OSU Network Based Computing

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

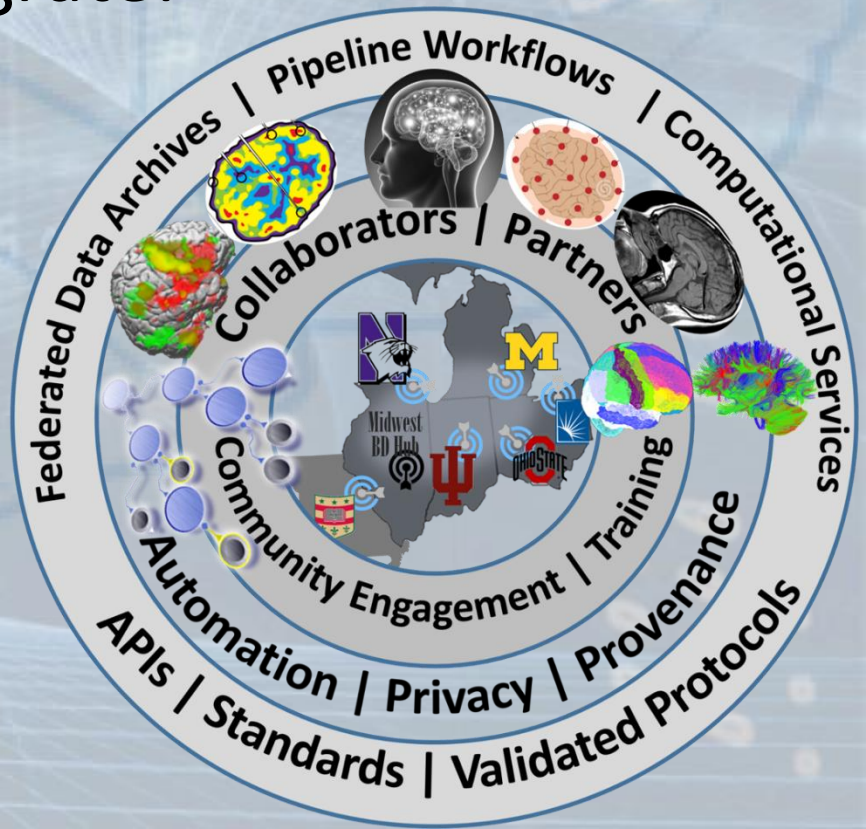
ACNN Objectives

1) Build foundations for modern, Big Data neuroscience technologies through community partnership
2) Leverage the expertise and technologies developed by the ACNN Spoke investigators and our partners to integrate:

1) **Data Sharing and Interoperability** using ontology-driven standardization, provenance metadata management, integrated into the most modern database and database-mediator technologies

2) **Analytics** leveraging upon the most agreed upon preprocessing pipelines (LONI and HCP) and advanced network science approaches to brain mapping

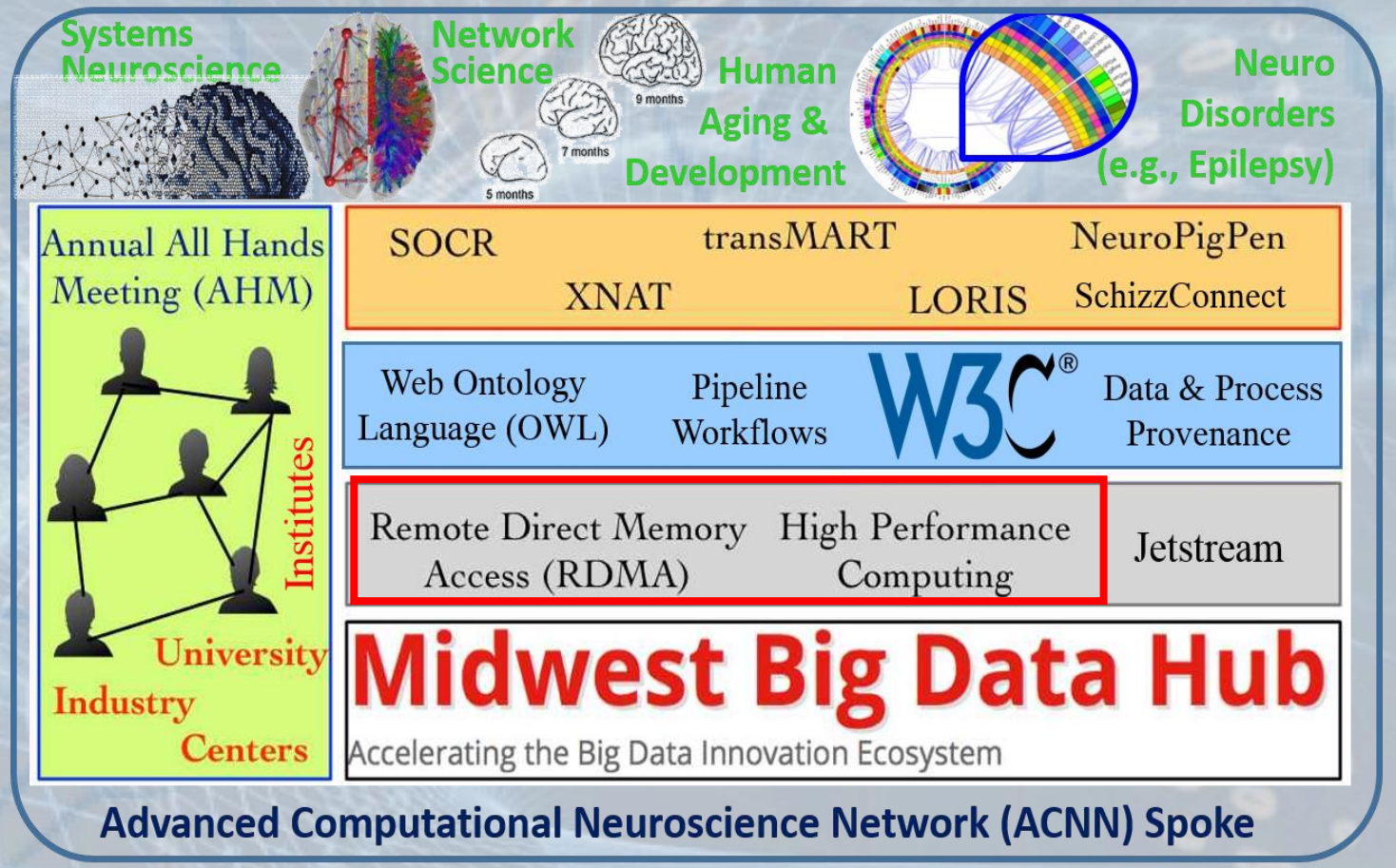
3) **Computing** approaches based on high performance clusters, MapReduce and Hadoop as well as canonical architectures will be deployed and connected to data and analytics



ACNN Technologies, Infrastructure, and Domain Applications

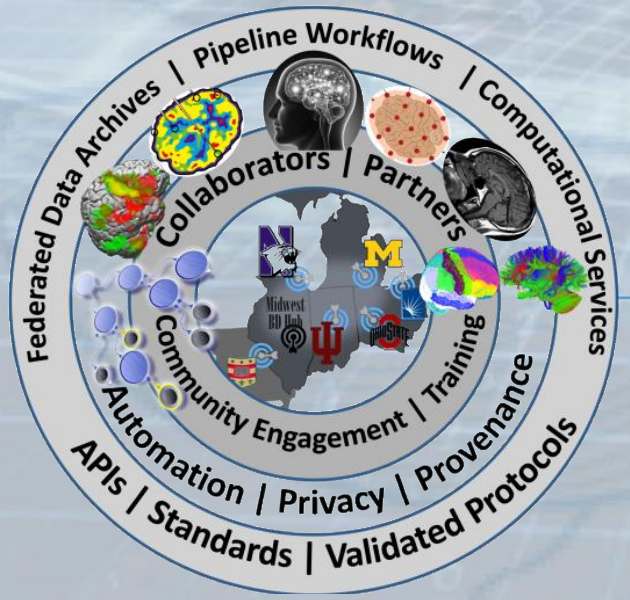
- Build a sustainable ecosystem of neuroscience community partners in both academia and industry using existing technologies for collaboration and virtual meeting together with face-to-face group meetings

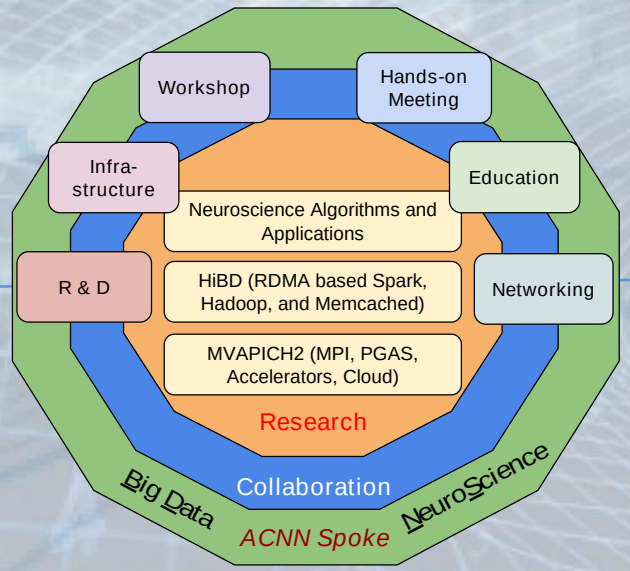
- OSU Role
  - Remote Direct Memory Access (RDMA) for Big Data
  - High-Performance Computing
  - Collaborations with other teams



Opportunities

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





OSU Team

- Strong expertise on HPC, Big Data, Deep Learning, Communication, I/O, etc.

Name	Role	Expertise
Dhabaleswar K. Panda	PI	High-Performance Computing, MPI, PGAS, High-Performance Networks
Khaled Hamidouche	Co-PI	Accelerators, NVDIA GPU, Intel MIC, Deep Learning
Xiaoyi Lu	Co-PI	Big Data Processing (Hadoop, Spark, HBase, Memcached) and Cloud Computing
Hari Subramoni	Co-PI	Communication and I/O, High-Performance Communication Protocols
Mark Arnold	Staff	System Management, Engineering

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)

