Opportunities for Accelerated Machine Learning Inference in Fundamental Physics

Markus Atkinson[1], Javier Duarte[2], Philip Harris[3], Alex Himmel[4], Burt Holzman[4], Wesley Ketchum[4], Jim Kowalkowski[4], Miaoyuan Liu[4], Mark Neubauer[4], Brian Nord[4], Gabriel Perdue[4], Kevin Pedro[4], Nhan Tran[4], and Mike Williams[3]

[1]University of Illinois Urbana Champaign, Champaign, IL 61820, USA
[2]University of California San Diego, La Jolla, CA 92093, USA
[3]Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Fermi National Accelerator Laboratory, Batavia, IL 60510, USA

ABSTRACT

In this brief white paper, we discuss the future computing challenges for fundamental physics experiments. The use cases for deploying machine learning across physics for simulation, reconstruction, and analysis is rapidly growing. This will lead us to many applications where exploring accelerated machine learning algorithm inference could bring valuable and necessary gains in performance. Finally, we conclude by discussing the future challenges in deploying new heterogeneous computing hardware.

This community report is inspired by discussions at the Fast Machine Learning Workshop[1] held September 10-13, 2019.

Contents

1 Introduction

Fundamental particle physics has pushed the bounds of computing for decades. As detectors become more sophisticated and granular, particle beams become more intense, and the datasets collected grow, the processing needs of the biggest fundamental physics experiments in the world are presented with massive computing challenges. In this short note, we discuss the upcoming challenges of a selection of current and future particle physics experiments, how they are intertwined with the development of machine learning algorithms, and where applications with heterogeneous computing for event processing (inference) can potentially provide breakthrough gains in performance.

1.1 Computing model in particle physics

Fundamental physics experiments provide uniquely massive datasets through exquisitely precise instruments and the need for large statistics of physical phenomena to study rare physics events. The basic unit of processing is the *event*, and the datasets comprise billions or trillions of events. Often, each event can be analyzed independently ("pleasingly parallel"), a good fit for *high throughput* computing. We typically process these large datasets multiple times.

[1]https://indico.cern.ch/event/822126

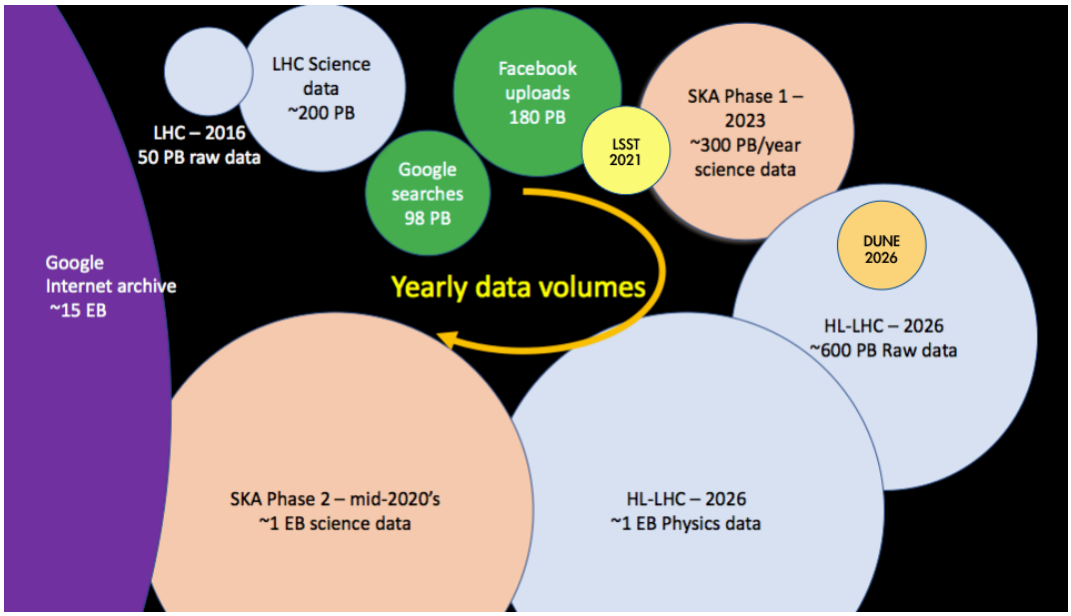

Figure 1. The projected dataset volumes for several big fundamental physics experiments in the mid-2020s.

