HVCMOS Pixel Sensor in 55nm Process: Readout Architecture Simulation, Hit Loss Analysis, and Data Transmission Optimization

Xiaoxu ZHANG[1,4], Xiaomin WEI[2], Yang CHEN[5], Anqi WANG[3], Yang ZHOU[1], Leyi LI[1,6], Yu ZHAO[2], Zexuan ZHAO[2], Huimin WU[2], Lei ZHANG[4], Jianchun WANG[1], Yiming LI[1]

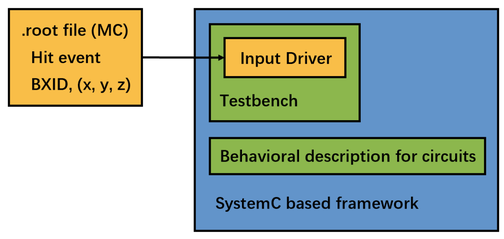
[1]Insitute of High Energy Physics, CAS, Beijing, China [2]Northwestern Polytechnical University, Xi'an, China
[3]University of Chinese Academy of Sciences [4]Nanjing University, Nanjing, China [5]Dalian Minzu University, Dalian, China [6]Shandong University, Qingdao, China

Introduction

Pioneering R&D in HVCMOS pixel sensors at the advanced 55 nm process, the COFFEE series prototypes are currently being developed for the CEPC inner tracker and Upstream Pixel tracker (UP) in the LHCb Upgrade II. COFFEE3, the latest prototype with two distinct readout architecture, was design and fabricated in 2025. Though featuring a small-scale prototype (3×4 mm^2), COFFEE3 is designed to match the final full-scale sensor (~2×2 cm^2), aiming for proof of concept. This poster shows the performance of the preferred readout architecture under the high-hit-density environment of LHCb, especially the efficiency loss. In response to the UP's data compression requirements and the readout link's bandwidth limitations, this poster also shows how simulation can be used to explore and iterate peripheral readout architectures that enable rational scheduling of transmission resources.

SystemC Based Framework using MC Input

By establishing a behavioral-level model of the pixel array and peripheral readout using SystemC, and inputting MC (Monte Carlo) hit data in the testbench, the advantages of this framework are:

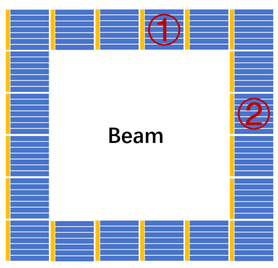


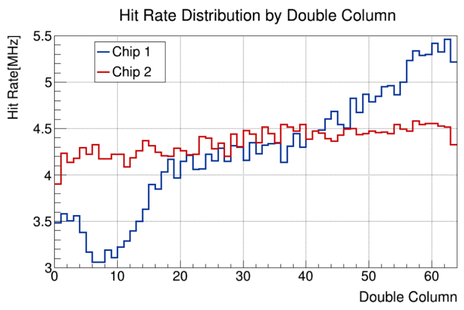
① Using MC data enables the simulation of real experimental environments by capturing hit density fluctuations in spacetime, which produce the non-uniform and bursty data traffic patterns, resulting in detailed and reliable simulation results.

② Compared to RTL-level implementations, behavioral-level models can more rapidly identify transmission bottlenecks for parameter optimization.

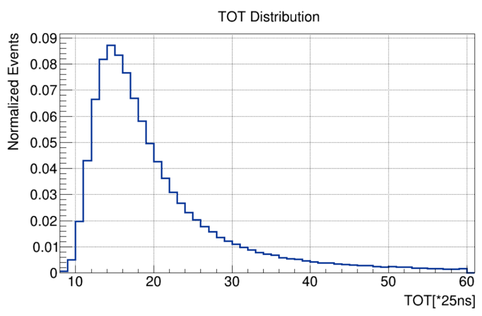
For the chips closest to the beam, two typical examples were selected, with their locations and hit density distributions illustrated below. An event sample of 50,000 BXIDs and a cluster size of 1.5 was used.

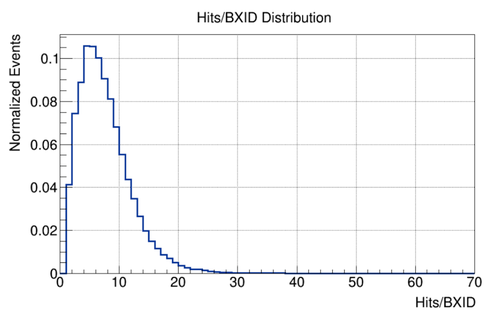
Innermost chip placement





A Landau distribution between 200 and 1500 ns is used for the TOT time. The Hits/BXID distribution shows that most values are below 30, but it has a long tail reaching up to 70.



