Contribution ID: **197** Type: **Parallel session talk**

PQuant: Streamlining ML Model Compression to Deployment for Next-Gen Detector Systems

Thursday 9 October 2025 15:20 (20 minutes)

Real-time machine learning is emerging as a key tool for next-generation detector systems, where strict latency and hardware constraints require highly efficient models. We present PQuant, a backend-agnostic Python library designed to unify and streamline pruning and quantization techniques for hardware deployment, supporting both PyTorch and TensorFlow. PQuant provides a comprehensive suite of methods, including unstructured pruning, structured pruning (PDP and ActivationPruning), and hardware-aware resource (DSP/BRAM) optimization, pattern compression for convolutional kernels, in FPGAs/ASICSs, through MDMM framework. PQuant also provides flexible quantization options, ranging from fixed-point to high-granularity schemes, with per-layer or per-weight bit control. Integration with hls4ml is ongoing, enabling compressed models to be deployed directly to FPGAs/ASICs. PQuant bridges advanced compression methods with implementation directly translating to resource optimization, providing a practical path to low-latency ML in triggers, DAQ, and online reconstruction for high-energy physics experiments.

Author: NIEMI, Roope Oskari

Co-authors: SUN, Chang (California Institute of Technology (US)); PETROVYCH, Anastasiia (CERN); DANOPOULOS, Dimitrios (CERN); LUPI, Enrico (CERN, INFN Padova (IT)); DAS, Arghya Ranjan (Purdue University (US)); DITTMEIER, Sebastian (Ruprecht-Karls-Universitaet Heidelberg (DE)); KAGAN, Michael (SLAC National Accelerator Laboratory (US)); LIU, Miaoyuan (Purdue University (US)); LONCAR, Vladimir (CERN)

Presenter: DAS, Arghya Ranjan (Purdue University (US))

Session Classification: SHARED SESSION II

Track Classification: RDC 5 Trigger & DAQ