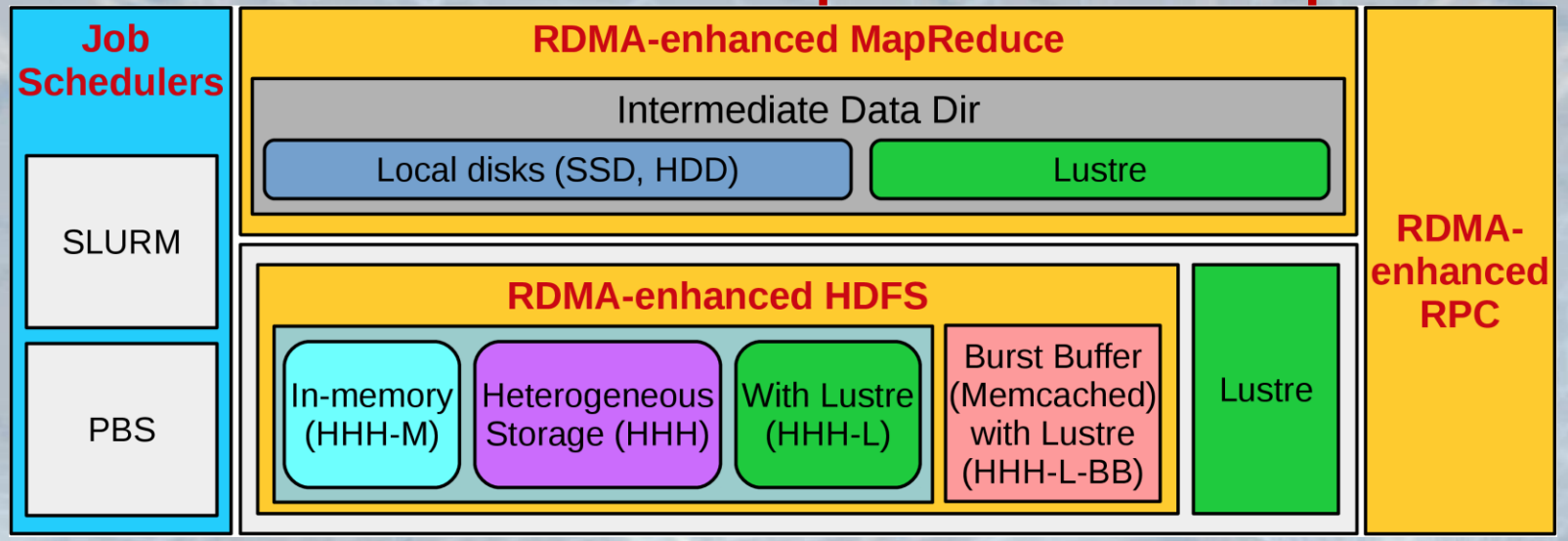
Available for InfiniBand and RoCE

HiBD
High-Performance Big Data

Network Based Computing Laboratory

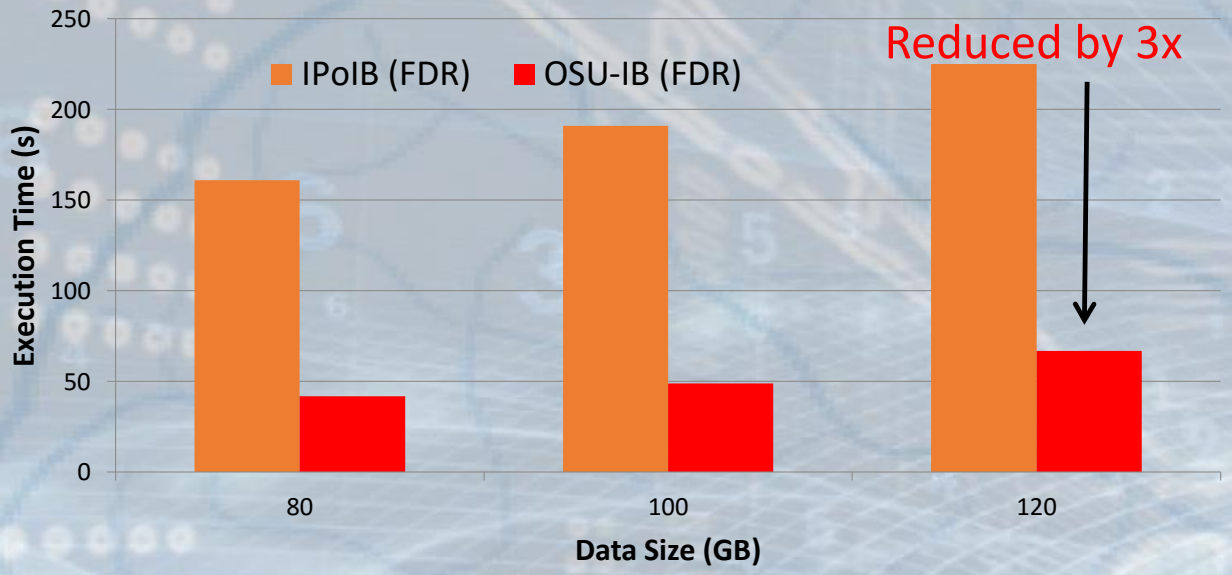
The Ohio State University

Different Modes of RDMA for Apache Hadoop 2.x

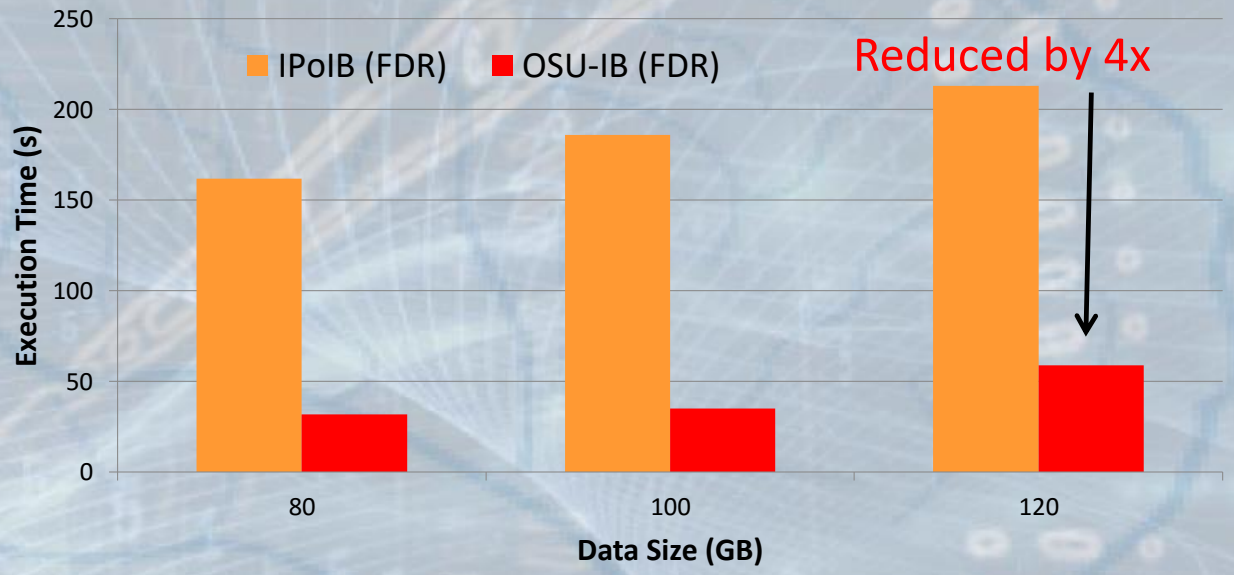


- **HHH**: Heterogeneous storage devices with hybrid replication schemes are supported in this mode of operation to have better fault-tolerance as well as performance. This mode is enabled by **default** in the package.
- **HHH-M**: A high-performance in-memory based setup has been introduced in this 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).

Performance Benefits – RandomWriter & TeraGen in TACC-Stampede



RandomWriter

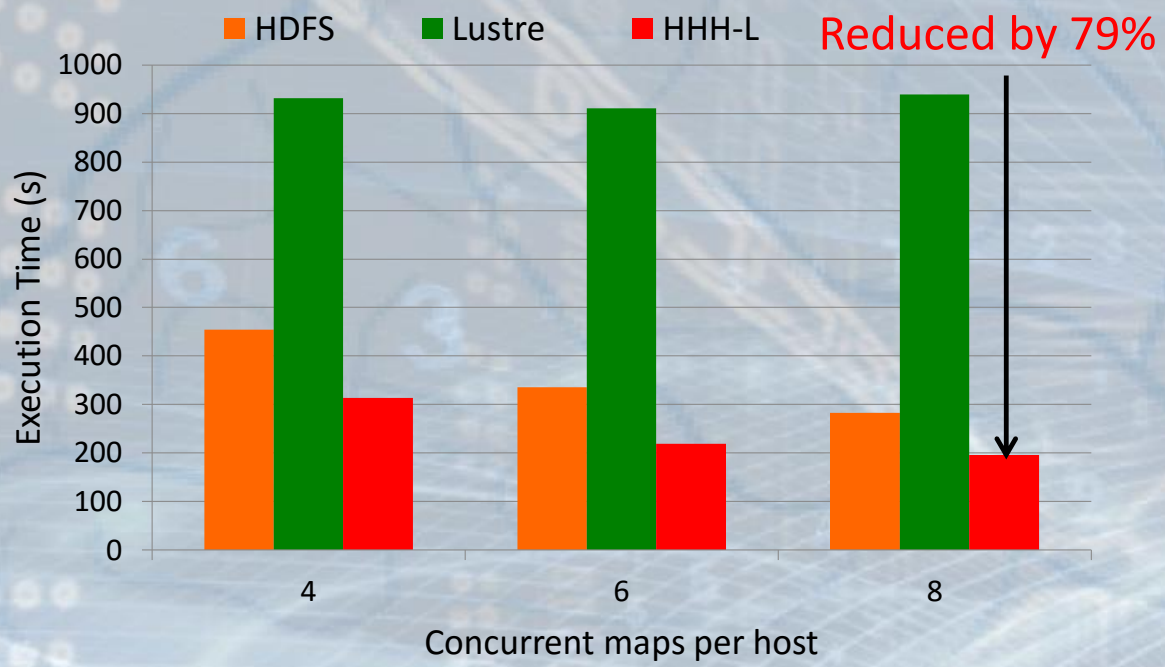


TeraGen

Cluster with 32 **Nodes with a total of** 128 **maps**

- RandomWriter
  - 3-4x improvement over IPoIB for 80-120 GB file size
- TeraGen
  - 4-5x improvement over IPoIB for 80-120 GB file size

Evaluation of HHH and HHH-L with Applications



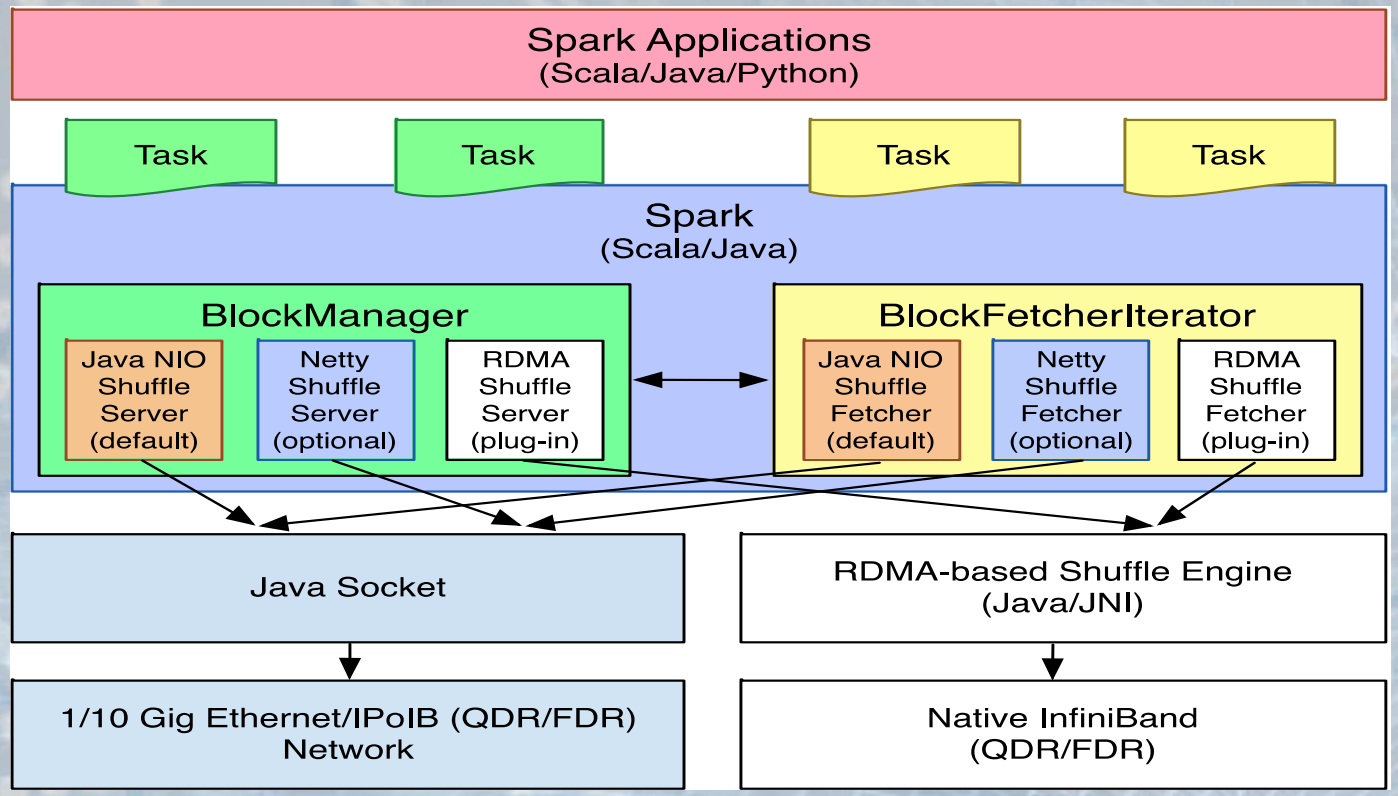
MR-MSPolyGraph

HDFS (FDR)	HHH (FDR)
60.24 s	48.3 s

CloudBurst

- MR-MSPolygraph on OSU RI with 1,000 maps
  - HHH-L reduces the execution time by **79%** over Lustre, **30%** over HDFS
- CloudBurst on TACC Stampede
  - With HHH: 19% improvement over HDFS

Design Overview of Spark with RDMA

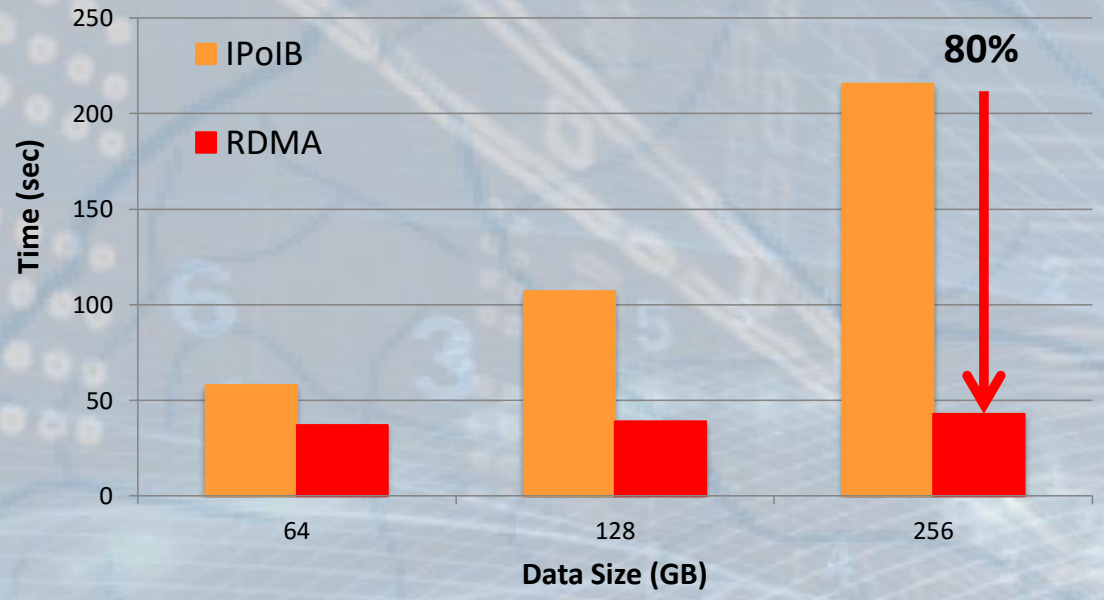


- Design Features
  - RDMA based shuffle
  - SEDA-based plugins
  - Dynamic connection management and sharing
  - Non-blocking data transfer
  - Off-JVM-heap buffer management
  - InfiniBand/RoCE support

