HPC and Big Data Convergence for Extreme Heterogeneous Systems

Andrew J. Younge[1], Shantenu Jha[2,3], and Geoffrey C. Fox[4]
[1]Sandia National Laboratories *
[2]Brookhaven National Laboratory
[3]Rutgers University
[4]Indiana University
POC: ajyoung@sandia.gov

***Abstract*—As the data deluge grows ever greater, large-scale data analytics workloads are quickly becoming critical computational tools within the scientific community. Recently, convergence efforts have focused on combining aspects HPC and "big data" analytics workloads together using a unified supercomputing system. This has the opportunity to bring advanced analytical tools to scientists which enable extreme data processing capabilities, some of which are complimentary to analyzing large HPC simulations themselves. Big data stacks often provide full-featured computational frameworks that can potentially lower the barrier of entry for scientists and make advanced computational tasks more amenable to a wider audience. As supercomputing and HPC resources represent a significant investment for the DOE, it is imperative to ensure such leadership computing capabilities will be most effectively utilized for advancing both simulations and analytics workloads.**

However, current big data stacks often struggle to maintain any reasonable performance on today's supercomputing systems. Far too often, system software from each ecosystem is incompatible and there is no conjoined effort for providing holistic resource management or interoperability. The gap between simulation and data analytics workload performance on supercomputers will only grow exponentially with the extreme heterogeneity expected in future supercomputing systems, unless action is taken soon.

I. Key Challenges

In just the past few years, large-scale data analytics tools and big data stacks have shown to be critical new tools for gaining new scientific knowledge and enabling novel discoveries derived from the deluge of data. This includes analytics operations not only as a processing component to large HPC simulations, but also as standalone scientific tools for knowledge discovery [1]. Many have predicted we are at the verge of convergence within scientific computing between High Performance Computing (HPC) and big data workloads [2]. However, reality has seen a different effect, with big data stacks still struggling with even basic usability on some of the DOE's largest supercomputing resources. Of the successes thus far, they are often accompanied with significant caveats and sub-optimal configurations that are unable to leverage significant advancements that are technically feasible with traditional MPI workloads [3]. This problem will be exacerbated as the supercomputing community embraces more heterogeneous systems, placing additional requirements on data science platform efforts.

The gap in performance and scalability on HPC systems with data science tools is not for a lack of effort. Instead, the challenges may be far more convoluted than many originally anticipated. For instance, the use of virtualization and container solutions within HPC have provided a first foothold for non-HPC software stacks, such as the with Apache frameworks. While container and virtualization technologies have enabled the use of some workloads on the Cori supercomputer [4], significant scalability and performance issues still exist. For instance, there are limitations with big data workloads to efficiently manage resources or leverage HPC storage systems. Big data software stacks assume they have a different level of control over their operating environment, often utilizing unique resource management tools such as YARN or Apache Mesos. While useful within clouds, these tools have no mechanisms to interact with existing HPC resource management tools, job schedulers, or batch queuing systems. As resource management becomes an evermore critical component as heterogeneous supercomputing environments are developed, this disconnect will need to be addressed.

Another example of the disconnect between HPC and big data analytics stacks include the interconnect utilization. As most big data stacks operate on commodity cloud system which utilize TCP/IP, their utility within a supercomputing environment today is often limited. Current HPC interconnects theoretically offer far more bandwidth and lower latencies, sometimes up to an order of magnitude improvements. However, these same HPC interconnects lose much of this advantage when TCP/IP is emulated or abstracted, leaving speedups for data analytics stacks on HPC resources limited. New HPC interconnect technologies spend considerable resources optimizing for message based communications with customized MPI implementations and tuned transport protocols. Yet such functionality is usually incompatible with big data stacks entirely, and little effort beyond basic functionality is provided. As interconnect heterogeneity increases, this problem continues to grow. When HPC-oriented big data stacks have re-written large components to leverage HPC interconnects directly (such as InfiniBand and Hadoop [5]), any notion of performance portability is lost when trying to utilize another

*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 Energys National Nuclear Security Administration under contract DE-NA0003525.

interconnect and a new implementation must be developed from scratch. HPC apps do not suffer form such limitations as they rely on vendor and laboratory investments in MPI itself.

The issue where HPC hardware heterogeneity is increasing the complexity for ensuring efficient operation of big data workloads can also be seen with memory and storage systems. Recent simulations have found considerable benefit to deeper levels of memory. This includes the use of High Bandwidth Memory (HBM) and some supercomputing systems now even include node-local storage (such as the Summit/Sierra systems). As new tools and system software mechanisms are designed to help HPC simulations better cope with such memory diversity, little thought is currently being given to the potential benefits or requirements of analytics platforms. As new system software and resource management tools emerge to aid current HPC applications in embracing hardware diversity, they too must also consider interoperability and usability within existing big data stack frameworks. The prudent, if not only approach, is to (re-)design the software stack for simulation *and* analytics capabilities.