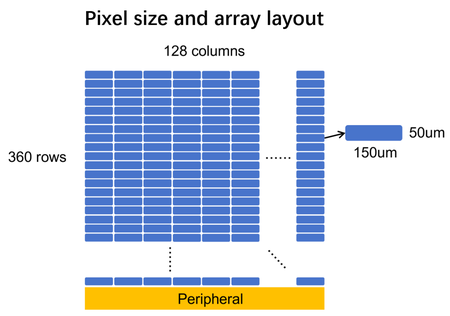


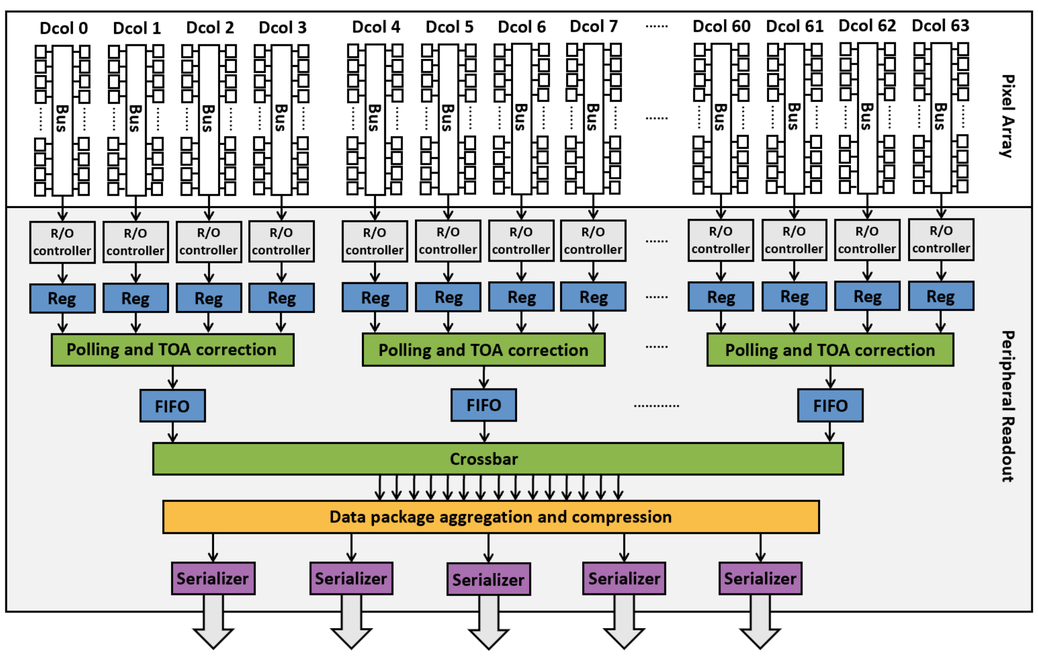
Simulation, Analysis and Optimization

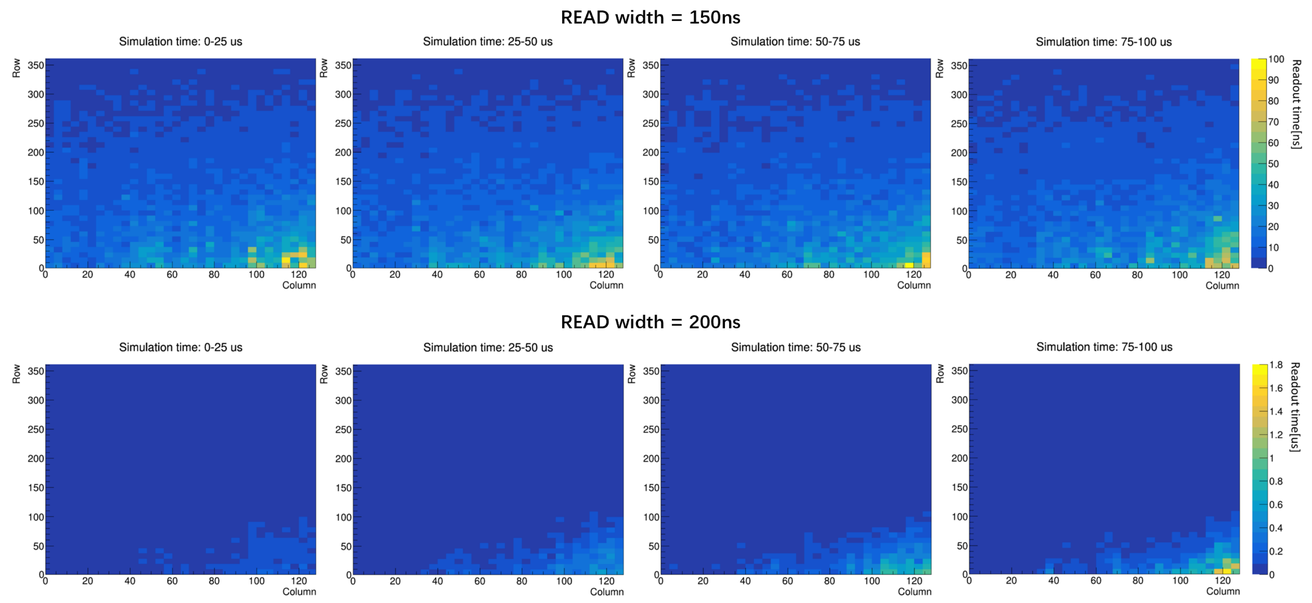
The row-column configuration of pixel array and the architecture of peripheral readout are shown below.