Our current computing model primarily relies on on-premises computing resources, datacenters which are sited on national laboratories and university campuses. Furthermore, in international collaborations, there are multiple datacenters sited in multiple countries and maintained by multiple funding agencies. In an era where single-threaded performance of CPUs has plateaued and datasets continue to grow by orders of magnitude, more efficient and specialized type of computing architectures are being explored. As an example, in Fig. 1 there is a graphic of the various anticipated dataset volumes for several next generation fundamental physics experiments planning to be online in the 2020s. The Large Hadron Collider experiments will have datasets surpassing an exabyte and other neutrino experiments (DUNE) and cosmology surveys (LSST) will be roughly within an order of magnitude of the HL-LHC datasets. As a result, heterogeneous computing resources—mixed architecture computing systems which can offer large gains in performance—are becoming increasingly popular to meet data processing demands.

1.2 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. Our focus here is on the synergy between machine learning techniques across physics and computing architectures which are specialized for such computations. Many physics applications are based on processing or simulating single 'events', a unit describing a particular physical phenomena, and performing often several operations on that event. In almost all cases, custom neural network architectures are built for a given application and so flexibility is an important aspect to deploying new compute architectures for machine learning in fundamental physics.

2 Challenges and Applications for Accelerated Machine Learning Inference

We now present various applications for accelerated ML inference at several experiments. These consider both real-time streaming "online" applications where processing is a part of the data acquisition chain as well as "offline" applications where the data has been stored and awaits further, and often multiple, processings.

2.1 CMS and ATLAS

CMS (2) and ATLAS (3) are the two multipurpose detectors operating at the Large Hadron Collider (LHC). They have a broad program which pushes the energy frontier of particle physics: from understanding the Higgs boson and its potential connections new physical phenomena to looking for dark matter particle candidates and other hints of new physics such as supersymmetry and warped extra dimensions.

The LHC is a proton-proton collider which collides bunches of protons at a rate of 40 MHz. The LHC will undergo a "high-luminosity" upgrade in the 2020s which will increase the number of collisions that we will process and save creating new online and offline computing challenges which cannot be resolved with current computing paradigms and resources. The data rates are such that not all events can be saved, and even for the data which is saved, the offline processing of the massive datasets expected in the future ($>$ 1 EB) present high-throughput computing challenges.

If we take the CMS experiment as an example[2], the experiment has several sub-detector technologies which measure particle properties from the collision debris. These detectors output data on hundreds of millions to billions of channels. During the online data-taking, the event processing and filtering begins at the earliest on-detector stages in ASICs and FPGAs with communication and readout via optical fibers. After this first stage of filtering (L1 triggering), the data is passed to a computing farm where the events are further processed. This second stage of triggering (high level trigger, HLT) is where accelerated computing will be extremely important in meeting strict throughput requirements in the data acquisition chain. The upgraded HLT system for online filtering will require more than a factor of 20 increase in performance for computing power and storage and network throughput. This is shown in Table 1 by comparing the requirements for the current and upgraded systems. Further as the collision environments get busier with more simultaneous interactions, the processing complexity increases and the need for more sophisticated and powerful algorithms, often machine learning, also grows. Therefore, the HLT is a system which would be ideal for deploying new accelerated computing hardware to meet the increasing demands.

CMS detector	LHC (current)	HL-LHC (upgraded)
Simultaneous interactions	60	200
L1 accept rate	100 kHz	750 kHz
HLT accept rate	1 kHz	7.5 kHz
Event size	2.0 MB	7.4 MB
HLT computing power	0.5 MHS06	9.2 MHS06
Storage throughput	2.5 GB/s	61 GB/s
Event network throughput	1.6 Tb/s	44 Tb/s

Table 1. Specifications for the CMS high level trigger in the current run and estimates for the upgraded detector (4).

Furthermore, as the HLT accept rate increases and with the amount of data collected integrated over a decade of operation both grow, the HL-LHC offline dataset will also present severe challenges for event processing. The CMS and ATLAS computing models rely on a highly distributed international grid of computing resources spread across many countries[3]. Unlike the online streaming HLT system, which is local to the experiments at the LHC, we have to consider how we could deploy heterogeneous computing resources to accelerate our event processing chain in a truly distributed system. The datasets, as mentioned above, will push 1 EB and the amount of compute power that is needed will be more than an order of magnitude more than current levels. Finally, it is important to note that thus far, we are considering only event processing. Often our measurements and searches rely on robust and precise simulation which requires additional computing resources. As the collision environments become more complex and our datasets grow, the simulation requirements grow commensurately and perhaps even more as the simulation uncertainties must be reduced as well.

The future computing challenges of CMS and ATLAS, both in online streaming applications and offline raw processing, are significantly greater than the resources currently being deployed. This presents an opportunity for transformative changes to the computing model and technology.