II. Research Direction

The convergence between HPC and big data has seen some progress from both respective fields. There are efforts that look to leverage the advancements provided by HPC resources to provide improved performance for big data analytics workloads. For instance, the High Performance Computing Enhanced Big Data Stack (HPC-ABDS) [6] examines the addition of high performance runtime and components to Apache systems, which have highlighted the importance of the big data systems associated with Apache Foundation, such as Hbase, Hadoop, Spark, Storm, etc. This is termed the Apache Big Data Stack (ABDS), even though important components such as MongoDB and Tensorflow are not Apache projects. Most of these technologies are in principle usable on both HPC and Cloud IaaS, though in practice many challenges remain. Independent of the hardware, there are stronger forces driving the adoption of ABDS technologies. They offer usability, functionality, and sustainability that are unavailable in HPC.

Such analytics frameworks are in some ways already accustomed to embracing heterogeneity. This can be seen in the diversity of computing resources within public cloud computing offerings in which such frameworks typically operate. Instance types large and small are often used together (driven by pricing demands). Public clouds have even offered GPUs, FPGAs, and other accelerators recently, indicating significant utility to analytics and machine learning applications.

Initial convergence efforts have also been made within the supercomputing community to better support analytics workloads. Virtualization and containerization efforts have enabled some analytics workloads to run atop current supercomputing systems. Containerization efforts on HPC, such as Shifter [7] have enabled Apache Spark workloads to leverage large allocations of the Cori supercomputer at NERSC. Host-level virtualization and the potential for virtual clusters on supercomputing resources also enables similar successes [8]. However, virtual clusters and containers have technological limitations to effectively utilizing heterogeneous resources.

Further research can determine how heterogeneous resources can be utilized efficiently in a way that also enhances data analytics. Meta scheduling and resource management tools are needed that transcend implementations from either HPC batch schedulers or Apache resource management tools, yet enable such sub-systems to operate within an infrastructure in tandem. A set of common, open-source utility libraries must be created for major classes of computational resources, including GPUs, interconnects, memory systems, or others. These system libraries should focus first on interoperability and ABI comparability. Moreover, OS and virtualization research is still needed to provide efficient access for portable ecosystems on all hardware without constraining performance.

III. State of the Art

Given the potential for data analytics tools to fundamentally change the mechanisms for scientific discovery, convergence between HPC and big data is underway to better leverage the investments of the HPC and supercomputing communities. This includes emerging virtual cluster and container-based system software mechanisms along with HPC-ABDS. When considering the inevitability of heterogeneity within HPC, **now is the time** to act to intercept such big data convergence with fundamental distributed system software research to enable a holistic supercomputing design. Big data stacks have shown to be a **mature** tool of key importance to knowledge discovery. While contributions in system software toward enabling effective utilization of heterogeneous resources may not be new for HPC, the additional requirements of big data analytics workloads presents a **unique** consideration to be addressed. Succeeding in such endeavors will result in a **novel** systems architecture that considers the requirements of both big data and HPC workloads in systems of the future that are both performant and heterogeneous in nature. Such an architecture will likely demonstrate a significant scientific impact to all workloads, as well as ensure the viability of supercomputing throughout the next decade and beyond.

References

[1] R. Leland, R. Murphy, B. Hendrickson, K. Yelick, J. Johnson, and J. Berry, "Large-Scale Data Analytics and Its Relationship to Simulation," Sandia National Laboratories, Tech. Rep., 2016.

[2] D. A. Reed and J. Dongarra, "Exascale computing and big data," *Communications of the ACM*, vol. 58, no. 7, pp. 56–68, 2015.

[3] S. Kamburugamuve, P. Wickramasinghe, S. Ekanayake, and G. C. Fox, "Anatomy of machine learning algorithm implementations in MPI, Spark, and Flink," in *The International Journal of High Performance Computing Applications*, 2017.

[4] N. Chaimov, A. Malony, S. Canon, C. Iancu, K. Z. Ibrahim, and J. Srinivasan, "Scaling spark on hpc systems," in *Proceedings of the 25th ACM International Symposium on High-Performance Parallel and Distributed Computing*. ACM, 2016, pp. 97–110.

[5] X. Lu, N. S. Islam, M. Wasi-Ur-Rahman, J. Jose, H. Subramoni, H. Wang, and D. K. Panda, "High-performance design of hadoop rpc with rdma over infiniband," in *Parallel Processing (ICPP), 2013 42nd International Conference on*. IEEE, 2013, pp. 641–650.

[6] J. Qiu, S. Jha, A. Luckow, and G. C. Fox, "Towards hpc-abds: An initial high-performance big data stack," *Building Robust Big Data Ecosystem ISO/IEC JTC 1 Study Group on Big Data*, pp. 18–21, 2014.

[7] D. M. Jacobsen and R. S. Canon, "Contain this, unleashing docker for hpc," *Proceedings of the Cray User Group*, 2015.

[8] A. J. Younge, K. Pedretti, R. E. Grant, and R. Brightwell, "Enabling Diverse Software Stacks on Supercomputers using High Performance Virtual Clusters," in *IEEE Cluster*, 2017.