Summary of Blueprint Workshop:

Fast Machine Learning and Inference

September 10–11, 2019
Fermilab
Meeting URL: https://indico.cern.ch/event/822126

Workshop Organizers:
Phil Harris (MIT)
Burt Holzman (Fermilab)
Shih-Chieh Hsu (University of Washington)
Sergo Jindariani (Fermilab)
Maurizio Pierini (CERN)
Mark Neubauer (University of Illinois at Urbana-Champaign)
Nhan Tran (Fermilab)

Summary prepared by:
Javier Duarte (University of California at San Diego)
Phil Harris (MIT)
Burt Holzman (Fermilab)
Sergo Jindariani (Fermilab)
Mark Neubauer (University of Illinois at Urbana-Champaign)
Kevin Pedro (Fermilab)
Gabriel Perdue (Fermilab)
Nhan Tran (Fermilab)

Major Goals

- Review the status of the Analysis Systems (AS) milestones and deliverables to inform the needs for a collaborative development and testing platform.
- Develop the Scalable Systems Laboratory (SSL) architecture and plans, using AS R&D activities as specific examples.
- Develop requirements on SSL to support the AS area, particularly the prototyping, benchmarking and scaling of AS deliverables toward production deployment.
- Increase the visibility of SSL and AS beyond IRIS-HEP to facilitate partnerships with organizations that might provide software and computing resources toward these objectives.
- Get informed on latest developments in open source technologies and methods important for the success of the SSL and AS R&D areas of the Institute.

Key Outcomes

- Communication of the AS area plans leading to a set of requirements to SSL team.
- Kubernetes identified as a planned *common denominator* technology for the SSL, increasing our innovation capability through flexible infrastructure.

About the Blueprint Activity: Designed to inform the development and evolution of the Institute's strategic vision. At its core, a series of workshops that bring together IRIS-HEP team members, key stakeholders and domain experts from disciplines of importance to the Institute's mission.

- Plans for a multi-site SSL *substrate project* that will federate SSL contributions from multiple resource providers (institutes and public cloud), offering the AS area a flexible platform for service deployment at scales needed to test the viability of system designs.
- Productive engagement of the AS/SSL team with representatives from NCSA, SDSC, NYU Research Computing, industry & cloud providers (Google, Redhat), generating actions and informing Year 2 planning of IRIS-HEP.
- A vision for an SSL that serves as an innovation space for AS developers and a testbed to prototype next generation infrastructure patterns for future HEP computing environments.

I Introduction

In search of answers to the most fundamental questions of our universe, particle physics has continually pushed the bounds of detector sensors and readout to higher granularities, better sensitivities, and larger rates. The incredible data rates at high energy physics (HEP) experiments are among the largest in the world. Machine learning techniques have the potential to greatly empower and expand the capabilities and science of the experiments at all stages of data processing, simulation, and analysis. In the Fast Machine Learning workshop held September 10–13, 2019 at Fermilab/LPC, we focused on *novel resource-constrained machine learning applications for HEP: from low-latency, low-power, and high throughput on-detector systems to massively accelerated compute*. We colloquially refer to this area as "fast machine learning." The workshop brought together domain experts from **microelectronics, photonics, computing, machine learning, and several areas of physics** to discuss cutting-edge techniques, their complementarity, and how they can be applied to solve important scientific problems. It is important to note that outside of physics, trends in microelectronics and computing are often driven by machine learning and its applications and therefore advances in the former provide natural benefits to machine learning algorithms.

The workshop agenda was divided into 3 parts: Challenges and Opportunities, Science Applications, and Techniques and Tools. In the "Challenges and Opportunities" section, we set the stage for the need for low-latency, low-power, high-throughput, and resource-constrained machine learning. We began by discussing the challenges facing real-time detector systems. In particular, we highlighted the whole LHC computing model and had a dedicated talk on one of the most difficult throughput and latency cases in HEP: the LHC trigger. We also heard about the need for new computing paradigms and cases beyond physics. Then we concluded with a talk on how machine learning techniques could also be further optimized through techniques like sparsity and quantization (i.e. reduced numerical precision). These talks set the stage for the core of the workshop, which focused on comparing and contrasting the applications in fundamental physics and the tools that we need to develop to find solutions to those applications.

