Big Neuroscience Infrastructure

Dhabaleswar K. (DK) Panda, Khaled Hamidouche,
Xiaoyi Lu and Hari Subramoni

The Ohio State University
Department of Computer Science
{panda,hamidouc,luxi,Subramoni}@cse.ohio-state.edu

THE OHIO STATE UNIVERSITY

OSU Expertise

- Big Data processing (Hadoop, Spark, HBase, and Memcached)
- Scientific computing (MPI and PGAS)
- Scalable Graph processing
- Virtualization and cloud
- High Performance Deep Learning

OSU Infrastructure

- Software
  - HiBD Software (Hadoop, Spark, HBase, and Memcached)
  - MVAPICH Software
    - MPI and PGAS
    - Virtualization and HPC Cloud
  - High Performance Deep Learning
- Hardware
  - OSU InfiniBand Cluster
  - Chameleon Cloud Computing Testbed

The High-Performance Big Data (HiBD) Project

- http://hibd.cse.ohio-state.edu
- RDMA for Apache Spark
- RDMA for Apache Hadoop 2.x (RDMA-Hadoop-2.x)
  - Plugins for Apache, Hortonworks (HDP) and Cloudera (CDH) Hadoop distributions
- RDMA for Apache HBase
- RDMA for Memcached (RDMA-Memcached)
- RDMA for Apache Hadoop 1.x (RDMA-Hadoop)
- OSU HiBD-Benchmarks (OHB)
  - HDFS, Memcached, and HBase Micro-benchmarks
- Users Base: 190 organizations from 26 countries
- More than 17,800 downloads from the project site
- RDMA for Impala (upcoming)