2.2 LHCb

LHCb (5) is another one of the experiments at the LHC which specializes in identifying and studying bottom quark production and decays in order to search for and understand very rare physics events in the bottom quark sector. The overall beam intensity is lower than CMS and ATLAS, but it has similar, if not greater, future computing challenges.

The event sizes at LHCb are $\mathcal{O}(100\,\mathrm{kB})$, which is much smaller than at ATLAS/CMS, allowing for much larger trigger accept rates. In the current run, the L1 trigger accept rate was limited to 1 MHz by the capabilities of the front-end electronics. The HLT was run in two stages: HLT1 partially reconstructed events and selected a subset for further processing by HLT2, which performed a more complete reconstruction then executed many selection algorithms to further reduce the rate at which data were written to permanent storage. The HLT1 accept rate was $\approx 120\,\mathrm{kHz}$ in Run 2, and the storage throughput out of HLT2 was about 0.7 GB/s. In the most recent data-taking period, instead of immediately processing the data selected by HLT1 in HLT2, the data were cached on a 10 PB buffer while the full calibration procedure was performed, and then HLT2 was run

[2]The ATLAS specifications are similarly challenging
[3]https://wlcg.web.cern.ch/

on the fully calibrated data. This permitted writing out some of the data in a reduced-size format, since no offline processing of the raw detector information is required.

In the upcoming data-taking period beginning in 2021, upgrades to LHCb will permit reading out every event, which due to a planned increase in luminosity, will be 5 TB/s; *i.e.*, there will be no L1 stage. In total, 25 EB each year will need to be processed online using high-level computing algorithms. On average, this will require analyzing events 100 times faster than in Run 2. One exciting option being tested now is running the entire first trigger-processing stage on GPUs. Performing all of HLT1 on the GPU results in the limited GPU-CPU data-transfer rates having no impact on the throughput. Furthermore, most of the LHCb reconstruction tasks are easily parallelized. (A prototype application named ALLEN exists which would be capable of running the full HLT1 processing on less than 500 current nvidia GPUs.) The HLT1 accept rate will be 0.5–1.0 MHz, while the storage throughput out of HLT2 will be about 10 GB/s. Data caching will again be used to perform the calibrations in real time, and no offline reconstruction is envisioned. Therefore, offline computing resources will predominantly be used to generate simulated data.

For future runs of LHCb which begin in the mid-2020s (HL-LHC), it is interesting to consider how this computing model may change under more complex collision environments and how it can incorporate machine learning algorithms; related, it is important to understand how else the HLT farm can be used for analysis and simulation. However, understanding the performance of the system in this upcoming run will provide a valuable benchmark for what is required from future computing technologies.

2.3 LSST

The investigation of dark energy and dark matter is a major driver of the DOE Cosmic Frontier program. To advance our understanding, large astronomical sky surveys that use large telescopes and digital cameras are now critical for understanding this "dark sector"that comprises most of the Universe. Sky surveys scan the sky over the course of many months or years to take 10^5–10^6 gigapixel-size images, containing data on relatively static objects like galaxies, as well as transient (moving) objects like planets and stars, and variable phenomena like exploding stars (supernovae).

Supernovae are critical for constraining the expansion rate of the universe, but they are challenging to capture and to classify accurately. After they erupt, they quickly increase in brightness to a peak, and then dim over the course of days or weeks. It is not yet possible to predict where a supernovae will occur on the sky, and therefore, the same field must be visited regularly. Once a supernova is identified, for example through detecting extreme brightness changes in the same pixel, more information is required to precisely classify the type of supernova. This then necessitates follow-up observations at another telescope as soon as possible, before the supernova becomes too dim to observe.

For example, the Dark Energy Survey (DES) (6) is a state-of-the-art experiment that recently finished six years of data collection in early 2019 — obtaining $\sim 10^8$ galaxies and thousands of supernovae. The next-generation Large Synoptic Survey Telescope (LSST) (7) is a time domain survey that will observe orders of magnitude more galaxies and supernovae, by taking a picture of the entire southern hemisphere once every four nights over the course of 10 years. This will produce over 10 million transient alerts per night, which need to be classified, cataloged, and communicated to the public immediately. LSST will produce sufficiently large and complex data sets that detecting supernovae and other variables with sufficient accuracy is becoming prohibitively time-consuming for standard non-AI algorithms; for AI algorithm implementations, heterogeneous computing hardware may greatly accelerate detection accuracy and timing within reach of science requirements.

