Accelerating offline computing with the Fast Machine Learning Lab

Contact PI: Philip Harris (MIT)
PI: Javier Duarte (UCSD)
PI: Kevin Pedro (FNAL)
PI: Nhan Tran (FNAL)
Proposal Theme : adapting reconstruction algorithms to run on heterogeneous computing resources
Research Project Term: June 1, 2020-May 31, 2021
Monetary request: $43,172 (total), $31,932 (salary) $11,240 (fringe)

Contents

1 Introduction

With the breakdown of Dennard scaling in the late 2000s and the resulting inability to significantly increase processor clock frequencies, there has been a progression toward computing models with larger degrees of parallelism. This trend started with multicore CPU processors, continued with general-purpose graphics processing units (GPUs), and now includes field programmable gate arrays (FPGAs) and application specific integrated circuits (ASICs). The rise of deep learning has further increased the pace of this trend since the backbone of deep learning is highly parallelizable large-scale matrix multiplications. The success of GPUs, ASICs, and FPGAs to accelerate deep learning has motivated the development of heterogeneous computing clusters, consisting of a mixture of CPUs, GPUs, FPGAs, and ASICs. In light of these developments, several groups within CMS have developed deep learning algorithms that replace and/or improve upon existing reconstruction algorithms, such as jet tagging and jet energy regression [1, 2, 3, 4], charged particle tracking [5], particle-flow association, and more. Additionally, various groups in HEP are considering GPU based algorithms for accelerating core reconstruction [6, 7]. In this proposal, we aim to increase the rate of such developments both by extending the scope of hardware developments beyond local GPUs to High Performance Computing centers (HPCs), FPGAs, and ASICs, and by incorporating deep learning into the heterogeneous workflow. We will accomplish these developments by ensuring members of CMS have full access to all possibilities of heterogeneous computing. To allow access we aim to build and support heterogeneous computing toolchains to allow for developers to integrate their new algorithms into CMSSW, test their algorithms on different heterogeneous platforms, and understand the limitations and benefits of different designs.

Within CMS, members at Fermilab, CERN, UCSD, MIT, and other institutes have formed a collaboration targeting the integration of accelerated machine learning (ML) into the software framework. This new collaboration, the Fast Machine Learning Lab [8], aims broadly at building a knowledge base and software toolkit to take advantage of ML speedups with specialized processors. In this context, we are building the knowledge and expertise required to (1) understand when heterogeneous computing architectures can lead to significant computing improvements and (2) deploy algorithms on these heterogeneous platforms in an efficient and non-disruptive way. Already, we have made several broad observations:

- Deep learning methods have the potential to replace a large number of traditional algorithms.
- Deep learning model inference can be significantly accelerated using GPUs/FPGAs/ASICs with minimal code development.
- Outsourcing heterogeneous computing architectures (GPU/FPGA/ASIC) to remote servers that can be utilized as needed ("as-a-service") allows for a natural transition to heterogeneous computing with some additional overhead in latency, but *minimal* cost in throughput.

Within the context of offline computing in CMS, we see that all three of these aspects would benefit from a dedicated R&D effort. The first item, machine learning algorithm development, is already underway. However, this development has been largely unconcerned with issues related to algorithm latency and throughput. The second item, the porting of algorithms to specialized coprocessors, requires a knowledge base and some hardware experience. For standard deep learning

algorithms, a large ecosystem of industry tools exists for training and inference on CPUs or GPUs. However, certain deep learning architectures may benefit more from certain co-processors, such as FPGAs, and many of the existing toolchains require additional expertise. The last item, utilizing "as-a-service" (aaS) tools, is a very natural way to gradually add heterogeneity, since the additional code needed to run algorithms on specialized processors can be standardized to set of asynchronous and non-blocking server calls independent of the processor technology. Whether it be GPU, ASIC, or FPGA the use of aaS works uder all paradigms. However, this requires additional software to set up and maintain the servers as well as the necessary development to integrate such calls into CMSSW. Here, we propose for the postdoc to maintain the heterogeneous computing lab so that members of the CMS collaboration can understand the benefit from the use of heterogeneous computing within CMSSW.

