Advance Program for International Workshop on High-Performance Big Data Computing (HPBDC 2015)

In conjunction with ICDCS 2015

June 29, 2015

- **8:50-9:00 Opening Remarks**

- **9:00-10:00 Keynote I: Fusing HPC and Big Data - Experiences with Design, Deployment and Usage of The Wrangler System at the Texas Advanced Computing Center**
  - **Dan Stanzione (Texas Advanced Computing Center)**
  - **Abstract**: Traditional cluster computing has always targeted the problems of data that are very large, as some of the largest datasets ever produced have been generated by clustered supercomputers and their associated parallel filesystems. But recent developments in technology and research practice known as "Big Data" address a class of problems that don't just have very large datasets, but have I/O requirements that traditional clusters are ill-suited to address. The Wrangler project at the Texas Advanced Computing Center is an ongoing attempt to rethink cluster computing around the needs of Big Data problems. Wrangler features a unique NAND flash-based storage system from DSSD, providing up to 1 TB/s bandwidth and a very high transaction rate for random accesses as a shared resource across the cluster. Rather than achieving these performance targets through massive scale, Wrangler reaches these performance targets with less than 100 nodes. As important as performance is flexibility - data on Wrangler can be a traditional file system, object store, relational store, or even Hadoop file system. Wrangler is now operational, and this new architecture is setting new performance benchmarks for applications that never quite "fit" on traditional HPC systems. This talk will provide an overview of the Wrangler project, the evolution of big data at TACC that led us to the Wrangler architecture, and the experiences of early users through the first few months of operations. Wrangler is supported by a grant from the US National Science Foundation (NSF).
  - **Bio**: Dr. Dan Stanzione is the Executive Director of the Texas Advanced Computing Center (TACC) at The University of Texas at Austin. A nationally recognized leader in high performance computing, Stanzione has served as deputy director since June 2009 and assumed the new post July 1, 2014. He is the principal investigator (PI) for several leading projects including a multimillion-dollar National Science Foundation (NSF) grant to deploy and support TACC's Stampede supercomputer over four years. Stanzione is also the PI of TACC's upcoming Wrangler system, a supercomputer designed specifically for data-focused applications. He served for six years as the co-director of the iPlant Collaborative, a large-scale NSF life sciences cyberinfrastructure in which TACC is a major partner. In addition, Stanzione was a co-principal investigator for TACC's Ranger and Lonestar supercomputers, large-scale NSF systems previously deployed at UT Austin. Stanzione previously served as the founding director of the Fulton High Performance Computing Initiative at Arizona State University and served as an American Association for the Advancement of

Science Policy Fellow in the NSF's Division of Graduate Education. He has served as acting director of TACC since January. Stanzione received his bachelor's degree in electrical engineering and his master's degree and doctorate in computer engineering from Clemson University, where he later directed the supercomputing laboratory and served as an assistant research professor of electrical and computer engineering.

- **10:00-10:30 Coffee Break**

- **10:30-11:30 Session I: High-Performance Big Data Systems, Session Chair: Xiaoyi Lu (The Ohio State University)**
  - **A Scalable Distributed Private Stream Search System, Peng Zhang** (Institute of Information Engineering, Chinese Academy of Sciences; National Engineering Laboratory for Information Security Technologies), **Yan Li** (National Computer Network Emergency Response Technical Team Beijing, China), **Qingyun Liu** (Institute of Information Engineering, Chinese Academy of Sciences; National Engineering Laboratory for Information Security Technologies), and **Hailun Lin** (Institute of Information Engineering, Chinese Academy of Sciences)
  - **KTV-TREE: Interactive Top-K Aggregation on Dynamic Large Dataset in Cloud, Yuzhe Tang** (Syracuse University), **Ling Liu** (Georgia Tech), **Junichi Tatemura** (NEC Labs America), and **Hakan Hacigumus** (NEC Labs America)

- **11:30-12:15 Invited Talk I: Understanding Big Data Workloads on Modern Processors using BigDataBench**
  - **Jianfeng Zhan (Institute of Computing Technology, Chinese Academy of Sciences, China)**
  - **Abstract**: BigDataBench is an open-source big data benchmark suite, and the current version 3.1 includes diverse data sets and workloads from five application domains: search engine, social networks, e-commerce, multimedia analytics, and bioinformatics. This talk presents the workload characterization of BigDataBench on modern processors. We found big data workloads have several subclasses of workloads, and exhibit disparate behaviors, e.g. IPC and pipeline front end stall. Our correlation analysis indicated that even though a part of big data analytics workloads own notable pipeline front end stalls, the main factors affecting the CPI performance are long latency data accesses rather than high front end stalls. Also, our evaluation shows the wimpy-core processor does not suit big data analytics workloads in most situations
  - **Bio**: Dr. Jianfeng Zhan is a Full Professor and Deputy Director at Computer Systems Research Center, ICT, Chinese Academy of Sciences (CAS) and University of CAS. His research work is driven by interesting problems. He enjoys building new systems, and collaborating with researchers with different backgrounds. He founded the BPOE workshop, focusing on big data benchmarks, performance optimization and emerging hardware.