Available for InfiniBand and RoCE

HiBD
High-Performance Big Data

Network Based Computing Laboratory

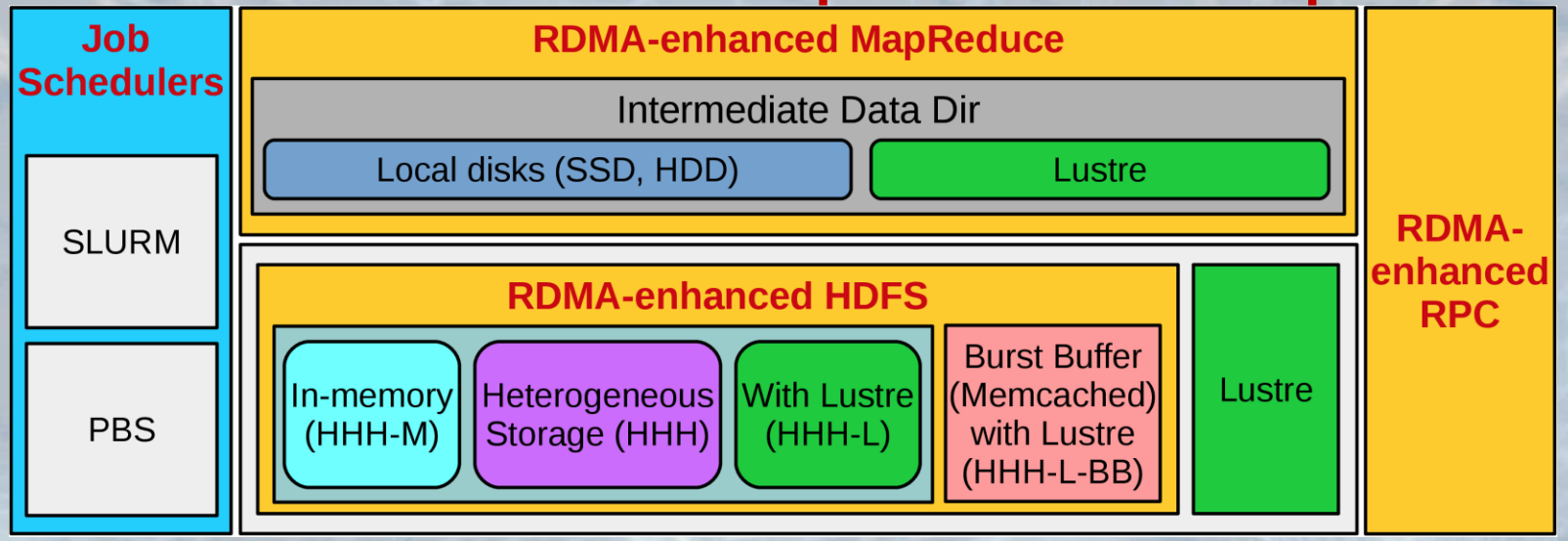
The Ohio State University

RDMA for Apache Hadoop 2.x Distribution

- High-Performance Design of Hadoop over RDMA-enabled Interconnects
  - High performance RDMA-enhanced design with native InfiniBand and RoCE support at the verbs-level for HDFS, MapReduce, and RPC components
  - Enhanced HDFS with in-memory and heterogeneous storage
  - High performance design of MapReduce over Lustre
  - Memcached-based burst buffer for MapReduce over Lustre-integrated HDFS (HHH-L-BB mode)
  - Plugin-based architecture supporting RDMA-based designs for Apache Hadoop, CDH and HDP
  - Easily configurable for different running modes (HHH, HHH-M, HHH-L, HHH-L-BB, and MapReduce over Lustre) and different protocols (native InfiniBand, RoCE, and IPoIB)
- Current release: 1.0.0
  - Based on Apache Hadoop 2.7.1
  - Compliant with Apache Hadoop 2.7.1, HDP 2.3.0.0 and CDH 5.6.0 APIs and applications
  - Tested with
    - Mellanox InfiniBand adapters (DDR, QDR, FDR, and EDR)
    - RoCE support with Mellanox adapters
    - Various multi-core platforms
    - Different file systems with disks and SSDs and Lustre

http://hibd.cse.ohio-state.edu

Different Modes of RDMA for Apache Hadoop 2.x



- **HHH**: Heterogeneous storage devices with hybrid replication schemes are supported in this mode of operation to have better fault-tolerance as well as performance. This mode is enabled by **default** in the package.
- **HHH-M**: A high-performance in-memory based setup has been introduced in this package that can be utilized to perform all I/O operations in-memory and obtain as much performance benefit as possible.
- **HHH-L**: With parallel file systems integrated, HHH-L mode can take advantage of the Lustre available in the cluster.
- **HHH-L-BB**: This mode deploys a Memcached-based burst buffer system to reduce the bandwidth bottleneck of shared file system access. The burst buffer design is hosted by Memcached servers, each of which has a local SSD.
- **MapReduce over Lustre, with/without local disks**: Besides, HDFS based solutions, this package also provides support to run MapReduce jobs on top of Lustre alone. Here, two different modes are introduced: with local disks and without local disks.
- **Running with Slurm and PBS**: Supports deploying RDMA for Apache Hadoop 2.x with Slurm and PBS in different running modes (HHH, HHH-M, HHH-L, and MapReduce over Lustre).

RDMA for Apache Spark Distribution

- High-Performance Design of Spark over RDMA-enabled Interconnects
  - High performance RDMA-enhanced design with native InfiniBand and RoCE support at the verbs-level for Spark
  - RDMA-based data shuffle and SEDA-based shuffle architecture
  - Non-blocking and chunk-based data transfer
  - Easily configurable for different protocols (native InfiniBand, RoCE, and IPoIB)
- Current release: 0.9.1
  - Based on Apache Spark 1.5.1
  - Tested with
    - Mellanox InfiniBand adapters (DDR, QDR and FDR)
    - RoCE support with Mellanox adapters
    - Various multi-core platforms
    - RAM disks, SSDs, and HDD
  - http://hibd.cse.ohio-state.edu

RDMA for Apache HBase Distribution

- High-Performance Design of HBase over RDMA-enabled Interconnects
  - High performance RDMA-enhanced design with native InfiniBand and RoCE support at the verbs-level for HBase
  - Compliant with Apache HBase 1.1.2 APIs and applications
  - On-demand connection setup
  - Easily configurable for different protocols (native InfiniBand, RoCE, and IPoIB)
- Current release: 0.9.1
  - Based on Apache HBase 1.1.2
  - Tested with
    - Mellanox InfiniBand adapters (DDR, QDR, FDR, and EDR)
    - RoCE support with Mellanox adapters
    - Various multi-core platforms
  - http://hibd.cse.ohio-state.edu

