IRIS-HEP Blueprint Meeting:
Fast Machine Learning and Inference

Live notes

Agenda: https://indico.cern.ch/event/820946/

Registered Attendees

Lots ;)

Pictures

https://www.facebook.com/pg/Fermilab/photos/?ref=page_internal

Presentations (notes from the discussion, questions, etc)

Tuesday Afternoon Session I

Welcome to the LPC (Sergo Jindariani)

The IRIS-HEP Blueprint: Concepts and Process (Mark Neubauer)

Overview of triggering and real-time systems (Isobel Ojalvo)

Q (Nhan) - what are the opportunities to have common/similar algorithms between the trigger levels?
Could be using ML output from L1 to information HLT -> could reduces HLT latency

Future computing architectures (Paolo Califiura)
(Nhan) HPCs for inference? (Paolo) For training mostly, for inference just translate tensor flow and put onto CPU/FPGA.

Energy Efficient Machine Learning (Andreas Moshovos)

Nhan: Are the energy efficient numbers (75-99% savings) for all of the neural networks? A; Yes, and novel networks.

Paolo: What kind of hardware do you use to test your sparsity algorithms? FPGAs? A: Not a good native solution (FPGAs), so they do early state processor design.

Sergo: Nice to see areas where this improve. Most of your talk was about classificaiton problems but we do regression problems in HEP. A: They are NLPs, LSTMs, ...

Tuesday Afternoon Session II

Fast Machine Learning at the LHC (Jean-Roch Vilmant)

Distributed training : [Mark] Reminded that there are two talks on this topic tomorrow, thinking how much HEP has to offer to technologies for distributed ML vs. industry that is putting so much effort into it.

Phil: What fraction of analyses will be ML-based in the future? A: hard to know

Fast Machine Learning is Cosmology (Brian Nord)

How many nights @ 1 TB / night ? → A: ~100 nights of observation per year.

Q: Is there any trigger in LSST or is there just difference images? A: Sorry, the differences images are part of the trigger system to tell LSST whether or not to follow some transient source

Fast Machine Learning in Accelerators (Auralee Edelen)

Phil: Do the diagnostics need to be “fast”? A: Rapid feedback from accelerator conditions to operators is very important (minutes -> seconds)

Q; Do you benefit from GANs? Yes, running 1 FEL simulations is a stochastic process and benefits

Fast Machine Learning in Neutrinos (Georgia Karagiorgi)

Phil: How long does it take to onload and offload data to the GPU? A: Yes, major challenge.
Phil: Would batching help overall? A: Yes, that’s an optimization that would help alot.

Fast Machine Learning in Gravity Wave Detection (Eliu Huerta)

Off to dinner...

Wednesday Morning Session I

Hls4ml status (Jennifer Ngaduba)

Real-time AI Systems in Academia (Giuseppe Di Guglielmo)

Real-time AI Systems in Industry (Jason Vidmar)

Wow, Versal new line looks cool.
Mark: Are there trade-offs relative to Ultrascale+ to fit this new AI-inspired functionality into the die? A: One thing is that there is less DSPs where some of those computations can be moved into the AI engine

Mark: How much will it cost relative to the Ultrascale+? A: Xilinx’s intent is that the Versal line will cost ~same as Ultrascale+ line

Neuromorphic Photonics (Charles Carnes)

Q: What is the largest matrix you have been able to calculate with this device? A: We haven’t yet built the device, but we’re working on it.

Q: What about calculations involving Nonlinearity? It only consumes 0.5% of consumed energy so not a big deal.

Q: How would on interface with this device. A: ADC interfeces, Also note that HEP already processes optical data, this would be natively processed in such an architecture.

