# NGT workshop: hls4ml HEP Community Forum

# Report of Contributions

https://indico.cern.ch/e/1426145

Contribution ID: **1** Type: **not specified**

# The hls4ml project: status and perspectives

*Friday 27 September 2024 09:30 (30 minutes)*

**Presenter:** LONCAR, Vladimir (Massachusetts Inst. of Technology (US))

Contribution ID: **2** Type: **not specified**

# The NGT WP1.2 and WP1.3

*Friday 27 September 2024 09:00 (20 minutes)*

**Presenter:** KAGAN, Michael (SLAC National Accelerator Laboratory (US))

Contribution ID: **3** Type: **Long talk (10min +5min questions)**

# Online track reconstruction with graph neural networks on FPGAs for the ATLAS experiment

For the HL-LHC upgrade of the ATLAS TDAQ system, a heterogeneous computing farm deploying GPUs and/or FPGAs is under study, together with the use of modern machine learning algorithms such as Graph Neural Networks (GNNs). We present a study on the reconstruction of tracks in the ATLAS Inner Tracker using GNNs on FPGAs for the Event Filter system. We explore each of the steps in a GNN-based tracking pipeline: graph construction, edge classification using an interaction network, and segmentation of the graph into track candidates. We investigate optimizations of the GNN approach that aim to minimize FPGA resources utilization and maximize throughput while retaining high track reconstruction efficiency and low fake rates required for the ATLAS Event Filter tracking system. These studies include model hyperparameter tuning, model pruning and quantization-aware training, and sequential processing of sub-graphs across the detector.

**Author:** PARAJULI, Santosh (Univ. Illinois at Urbana Champaign (US))

**Presenter:** PARAJULI, Santosh (Univ. Illinois at Urbana Champaign (US))

Contribution ID: **4** Type: **Short talk (5min +3min questions)**

# Machine learning evaluation in the Global Event Processor FPGA for the ATLAS trigger upgrade

The Global Event Processor (GEP) FPGA is an area-constrained, performance-critical element of the Large Hadron Collider's (LHC) ATLAS experiment. It needs to very quickly determine which small fraction of detected events should be retained for further processing, and which other events will be discarded. This system involves a large number of individual processing tasks, brought together within the overall Algorithm Processing Platform (APP), to make filtering decisions at an overall latency of no more than 8ms. Currently, such filtering tasks are hand-coded implementations of standard deterministic signal processing tasks.

In this work we present methods to automatically create machine learning based algorithms for use within the APP framework, and demonstrate several successful such deployments. We leverage existing machine learning to FPGA flows such as HLS4ML and fwX to significantly reduce the complexity of algorithm design. These have resulted in implementations of various machine learning algorithms with latencies of 1.2 $\mu$s and less than 5% resource utilization on an Xilinx XCVU9P FPGA. Finally, we implement these algorithms into the GEP system and present their actual performance.

Our work shows the potential of using machine learning in the GEP for high-energy physics applications. This can significantly improve the performance of the trigger system and enable the ATLAS experiment to collect more data and make more discoveries. The architecture and approach presented in this paper can also be applied to other applications that require real-time processing of large volumes of data.

**Authors:** CARLSON, Ben (Westmont College); PARAJULI, Santosh (Univ. Illinois at Urbana Champaign (US))

**Presenters:** CARLSON, Ben (Westmont College); PARAJULI, Santosh (Univ. Illinois at Urbana Champaign (US))

Contribution ID: **5** Type: **Long talk (10min +5min questions)**

# Reusability in emulation and firmware

With hls4ml now deployed online at CMS, we may look back on the deployment experience with a view to improvements for the years ahead. In particular, the compile-time fixing of network architecture and parameters has some implications. In emulation, model updates are tied to CMS Software release schedules and cannot be updated on changing conditions. Likewise in firmware, models are baked into the binaries that are running at Point 5. In this talk, I will present how additional flexibility or reprogrammability in both emulation and firmware of hls4ml would allow new workflows.

**Author:** SUMMERS, Sioni Paris (CERN)

**Presenter:** SUMMERS, Sioni Paris (CERN)

Contribution ID: **6** Type: **Short talk (5min +3min questions)**

# hls4ml web interface

Quickly understanding the resource usage of a model is a common workflow for anyone designing an FPGA targeted model. Often the developer won't have access to vivado or a PC capable of running it. A website where the user can upload a trained model, hls4ml can run on the model, produce firmware estimates and a package of the associated firmware could solve this issue. A web interface GUI could allow the user to change simple parameters of the firmware build; bitwidth, FPGA targeted, clock speed etc. and developers could easily determine if a given model is feasible in FW and have some prototype FW to experiment with. This would be a useful tool for the community allowing novices to experiment with the process of designing fastML models.