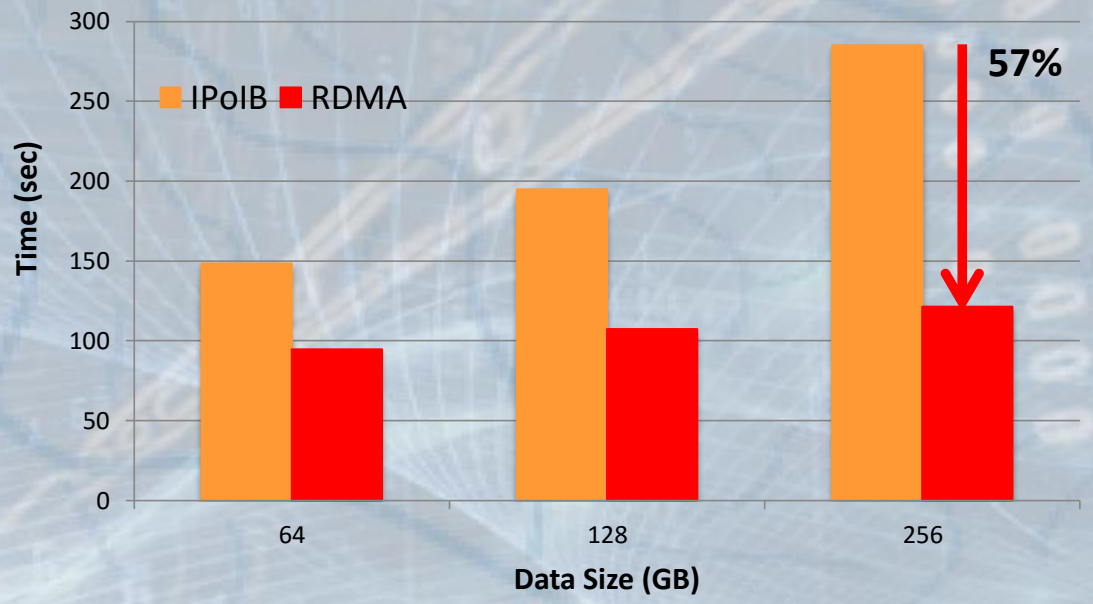
- Enables high performance RDMA communication, while supporting traditional socket interface
- JNI Layer bridges Scala based Spark with communication library written in native code

X. Lu, M. W. Rahman, N. Islam, D. Shankar, and D. K. Panda, Accelerating Spark with RDMA for Big Data Processing: Early Experiences, Int'l Symposium on High Performance Interconnects (HotI'14), August 2014

Performance Evaluation on SDSC Comet – SortBy/GroupBy



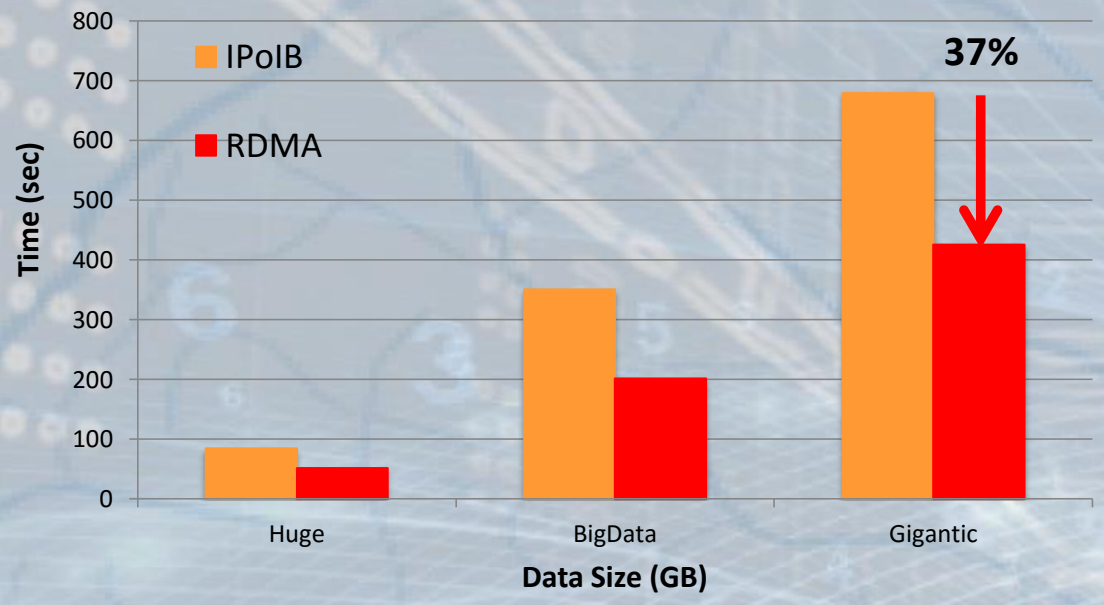
64 Worker Nodes, 1536 cores, SortByTest Total Time



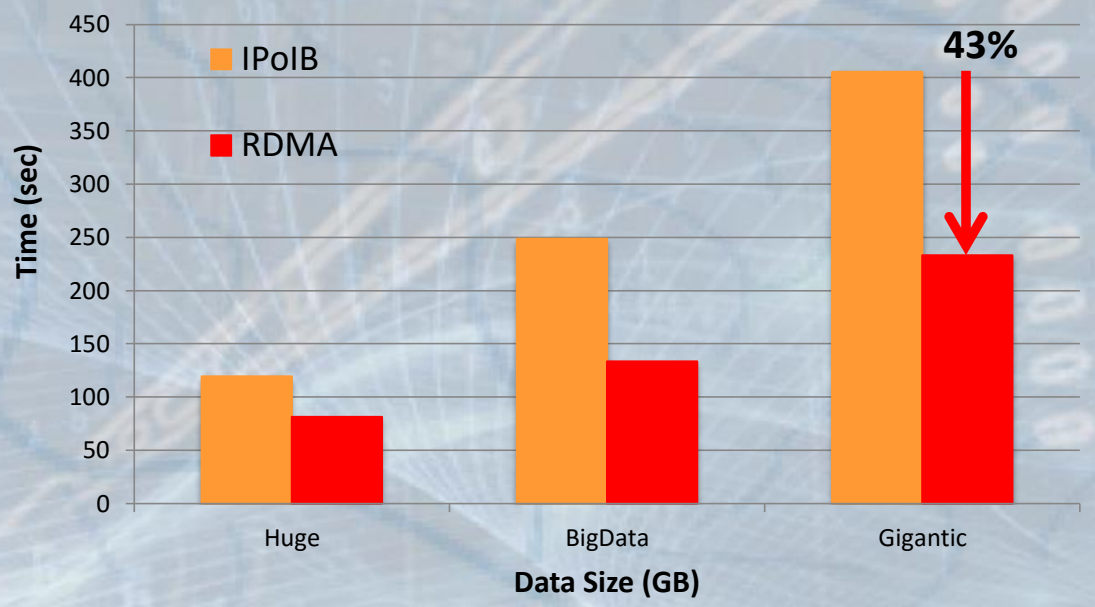
64 Worker Nodes, 1536 cores, GroupByTest Total Time

- InfiniBand FDR, SSD, 64 Worker Nodes, 1536 Cores, (1536M 1536R)
- RDMA-based design for Spark 1.5.1
- RDMA vs. IPoIB with 1536 concurrent tasks, single SSD per node.
  - SortBy: Total time reduced by up to 80% over IPoIB (56Gbps)
  - GroupBy: Total time reduced by up to 57% over IPoIB (56Gbps)

Performance Evaluation on SDSC Comet – HiBench PageRank



32 Worker Nodes, 768 cores, PageRank Total Time

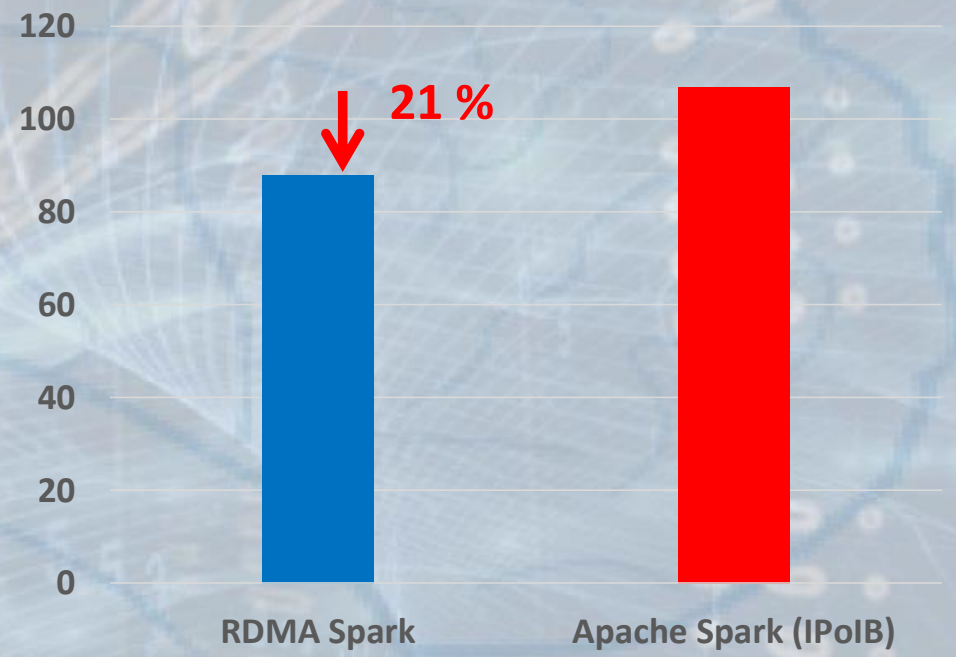


64 Worker Nodes, 1536 cores, PageRank Total Time

- InfiniBand FDR, SSD, 32/64 Worker Nodes, 768/1536 Cores, (768/1536M 768/1536R)
- RDMA-based design for Spark 1.5.1
- RDMA vs. IPoIB with 768/1536 concurrent tasks, single SSD per node.
  - 32 nodes/768 cores: Total time reduced by 37% over IPoIB (56Gbps)
  - 64 nodes/1536 cores: Total time reduced by 43% over IPoIB (56Gbps)

Performance Evaluation on SDSC Comet: Astronomy Application

- **Kira Toolkit[1]:** Distributed astronomy image processing toolkit implemented using Apache Spark.
- Source extractor application, using a 65GB dataset from the SDSS DR2 survey that comprises 11,150 image files.
- Compare RDMA Spark performance with the standard apache implementation using IPoIB.



Execution times (sec) for Kira SE benchmark using 65 GB dataset, 48 cores.

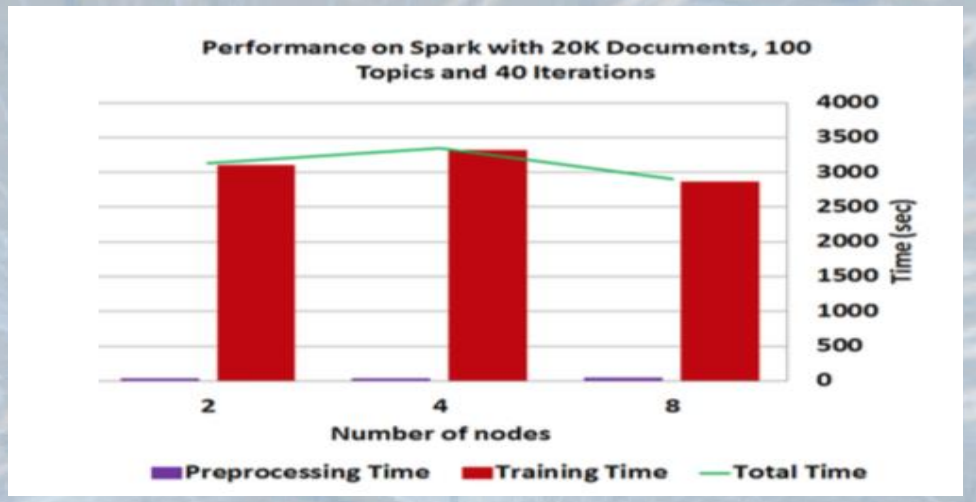
1. Z. Zhang, K. Barbary, F. A. Nothaft, E.R. Sparks, M.J. Franklin, D.A. Patterson, S. Perlmutter. Scientific Computing meets Big Data Technology: An Astronomy Use Case. *CoRR, vol: abs/1507.03325*, Aug 2015.

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

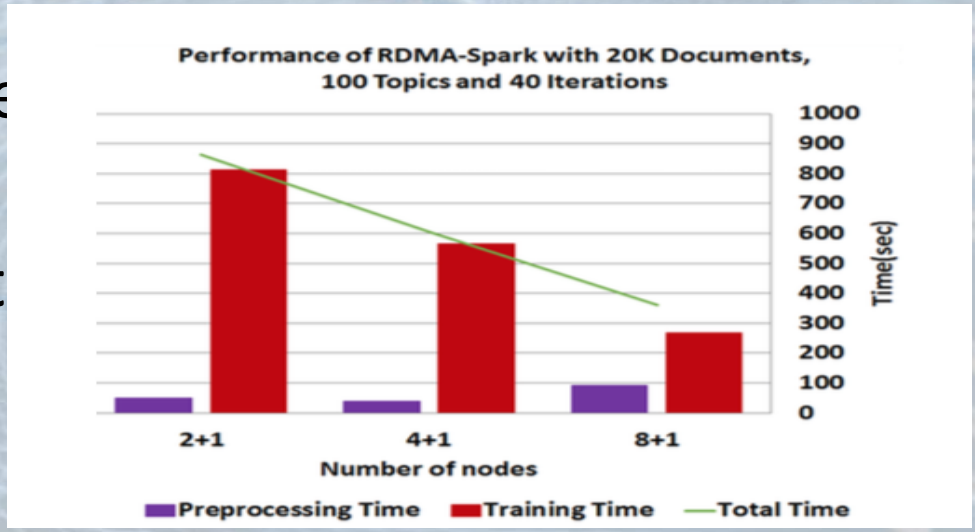
Performance Evaluation on SDSC Comet: Topic Modeling Application

- Application in Social Sciences: Topic modeling using big data middleware and tools*.
- Latent Dirichlet Allocation (LDA) for unsupervised analysis of large document collections.
- Computational complexity increases as the volume of data increases.
- RDMA Spark enabled simulation of largest test cases.