Wednesday Morning Session II

Heterogeneous Computing as a Service in Industry (Andrew Putnam, Kalin Ovtcharov)

“At one point in time, Pokemon Go was the largest cloud computing workflow on the planet”

How do we use this giant network of interconnected FPGAs? A: some security considerations of giving that level of functionality to end-users on Brainwave

[Mark]: How far are we from being limited by speed-of-light latencies for accelerated inference at FPGA-based cloud resources. A: It varies based on data location, but even if local still have extra latency associated with network packet handling (not SFP-to-SFP)

SONIC Status (Miaoyuan Liu)

Heterogeneous Computing as a Service in Academia (Naif Tarafdar)

Q: Is the network limiting your scalability? A: No scalability limits and 10gb SFPs run at line rate

Wednesday Afternoon Session I

Training over Distributed Computing Systems (Jean-Roch Vlimant)

Distributed Training on HPCs (Aaron Saxton)

Scalable Systems Laboratory (Rob Gardner)

Wednesday Afternoon Session II (Lightning Talks)

Real-time RF Processing (EJ Kreinar)

Reconstruction at Nova (Thomas Warburton)

Searches for Black Holes with ML (Sean Condon)
Deep Cleaning for Gravitational Wave Data(Tri Nguyen)

Graph NNs for LHC reconstruction (Lindsey Gray)

IRIS-HEP & Community Planning Breakout

Recollection of the Goals and Deliverables from the Blueprint Overview talk

Major Goals

- Summarize the current status of R&D in the field

→ The many excellent workshop topical presentations!

- Get informed by experts on the latest technologies

→ Same as above with lots of IRIS-HEP people in attendance

- Communicate and contrast the various technological choices available for both hardware trigger algorithms and heterogeneous computing with accelerated machine learning

→ Same as above with lots of IRIS-HEP people in attendance

- Build and educate the community for developing a wide range of accelerated machine learning use-cases

→ Same as above with lots of IRIS-HEP people in attendance. There were many applications presented in neutrino, GW and collider physics. In this IRIS-HEP breakout session, we should boil this down to a few early HEP use cases

- Review the IRIS-HEP IA Milestones & Deliverables in Fast ML and plan for Y2 and the 18-month review

→ We should have a look as a group at the milestones and deliverables in the IA area for Fast ML. Some more of this will be discussed in the IRIS-HEP retreat.

Desired Outcomes/Deliverables

- Develop a roadmap of hardware trigger applications to pursue from near term to blue sky
- Develop a roadmap of accelerated machine learning computing applications from near term to blue sky
- Establish a hardware demonstration platform work plan at multiple sites
- Establish partners for possibilities for connecting with use-cases outside of LHC

In the context of the desired workshop outcomes & deliverables, we should work to answer a few questions by discussing:

Topic #1: ***Alignment***
Q: Given the mission of IRIS-HEP in addressing challenges for the HL-LHC era, what role should R&D in accelerated machine learning play in IRIS-HEP?

Topic #2: ***Opportunities***
Q: Given what we heard at the workshop over the last two days, are there new opportunities that we should consider pursuing within HEP? Specifically IRIS-HEP and/or FastML?

Topic #3: ***Use Cases***
Q: What are the primary use cases for Fast ML in HEP trigger and reconstruction (also considering tracking in both)?

- Can we identify a small number (maybe 1) of early use cases that are priorities for the stakeholder (experiments), allow us to work together on FastML within IRIS-HEP and across experiments to some (hopefully large) degree? What is the roadmap and the action items?

Topic #4: ***Infrastructure***
Q.: How can we establish a hardware demonstration platform for accelerated ML and inference at multiple sites? What role can/should SSL play in managing these demonstrators? What could be the role of cloud providers?

j

--- Topic #1: ***Alignment*** ------------------------------------

Q: Given the mission of IRIS-HEP in addressing challenges for the HL-LHC era, what role should R&D in accelerated machine learning play in IRIS-HEP?

Phil: Want infrastructure within SSL

Well-documented HLT test environment/demonstrator?

SSL can host existing knowledge/expertise on tools (how to set up / use) - git repo?

Survey what already exists for Run3/Run4 HLT

GitHub project with curated accelerated ML examples (with instructions) like https://github.com/iris-hep/awesome-hep

FastML group: R&D + IRIS-HEP: sharing developments with community

With regards the footprint/scope of the project, it may be beneficial to focus in a couple of use cases, rather than try to add as many heterogeneous resources as possible.

FastML: bleeding-edge R&D, small rough prototypes

IRIS-HEP can facilitate community adoption of ML tools
Smaller communities tend to follow LHC?

-- Topic #2: ***Opportunities*** -----------------------------------

Q: Given what we heard at the workshop over the last two days, are there new opportunities that we should consider pursuing within HEP? Specifically IRIS-HEP and/or FastML?

