However, this requires reducing the minimum detection threshold of the frontend and overcoming pileup errors.

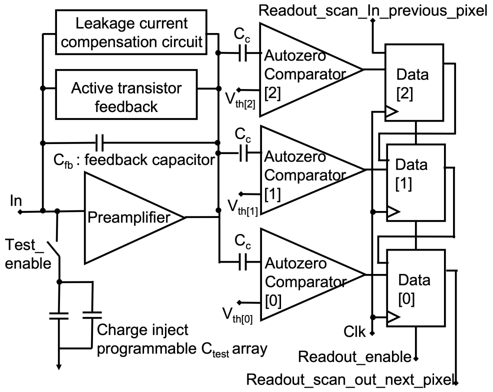
Leakage current compensation circuit
Active transistor feedback
C_{fb} : feedback capacitor
In
Preamplifier
Test_ enable
Charge inject programmable C_{test} array
Readout_scan_In_previous_pixel
C_c
C_c
C_c
Autozero Comparator [2]
Autozero Comparator [1]
Autozero Comparator [0]
$V_{th[2]}$
$V_{th[1]}$
$V_{th[0]}$
Data [2]
Data [1]
Data [0]
Clk
Readout_enable
Readout_scan_out_next_pixel

Fig. 1: The pixel analog frontend for hit registration in a particle detector containing a preamplifier with leakage current compensation and AC coupled synchronous comparators.

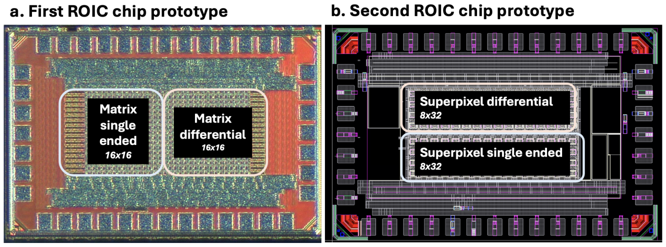
a. First ROIC chip prototype
Matrix single ended 16x16
Matrix differential 16x16
b. Second ROIC chip prototype
Superpixel differential 8x32
Superpixel single ended 8x32

Fig. 2: a. Picture of the first 1.5mm2 ROIC chip. The chip consists of two 16 × 16-pixel matrix variants with just the analog frontend. b. Layout of the second prototype chip. The chip consists of two 8 × 32 - pixel matrix variants with the analog frontend and the digital NN

A. Architecture

The pixel architecture for in-pixel detection and digitization of collision events is depicted in Fig. 1.

The charge collected at the sensor's electrode is integrated, amplified, and converted to voltage using a charge-sensitive preamplifier. An AC-coupled 2-bit flash-type ADC digitizes the signal. Due to the thermometric nature of the flash ADC in our design, analog-to-digital conversion begins as soon as the integrated charge output is above the first threshold, and continues until the signal reaches its maximum value or the time for conversion runs out (since all comparators are reset at the end of the bunch crossing cycle). The first ROIC prototype was designed and fabricated in 2023. The topology, implementation and simulation results are discussed in detail in [2].

B. Test Results Of First ROIC Prototype

Fig. 2a shows the first ASIC chip, 1.5 mm^2 in size. It is composed of a 512 pixels matrix divided into two halves

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

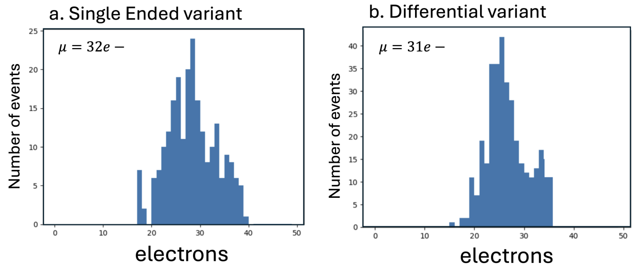

Fig. 3: Total pixel ENC for both variant with 400e- equivalent threshold voltages and no sensor capacitance connected to the ROIC. a. ENC histogram of the single ended variant. b. ENC histogram of the differential variant

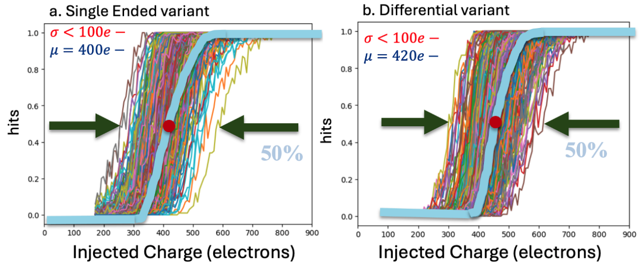

Fig. 4: S-curves of total dispersion for all Hit Comparators in the Array for both variants with 400e- equivalent threshold voltages and no sensor capacitance connected to the ROIC

with a single-ended and differential frontend variants that were implemented. The ROIC was characterized without connecting any sensor and all the comparators in the pixel were set at 400e- equivalent threshold voltages. The ASIC incorporates programmable front-end charge injection circuitry to generate pixel cluster charges during characterization (bottom left in Fig. 1). The programming didn't work as reliably as expected which added a stochastic error to our results. The total equivalent noise charge (ENC) of the preamplifier and the comparator is presented in Fig. 3. The histograms were extracted by scanning the entire input dynamic range in steps of 2 e- and injecting charges 200 times for each step at every pixel. After removing the incorrectly programmed pixels, we find the ENC to be 32e- for both variants. The total dispersion for each matrix variant is presented in Fig. 4. For both variants, the effective threshold dispersion of the hit comparator across the entire ASIC is 100 e^{-}. This result includes gain dispersion in the preamplifier and threshold dispersion of the comparators.

