Emerging Computational Workloads Demand Virtual Clusters for Extreme-scale Systems

Andrew J. Younge, Kevin Pedretti, and Ron Brightwell
Sandia National Laboratories
P.O. Box 5800, MS-1319
Albuquerque, NM 87185-1110
Email: {ajyoung,ktpredre,rbbrigh}@sandia.gov

***Abstract*—Large Scale Data Analytics workloads are gaining attention within the scientific community not only as a processing component to large HPC simulations, but also as standalone scientific tools for knowledge discovery. However, system software for such capabilities on the latest extreme-scale DOE supercomputing systems has failed to appropriately support these types of application ecosystems. Static batch job schedulers with vendor-specific OS and library services are tuned only for MPI-based execution models. Instead of adapting analytics workloads to such a specific software environment, we need new mechanisms to provision Virtual Clusters on advanced supercomputing resources to enable diverse software ecosystem support at scale. Emerging analytics workloads will not only be able to create user-defined environments suitable to their computational needs independent of facilities software stacks, but also leverage advanced HPC hardware resources often unavailable for use today.**

I. Extended Abstract

Currently, we are at the forefront of a convergence within scientific computing between High Performance Computing (HPC) and Large Scale Data Analytics (LSDA) [1], [2]. This amalgamation of differing viewpoints in distributed systems looks to force the combination of performance characteristics of HPC's pursuit towards Exascale with data and programmer oriented concurrency models found in Big Data analytics platforms. Capitalizing upon the community's existing intellectual investments in advanced supercomputing systems and leveraging the economies of scale in hardware infrastructure could benefit far more computational methods beyond what is possible as disjoint environments. Current software efforts in each area have become specialized with the gap growing rapidly, making concurrent ecosystem support within a single architectural system increasingly intractable.

Instead, we postulate embracing this software diversity on advanced supercomputing platforms through the use of Virtual Clusters. Virtual Clusters enable disjoint software ecosystems consisting of many underlying node deployments, to provision and operate independent software stacks deployed concurrently on extreme-scale distributed memory systems. This allows for the classic HPC system software stacks to continue to operate in the same environment on the same hardware, yet also enable emerging analytics and visualization workloads such as those in the Apache Big Data Stack [3], among others to deploy their custom software ecosystems specific to application needs, rather than site-specific software. Furthermore, such virtualized clusters enable new methods for non-standard workflow composition, such as the in-situ coupling of parallel MPI simulations with emerging data analytics and visualization tools for real-time experimental control, either across or within clusters.

We expect virtualization to be a key aspect to providing Virtual Clusters, however the type and level of virtualization and the interactions with the underlying OS & runtime environment are still an unknown. Work is needed to determine the most effective way to provide the level of system abstraction, and what trade-offs are necessary in regards to performance considerations, cluster deployment efficiency, OS type and flexibility, workload reproducibility, and advanced hardware accessibility. Host-level virtualization efforts as the Hobbes project [4] is one example of a conjoined OS and virtualization effort that provides building blocks necessary for constructing Virtual Clusters. OS-level virtualization, or container solutions such as Shifter [5] and Singularity [6] also offer a new potential component by extending the notion of Docker and integrate within existing HPC environments. While these various virtual abstractions as well as bare-metal OS provisioning [**?**] are important building blocks for Virtual Clusters within advanced technology systems, we still require extensive distributed systems software research and development to overcome the issues brought by batch processing mechanisms. This includes challenges including meta scheduling & resource management, user-defined image creation & orchestration, network segmentation & isolation, and performance considerations at extreme scale, to name a few. While the concept of Virtual Clusters originates from commodity cloud infrastructure development such tools like OpenStack [7], [8], their successful application within an advanced supercomputing infrastructure as described herein will likely require a new software architecture largely independent of existing cloud solutions.

Virtual Clusters enable users to focus on application ecosystem composition matching scientific endeavors rather than forcing new development environments to adapt to platforms that were never made to support such designs. Effectively, this can lower the barrier of entry to extreme-scale computing for many emerging computational tools embodied by the 4th paradigm of science [9]. Additionally, we can construct a framework of scientific experiment management where Virtual Clusters and their environments can be rebuilt, shared, rerun, or archived upon demand across the greater scientific community.

ACKNOWLEDGMENT

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energys National Nuclear Security Administration under contract DE-AC04-94AL85000.

REFERENCES

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

[2] 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.

[3] G. C. Fox, J. Qiu, S. Kamburugamuve, S. Jha, and A. Luckow, "Hpc-abds high performance computing enhanced apache big data stack," in *2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, May 2015, pp. 1057–1066.

[4] R. Brightwell, R. Oldfield, A. B. Maccabe, and D. E. Bernholdt, "Hobbes: Composition and virtualization as the foundations of an extreme-scale os/r," in *Proceedings of the 3rd International Workshop on Runtime and Operating Systems for Supercomputers*. ACM, 2013, p. 2.

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

[6] V. Sochat, "Singularity: Containers for Scientific Compute," Stanford University Presentation, February 2017.

[7] S. Telfer, "The Crossroads of Cloud and HPC: OpenStack for Scientific Research," Cambridge University, Tech. Rep., 2016.

[8] A. J. Younge, J. P. Walters, S. P. Crago, and G. C. Fox, "Supporting High Performance Molecular Dynamics in Virtualized Clusters Using IOMMU, SR-IOV, and GPUDirect," in *Proceedings of the 11th ACM SIGPLAN/SIGOPS Conference Virtual Execution Environments*. ACM, 2015.

[9] T. Hey, S. Tansley, K. M. Tolle *et al.*, *The fourth paradigm: data-intensive scientific discovery*. Microsoft research Redmond, WA, 2009, vol. 1.