# Large Scale Data Analytics Demand Virtual Clusters for HPC systems

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

***Abstract*—**

## 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) [**?**], [**?**]. This amalgamation of historically differing viewpoints of 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 investiment in advanced supercomputing systems and economiges of scale could benefit both areas beyond what is current possible as disjoint environments. However, current software efforts in each area have become extremely specialized and the gap only continues to grow, making the concurrent support with a single architectural model increasingly intractable.

Instead, we postulate the embrasing of software diversity on advanced supercomputing platforms through the use of Virtual Clusters. *TODO: Describe the notion of virtual clusters*

We expect virtualization to be a key aspect to providing Virtual Clusters, however the type and level of virtualization and its interactions with the underlying OS environment are still unknown. Work is need to determine the most effective way to provide this level of abstraction necessary, and what tradeoffs are necessary in regards to performance considerations, cluster deployment efficiency, OS type and flexibility, workload reproducability, hardware accessability, and others. Host virtualization efforts as the Hobbes project [**?**] provide one example of an OS and virtualization effort that coudl enable the underpinnings necessary for Virtual Clusters. OS-level virtualization, or container solutions such as Shifter and Singularity [**?**], [**?**] extend the notion of Docker towards an HPC environment by integrating within an existing HPC environment. With either solution, there is software ecosystem research and design that needs to support Virtual Clusters, including image management, network segmentation,

Virtual Clusters will enable focus more on application ecosystem compoistion to fit scientific products rather than adapting development environemtns into a running space that was never designed for such considerations. VCs can also

*TODO: In situ analysis, on-node privisoning with support for intra and inter VC coupling, etc*

Prpose virtual clusters to enable custom ecosystems independent of application types on the same hardware, rather than forcing one or the other ecosystems on applications. Still share the same hardware.

## 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.