*Investigating Topic Models for Big Data Analysis in Social Science Domain Nitin Sukhija, Nicole Brown, Paul Rodriguez, Mahidhar Tatineni, and Mark Van Moer, XSEDE16 Poster Paper



Default Spark with IPoIB does not scale!

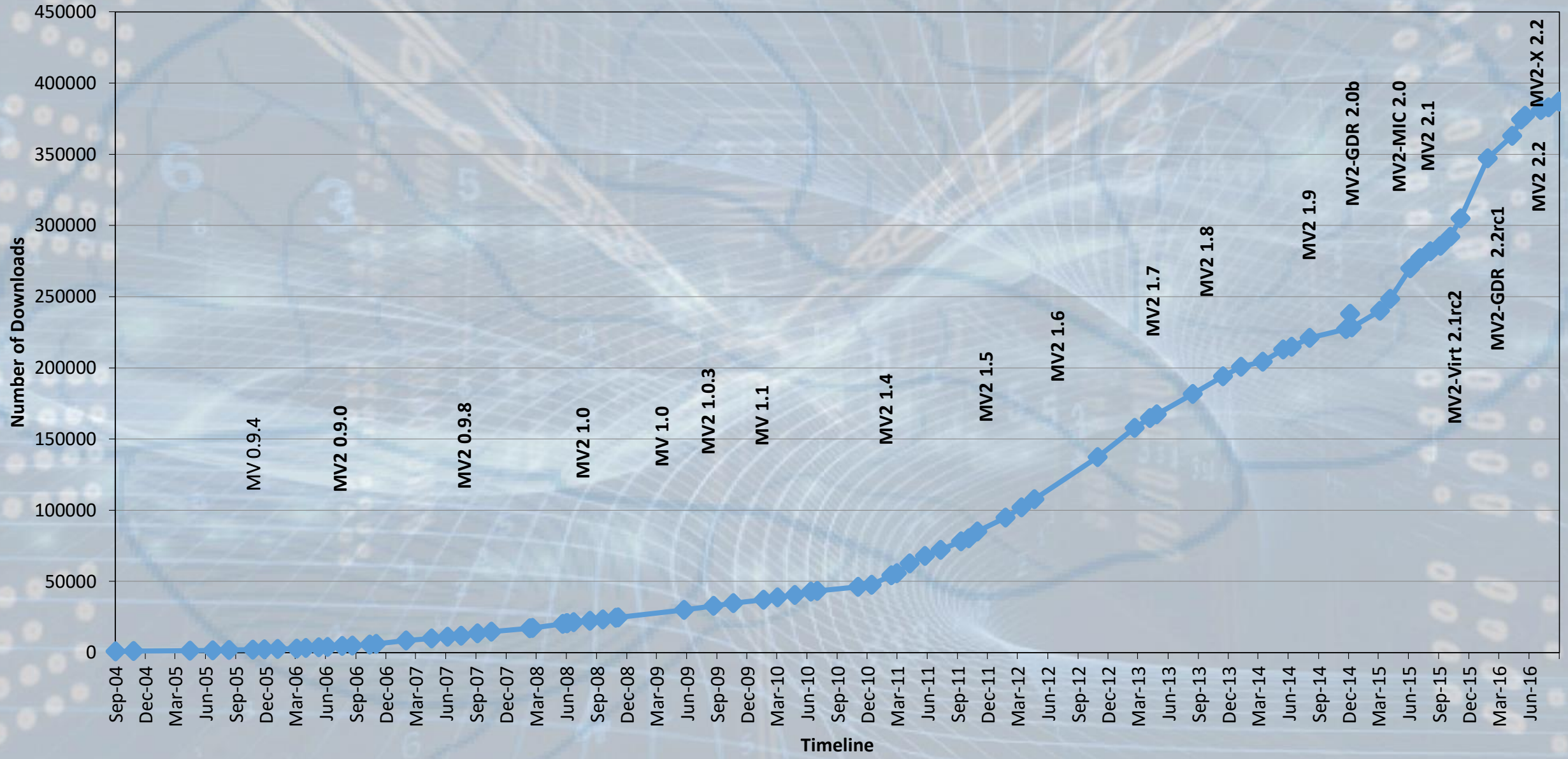


RDMA-Spark can scale very well!

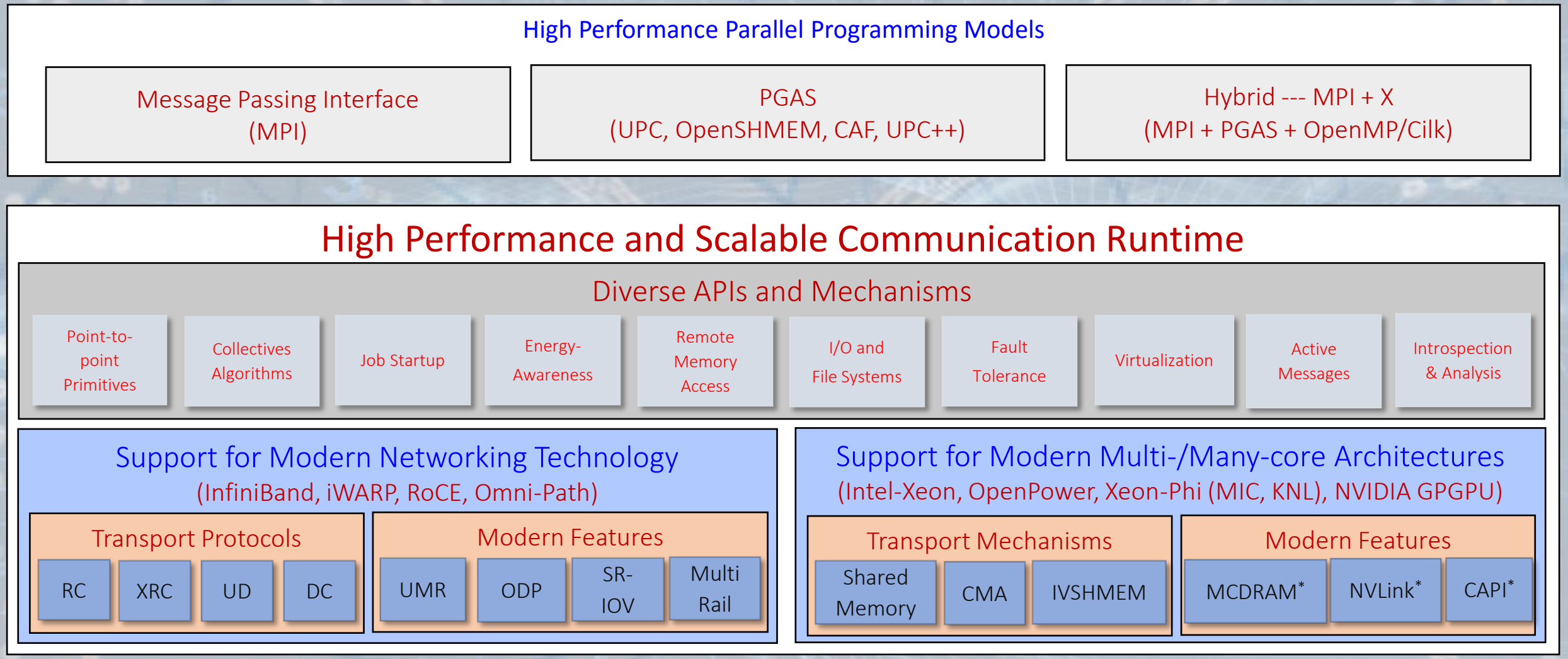
Overview of the MVAPICH2 Project

- 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)