III. On-chip Neural Network Implementation

A neural network classifier was designed and implemented on the second ROIC prototype (Fig. 2b) to identify clusters associated with high p_T charged particles. The high p_T signal class contains tracks with $p_T > 200$ MeV. Oppositely charged particle tracks curve in opposite directions in the magnetic field, resulting in clusters of very different in shape, therefore two background classes are defined, corresponding to positively and negatively charged particles with $p_T < 200\,\text{MeV}$. The choice of p_T threshold in the training class definition affects the physics performance and could be adjusted depending on the physics goals. The details about the training datasets and the algorithm development can be found in this reference [3].

We adopted `hls4ml` to translate the neural network classifier into optimal hardware implementations [4]. `hls4ml` is an open-source Python framework that facilitates the co-design of machine learning algorithms for hardware deployment, supporting models from quantized models from QKeras and other formats [5]. hls4ml allowed us to fine-tune the numerical precision and the hardware parallelism to optimize area, performance, and power consumption according to our system constraints. The conversion process started with the quantized model of the classifier, which `hls4ml` translated into HLS-ready C++ code for Siemens Catapult HLS [6] that generates a hardware description at the register-transfer level (RTL) suitable for the ASIC flow. We chose to fully parallelize the hardware logic to minimize the latency of the neural network, integrating the HLS-generated RTL design with system registers and data movers for efficient operation shown in Fig. 5.

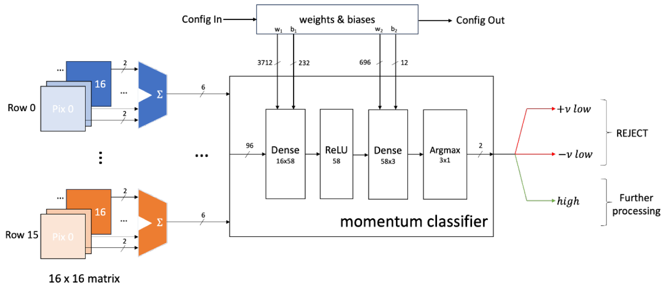

Fig. 5: Data flow through the digital implementation of the algorithm from the summed ADC bits (on the left) through the neural network and the final classification layer. At the top of the diagram we illustrate the reconfigurability of the weights and biases in the algorithm stored in memory.

IV. Conclusions and Future Works

We have presented the first prototype test results and showed the analog frontend based on a synchronous ADC architecture is capable of processing and digitizing input signals within a 25 ns bunch crossing period. The compact area of the analog part of the pixel (<210µm^2) results in $\times 4$ improved granularity while offering a $\times 2.5$ power reduction (~ 4 µW/pixel) and an $\times 2.5$ improvement in the minimum threshold detection ($\approx$ 475e^{-}) compared to state of the art RD53B designs [7] such as CROCv1 [8] and ITKv1 [9]. We also showed the implementation of machine learning-based approaches for data processing in the pixelated region, such as analyzing cluster shapes to determine track parameters and/or suppressing noise hits. By leveraging the proposed architecture, only useful clusters will be transmitted to the periphery for further processing and off-chip data transfer, thus minimizing the necessary data bandwidth for data collection under much higher collision rates for future upgrades. The second prototype chip is expected to be received in May 2024 and test results will be presented at the conference.

References

[1] E. A. A. Bergamaschi, M. Andrä1, "The mÖnch detector for soft x-ray, high-resolution, and energy resolved applications," *Synchrotron Radiation News*, vol. 36, no. 6, pp. 11–15, 2018.

[2] B. Parpillon, A. Trivedi, and F. Fahim, "Readout IC with 40 MSPS in-pixel ADC for future vertex detector upgrades of Large Hadron Collider," *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10182033

[3] J. Yoo, J. Dickinson, M. Swartz, G. Di Guglielmo, A. Bean, D. Berry, M. B. Valentin, K. DiPetrillo, F. Fahim, L. Gray *et al.*, "Smart pixel sensors: towards on-sensor filtering of pixel clusters with deep learning," *arXiv preprint arXiv:2310.02474*, 2023.

[4] F. Fahim, B. Hawks, C. Herwig, J. Hirschauer, S. Jindariani, N. Tran, L. P. Carloni, G. D. Guglielmo, P. Harris, J. Krupa, D. Rankin, M. B. Valentin, J. Hester, Y. Luo, J. Mamish, S. Orgrenci-Memik, T. Aarrestad, H. Javed, V. Loncar, M. Pierini, A. A. Pol, S. Summers, J. Duarte, S. Hauck, S.-C. Hsu, J. Ngadiuba, M. Liu, D. Hoang, E. Kreinar, and Z. Wu, "hls4ml: An open-source codesign workflow to empower scientific low-power machine learning devices," 2021.

[5] C. N. Coelho, A. Kuusela, S. Li, H. Zhuang, J. Ngadiuba, T. K. Aarrestad, V. Loncar, M. Pierini, A. A. Pol, and S. Summers, "Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors," *Nature Machine Intelligence*, vol. 3, no. 8, pp. 675–686, 2021.

[6] Siemens, "Catapult HLS," https://eda.sw.siemens.com/en-US/ic/ic-design/high-level-synthesis-and-verification-platfor.

[7] L. Gaioni, "Cms analog front-end: simulations and measurements," https://cds.cern.ch/record/2746420/files/Linear%20AFE%20report.pdf.

[8] A. Papadopoulos and on the CMS Tracker Group, "Analog performance of the crocv1 pixel readout chip for the cms phase-2 tracker upgrade," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 18, 2023.

[9] E. A. Stefano Terzo, Maurizio Boscardin, "Novel 3d pixel sensors for the upgrade of the atlas inner tracker," *Frontiers in Physics*, vol. 9, 2021.