**Author:** BROWN, Christopher Edward (CERN)

**Presenter:** BROWN, Christopher Edward (CERN)

Contribution ID: **7** Type: **Long talk (10min +5min questions)**

# Smart pixels with data reduction at the source

Highly granular pixel detectors allow for increasingly precise measurements of charged particle tracks. At the High Luminosity Large Hadron Collider, data rates from the pixel detector exceed those feasible for integration in the Level 1 trigger. However, the shape of charge clusters deposited in the pixel sensors can be used to determine the physical properties of the traversing particle, such as pT, by on-detector locally customized neural networks. This technique can allow for smart data reduction at the source to reject low-momentum charged particles within the pixelated region of the detector - potentially enabling the use of pixel data at the LHC bunch crossing frequency of 40 MHz in the Level-1 trigger. To achieve this goal, the hls4ml framework has been extended to include a Siemens Catapult HLS backend, which allows for the translation of HLS-ready C++ to ASICs instead of FPGAs. This extension broadens the applications of hls4ml to novel detector applications and demanding radiation-hard environments. Hardware tests on the first tape-outs produced with TSMC using the hls4ml + Catapult HLS pipeline are presented.

**Authors:** YOUNG, Aaron (Oak Ridge); BEAN, Alice (The University of Kansas (US)); BADEA, Anthony (University of Chicago (US)); PARPILLON, Benjamin (Fermi National Accelerator Lab. (US)); SYAL, Chinar (Fermi National Accelerator Lab. (US)); MILLS, Corrinne (University of Illinois at Chicago (US)); WEN, Dahai (Nanjing Normal University (CN)); BERRY, Douglas Ryan (Fermi National Accelerator Lab. (US)); FAHIM, Farah (Fermilab); Ms PRADHAN, Gauri; DI GUGLIELMO, Giuseppe (Fermilab); PEARKES, Jannicke (University of Colorado Boulder (US)); DICKINSON, Jennet Elizabeth (Cornell University (US)); Ms YOO, Jieun (UIC); HIRSCHAUER, Jim (Fermi National Accelerator Lab. (US)); DI PETRILLO, Karri Folan (University of Chicago); GRAY, Lindsey (Fermi National Accelerator Lab. (US)); VALENTIN, Manuel; NEUBAUER, Mark Stephen (Univ. Illinois at Urbana-Champaign); SWARTZ, Morris (Johns Hopkins University (JHU)); TRAN, Nhan (Fermi National Accelerator Lab. (US)); MAKSIMOVIC, Petar (Johns Hopkins University (US)); LIPTON, Ronald (Fermi National Accelerator Lab. (US)); KULKARNI, Shruti R (Oak Ridge National Laboratory)

**Presenter:** PEARKES, Jannicke (University of Colorado Boulder (US))

Contribution ID: **8** Type: **not specified**

# ATLAS 2.1: L0 Global

*Friday 27 September 2024 10:30 (25 minutes)*

**Presenter:** XIOTIDIS, Ioannis (CERN)

**Session Classification:** User contributions

Contribution ID: **9** Type: **not specified**

# ATLAS 2.2: L0 Muon

*Friday 27 September 2024 11:00 (25 minutes)*

**Presenter:** ROJAS CABALLERO, Rimsky Alejandro (University of Massachusetts (US))

**Session Classification:** User contributions

Contribution ID: **10** Type: **not specified**

# ATLAS 2.4: Event Filter Tracking

*Friday 27 September 2024 11:30 (20 minutes)*

**Presenter:** WOLLRATH, Julian (CERN)

**Session Classification:** User contributions

Contribution ID: **11** Type: **not specified**

# CMS 3.5: L1 scouting

*Friday 27 September 2024 13:30 (25 minutes)*

**Presenter:** PETRUCCIANI, Giovanni (CERN)

**Session Classification:** User contributions

Contribution ID: **12** Type: **not specified**

# CMS 3.6: ML objects in L1 trigger

*Friday 27 September 2024 14:00 (25 minutes)*

**Presenter:** SUMMERS, Sioni Paris (CERN)

**Session Classification:** User contributions

Contribution ID: **13** Type: **not specified**

# CMS 3.7: Compression/Anomaly detection

*Friday 27 September 2024 14:30 (25 minutes)*

**Presenters:** NGADIUBA, Jennifer (FNAL); GLOWACKI, Maciej Mikolaj (CERN)

**Session Classification:** User contributions

Contribution ID: **14** Type: **not specified**

# Online track reconstruction with graph neural networks on FPGAs for the ATLAS experiment

*Friday 27 September 2024 15:30 (10 minutes)*

**Presenters:** PARAJULI, Santosh (Univ. Illinois at Urbana Champaign (US)); PARAJULI, santosh

**Session Classification:** User contributions

Contribution ID: **15** Type: **not specified**

# Smart pixels with data reduction at the source

*Friday 27 September 2024 15:45 (10 minutes)*

**Presenter:** PEARKES, Jannicke (University of Colorado Boulder (US))

**Session Classification:** User contributions

Contribution ID: **16** Type: **not specified**

# Reusability in emulation and firmware

**Presenter:** SUMMERS, Sioni Paris (CERN)

**Session Classification:** User contributions

Contribution ID: **17** Type: **not specified**

# Machine learning evaluation in the Global Event Processor FPGA for the ATLAS trigger upgrade

*Friday 27 September 2024 16:00 (10 minutes)*

**Presenter:** CARLSON, Ben (Westmont College)

**Session Classification:** User contributions

Contribution ID: **18** Type: **not specified**

# hls4ml web interface

*Friday 27 September 2024 16:15 (10 minutes)*

**Presenter:** BROWN, Christopher Edward (CERN)

**Session Classification:** User contributions

Contribution ID: **19** Type: **not specified**

# FastML general meeting

*Friday 27 September 2024 17:00 (1 hour)*

Discussion on the formation of FastML Foundation