2 Previous Experience

The Fast Machine Learning Lab is already working to develop a dedicated hardware setup and software stack. With the lab, we have already set up a central code repository along with hardware from industry donors, and we are building institutional knowledge for accelerated machine learning within CMS. Our current machines, `ailab01.fnal.gov` and `prp-gpu-1.t2.ucsd.edu`, are equipped with multiple NVIDIA Tesla T4 GPUs, NVIDIA GeForce GTX 1080 GPUs, and Xilinx Alveo U250 FPGAs. They have already been used as a GPU service engine to run CMSSW reconstruction, CMS HLT, and several FPGA-based algorithms. Additionally, through an NSF/Internet2-funded project "Exploring Clouds for Acceleration of Science," we have been allocated $75k cloud credits distributed across Google cloud platform and Amazon Web Services to test large scale "production"-sized tests of deep learning acceleration. With our existing cloud credits, we have been able to dynamically scale up the GPU and FPGA usage to test the ML accelerator infrastructure and to measure latency speed ups, throughput and understand bottlenecks. The advantage of the cloud is that it allows for quick profiling of different hardware architectures on scales that are comparable to the CMS computing infrastructure [9]. In the future, we hope to extend this work towards equivalent usage in DOE administered High Performance Computing (HPC) centers.

The algorithms we have run in CMSSW rely on the ExternalWork feature in CMSSW. Here, we have built an interface, which we refer to as Services for Optimized Network Inference on Coprocessors (SONIC), to perform the gRPC calls for both FPGAs and GPUs. For FPGAs, using the Microsoft Azure cloud and SONIC, we have demonstrated improvements in computing a top tagger using the ResNet-50 architecture of over a factor of 20 in latency and a factor of several hundred in total throughput [10]. With GPUs and SONIC, we have mirrored these studies and we observe comparable improvements with roughly a factor of 100 increase in throughput. Recently, we have been working on developing an ML based replacement of the HCAL local energy reconstruction algorithm. With SONIC, we have already run this algorithm while running the full HLT reconstruction chain on a GPU both as a service and locally, and locally on an FPGA. We see greater than a factor of three improvements in reconstruction latency switching to ML. Going further to GPU we see an additional factor of two; with an FPGA we obtain a further reduction of 2 [11]. While the HCAL algorithm is very simple the current non-ML algorithm comprises roughly 15 percent of the total HLT budgetm thus the adoption of an accelerated algorithm would already lead to significant performance improvements; more complicated algorithms will

lead to substantially larger improvements when accelerated.

In the interest of improving the whole stack, our team is also pursuing next-generation deep learning accelerator approaches. In particular, with FPGAs, the FML team has developed a deep learning compiler called `hls4ml` [12]. In collaboration with computer scientists at the University of Toronto, we are working on a heterogeneous abstraction stack with optimized protocols so that we can natively integrate FPGAs into the CMS software stack [13]. This work has the potential to allow for fast integration of "bump-in-the-wire" technology that will be available in next-generation heterogeneous clusters such as those being developed by Nimbix.

3 Proposal

In this proposal, we request 50% remuneration for a postdoc to facilitate the acceleration lab. To perform this, the postdoc will maintain an inventory of accessible hardware along with the software tools to use them. The idea for the hardware component of the lab follows from equivalent setups in other parts of CMS, such as the mini-DAQ used within TriDAS [14] or the cosmic rack used within the tracker project [15]. The postdoc will play a leading role in the gRPC CMSSW developments and will maintain the software stack so that fast integration and testing of accelerator based algorithms can be performed. As has already been the case, the software stack will be developed by a broader group of people, including dedicated computing experts. The goal of this position is to integrate these developments into CMSSW and to facilitate testing of new algorithms. While we do not envision this position to focus heavily on software development, we expect the postdoc to work with software experts and other physicists to help develop new features within the software stack.