RDMA for Memcached Distribution

- High-Performance Design of Memcached over RDMA-enabled Interconnects
  - High performance RDMA-enhanced design with native InfiniBand and RoCE support at the verbs-level for Memcached and libMemcached components
  - High performance design of SSD-Assisted Hybrid Memory
  - Non-Blocking Libmemcached Set/Get API extensions
  - Support for burst-buffer mode in Lustre-integrated design of HDFS in RDMA for Apache Hadoop-2.x
  - Easily configurable for native InfiniBand, RoCE and the traditional sockets-based support (Ethernet and InfiniBand with IPoIB)
- Current release: 0.9.5
  - Based on Memcached 1.4.24 and libMemcached 1.0.18
  - Compliant with libMemcached APIs and applications
  - Tested with
    - Mellanox InfiniBand adapters (DDR, QDR, FDR, and EDR)
    - RoCE support with Mellanox adapters
    - Various multi-core platforms
    - SSD
  - http://hibd.cse.ohio-state.edu

HiBD Packages on SDSC Comet

- RDMA for Apache Hadoop 2.x and RDMA for Apache Spark are installed and available on SDSC Comet.
- Examples for various modes of usage are available in:
  - RDMA for Apache Hadoop 2.x: /share/apps/examples/HADOOP
  - RDMA for Apache Spark: /share/apps/examples/SPARK/
- Any user with XSEDE account on SDSC Comet should be able to use it
- Please email help@xsede.org (reference Comet as the machine, and SDSC as the site) if you have any further questions about usage and configuration.

M. Tatineni, X. Lu, D. J. Choi, A. Majumdar, and D. K. Panda, Experiences and Benefits of Running RDMA Hadoop and Spark on SDSC Comet, XSEDE'16, July 2016

OSU Infrastructure

- Software
  - HiBD Software (Hadoop, Spark, HBase, and Memcached)
  - MVAPICH Software
    - MPI and PGAS
    - Virtualization and HPC Cloud
  - High Performance Deep Learning
- Hardware
  - OSU InfiniBand Cluster
  - Chameleon Cloud Computing Testbed

Overview of the MVAPICH2 Project

- High Performance open-source MPI Library for InfiniBand, Omni-Path, Ethernet/iWARP, and RDMA over Converged Ethernet (RoCE)
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.0), Started in 2001, First version available in 2002
  - MVAPICH2-X (MPI + PGAS), Available since 2011
  - Support for GPGPUs (MVAPICH2-GDR) and MIC (MVAPICH2-MIC), Available since 2014
  - Support for Virtualization (MVAPICH2-Virt), Available since 2015
  - Support for Energy-Awareness (MVAPICH2-EA), Available since 2015
  - Support for InfiniBand Network Analysis and Monitoring (OSU INAM) since 2015
  - Used by more than 2,650 organizations in 81 countries
  - More than 389,000 (> 0.38 million) downloads from the OSU site directly
  - Empowering many TOP500 clusters (Jun '16 ranking)
    - 12th ranked 519,640-core cluster (Stampede) at TACC
    - 15th ranked 185,344-core cluster (Pleiades) at NASA
    - 31st ranked 76,032-core cluster (Tsubame 2.5) at Tokyo Institute of Technology and many others
  - Available with software stacks of many vendors and Linux Distros (RedHat and SuSE)
  - http://mvapich.cse.ohio-state.edu
- Empowering Top500 systems for over a decade
  - System-X from Virginia Tech (3rd in Nov 2003, 2,200 processors, 12.25 TFlops) ->
  - Stampede at TACC (12th in Jun'16, 462,462 cores, 5.168 Plops)

MVAPICH2 Software Family

High-Performance Parallel Programming Libraries	
MVAPICH2	Support for InfiniBand, Omni-Path, Ethernet/iWARP, and RoCE
MVAPICH2-X	Advanced MPI features, OSU INAM, PGAS (OpenSHMEM, UPC, UPC++, and CAF), and MPI+PGAS programming models with unified communication runtime
MVAPICH2-GDR	Optimized MPI for clusters with NVIDIA GPUs
MVAPICH2-Virt	High-performance and scalable MPI for hypervisor and container based HPC cloud
MVAPICH2-EA	Energy aware and High-performance MPI
MVAPICH2-MIC	Optimized MPI for clusters with Intel KNC
Microbenchmarks	
OMB	Microbenchmarks suite to evaluate MPI and PGAS (OpenSHMEM, UPC, and UPC++) libraries for CPUs and GPUs
Tools	
OSU INAM	Network monitoring, profiling, and analysis for clusters with MPI and scheduler integration
OEMT	Utility to measure the energy consumption of MPI applications