The pixel array uses a column-drain readout architecture. Every two columns serve as a unit sharing one EoC (End of Column). The single read period of EoC (the pulse width of READ signal) and the priority logic determine the readout time required for the process from the generation of hit information within pixel to its readout to EoC. Taking Chip1 as an example, positions with higher hit density or lower priority demonstrate longer readout time. When READ width = 200 ns, the readout time gradually increases as the simulation progresses, indicating that data congestion occurs within the pixel array.

Pixel size and array layout





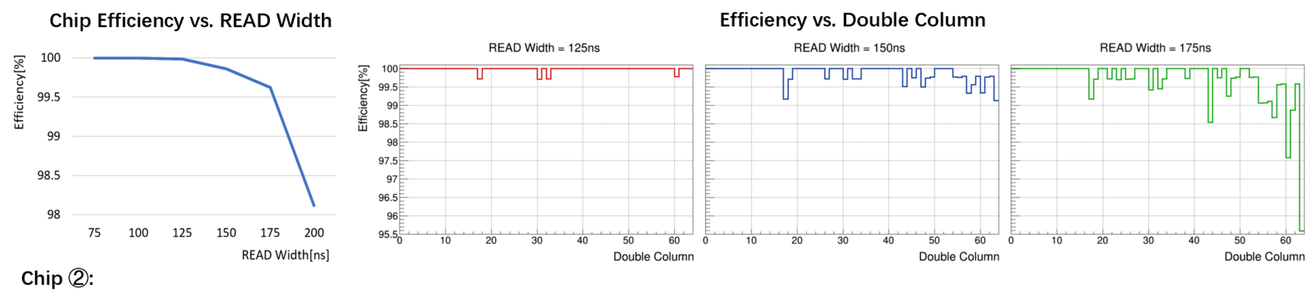


The READ width affects the hit loss as the simulation results indicate:

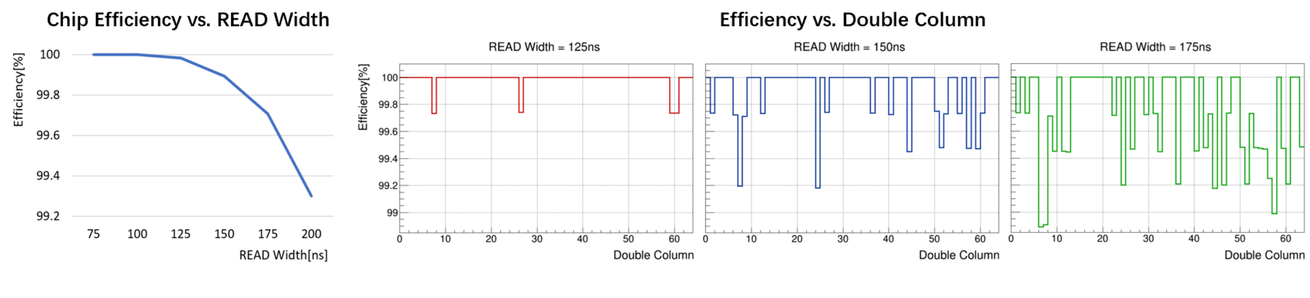
① The chip efficiency remains near 100% for READ width ≤ 125 ns but exhibits a significant decrease for READ widths > 125 ns.

② For READ width > 125 ns, the efficiency of chip1 not only decreases but also exhibits significant variation across different double columns.