MVAPICH/MVAPICH2 Release Timeline and Downloads



Architecture of MVAPICH2 Software Family

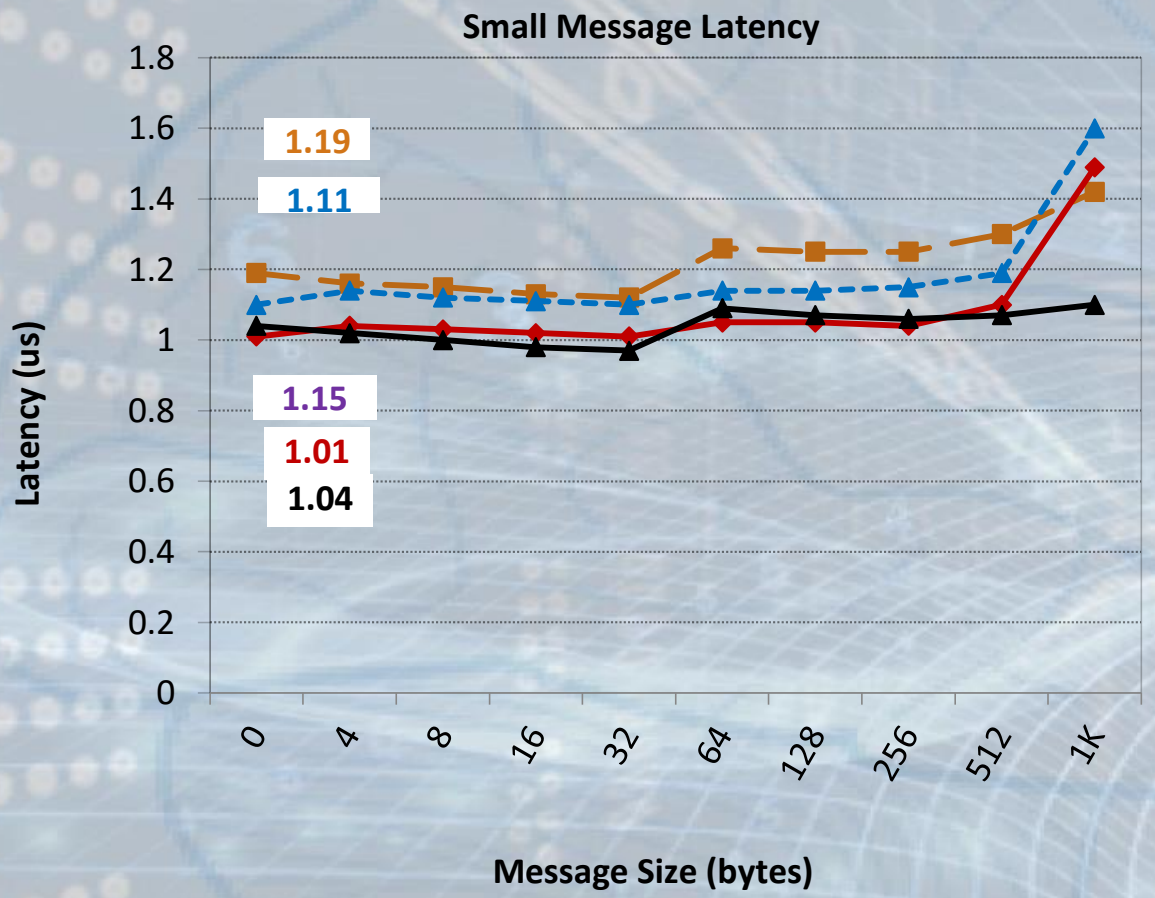


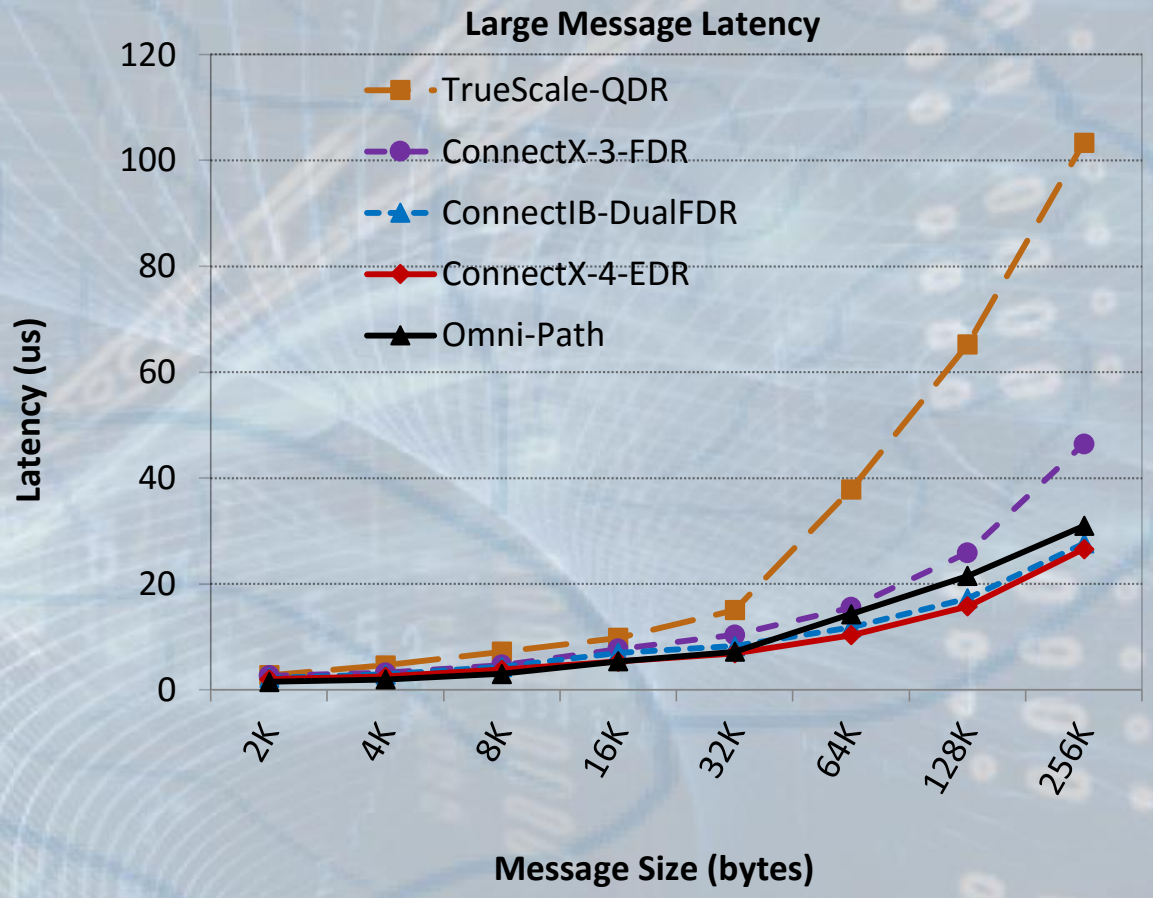
* Upcoming

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

One-way Latency: MPI over IB with MVAPICH2





TrueScale-QDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

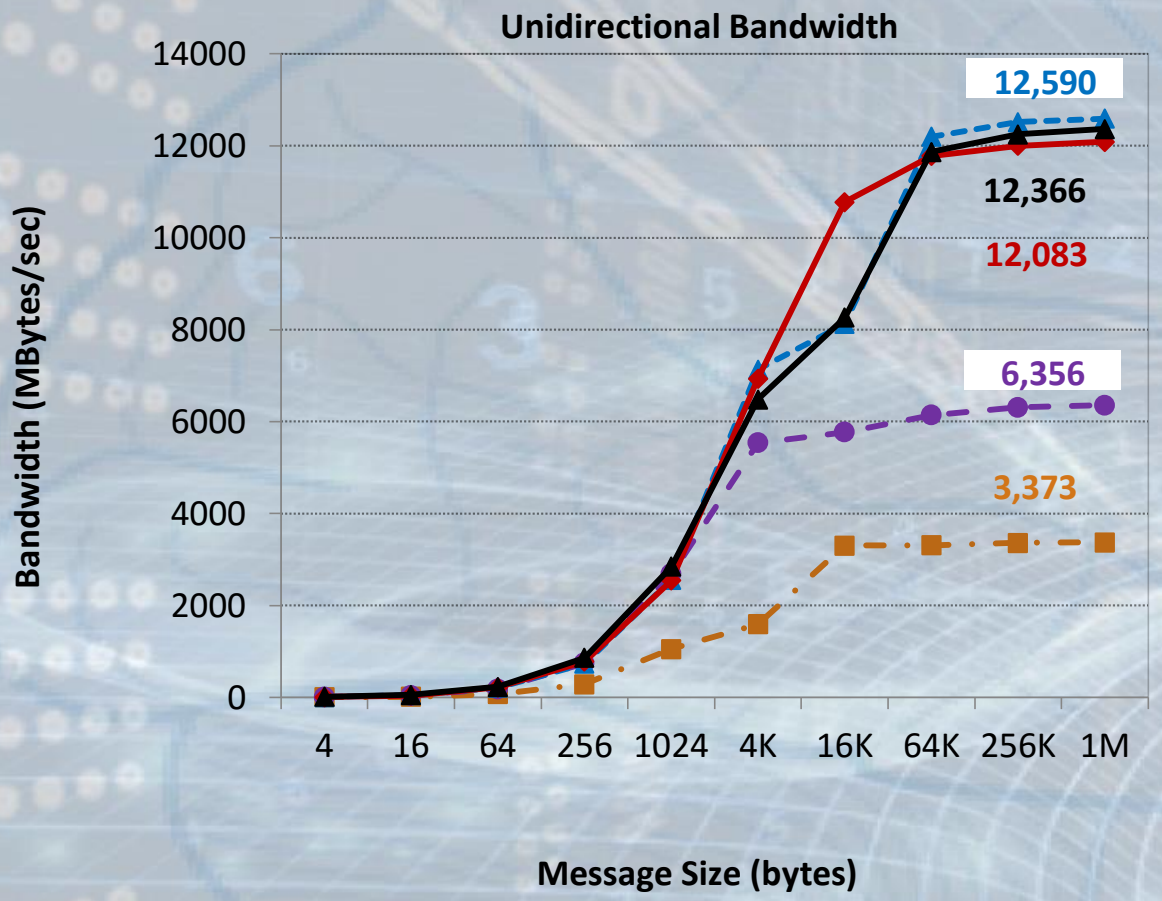
ConnectX-3-FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch

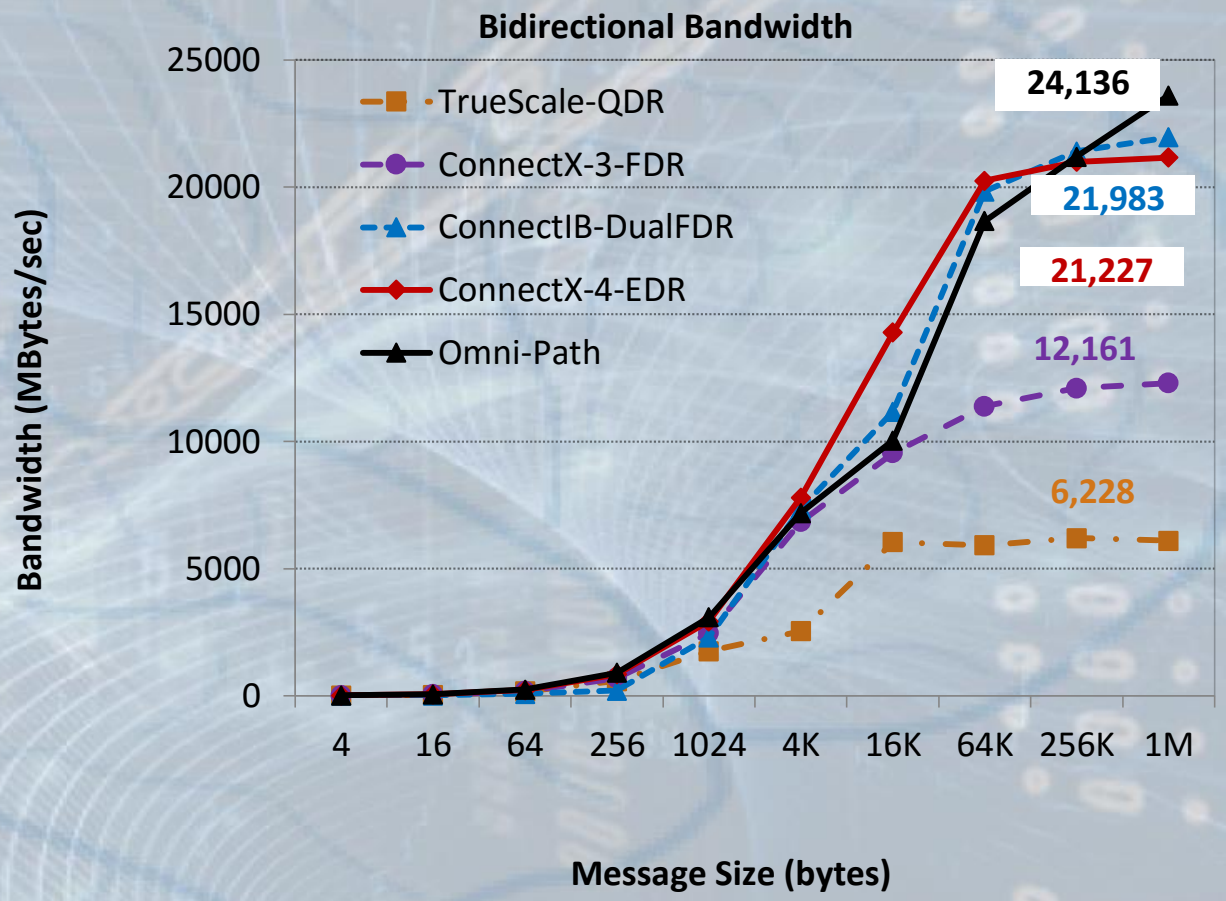
ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

ConnectX-4-EDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB Switch

Omni-Path - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with Omni-Path switch

Bandwidth: MPI over IB with MVAPICH2





TrueScale-QDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

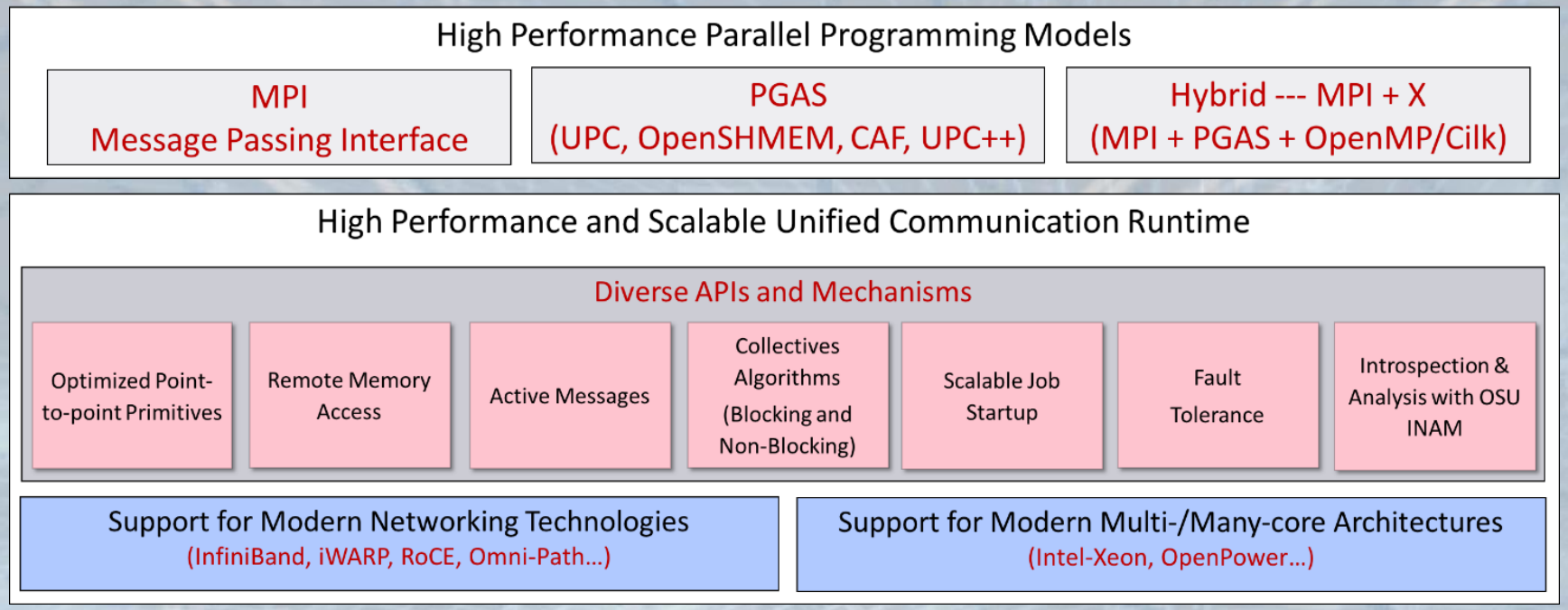
ConnectX-3-FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch

ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

ConnectX-4-EDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 IB switch

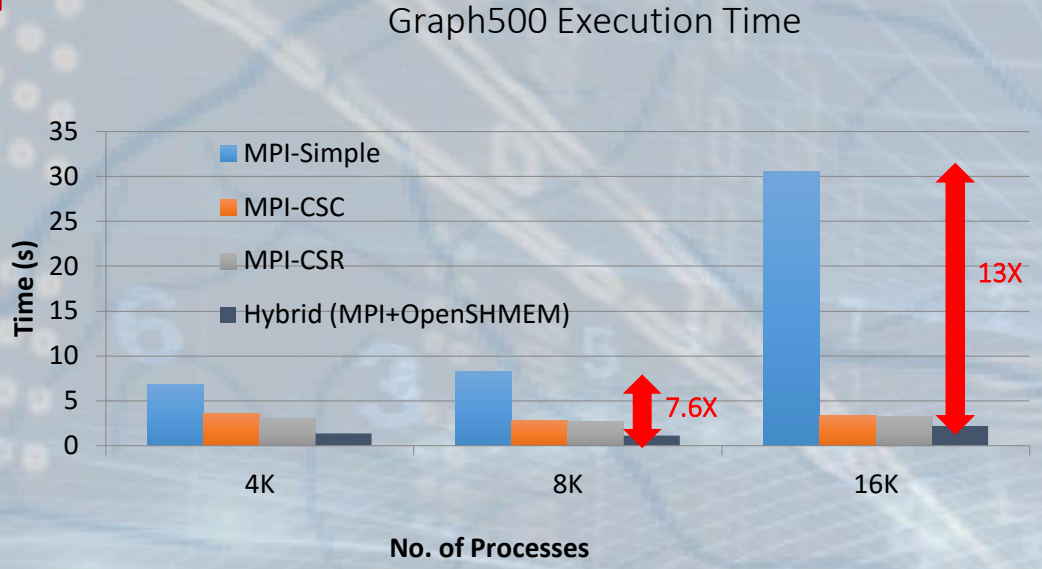
Omni-Path - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with Omni-Path switch