Following the talks, at the workshop, two days of hands-on tutorials and demonstrations were given. The tutorials relied on cloud resources and showed people how to use FPGAs in the Amazon Web Services cloud, and with FPGA cluster in Microsoft Azure for inference as a service. This provided a window for new members to the community and showed them that FPGA-based accelerator development can be made tractable.

II Applications for "Fast Machine Learning"

The history and motivation for machine learning across HEP are broad and have been discussed in detail elsewhere. There are many reviews (e.g., Ref. [1]) which have laid out the impact from detector/accelerator controls and operation to data simulation, reconstruction, and analysis. We focus here on the scientific opportunities of "fast machine learning." Rather than categorizing by scientific domain as is typically done, we categorize the opportunities below by their techniques and methodologies. This allows us to find common threads across different domains of physics and industry. Furthermore, we aim to exploit this style of organization as we continue to pursue developments in each area.

II.1 Low-latency on-detector processing

We can broadly define the low-latency on-detector processing as machine learning applications which have "no time for a CPU or GPU." We focused on detector systems where analog or digital signals have a latency requirement less than (very roughly 1 ms) or throughput requirements at bandwidths greater than modern Ethernet speeds provided by a global infrastructure. In this workshop, we discussed:

- LHC: anywhere in the early processing stages from the front-end analog-to-digital signal conversion to the hardware trigger, latencies are of the order of 1 μs or smaller.
- Particle accelerators: real-time feedback control of MHz RF signals of the beam require on-system diagnostics and embedded machine learning implementations.
- Neutrinos: for next-generation experiments such as the Deep Underground Neutrino Experiment (DUNE), there is a class of physics processes (e.g. supernovae, proton decay) that are not associated with the beam structure; these require challenging pattern recognition algorithms to be performed continuously on the detector readout and thus require timescales of <1 ms.

II.2 High-throughput computing (inference)

Several of the experimental physics programs that were discussed have needs for high throughput computing to process petabyte to exabyte scale datasets (or even larger, when real-time online processing is required before storage). Therefore, the event processing and simulation of these massive datasets are a natural application for very high throughput inference of complex machine learning models. Many groups have started to study co-processors, that is CPU hardware augmented by specialized machine learning hardware like GPUs, FPGAs, and ASICs.

The LHC has traditionally been a major consumer here with high-level trigger computing farms having to process hundreds of kHz of data with latencies on the order of 100 ms. Once data is stored offline, future exabyte datasets for the HL-LHC will require being simulated and reconstructed with sophisticated machine learning algorithms. Machine learning algorithms already exist within LHC reconstruction; recently, a significant uptick in machine learning algorithms for core reconstruction has occurred due to the extended capabilities available with new machine learning tools.

Similar offline computing throughput and latency bottlenecks exist for neutrino and cosmology experiments. Further, because the reconstruction algorithms used for both neutrino detectors and astrophysical data employ image-based processing of convolutional neural networks — a mainstay of industry-based co-processor developments — the technology is even closer to maturity. One interesting scenario within cosmology that would benefit from improvements in high-speed computing is the identification of transient objects. Currently, the online filtering of potential transient objects requires a simplistic difference-based imaging algorithm performed with a short lifetime through millions of images in less than an hour.

Lastly, with gravitational waves, the machine learning-based algorithmic processing of gravitational signals — going all the way to the properties of the merger — can lead to significant speed-ups in processing. Rapid processing of LIGO data can significantly reduce the time window for the identification of neutron stars and other interesting mergers, thereby allowing for enhancements in the use of multi-messenger astronomy. A similar challenge exists for the processing of telescope images in other multi-messenger astronomy experiments.