Beyond the specific challenges for supernovae and variables, the LSST pipeline for cleaning and processing images into object catalogs will only run once per year, which severely limits opportunities to implement improved analysis techniques: in addition, identifying and modeling objects requires 0.1 seconds and 60 seconds per object, respectively. For over a billion objects in the survey, this requires over 15 million CPU hours. AI algorithms accelerated by heterogeneous compute architectures could dramatically change our capabilities for the continuous development of modeling and analysis tools.

Finally, large n-body simulations of cosmological volumes are critical for testing our analysis tools. However sufficiently detailed and large simulations require millions of CPU hours to generate. AI tools are emerging as important methods for creating surrogate representations of simulations with equivalent physical fidelity. Generative models run on GPUS, FPGAs, and ASICs have the potential to increase access to large simulations with a variety of underlying fundamental theories.

2.4 LIGO

The Laser Interferometer Gravitational Wave Observatory (LIGO) (8) is capable of detecting gravitational waves (GW) from black hole and neutron star mergers. These gravitational waves can often be correlated with additional astrophysical signatures coming from broad electromagnetic (EM) signatures. In 2017, the first correlated electromagnetic and gravitational wave incident was observed. This had a profound scientific impact in many fields, and in particular showed that neutron star mergers do not create gamma ray bursts. These correlated events have heralded a scientific field referred to as Multi-Messenger

Astronomy (MMA) where by gravitational signatures are directly correlated with signatures from telescopes and neutrino detectors.

To maximize the scientific output of MMA, we aim to reconstruct gravitational waves within the minimum amount of latency. Gravitational waves will typically arrive first, with EM signatures coming as short a seconds after the arrival of the GW signature. Telescopes with limited amount of directional resolution have to react to the parameters of the gravitational waves to allow for near-immediate correlated detection of electromagnetic sources. Current reconstruction analysis of the algorithms follow a systematic procedure that takes about minutes to reconstruct wave forms. Ideal times for reconstruction and messaging would benefit from analysis that can be done in seconds or fractions of a second.

Neural network models that characterize a 4-D signal manifold identical to the parameter space covered by established algorithms that are used in actual low-latency GW searches have been developed (9). These networks have already demonstrated significant speedups for an inference. Already, these models can process GW data faster than real-time using a single GPU. They have been applied to characterize the astrophysical properties of BBH mergers, and of the post-merger object, demonstrating for the first time that deep learning can perform tests of general relativity in real-time. Other aspects of the reconstruction, in particular noise characterization, are currently being approached with machine learning techniques with evidence that these algorithms can also be ported to ML. Full acceleration and integration of this system with the multi messenger framework is still to be done, but there are encouraging possibilities.

Evidence for significant speed-ups with heterogeneous hardware compounded with the fact that LIGO is being upgraded, and additional gravitational wave observatories are coming online. By 2025 the rate of gravitational waves is expected to increase by two orders of magnitude. The increased throughput presents a further challenged to ensure high speed inference and signal processing is needed to detect gravitational waves.

2.5 DUNE

The Deep Underground Neutrino Experiment (DUNE) will conduct a rich program in neutrino and underground physics, including determination of the neutrino mass hierarchy and measurements of CP violation in neutrino mixing using a long baseline accelerator-based neutrino beam, detection and measurements of atmospheric and solar neutrinos, searches for supernova-burst neutrinos and other neutrino bursts from astronomical sources, and searches for GUT-scale physics in proton decay.

The detectors will consist of 4 modules, of which at least three are planned to be 10 kton Liquid Argon Time Projection Chambers (LArTPCs). Charged particles produced from neutrino or other particle interactions will travel through and ionize the argon, with ionization electrons drifted over many meters in a high electric field, and detected on planes of sensing wires or printed-circuit-board charge collectors. What results is essentially a high-definition image of a neutrino interaction, which naturally lends itself to applications of machine learning techniques designed for image classification, object detection, and semantic segmentation. Machine learning can also aid in other important applications, like noise reduction and anomaly/region-of-interest detection.

Due to the size and long readout times of the detectors, the data volume produced by the detectors will be very large: uncompressed continuous readout of a single module will be nearly 1.5 PB per second. Because that amount of data is impossible to collect and store (not to mention process), and because most of that data will not contain interactions of interest, a real-time data selection scheme must be employed to identify and store data containing neutrino interactions. With a limit on total bandwidth of 30 PB of data per year for all DUNE modules, that data selection scheme (and any accompanying compression) must effectively reduce the data rate by a factor of around 10^6.