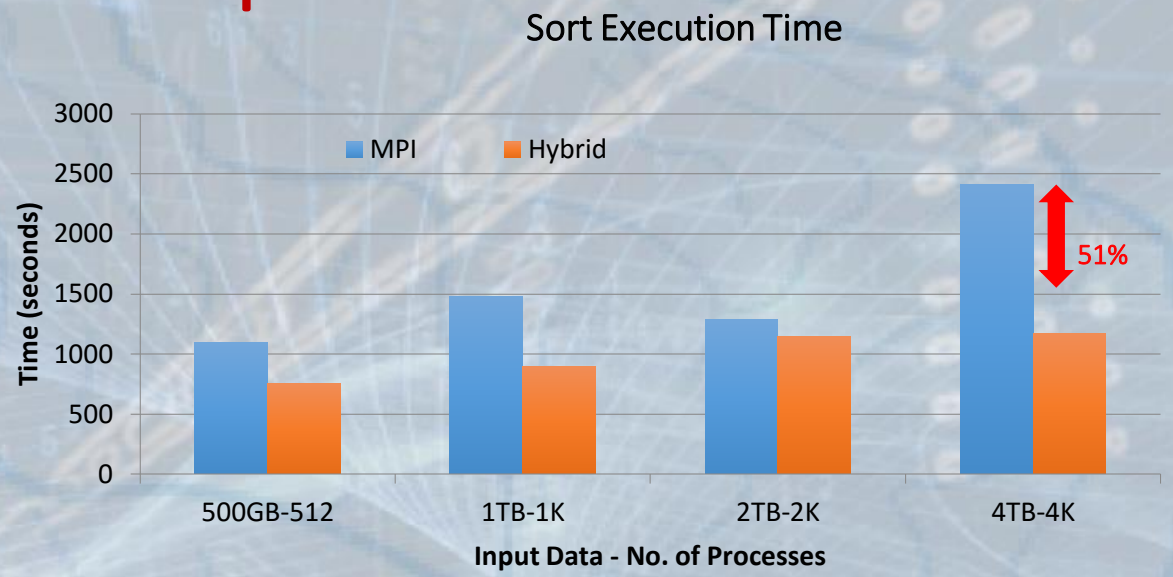
MVAPICH2-X for Hybrid MPI + PGAS Applications



- Current Model – Separate Runtimes for OpenSHMEM/UPC/UPC++/CAF and MPI
  - Possible deadlock if both runtimes are not progressed
  - Consumes more network resource
- Unified communication runtime for MPI, UPC, UPC++, OpenSHMEM, CAF
  - Available with since 2012 (starting with MVAPICH2-X 1.9)
  - http://mvapich.cse.ohio-state.edu

Application Level Performance with Graph500 and Sort





- Performance of Hybrid (MPI+ OpenSHMEM) Graph500 Design
  - 8,192 processes
    - 2.4X improvement over MPI-CSR
    - 7.6X improvement over MPI-Simple
  - 16,384 processes
    - 1.5X improvement over MPI-CSR
    - 13X improvement over MPI-Simple

- Performance of Hybrid (MPI+OpenSHMEM) Sort Application
  - 4,096 processes, 4 TB Input Size
    - MPI – 2408 sec; 0.16 TB/min
    - Hybrid – 1172 sec; 0.36 TB/min
    - 51% improvement over MPI-design

J. Jose, K. Kandalla, S. Potluri, J. Zhang and D. K. Panda, Optimizing Collective Communication in OpenSHMEM, Int'l Conference on Partitioned Global Address Space Programming Models (PGAS '13), October 2013.

J. Jose, S. Potluri, K. Tomko and D. K. Panda, Designing Scalable Graph500 Benchmark with Hybrid MPI+OpenSHMEM Programming Models, International Supercomputing Conference (ISC'13), June 2013

J. Jose, K. Kandalla, M. Luo and D. K. Panda, Supporting Hybrid MPI and OpenSHMEM over InfiniBand: Design and Performance Evaluation, Int'l Conference on Parallel Processing (ICPP '12), September 2012

GPU-Aware (CUDA-Aware) MPI Library: MVAPICH2-GPU

- Standard MPI interfaces used for unified data movement
- Takes advantage of Unified Virtual Addressing (>= CUDA 4.0)
- Overlaps data movement from GPU with RDMA transfers

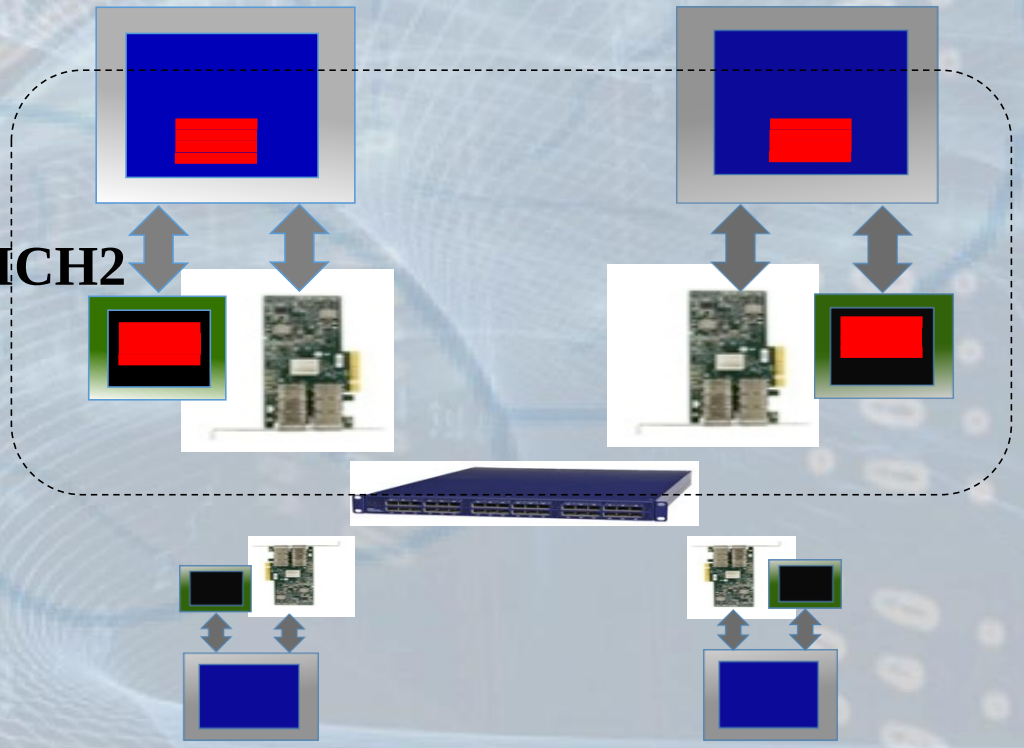
At Sender:

```
MPI_Send(s_devbuf, size, …);
```

At Receiver:

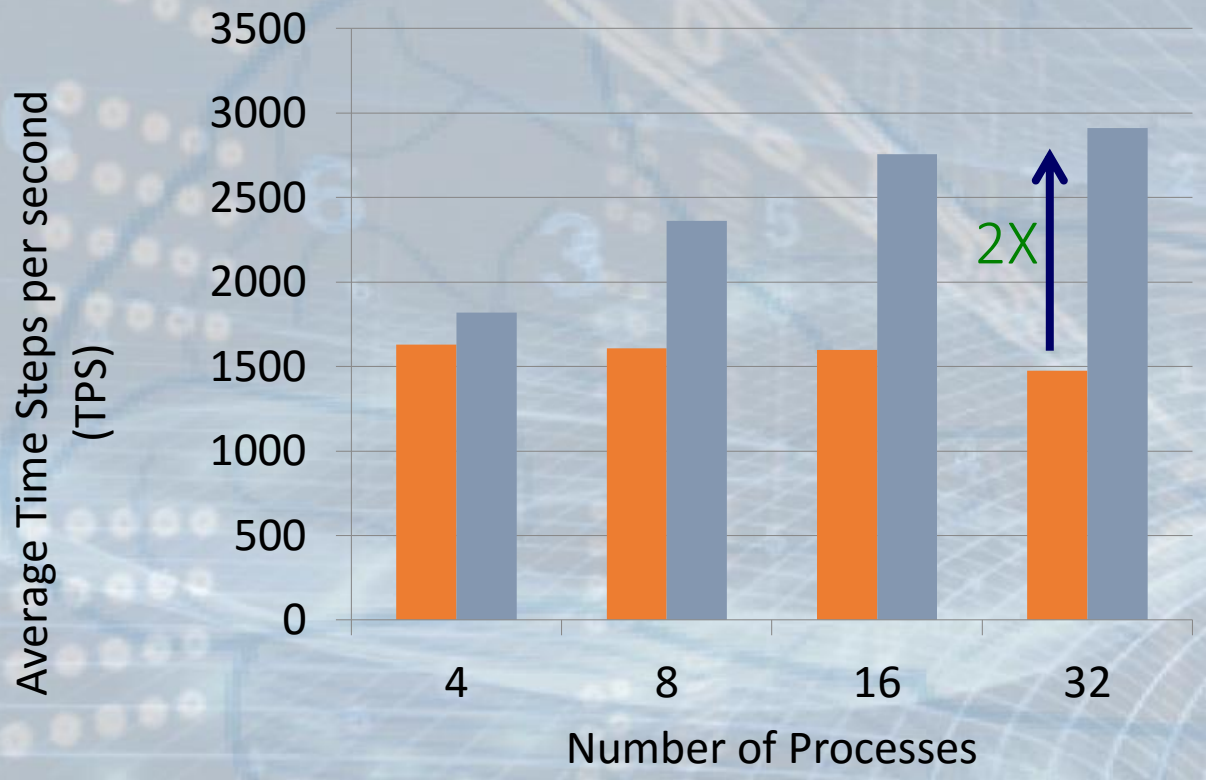
```
MPI_Recv(r_devbuf, size, …);
```

High Performance and High Productivity

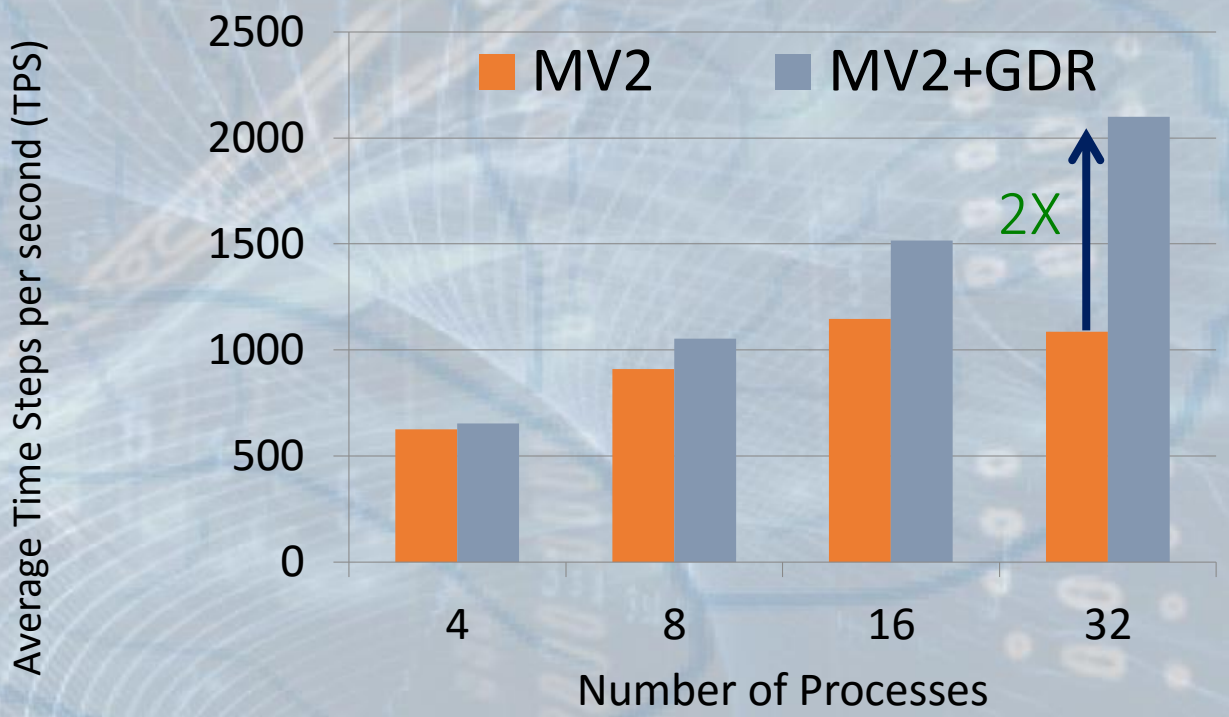


Application-Level Evaluation (HOOMD-blue)