II.3 Large scale training and modeling

In addition to inference challenges, training more and more sophisticated models and optimizing network hyperparameters is a growing challenge for the field as well. As machine learning algorithms become more sophisticated, there are many compelling applications which will potentially go beyond a single-to-few GPU training run in a reasonable amount of time. Beyond the simple growth of training samples and model size, applications such as exploring new neural network architectures (like graph neural networks), quantifying uncertainties, and developing sophisticated models for simulation will require large-scale machine learning workflows. In the workshop, we have heard about explorations of distributed training using open-source tools; we are interested to try more cutting-edge hardware as well.

III Cross-Cutting Initiatives and Outlook

Through both presentations and discussions, we continue to identify interesting areas of cross-cutting applications across the HEP physics program and with industry partners. Industry is pursuing their efforts, and they are interested to understand where their developments best overlap with science-based demands and how they can further direct their research to further enhance scientific output.

We highlight a few of the most interesting opportunities:

- Programmable hardware (FPGAs) in real-time systems: `hls4ml` [2] is an open-source tool being developed by physicists and engineers to provide hardware solutions across orders of magnitude of latency — from sub-microsecond constraints at the LHC and for RF accelerator signals to millisecond timescales of neutrino experiments. The advantage of `hls4ml` is that it provides a hardware solution for a dedicated machine learning algorithm under constrained resources that can be embedded into existing processing systems. Such embedded systems are interesting to companies like Xilinx (FPGA manufacturer) and HawkEye360 (RF signal processing) who also operate across a similar range of latencies, from satellite communications to self-driving cars. An interesting spin-off of this category is weight-programmable ASICs, which offer advantages over FPGAs for certain experimental requirements that are common in HEP including low-power, high radiation, and low temperature environments. Such readout systems could generally also be important for improvement in the readout of quantum computing systems.

- An emerging technology with the potential for ultra-low power and very low latencies is neuromorphic photonics. We discussed where such systems could be used, such as at analog front-ends, and also the challenges of this technology to be scaled up for practical use. Development is very rapid, and this technology could potentially become realistic for computing clusters within a few years.

- The area of heterogeneous computing, both for training and inferences, is very active both in academia and industry. However, HEP is uniquely positioned to be a super-user as a partner to the computing community. With the LHC and the future SKA, physics continues to have the largest datasets in sciences. These large datasets present unique challenges for distributed heterogeneous computing. This includes pushing computing architectures (silicon technologies) for both inferences and training, datacenter infrastructure including node orchestration and communication, and network bandwidths across the internet. Additionally, the long history of ASIC/FPGA use within HEP allows for existing trigger and data acquisition (DAQ) experience to be reutilized. The workshop included contributions from industry (Microsoft) using FPGAs across the entire data center stack and academia (University of Toronto heterogeneous computing stack). While FPGAs were highlighted in this workshop, GPUs and ASICs provide complementary capabilities that need to be further explored in more detail across HEP.

The closing part of the workshop addressed coordination with existing computing efforts within HEP. In particular, the meeting also served as a blueprint workshop for IRIS-HEP, an NSF project targeting software tools needed for the future HL-LHC program. Because heterogeneous computing hardware is not firmly within the scope of IRIS-HEP, but software interface tools are, one solution would be to collaborate and interface through the IRIS-HEP effort using their Scalable Systems Laboratory (SSL) where a broad community of LHC users could access and test available hardware from partnering institutions. Because IRIS-HEP focuses on software, machine learning algorithms will continue to be part of their scope; hardware integration is an opportunity for collaboration.

There are exciting challenges in "fast machine learning" for fundamental physics. While we are beginning to develop the technology needed for current and future physics experiments, this workshop has only started to explore the cross-cutting implications both across physics and also with industry.

References

[1] Kim Albertsson et al. Machine Learning in High Energy Physics Community White Paper. *J. Phys. Conf. Ser.*, 1085(2):022008, 2018.

[2] hls4ml. https://github.com/hls-fpga-machine-learning/hls4ml.

A Revision History

- Version 0.1
  - Initial version from Nhan and Co.