Availability of Systems with MVAPICH2 Libraries

- InfiniBand networking technology is very common these days
  - 41% of TOP500 supercomputers use InfiniBand
- Many InfiniBand clusters Nationally and Internationally
- Many of these clusters (including NSF XSEDE systems) use MVAPICH2
- Take a look at the User list of MVAPICH2 (http://mvapich.cse.ohio-state.edu/users) to see if some clusters in your organization use InfiniBand and MVAPICH2
  - List is based on voluntary registration

MVAPICH2-Virt 2.2rc1

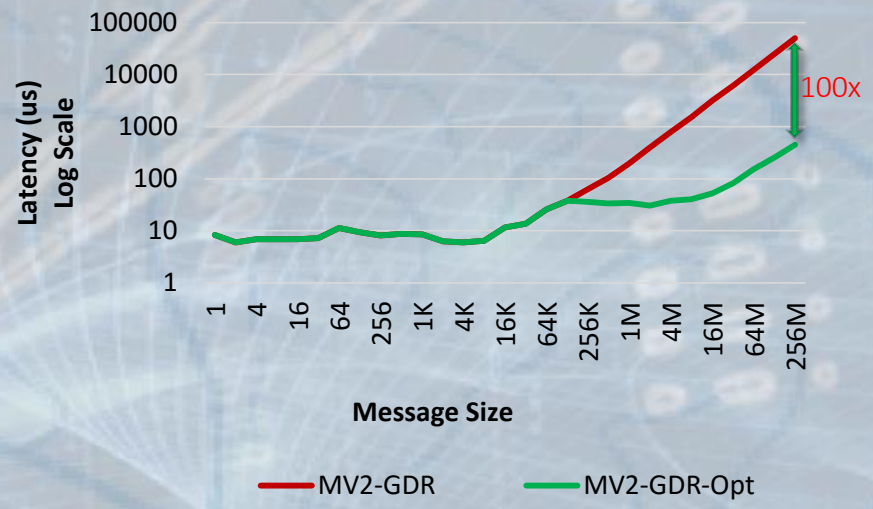
- Released on 07/12/2016
- Major Features and Enhancements
  - Based on MVAPICH2 2.2rc1
  - High-performance and locality-aware MPI communication with IPC-SHM and CMA for containers
  - Support for locality auto-detection in containers
  - Automatic communication channel selection among IPC-SHM, CMA, and HCA
  - Support for easy configuration through runtime parameters
  - Tested with
    - Docker 1.9.1 and 1.10.3
    - Mellanox InfiniBand adapters (ConnectX-3 (56Gbps))

OSU Infrastructure

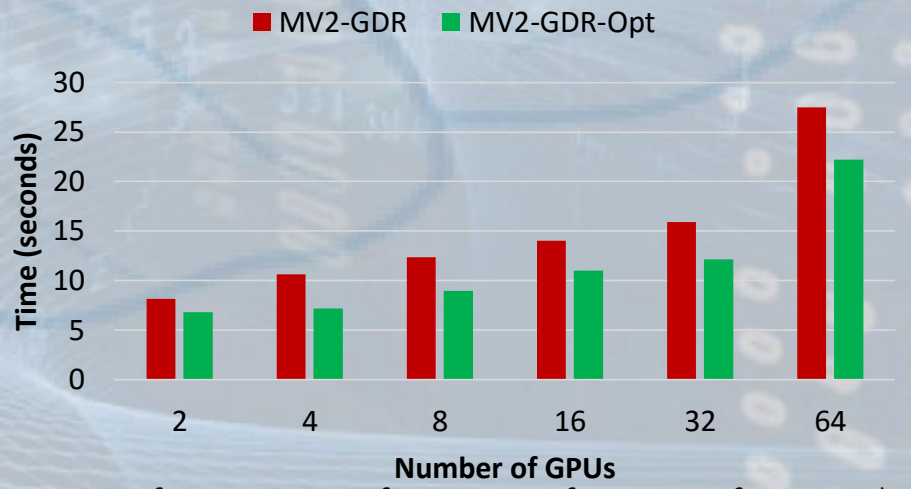
- Software
  - HiBD Software (Hadoop, Spark, HBase, and Memcached)
  - MVAPICH Software
    - MPI and PGAS
    - Virtualization and HPC Cloud
  - High Performance Deep Learning
- Hardware
  - OSU InfiniBand Cluster
  - Chameleon Cloud Computing Testbed

