SAND2020-1768C

ENABLING 5G DISTRIBUTED SCIENTIFIC DATA ANALYSIS THROUGH HIERARCHICAL ARCHITECTURES AND ALGORITHMS

ERIC C. CYR
CENTER FOR COMPUTING RESEARCH
COMPUTATIONAL MATHEMATICS DEPARTMENT
SANDIA NATIONAL LABORATORIES
ALBUQUERQUE, NM 87185-1320

The increased bandwidth and lower latencies associated with distributed 5G devices will enable scientific sensor and feedback networks that yield unprecedented data acquisition and control of experiments. In addition, to the 5G enabled devices, these networks will contain a hierarchy of computational devices, from edge compute to large scale HPC potentially coupled by line of sight fast 5G transfer. These computational resources present the intriguing possibility of achieving real time data analysis coupled to control of streaming data sources.

The edge computational resources will enable local fast (millisecond) analysis of streaming data using edge computing. This analysis is two-fold, first a data reduction algorithm consolidating multiple streaming sources to ship to larger HPC capabilities over the fast line-of-sight 5G wireless transfer for storage or further computation. The second task will incorporate newly acquired data into machine learning models that provide the opportunity for realtime local analysis and feedback for control of the end-point 5G devices. Multiple edge accelerators will be required to integrate and control the high spatial device density possible (a million per square kilometer!). The role of HPC computing in this architecture is then to integrate and coordinate the whole network using the edge devices as the mediator to the 5G end-points. The HPC resource is the only architecture to have the computational horsepower to deliver the type of analysis and control required for the whole network. Holistically, this multi-layered network will provide milli-to-second time scale robust local control and analysis through the edge, while global integration revolves around the HPC hub and is delivered on the time scales of seconds to tens of seconds.

This vision will require substantial algorithmic advances that will mirror the heterogeneous hierarchy of data sources and analysis capabilities. At the level of the edge, new algorithms capable of compressing multiple O(100Mbps) streams of data from the 5G devices will be required. Further consolidation, of these streams into a coherent local model will be required for effective control. All of this must occur quickly enough to take advantage of the short latencies enjoyed by the 5G network. This suggest that previously trained machine learning algorithms will be used to guarantee performance. A similar problem exits at the HPC level, where multiple faster streams on the order of 1 Gbps must be integrated and analyzed. However, at this level the resources will be substantial enough to improve on the machine learning model through additional training.

One approach to doing this is to have the HPC resource train a hierarchy of models. This has two benefits, first the high resolution machine learning model on the HPC machine can be accelerated by using parallel multigrid algorithms that achieve scalability through a coarse (or reduced) problem. Second, the reduced problem itself could be directly utilized at the edge thus using the multigrid algorithm to naturally integrate all the disparate streams of data. The groundwork for this kind of algorithmic architecture has already been laid in the machine learning community in the form of "cascade learning" [1, 2] and in the multigrid community in the form of layer-parallel training [3]. Further innovations arising in the computational neuroscience community like "local learning" [4] may enable incorporation of new data directly on the edge devices, with transfers back to the HPC platforms providing a feedback loop to realize global integration of recent data throughout the network.

Acknowledgment. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

[1] S.E. Fahlman and C. Lebiere. The cascade-correlation learning architecture. In *Advances in neural information processing systems*, pages 524–532, 1990.

[2] E.S. Marquez, J.S. Hare, and M. Niranjan. Deep cascade learning. *IEEE transactions on neural networks and learning systems*, 29(11):5475–5485, 2018.

[3] S. Günther, L. Ruthotto, J.B. Schroder, E.C. Cyr, and N.R. Gauger. Layer-parallel training of deep residual neural networks. *arXiv preprint arXiv:1812.04352, Accepted to SIMODs*, 2019.

[4] D. Krotov and J.J. Hopfield. Unsupervised learning by competing hidden units. *Proceedings of the National Academy of Sciences*, 116(16):7723–7731, 2019.