64K Particles



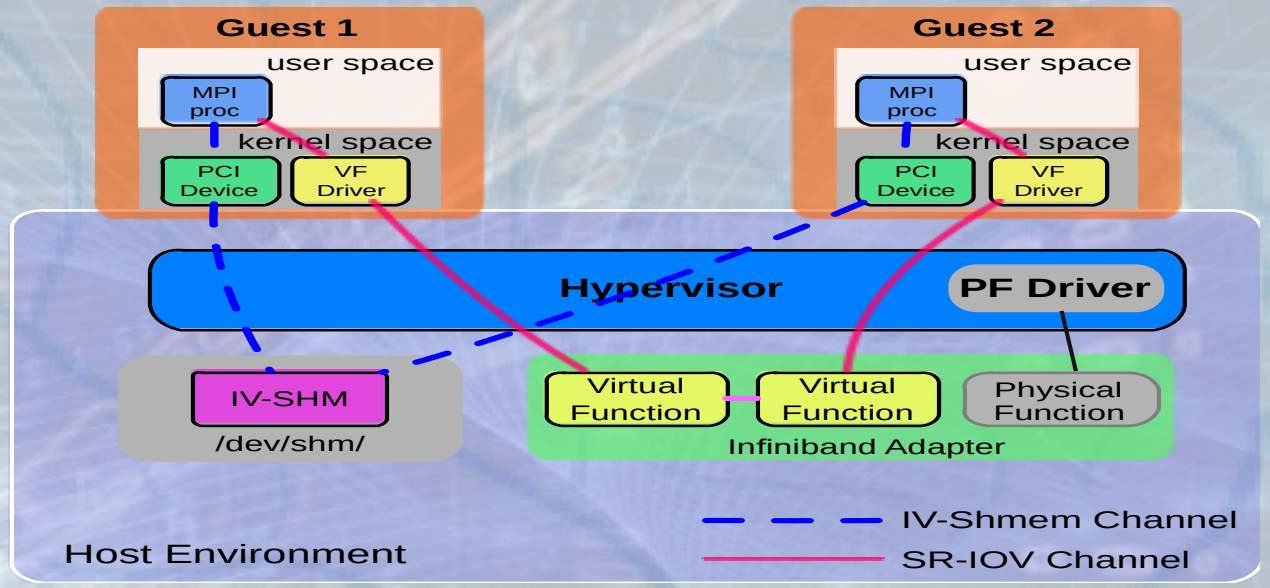
256K Particles



- Platform: Wilkes (Intel Ivy Bridge + NVIDIA Tesla K20c + Mellanox Connect-IB)
- HoomdBlue Version 1.0.5
  - GDRCOPY enabled: MV2_USE_CUDA=1 MV2_IBA_HCA=mlx5_0 MV2_IBA_EAGER_THRESHOLD=32768 MV2_VBUF_TOTAL_SIZE=32768 MV2_USE_GPUDIRECT_LOOPBACK_LIMIT=32768 MV2_USE_GPUDIRECT_GDRCOPY=1 MV2_USE_GPUDIRECT_GDRCOPY_LIMIT=16384

MVAPICH2-Virt with SR-IOV and IVSHMEM for HPC Cloud

- Redesign MVAPICH2 to make it virtual machine aware
  - SR-IOV shows near to native performance for inter-node point to point communication
  - IVSHMEM offers shared memory based data access across co-resident VMs
  - Locality Detector: maintains the locality information of co-resident virtual machines
  - Communication Coordinator: selects the communication channel (SR-IOV, IVSHMEM) adaptively

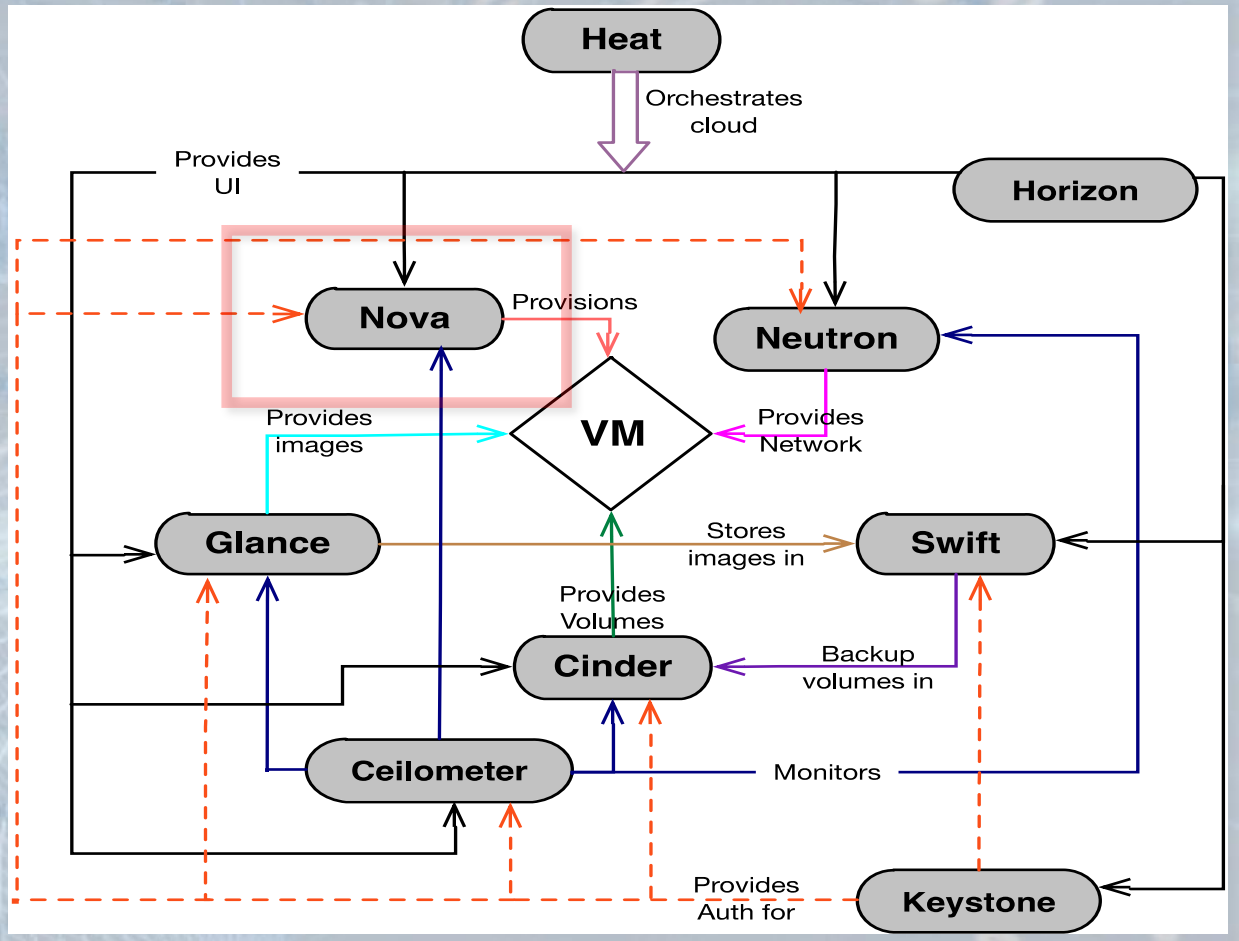


J. Zhang, X. Lu, J. Jose, R. Shi, D. K. Panda. Can Inter-VM Shmem Benefit MPI Applications on SR-IOV based Virtualized InfiniBand Clusters? Euro-Par, 2014

J. Zhang, X. Lu, J. Jose, R. Shi, M. Li, D. K. Panda. High Performance MPI Library over SR-IOV Enabled InfiniBand Clusters. HiPC, 2014

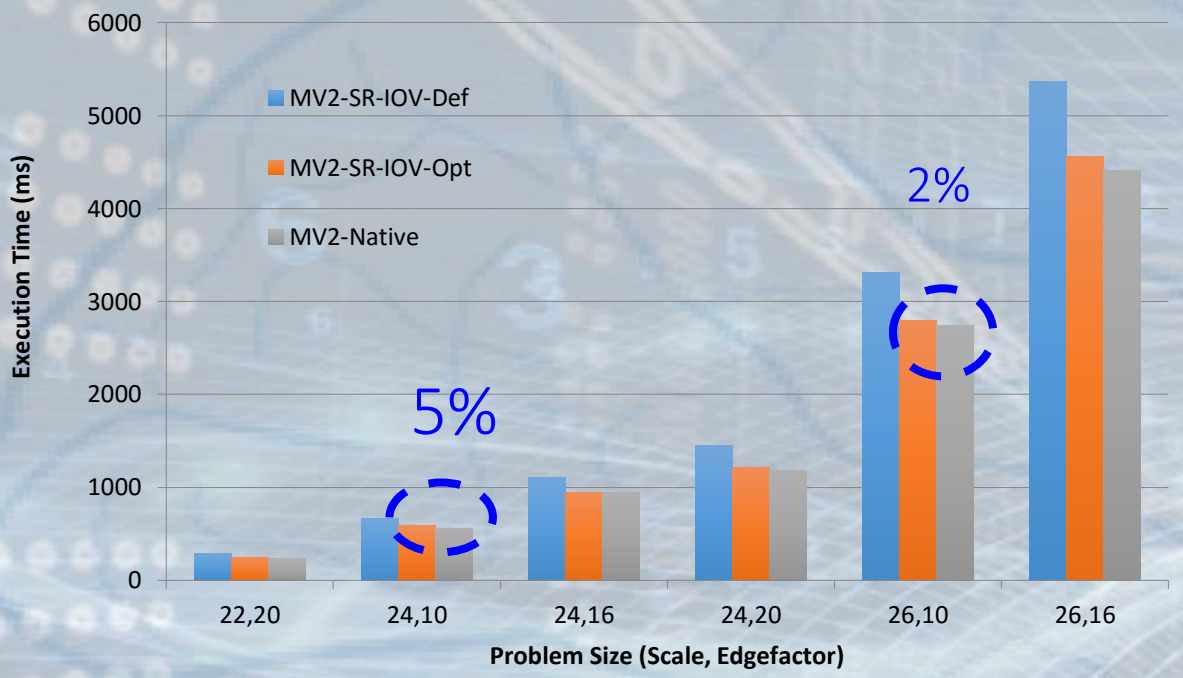
MVAPICH2-Virt over OpenStack for HPC Cloud

- OpenStack is one of the most popular open-source solutions to build clouds and manage virtual machines
- Deployment with OpenStack
  - Supporting SR-IOV configuration
  - Supporting IVSHMEM configuration
  - Virtual Machine aware design of MVAPICH2 with SR-IOV
- An efficient approach to build HPC Clouds with MVAPICH2-Virt and OpenStack

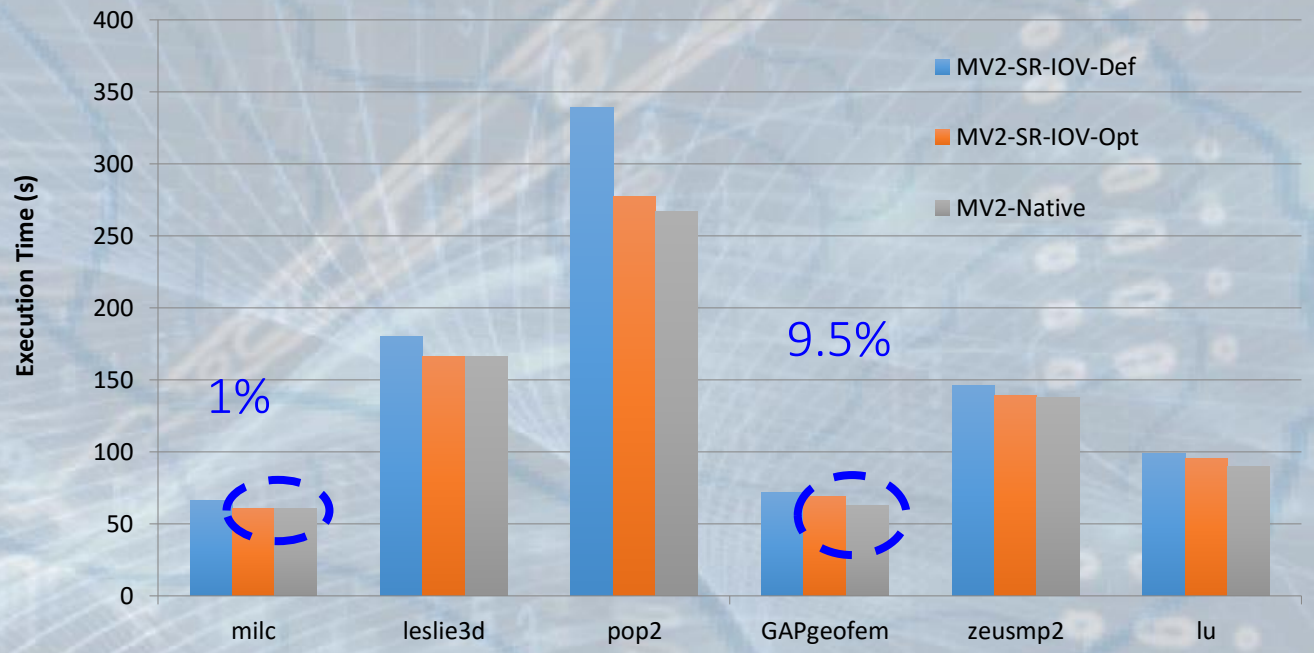


J. Zhang, X. Lu, M. Arnold, D. K. Panda. MVAPICH2 over OpenStack with SR-IOV: An Efficient Approach to Build HPC Clouds. CCGrid, 2015

Application-Level Performance on Chameleon (KVM)



