Contribution ID: **149** Type: **Oral**

In-pixel AI for lossy data compression at source for X-ray detectors

Wednesday 4 October 2023 11:40 (20 minutes)

This work introduces AI-In-Pixel-65, an ROIC test chip designed for pixelated X-ray detectors using a 65nm Low Power CMOS process. The study compares two data compression techniques, Principal Component Analysis (PCA) and AutoEncoder (AE), implemented within the chip's pixelated area to address I/O bottlenecks. Our design methodology utilizes high-level synthesis (HLS) and hls4ml, offering shorter design cycles and similar quality results compared to register-transfer level (RTL) flows. Results show PCA achieves 50-fold compression with a 21% pixel area increase, while AE offers 70-fold compression and similar area increase.

Summary (500 words)

Integrating data compression neural networks into Read-Out Integrated Circuits (ROICs), specifically within the pixelated front-end, has the potential to significantly decrease off-chip data transfer, thereby overcoming the input/output bottleneck. We have developed an ROIC test chip called AI-In-Pixel-65, which is designed using a 65nm Low Power CMOS process to read pixelated X-ray detectors.

Each pixel in the chip comprises an analog front-end responsible for signal processing and a 10-bit analog-to-digital converter that operates at a speed of 100,000 samples per second (100KSPS). In our study, we have compared two non-reconfigurable techniques for data compression, namely Principal Component Analysis (PCA) and AutoEncoder (AE), which are implemented within the pixelated area of the chip.

Our design methodology leverages high-level synthesis (HLS) and hls4ml to provide a shorter design cycle and similar quality of results compared to register-transfer level (RTL) flows. hls4ml is an open-source framework for the hardware codesign of neural networks. At its core, hls4ml translates machine-learning models from common open-source software frameworks such as TensorFlow into an RTL implementation using HLS tools. In particular, we adopted the industry standard Siemens Catapult HLS. In hls4ml, a designer can trade off the performance and area utilization for a model by varying the parallelization of the algorithm via several configuration parameters. For example, the reuse factor parameter controls how many times each multiplier resource is used in the final hardware implementation: a designer with the goal of low latency will choose a lower RF value.

Additionally, for the model training, we adopted quantization-aware training (QAT) with QKeras. QAT is a machine learning technique to train models optimized for deployment on hardware with low-precision arithmetic. QAT simulates the effects of quantization during the training process, mapping a continuous range of values to a smaller set of discrete values. As a result, the model learns to be more robust to the effects of quantization, which can lead to improved accuracy and reduced memory and computation requirements at inference time.

In our chip, the PCA algorithm demonstrates a 50-fold compression, introduces a latency of one clock cycle, and results in a 21% increase in the pixel area. On the other hand, the AE method achieves a higher compression rate of 70 times, adds a latency of 30 clock cycles, and leads to a similar increase in the pixel area compared to PCA. By evaluating the performance of these two data compression techniques within the pixelated region, our study offers insights into their respective merits and limitations in addressing the I/O bottleneck in ROICs.

Authors: QUINN, Adam (Fermi National Accelerator Laboratory); JACOBSEN, Chris (Northwestern University); NOONAN, Danny (Fermi National Accelerator Lab. (US)); BRAGA, Davide (FERMILAB); FAHIM, Farah (Fermilab); DI GUGLIELMO, Giuseppe (Fermilab); VALENTIN, Manuel; TRAN, Nhan (Fermi National Accelerator Lab. (US)); HUANG, Panpan (Northwestern University); DILIP, Priyanka (Stanford University / Fermilab); OGRENCI, Seda (Northwestern University); ZIMMERMAN, Thomas (Fermi National Accelerator Lab. (US))

Presenter: NOONAN, Danny (Fermi National Accelerator Lab. (US))

Session Classification: ASIC

Track Classification: ASIC