Contribution ID: **43** Type: **talk**

Attention at Silicon Speed: Towards Efficient Transformers on FPGAs

Monday 14 October 2024 13:55 (10 minutes)

The High-Luminosity Large Hadron Collider (HL-LHC), anticipated to begin operations in 2029, will generate data at an astounding rate on the order of 100 terabits per second. To efficiently process and filter these data, the Compact Muon Solenoid (CMS) experiment
relies on the extremely low-latency Level-1 trigger, which uses Field-Programmable Gate Arrays (FPGAs). My project focuses on further optimizing this process by adapting efficient transformer models for implementation on FPGAs using the hls4ml package. This
talk will highlight my contributions, ongoing work to optimize FPGA implementations of transformer models, and my experiences as an A3D3 Postbaccalaureate Fellow.

Author: FLYNN, Rian (Purdue University (US))

Presenter: FLYNN, Rian (Purdue University (US))

Session Classification: Afternoon Plenary