That data selection scheme must also operate on a time-scale commensurate with the available data buffers in the readout electronics and data acquisition system: in current designs, those buffers may hold up to 10 s worth of data, and so decisions as to when and what parts of the detector data to extract and store must be made on the order of a few seconds.

As mentioned above, machine learning has many potential applications to the basic data preparation and processing of LArTPC data, and so fast inference that can compress/decompress data, apply noise filtering and region-of-interest detection, and perform basic signal (neutrino interaction) identification in the presence of backgrounds (largely noise and radiological backgrounds) may be critical parts of the DUNE data selection process.

Further details on the detector, data acquisition, and data selection can be found in Ref. (10).

In addition to applications in real-time data selection, accelerated machine learning inference that can scale to processing of large data volumes will be important for offline reconstruction and selection of neutrino interactions. A total data volume of 30 PB of raw data is anticipated to be collected per year, with individual event sizes of the order of a few GB, and extended readout events (associated, for example, with supernova burst events) that may be around 100 TB per module. It will be a challenge to efficiently analyze that dataset without transformations in computing models and technology that can handle data retrieval, transport, parrelelized processing, and storage in a cohesive manner.

3 Outlook and Opportunities

In the previous section, we outlined a number of the present and upcoming computing challenges for a several fundamental physics experiments. Furthermore, this is just a sampling and there are many other experiments with similar challenges. This includes both real-time streaming computing applications as well as large dataset processing and simulation performed offline.

Given these challenges, we believe there is the potential for great impact in fundamental physics in deploying new technologies. Across the various experimental applications, there are some generic considerations when approaching these opportunities.

- **Generalizable models**: while machine learning is a powerful tool, the exploration of such techniques is constantly evolving and the types of network architectures (and thus specialized computations) will also necessarily change; physicists are often creating custom network architectures which presents challenges for generalizability of hardware
- **Elastic, non-disruptive**: a considerable amount of the current infrastructure and software has been developed around the (event-based) computing model for fundamental physics; being able to incorporate and elastically extend that computing model, for example through services and container orchestration, is an important consideration in how best to deploy emerging hardware
- **Global infrastructure**: related to the previous point, often for large international experiments, the computing resources are globally distributed across many countries (certainly true for the LHC experiments); this will challenge not only raw computing power but also networking bandwidth, data storage, and datacenter orchestration and operation.

The intersection of **machine learning inference and accelerated compute is a very exciting opportunity** with rapid developments in both academia and industry, and fundamental physics is an area which presents interesting applications to **explore emerging technologies**.

References

1. Albertsson, K. *et al.* Machine Learning in High Energy Physics Community White Paper. *J. Phys. Conf. Ser.* **1085**, 022008, DOI: 10.1088/1742-6596/1085/2/022008 (2018). 1807.02876.
2. Chatrchyan, S. *et al.* The CMS Experiment at the CERN LHC. *JINST* **3**, S08004, DOI: 10.1088/1748-0221/3/08/S08004 (2008).
3. Aad, G. *et al.* The ATLAS Experiment at the CERN Large Hadron Collider. *JINST* **3**, S08003, DOI: 10.1088/1748-0221/3/08/S08003 (2008).
4. Collaboration, C. The Phase-2 Upgrade of the CMS DAQ Interim Technical Design Report. Tech. Rep. CERN-LHCC-2017-014. CMS-TDR-018, CERN, Geneva (2017). This is the CMS Interim TDR devoted to the upgrade of the CMS DAQ in view of the HL-LHC running, as approved by the LHCC.
5. Alves, A. A., Jr. *et al.* The LHCb Detector at the LHC. *JINST* **3**, S08005, DOI: 10.1088/1748-0221/3/08/S08005 (2008).
6. Flaugher, B. *et al.* The Dark Energy Camera. **150**, 150, DOI: 10.1088/0004-6256/150/5/150 (2015). 1504.02900.
7. LSST Science Collaboration *et al.* LSST Science Book, Version 2.0. *ArXiv e-prints* (2009). 0912.0201.
8. Abbott, B. P. *et al.* LIGO: The Laser interferometer gravitational-wave observatory. *Rept. Prog. Phys.* **72**, 076901, DOI: 10.1088/0034-4885/72/7/076901 (2009). 0711.3041.
9. George, D. & Huerta, E. A. Deep Learning for Real-time Gravitational Wave Detection and Parameter Estimation: Results with Advanced LIGO Data. *Phys. Lett.* **B778**, 64–70, DOI: 10.1016/j.physletb.2017.12.053 (2018). 1711.03121.
10. Abi, B. *et al.* The DUNE Far Detector Interim Design Report, Volume 2: Single-Phase Module. (2018). 1807.10327.