- **12:15-13:15 Lunch Break**

- **13:15-14:15 Keynote II: Efficiency + Scalability = High-Performance Big Data Computing**
  - **Zhiwei Xu (Institute of Computing Technology, Chinese Academy of Sciences, China)**
  - **Abstract**: We are entering a ZB data era, which needs scalable and efficient capabilities of sensing, communicating, and processing big data. In the past decade, great strides were made in scalability, where map-reduce based systems offer good examples. However, efficiency is still low for big data systems, at less than 1 Giga operation per Joule. In this talk, we first argue that the research community should set a bold efficiency goal of 1 Tera operation per Joule. We then present some encouraging initial results towards this objective from three directions of research: functional sensing, elastic processing, and high-performance data computing.
  - **Bio**: Dr. Zhiwei Xu is a Professor and CTO of the Institute of Computing Technology (ICT) of the Chinese Academy of Sciences. His prior industrial experience included chief engineer of Dawning Corp., a leading high-performance computer vendor in China. He currently leads "Cloud-Sea Computing Systems", a 10-year research project of the Chinese Academy of Sciences that aims at handling ZB of data with billion-thread computers and elastic processors. Dr. Zhiwei Xu holds a Ph.D. degree from the University of Southern California.

- **14:15-15:00 Invited Talk II: Benchmarking Big Data Systems**
  - **Raghunath Nambiar (Cisco)**
  - **Abstract**: Benchmarking standards matter for end-users, vendors and researchers, from fair comparisons of technologies and products to drive innovations. This session will cover some of the defining characteristics and recent developments in the area of performance evaluation and benchmarking of Big Data Systems.
  - **Bio**: Mr. Raghunath Nambiar is a distinguished engineer and chief architect of big data and analytics solutions. He is responsible for emerging technologies and data center solution strategy. He is the chairman of TPC Big Data standards committee and general chair of International Conference Series on Performance Evaluation and Benchmarking. More details about Mr. Raghunath Nambiar are available at http://blogs.cisco.com/author/raghunathnambiar.

- **15:00-15:30 Coffee Break**

- **15:30-16:30 Session II: Performance Studies of Big Data Systems and Applications, Session Chair: Xiaoyi Lu (The Ohio State University)**
  - **A Tiny GPU Cluster for Big Spatial Data: A Preliminary Performance Evaluation, Jianting Zhang** (Dept. of Computer Science, The City College of New York), **Simin You** (Dept. of Computer Science CUNY Graduate Center), and **Le Gruenwald** (Dept. of Computer Science, The University of Oklahoma Norman)
  - **Optimising Bootstrapping Algorithms using R and Hadoop, Shicai Wang** (Data Science Institute, Imperial College London, UK), **Mihaela A. Mares** (Data Science Institute, Imperial College London, UK), and **Yike Guo** (Data Science Institute, Imperial College London, UK; School of Computer Science, Shanghai University, China)

- **16:30-18:00 Panel: Wide Adoption of HPC Techniques in Big Data: Hype or Reality?**
  - **Panel Moderator: Jianfeng Zhan** (Institute of Computing Technology, Chinese Academy of Sciences, China)
  - **Panel Members:**
    - **D. K. Panda** (The Ohio State University)
    - **Dan Stanzione** (Texas Advanced Computing Center)
    - **Zhiwei Xu** (Institute of Computing Technology, Chinese Academy of Sciences, China)
    - **Xiaodong Zhang** (The Ohio State University)
  - **The panel will discuss on the following three important questions the Big Data and HPC communities are facing today:**
    - What is the precondition of wide adoption of HPC technologies in Big Data?
    - Do you have any prediction on merging HPC and Big Data technologies? When and how will this merge happen?
    - What are the differences between Big Data in HPC and Big Data in the other domains?

- **18:00-18:10 Closing Remarks**