In developing the software stack, the postdoc will ensure SONIC is functional with existing hardware and has a generic "fallback" option for when heterogeneous resources are not available (SwitchProducers) or there are hardware or network failures. We aim for "full stack" competency, including setting up services/servers for different resources (such as cloud and HPC) and running workflows in CMSSW. The postdoc will develop standard tools for integration of "generic" ML algorithm inference in CMSSW based on ONNX specification and tools for fast prototyping of ML algorithms in CMSSW using SONIC w/ Apache IPC. To tie this in with active R&D efforts, the postdoc will help to pursue the next generation of developments on the HCAL algorithm targeting ECAL and neural network clustering. This work parallels other developments using more complicated graph network architectures for HGCal reconstruction [16, 17] within the ML4RECO working group, with whom we actively collaborate. A detailed timeline of the goals is outlined in table 1. The project is split into first establishing the accelerator lab as a CMS resource that is accessible to the collaboration, and to provide a few examples of accelerated machine learning with existing models. Then, we propose a full investigation of the HPCs so that we understand the benefits and limitations of this paradigm. In particular, we aim to exploit the GPUs, and potentially FPGAs/ASICs (if present) at the HPCs within the LHC reconstruction workflow under the as-a-service paradigm. In the last two quarters, we hope the lab becomes a well established workbench and we expect the postdoc to work with various teams who are working on ML based algorithms. To give a specific example, the postdoc will work with the ongoing HGCal ML reconstruction effort to benchmark latency and throughput. While the normal operation of the accelerators is occurring, the postdoc will also work on an ML based calorimeter reconstruction algorithm, and ensure that Apache IPC is available within the software stack.

Lastly, the postdoc will ensure that any other accelerator developments including FPGAs and ASICs (such as the Intel Habana chip) are benchmarked. Explicitly, the postdoc will document the end-to-end procedure for how to take an existing algorithm, run it on a Xilinx Alveo FPGA, and then establish this algorithm within the CMSSW workflow.

Quarter of Year	Milestone	Deliverables
1	Integration of existing ML algorithms to the SONIC accelerator setup, testing of latency and throughput and fallback producer with GPUs	Postdoc will integrate gRPC and related packages(tensorrt-inference-server...) as a official supported CMS externals. Additionally postdoc will integrate current NN b-tagging, substructure tagging (deep-AK8) workflows into the SONIC environment, and test reconstruction with GPU. The postdoc will integrate a fallback producer framework (when GPU/... is not available) and develop a series of tests to ensure robust production. Testing will be performed at the FML GPU accelerator lab and in the cloud.
2	Investigation of GPUs in HPC centers for offline production.	Postdoc will take existing GPU accelerator infrastructure and understand the pros and cons of using HPCs as a way to accelerate the reconstruction. The postdoc will incorporate HPC tools within the software stack.
3	Incorporation of first HGCal models within accelerator framework. Testing of Apache IPC tools. Work on clustering algorithms.	Postdoc will work as a contact for the integration of new algorithms. In particular, the postdoc will work with the existing HGCal effort to benchmark reconstruction. Postdoc will additionally develop infrastructure for Apache IPC. The postdoc will additionally help with Calorimeter (non-HGCAL) ML based clustering effort.
4	Testing of and inclusion of FPGAs/ASICs to software stack. Potential testing of accelerated simulation.	Provided funding for GEANT based accelerator studies (K. Pedro's proposal), the postdoc will work with the team to benchmark speed ups with GPUs both locally and within the HPCs. The postdoc will work with the team to further integrate FPGA accelerators into the Fast Machine Learning Lab. A survey of cloud-based ASICs will be performed. Full documentation and presentation at the Fast Machine Learning conference.

Table 1: Quarter by Quarter outline of proposal

4 Mentoring plan

The PIs maintain regular group meetings with everyone involved in this effort. The meetings are currently held every Friday at 10 am central US time and include contributors from the ATLAS collaboration, industry partners, neutrino physicists, and astrophysicists. The postdoc is expected to attend these meetings and give regular updates on project status. While such meetings mostly focus on technical aspects, the PIs will hold separate, regular meetings with the postdoc focused on career planning questions and actions. Following MIT official university procedures, we maintain a yearly review system for all postdocs, a formal process that assesses progress toward the postdoc's career goals, their impact in research being carried out; it also

sets the goals for the year to come. Such a review will largely build on our frequent meetings with the postdoc over the year. As part of that, we will work with the postdoc in identifying key conferences and meetings where to participate and ensure the postdoc receives excellent visibility both locally and across the scientific community in the US and internationally. We will aim in preparing them accordingly for them to be in a position to receive several invitations for talks in academia and industry. We will provide them with opportunities to participate in and lead public outreach events. The project involves state of the art understanding of ML and heterogeneous computing, which will require active collaboration with industry, and HPC centers. The PIs have already established relationships with Xilinx, NVIDIA, Microsoft, Intel, and academic computing experts. The postdoc will be introduced to this network and will contact them when appropriate.

