CERN openlab

# Status Report

Alberto Di Meglio – CERN openlab Head

02/07/2019

# Micron

- Agreement signed in October 2018 for 36 months
- Two main use cases (nominally 1 Fellow each + overhead)
  - CMS Real-Time Streaming Matching Inference Engine Prototype
  - Prototyping of a DL-based Particle Identification System for the Dune Neutrino Detector
- Recruitment started, both projects are currently staffed for the first year of the project, situation will be revised for Y2 and Y3
- Hardware received from Micron
  - 2 x Micron Pico SB-852 Co-Processor, a Hybrid-Memory-Cube-based co-processor including DDR4 memory and a Xilinx FPGA application engine
- Both projects are progressing as planned
- Check-point meeting took place on June 21st at CERN

# Micron - CMS

- The goal of the project is to prototype a Real-Time Streaming Matching Inference Engine to use in the Level-1 trigger system of an LHC experiment
- Work progress (extracted from Dejan Golubovic's slides)
  - Installation and set-up: done, all fine
  - Defined scouting objectives (increase trigger rates ~ x10, >= 500k inf/sec, parallelisation)
  - Run benchmarks and extracted required results
  - Achieved promising latency performance
  - Demonstrated parallelisation on 2 and 4 clusters
  - Achieved 2M inf/sec on 4 clusters with the "extended scouting model"
  - Started work on Graph Neural Networks and Convolutional Neural Networks
    - Summer student projects

# Micron - DUNE

- Two sub use-cases
  - Neutrino interactions classification (from generated images)
  - Detector raw data collection and denoising (to generate images)
- Work Progress (extracted from Manuel J. Rodriguez Alonso's slides)
  - Use case 1:
    - Get used to the Inference Engine ✓
    - Run existing ResNet-18 on the FPGA ✓
    - Compare results (existing GPU and Micron FPGA result match always) ✓
    - Check performance improvement: work in progress with summer students to prepare a benchmark over 10M images
  - Use case 2:
    - Not started yet

# Google

- First contact in October 2018 in the context of the CERN openlab Quantum Computing Initiative
- Moved on to define a collaboration programme over a broader set of use cases. Various meetings took place at CERN, Google, and computing events (SC, ISC) in the past 8 months
- First iteration of the proposal included
  - Kubernetes performance and scalability  (IT-CM)
  - Composable data centers for efficient HEP computing workflows (CMS)
  - Generative models for detector simulation (IT-DI-OPL)
  - Deep Learning Inference on TPU for LHC real-time collision processing (CMS)
  - Deep Learning data pipelines for high energy physics and event classification (IT-DB)
  - Deep Learning for Earth Observation (UNOSAT)
  - RICH reconstruction using Google TPUs (LHCb)
  - Quantum generative models (IT-DI-OPL)
  - Optimisation of GRID job placement (ALICE)

# Google

- Second iteration just started, projects being consolidated in three separate proposals with priorities to be discussed
- Proposal 1:
  - Kubernetes performance and scalability (IT-CM)
  - Deep Learning data pipelines for high energy physics and event classification (IT-DB)
  - Composable data centers for efficient HEP computing workflows (CMS)
- Proposal 2
  - Generative models for detector simulation (IT-DI-OPL)
  - Deep Learning Inference on TPU for LHC real-time collision processing (CMS)
  - RICH reconstruction using Google TPUs (LHCb)
- Proposal 3
  - Quantum generative models (IT-DI-OPL)
- ATTRACT QUOG-DP
  - Optimisation of GRID job placement (ALICE)

# Google

- Activities already started
- Kubernetes workshop on June 11th
- Many discussions about storage
- TPU credits

# Update on Hardware Resources

| Type | Vendor | Notes |
|---|---|---|
| X86 CPUs<br>various configurations of CPU family, storage, memory (including 3D-Xpoint). Two Cascade Lake CPUs with full Optane (3D-Xpoint) support have just been received, they will be installed in two dedicated nodes | Intel | Used for benchmarks or investigations of applications with high memory requirements (mainly DL/ML, simulation, data analytics)<br>Two nodes dedicated to Fast K-V Storage project in EP-DT |
| Movidius 2 NN-optimized chips | Intel | The chip has been tested by CERN for radiation resistance as part of a collaboration among CERN, ESA and Intel. Now being evaluated for performance compared to more traditional platforms (GPUs). Potential applications in HEP data acquisition or space applications (inference) |
| 4 x Nvidia T4 GPU (PCIe)<br>4 x Nvidia V100 GPU (Nvlink) | E4/Nvidia | Formally dedicated to the use cases of the E5/Nvidia collaboration, but being also integrated in the standard IT GPU provisioning pilot with IT-CM |
| 1 IBM Minsky Cluster<br>Power 8+<br>2 x Nvidia Tesla P100 (Nvlink) | IBM | Addition of a second cluster based on Power 9 being discussed |
| *EPYC Rome* | *AMD* | *First discussions for a possible evaluation project after Q3 2019* |