**HPBDS: The Case for an Integrating Apache Big Data Stack with HPC**

*Shantenu Jha, Judy Qiu and Geoffrey Fox, Rutgers and Indiana University*

We analyze the ecosystems of the two prominent paradigms for data-intensive applications, referred to as the high-performance computing and the Apache-Hadoop paradigm. We propose a basis, common terminology and functional factors upon which to analyze the two approaches of both paradigms.

In related work, we have  proposed a set of BigData Ogres in analogy with parallel computing with the Berkeley Dwarfs, NAS benchmarks and linear algebra templates. The purpose of Big Data Ogres is to discern commonalities and patterns across a broad range of seemingly different Big Data applications, propose an initial structure to classify them, and help cluster some commonly found applications using the derived structure.  We propose that the set of Ogres will serve as a benchmark to evaluate the two paradigms along different dimensions.

The HPC stack is limited in its data-capability and somewhat rigid models of communication and data abstractions. From the other end, even though a vibrant ecosystem has evolved to support the Apache-Hadoop objective of providing affordable, scale-out storage and compute on commodity hardware, given the increasing processing/computational requirements, there is a need for convergence. We use a simple and broadly used Ogre (K-means clustering), characterize its performance on a range of representative platforms, covering several implementations from both paradigms.  Our experiments are performed on different infrastructure (clusters, HPC and EC2) using multiple technologies -- Mahout, MPI, Pig, Python Scripting, Spark, amongst others and are ongoing. Our experiments provide an insight into the relative strengths of the two paradigms.  Using both qualitative and quantitative analytical methods, we propose High-Performance Big Data Stack (HPBDS) as a  potential integration of different paradigms, bringing the best of both to the HPC and Big Data communities.