Photonics! Existing chip does 16x16 matrix multiplication, larger chips are planned

Xilinx new AI-inspired architecture

On-detector AI devices (radiation tolerant, process signals upstream of L1T)

Galapagos - run on CPU or FPGA interchangeably

Understand different kinds of scaling (single event batch vs. many event batch, shape of data)

Distributed training?

Online learning (a la prompt calibration loop)?

-- Topic #3: ***Use Cases*** -----------------------------------

Q: What are the primary use cases for Fast ML in HEP trigger and reconstruction (also considering tracking in both)?

- Can we identify a small number (maybe 1) of early use cases that are priorities for the stakeholder (experiments), allow us to work together on FastML within IRIS-HEP and across experiments to some (hopefully large) degree? What is the roadmap and the action items?

Accelerate graph NN for tracking in trigger, and/or pixel seeding, road finding, ...
Also jets (replace fastjet w/ NN), calorimetry (clustering), …
Taus, lepton isolation (existing NNs)

Get more physics (lower thresholds or more triggers if HLT latency can be reduced)

Pileup suppression? (ML PUPPI has been piloted)

Particle flow (could be cross-cutting)

Improving Online (L1) MET?

Various tagging algorithms (don't dominate computing currently)

CMS HCAL ML reco?

Need to understand data flow, physics validation, comparison to classical algorithms for any new network (especially blue-sky exploratory efforts) - need expert feedback

Fast Kalman filter using NNs (separate from graph NN), can be accelerated
-> doesn't help with combinatorics

Data compression -> autoencoders

--- Topic #4: ***Infrastructure*** -----------------------------------

Q: How can we establish a hardware demonstration platform for accelerated ML and inference at multiple sites? What role can/should SSL play in managing these demonstrators? What could be the role of cloud providers?

(Ted): Suggest looking into something like mlperf.org. You provide the data and model, and then let the industry do their best work and optimize their system. Then you have a benchmark and apples-to-apples comparison of what's out there, and the burden of evaluation and optimization is not on you. The motivator for industry is usage of their product.

Google doc from Rob Gardner's talk on SSL

Kevin : Is there a way to use SSL for R&D? And is there a way to scale SSL from R&D to production?
=> Production resources are static at the moment, but on HL-LHC timescale an aim is for adiabatic convergence towards an SSL motivated model.

E.g. HPC in a box? Can't virtualize HPC with same level of flexibility as institutional resources
Small-scale HPC-like architecture & constraints
Spack is large-scale management of software, also a tool exists for per-job management from VC3 to inspect environment and install dependencies as needed.

Can manage cloud resources with associated credits
-> need to avoid using up credits too quickly… some solutions exist (e.g. detection (& public shaming) for people who leave inactive nodes online)

Heterogeneity of resources requires more management, role for SSL

Can we settle on an asyncrhonoous service based protocol? (gRPC is broadly supported)

Document APIs/interfaces for accessing accelerators
E.g. gRPC (Microsoft Brainwave, Google TPU, etc.), REST API (Nvidia), etc.
Can be used in HLT, offline, analysis…
Hard part is asynchronous requests (CPU does other work while it waits for response) -> SONIC approach in CMS

Use of directly connected GPU/FPGA via CUDA/OpenCL (e.g. AWS F1 instance) has some similarities, but not really accelerator as a service (not using a standard network communication protocol)

https://mlperf.org/inference-overview/#benchmarks
Microsoft/Xilinx/…. Industry has something called MLPerf. You provide the data, you provide the model and then everybody reports their best result and it allows for a comparison for industry.

Target standardized FPGA board e.g. Alveo from Xilinx
AWS F1s can have 8 FPGAs per CPU, but currently (artificially?) limited to 1

Attendance

Kevin Pedro
Lindsey Gray
Nhan Tran
Mark Neubauer
Peter Elmer
Ben Tovar
Robert Gardner
Javier Duarte
Phil Harris
Jean-Roch Vlimant
Isobel Ojalvo
Savannah Thais
Ted Way
Sudhir Malik
Zhenbin Wu
Dylan Rankin
Gordon Watts