Contribution ID: **32** Type: **Workshop/BoF proposal**

Big Data Meets HPC: Exploiting HPC Technologies for Accelerating Big Data Processing and Management

Sunday, 3 December 2017 09:00 (20 minutes)

The convergence of HPC, Big Data, and Deep Learning is the next game-changing business opportunity. Apache Hadoop, Spark, gRPC/TensorFlow, and Memcached are becoming standard building blocks in handling Big Data oriented processing and mining. Recent studies have shown that default designs of these components can not efficiently leverage the features of modern HPC clusters, like Remote Direct Memory Access (RDMA) enabled high-performance interconnects, high-throughput parallel storage systems (e.g. Lustre), Non-Volatile Memory (NVM). In this tutorial, we will provide an in-depth overview of the architecture of Hadoop, Spark, gRPC/TensorFlow, and Memcached.

We will examine the challenges in re-designing networking and I/O components of these middleware with modern interconnects, protocols (such as InfiniBand, RoCE) and storage architectures. Using the publicly available software packages in the High-Performance Big Data project (HiBD, http://hibd.cse.ohio-state.edu), we will provide case studies of the new designs for several Hadoop/Spark/gRPC/TensorFlow/Memcached components and their associated benefits. Through these, we will also examine the interplay between high-performance interconnects, storage (HDD, NVM, and SSD), and multi-core platforms to achieve the best solutions for these components and applications on modern HPC clusters.

We also present in-depth case-studies with modern Deep Learning tools (e.g., Caffe, TensorFlow, BigDL) with RDMA-enabled Hadoop, Spark, and gRPC.

HPC content

The convergence of HPC, Big Data, and Deep Learning is the next game-changing business opportunity. Apache Hadoop, Spark, gRPC/TensorFlow, and Memcached are becoming standard building blocks in handling Big Data oriented processing and mining. Recent studies have shown that default designs of these components can not efficiently leverage the features of modern HPC clusters, like Remote Direct Memory Access (RDMA) enabled high-performance interconnects, high-throughput parallel storage systems (e.g. Lustre), Non-Volatile Memory (NVM). This tutorial will provide an in-depth overview of the architecture and accelerations of Hadoop, Spark, gRPC/TensorFlow, Memcached, and others.

Primary author: Dr LU, Xiaoyi (The Ohio State University)

Presenter: Dr LU, Xiaoyi (The Ohio State University)

Session Classification: Sunday Workshop: Big Data Meets HPC

Track Classification: Workshops