R&D steering group meeting, CERN, 13th April 2018

Access policy

- **Filling up the form raises the bar**
  - All users must provide a clear description of the goals and work

- **Requests are typically categorised in three categories**
  - “Official” request for non-GPU hardware
  - “Official” request for GPU hardware
    - Will be mapped to best available GPU, ideally dedicated
  - “Grassroots” request for GPU hardware
    - Random requests, early testing without specific long term planning
    - 80-90% of recent requests
    - Will be allocated to a shared GPU
    - Lower priority

- **Openstack and vGPUs should be a huge improvement**
  - Better service to the user, better integration with CERN IT
  - Lower cost to operate, reduce load on Techlab
  - Suits perfectly the “Grassroots” use cases

Use cases

- **CMS Data Acquisition:**
  - Goal: "**Initially only to familiarise with CUDA**"
  - Plans: "We are interested in the future possible use of GPUs as offload for reconstruction tasks in our High Level Trigger. From our standpoint, **we would like to understand better the software infrastructure that will be required**."

  - "In the medium term, we will decide whether to acquire hardware of our own to test the integration into our computing platform. In order to make an informed decision on exactly what hardware to acquire, if any, we would like to understand better the use of the GPU so we can have better exchanges with the algorithm developers"

Use cases

- **CMS**
  - Goal:
    - "Understand features of **Intel FPGAs** and associated programming tools (e.g. differences wrt Xilinx solutions)
    - Test usage of FPGAs as accelerators for machine-learning and classical algorithms."
  - Plan: "Evaluation of possible architectures for the Phase 2 upgrade of the CMS HLT or L1"

Use cases

- **LHCb (2 requests)**
  - Goal: run benchmarks with TensorFlow for amplitude fits (TensorFlowAnalysis package, code at https://gitlab.cern.ch/poluekt/TensorFlowAnalysis and presentation at https://indico.cern.ch/event/686641/contributions/2867259/attachments/1608597/2553481/tfa.pdf) on a **Tesla architecture**.
  - Plan: *"University of Warwick aims to purchase GPU hardware to be used in physics analyses, and the results of the benchmarks would allow us to better understand our needs (e.g. multicore CPU vs. Tesla GPU, amount of VRAM needed, effect of memory bandwidth)."*

  - Goal: "Validate **ARM64 as possible production platform** for LHCb. Include in nightly builds of the experiment software stack"
  - Plan: "This work might influence future procurements"

Use cases

- **ATLAS**
  - Goal: “Deep Neural Network trainning forf Higgs analysis” “TMVA with Deep Neural Network (DNN)”
  - Plan: “Improve performance”
  - Requests was finally accepted as low prioritiy after discussions, but required about ~25 emails to setup the development environment requested :-(

Use cases

- **TH department (2 requests)**
  - Goal: “The idea is to test some software that should be released open source to perform integrations numerically in an improved way. **Testing to see if the code we have compiles on different architectures**”

  - Goal: “We're training deep learning networks for jet substructure analyses. So far the postdoc I'm working with (XXX XXX) has managed training with the help of small desktop GPU unit, but to further pursue the research **we need something more powerful**. It will enable us to train our deep networks with tools that are (at least partially) comparable to those being used by other groups. The question that we want to establish is whether by bringing better physical insight into the choice of inputs we can improve the performance of deep learning for quark/gluon discrimination and W-boson/normal-jet discrimination. It's important that our comparisons not be biased by limitations in the training that we're able to achieve.”

Use cases

- **BioDynaMo (2 requests)**
  - Goal: " Besides scaling in the number of nodes, we want to investigate the performance improvement of using hardware accelerators (i.e. GPUs, FPGAs) for our most compute intensive operations (this is currently mechanical interactions between cells). In a distributed environment there is nowadays access to more than one hardware accelerator to offload these computations to."
  - Plan: "initially interested in driving the **Altera FPGA with OpenCL**"

  - Goal: "Investigate **GPUs** for acceleration of biological simulations"

Use cases

- **Technical Infrastructure (Electricity)**
  - Goal: **“Image processing** with TensorFlow, currently it takes too much time to run this on cpu's. We could only train on 2000 images and the accuracy is far from what we need. Hence the request for the GPU resources.”

  - Plan: “Understanding if machine learning model accuracy increases by training them large dataset “

Use cases

- **IT storage group**
  - Goal: “Measure the achievable IO rates during NN learning with different IO methods with keras/tensoflow/theano.”
  - “It would allow me to compare some earlier measurement (on a desktop system under macos) with more a standard environment.”

Use cases

- **WA105**
  - Goal: “We would like to **train tensorflow models on liquid argon data taken with the WA105 detector**. Events data are 2D images of particle interaction. The model Convolutional NN to classify different type.”

Use cases

- **EP/SFT ROOT**
  - Goal: “Run machine learning algorithms (Deep Learning) using GPU. The idea is to run GPU DNN with cuda using TMVA and Keras with tensorflow. We need to compare models and performance in the statistical results.”

Discussion

http://cern.ch/techlab