5 Conclusions

With the upgrade to the High Luminosity LHC, the demands for offline computing resources in CMS will grow immensely [18, 19, 20]. Both additional computing resources and new ideas for how to process the data are needed to ensure that CMS can continue to produce high quality scientific results. In this proposal, we ask for 50% remuneration for a postdoc to facilitate a heterogeneous acceleration lab to integrate heterogeneous computing, specifically accelerated deep learning, into the CMS offline production workflow. Having already taken several initial steps and with this additional support, our team can turn heterogeneous computing into a major research effort within CMS.

A References

[1] CMS Collaboration, "Machine learning-based identification of highly Lorentz-boosted hadronically decaying particles at the CMS experiment", Technical Report CMS-PAS-JME-18-002, CERN, Geneva, 2019.

[2] CMS Collaboration, "A deep neural network to search for new long-lived particles decaying to jets", arXiv:1912.12238.

[3] CMS Collaboration, "A deep neural network for simultaneous estimation of b jet energy and resolution", arXiv:1912.06046.

[4] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", *JINST* **13** (2018), no. 05, P05011, doi:10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.

[5] S. Farrell et al., "Novel deep learning methods for track reconstruction", in *4th International Workshop Connecting The Dots 2018 (CTD2018) Seattle, Washington, USA, March 20-22, 2018*. 2018. arXiv:1810.06111.

[6] R. Aaij et al., "Allen: A high level trigger on GPUs for LHCb", arXiv:1912.09161.

[7] A. Bocci, "Heterogeneous online reconstruction at CMS", 2019. https://indico.cern.ch/event/773049/contributions/3474336/.

[8] "Fast Machine Learning Lab", 2020. https://fastmachinelearning.org/.

[9] B. Holzman et al., "HEPCloud, a new paradigm for HEP facilities: CMS Amazon Web Services investigation", *Comput. Soft. Big Sci.* **1** (Sep, 2017) doi:10.1007/s41781-017-0001-9.

[10] J. Duarte et al., "FPGA-accelerated machine learning inference as a service for particle physics computing", *Comput. Softw. Big Sci.* **3** (2019), no. 1, 13, doi:10.1007/s41781-019-0027-2, arXiv:1904.08986.

[11] P. Harris, "ML acceleration with heterogeneous computing for big data physics experiments", 2019. https://sc19.supercomputing.org/presentation/?id=pec408&sess=sess110.

[12] J. Duarte et al., "Fast inference of deep neural networks in FPGAs for particle physics", *JINST* **13** (2018), no. 07, P07027, doi:10.1088/1748-0221/13/07/P07027, arXiv:1804.06913.

[13] N. Tarafdar et al., "Galapagos: A full stack approach to FPGA integration in the cloud", *IEEE Micro* **38** (2018), no. 6, 18–24, doi:10.1109/MM.2018.2877290.

[14] L. Agostino et al., "Commissioning of the CMS high level trigger", *Journal of Instrumentation* **4** (Oct, 2009) P10005–P10005, doi:10.1088/1748-0221/4/10/p10005.

[15] C. T. Collaboration, "Stand-alone cosmic muon reconstruction before installation of the CMS silicon strip tracker", *Journal of Instrumentation* **4** (May, 2009) P05004–P05004, doi:10.1088/1748-0221/4/05/p05004.

[16] S. R. Qasim, J. Kieseler, Y. Iiyama, and M. Pierini, "Learning representations of irregular particle-detector geometry with distance-weighted graph networks", *Eur. Phys. J.* **C79** (2019), no. 7, 608, doi:10.1140/epjc/s10052-019-7113-9, arXiv:1902.07987.

[17] X. Ju et al., "Graph Neural Networks for Particle Reconstruction in High Energy Physics Detectors", in *ML4PS Workshop, NeurIPS*. 2019.

[18] J. Albrecht et al., "HEP Community White Paper on Software trigger and event reconstruction: Executive Summary", arXiv:1802.08640.

[19] HEP Software Foundation Collaboration, "HEP Software Foundation Community White Paper Working Group - Data Analysis and Interpretation", arXiv:1804.03983.

[20] HEP Software Foundation Collaboration, "A Roadmap for HEP Software and Computing R&D for the 2020s", *Comput. Softw. Big Sci.* **3** (2019), no. 1, 7, doi:10.1007/s41781-018-0018-8, arXiv:1712.06982.