Deep Learning: Accelerating CNTK with MVAPICH2-GDR and NCCL

- NCCL has some limitations
  - Only works for a single node, thus, no scale-out on multiple nodes
  - Degradation across IOH (socket) for scale-up (within a node)
- We propose optimized MPI_Bcast
  - Communication of very large GPU buffers (order of megabytes)
  - Scale-out on large number of dense multi-GPU nodes
- Hierarchical Communication that efficiently exploits:
  - CUDA-Aware MPI_Bcast in MV2-GDR
  - NCCL Broadcast primitive



Performance Benefits: OSU Micro-benchmarks

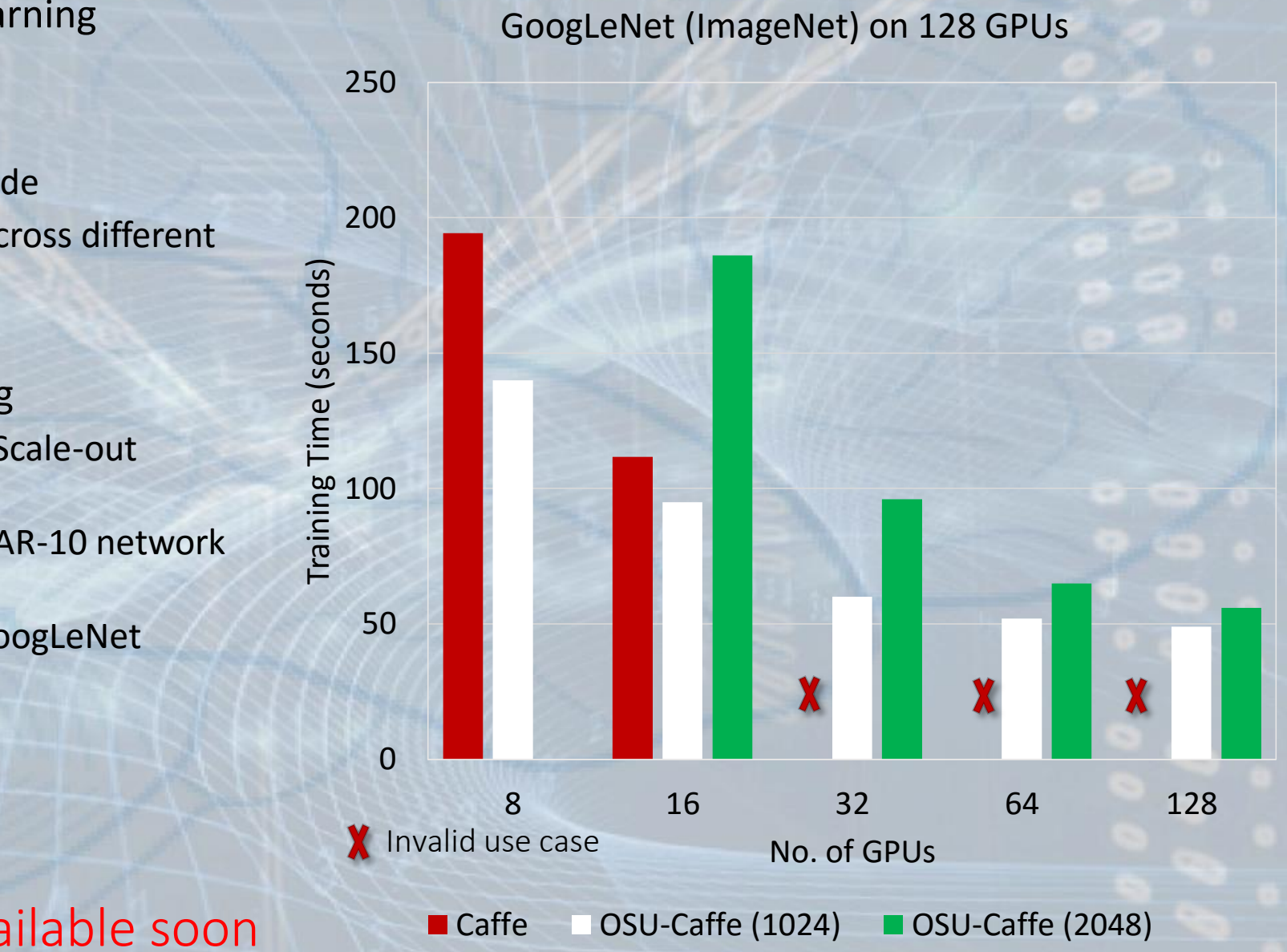


Performance Benefits: Microsoft CNTK DL framework
(25% avg. improvement)