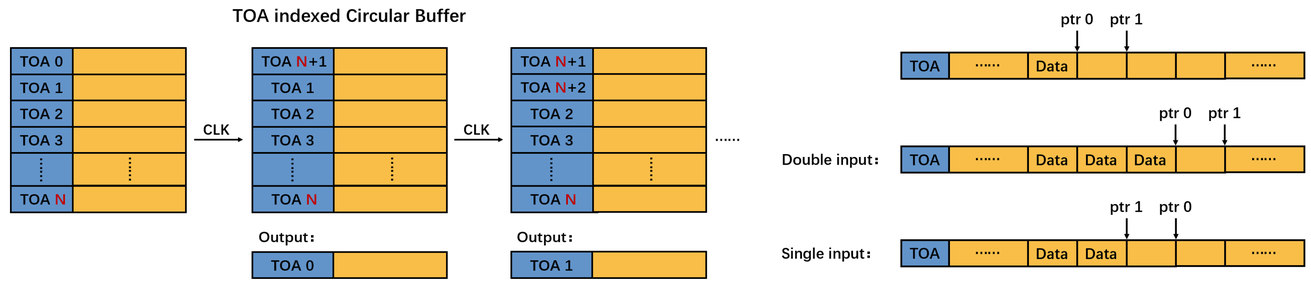
Chip ①:



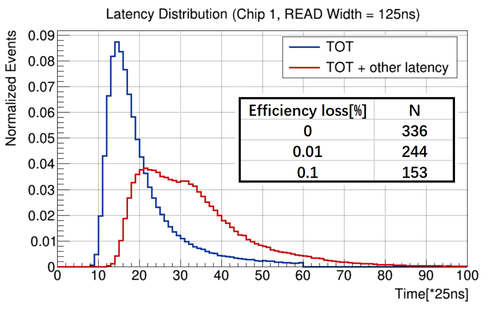
Chip ②:



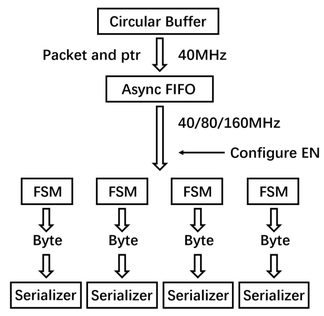
The core of the peripheral readout architecture lies in the data compression format. Aggregating data packets with identical TOA (Time of Arrival) can achieve approximately 30% bit saving. The key to realizing this lies in a globally shared multi-bank circular buffer. Double pointers are used in each bank.



Due to the priority queuing scheme, the latency from hit generation to packet arrival at the circular buffer follows a distribution. The number of banks, N, in the circular buffer is equal to the range of this latency distribution, expressed in units of main clock cycles. Most packets arriving in 11-90 cycles, which primarily influenced by the TOT distribution. However, a few packets suffer from significantly longer latency of up to 346 cycles, creating a long tail in the distribution. Using double pointers in each bank and a CIOQ (Combined Input and Output Queued) queuing strategy in crossbar, the long tail primarily originates from the queuing latency in pixels. The strategy for cutting this tail is driven by the trade-off between silicon area and efficiency loss. The 55nm HVCMOS process is expected to facilitate 100% efficiency in the peripheral readout.



The circuit composition of the circular buffer backend includes an asynchronous FIFO, some FSMs (Finite State Machine) and the same number of 8-bit to 1-bit serializers. The frequency of the final output link can be configured by setting the read clock of the asynchronous FIFO. The number of final output links can be configured by the ENABLE switches before the FSMs. These two configurations enable the chip to be compatible with different hit density. The FSMs split long packets into individual bytes and send them to serializers. The FSMs are designed carefully so that there is no bandwidth waste when FSMs handshake with the asynchronous FIFO. For the condition with four 1.28 GHz output links, the bandwidth utilization was 99.3%, 98.6%, 96.9%, and 95.0% respectively, enabling timely data transmission without congestion. In practical use, as a precautionary measure, the maximum number of output links is set to five.



Conclusion

Using the framework of SystemC behavioral modeling combined with Monte Carlo data, we simulate and analysis the chip performance in the real experimental environments and optimized the peripheral readout architecture to enable rational scheduling of transmission resources, addressing LHCb UP's data compression format requirements and the bandwidth limitations of the chip's readout links. The outcomes of this work are intended to identify bottlenecks in on-chip data processing and transmission under the high hit-density environment, and to further optimize the chip architecture to satisfy the comprehensive performance requirements of applications.

Email: xiaoxuzhang@smail.nju.edu.cn