Efficient Large Message Broadcast using NCCL and CUDA-Aware MPI for Deep Learning, A. Awan , K. Hamidouche , A. Venkatesh , and D. K. Panda, The 23rd European MPI Users' Group Meeting (EuroMPI 16), Sep 2016 [Best Paper Runner-Up]

OSU-Caffe: Scalable Deep Learning

- Caffe : A flexible and layered Deep Learning framework.
- Benefits and Weaknesses
  - Multi-GPU Training within a single node
  - Performance degradation for GPUs across different sockets
  - Limited Scale-out
- OSU-Caffe: MPI-based Parallel Training
  - Enable Scale-up (within a node) and Scale-out (across multi-GPU nodes)
  - Scale-out on 64 GPUs for training CIFAR-10 network on CIFAR-10 dataset
  - Scale-out on 128 GPUs for training GoogLeNet network on ImageNet dataset



OSU-Caffe will be publicly available soon

OSU Infrastructure

- Software
  - HiBD Software (Hadoop, Spark, HBase, and Memcached)
  - MVAPICH Software
    - MPI and PGAS
    - Virtualization and HPC Cloud
  - High Performance Deep Learning
- Hardware
  - OSU InfiniBand Cluster
  - Chameleon Cloud Computing Testbed

OSU InfiniBand Cluster

- NSF funded department-level cluster
- Supports faculty and students of the OSU CSE department
- 1,200 cores (Intel Broadwell)
- InfiniBand EDR (100 Gbps)
- Specialized nodes
  - Data-intensive computing (with SSDs and larger memory)
  - GPU nodes (NVIDIA K80s)
- We should be able to carry out initial experimentation on this cluster

NSF Chameleon Cloud: A Powerful and Flexible Experimental Instrument

- Large-scale instrument
  - Targeting Big Data, Big Compute, Big Instrument research
  - ~650 nodes (~14,500 cores), 5 PB disk over two sites, 2 sites connected with 100G network
  - Virtualization technology (e.g., SR-IOV, accelerators), systems, networking (InfiniBand), infrastructure-level resource management, etc.
- Reconfigurable instrument
  - Bare metal reconfiguration, operated as single instrument, graduated approach for ease-of-use
- Connected instrument
  - Workload and Trace Archive
  - Partnerships with production clouds: CERN, OSDC, Rackspace, Google, and others
  - Partnerships with users
- Complementary instrument
  - Complementing GENI, Grid'5000, and other testbeds
- Sustainable instrument
  - Industry connections

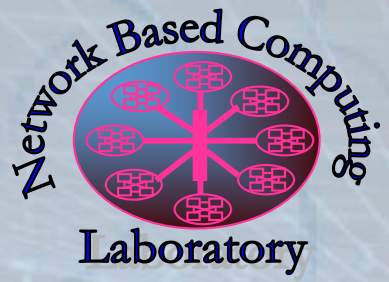
http://www.chameleoncloud.org/

Capabilities of the Chameleon Cloud

- Available in bare-metal form
  - Can be configured for different environments and workflows
- Several appliances already available from OSU
  - KVM SR-IOV
  - MVAPICh2-Virt
  - RDMA-Hadoop
  - Can be directly used
- Obtaining resources on this cluster is free
  - Requires a short (one paragraph) justification
- Had some discussion yesterday evening to explore this for putting together a `sandbox' consisting of all activities within ACNN
- Dr. Xiaoyi Lu (OSU) will provide a short demo today afternoon (scheduled for 1:40-2:10pm) about using the Chameleon Cloud

Thank You!

panda@cse.ohio-state.edu



Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

MVAPICH

The MVAPICH2 Project
http://mvapich.cse.ohio-state.edu/

HiBD
High-Performance
Big Data

The High-Performance Big Data Project
http://hibd.cse.ohio-state.edu/

Computational Neuroscience Network (ACNN)

http://www.NeuroscienceNetwork.org/ACNN_Workshop_2016.html