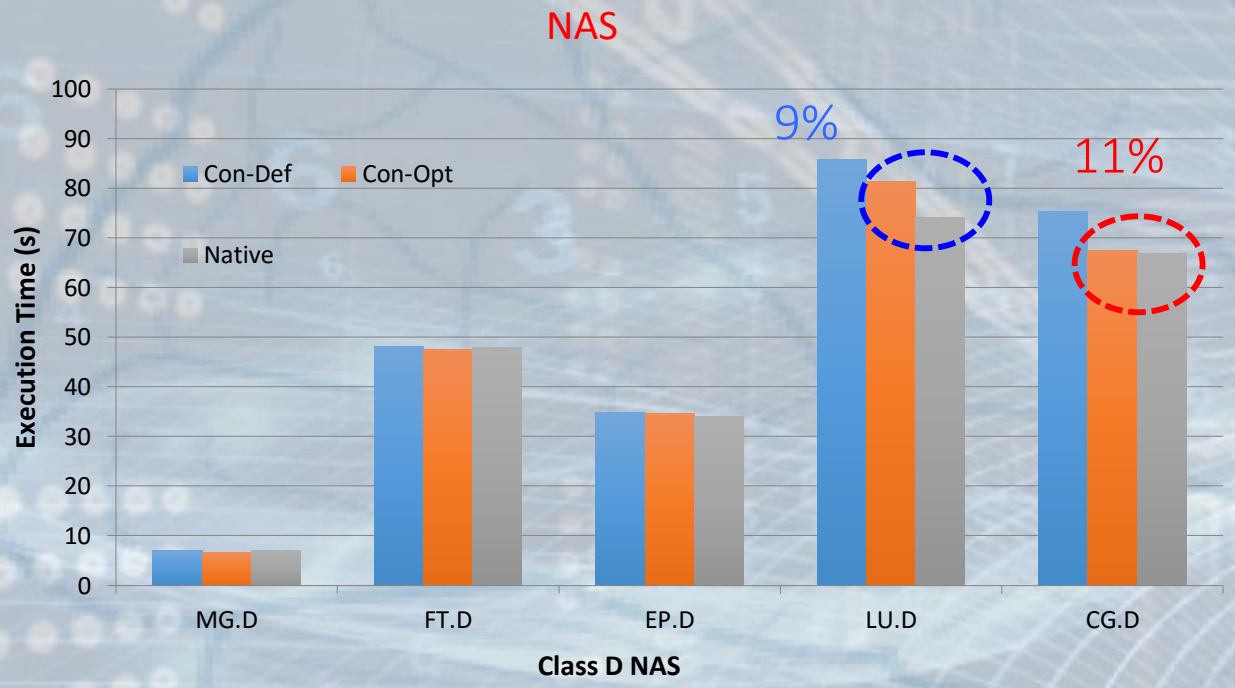
Graph500

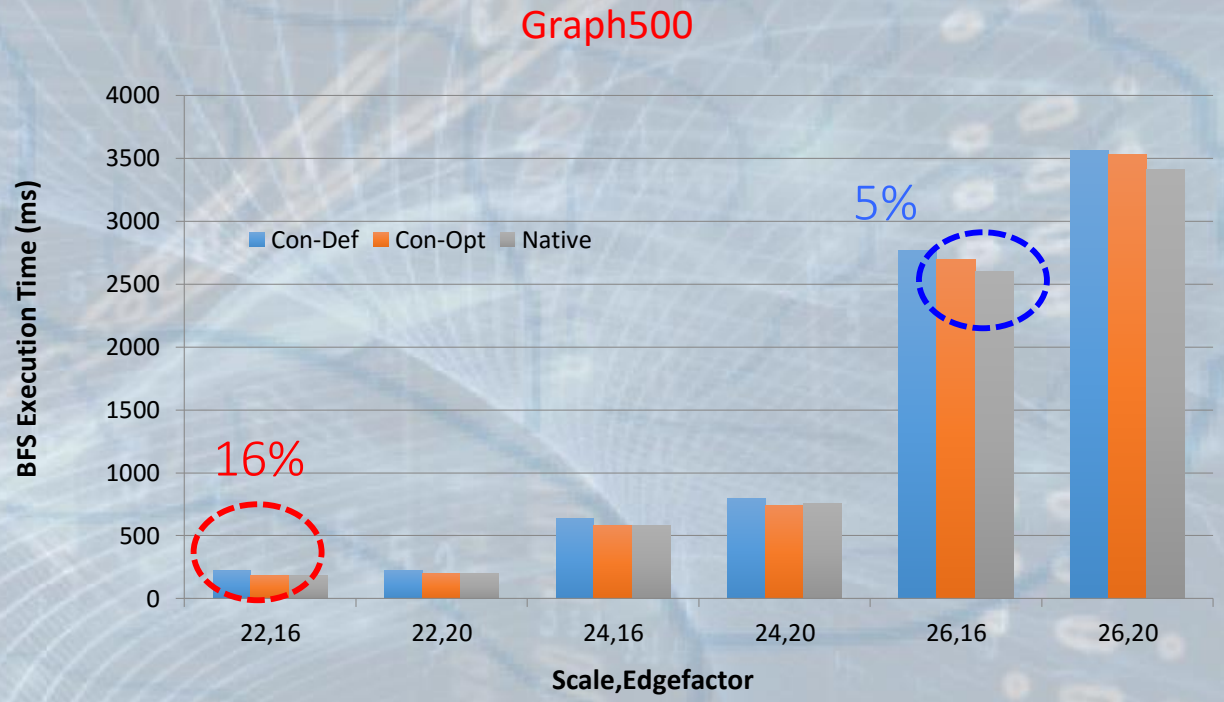


SPEC MPI2007

- 32 VMs, 6 Core/VM
- Compared to Native, 2-5% overhead for Graph500 with 128 Procs
- Compared to Native, 1-9.5% overhead for SPEC MPI2007 with 128 Procs

Application-Level Performance on Chameleon (Docker-based Container)

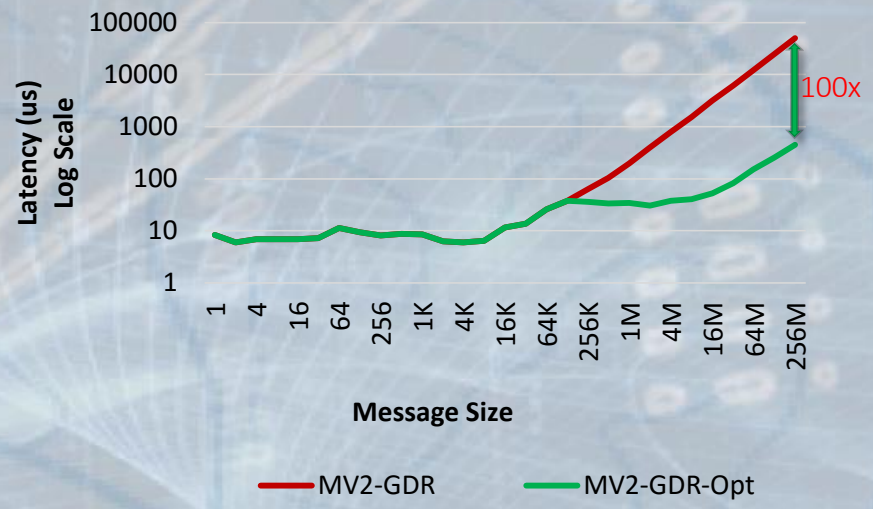




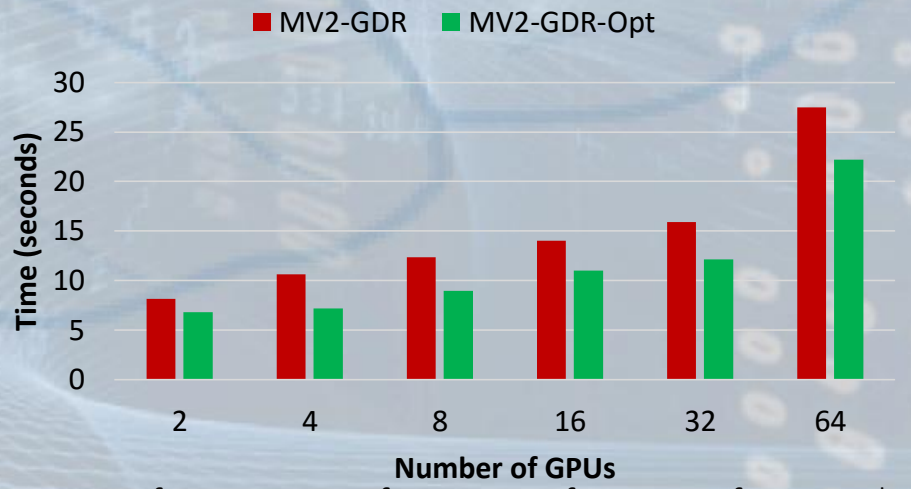
- 256 Procs (4 procs * 4 containers * 16 nodes)
- Reduces the execution time by up to 11%(CG) and 16%(22,16) on Class D NAS and Graph500 (Cont-Opt vs. Cont-Def)
- Only has up to 9%(LU) and 5%(26,16) overhead, compared with native performance

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

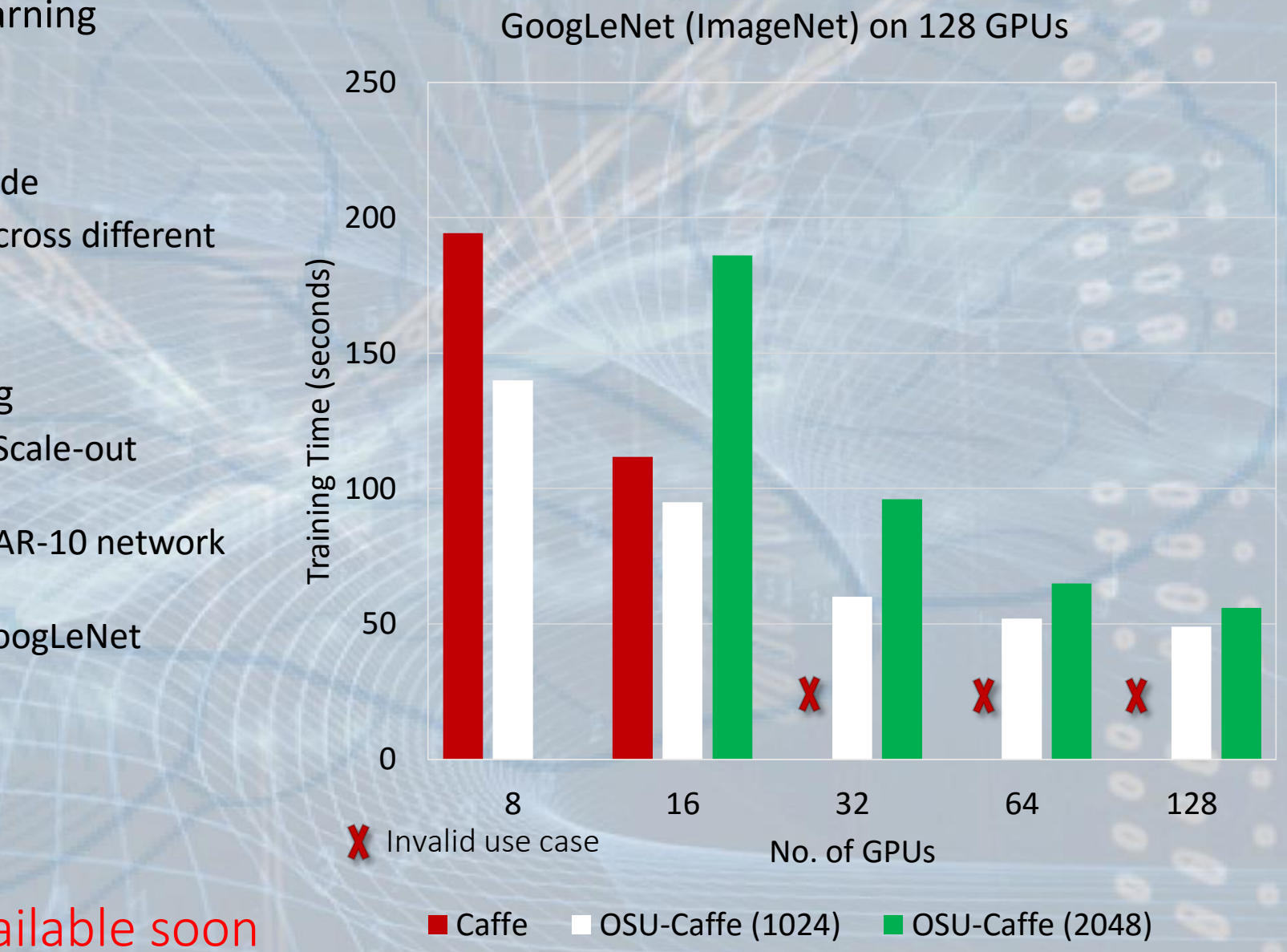


Performance Benefits: Microsoft CNTK DL framework
(25% avg. improvement)

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



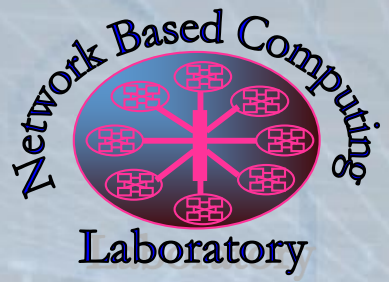
OSU-Caffe will be publicly available soon

Conclusions

- Multiple on-going projects at OSU provide robust solutions for HPC, Big Data, Cloud Computing and Deep Learning
- These solutions can be used by ACNN team members to exploit the latest technologies for processing Neuroscience data
- Multiple posters, lightning sessions and breakout sessions by OSU team members to explore collaboration opportunities and bridge the existing gaps in the community
- OSU team is looking forward to working together with all ACNN team members to push the frontier of Neuroscience

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)

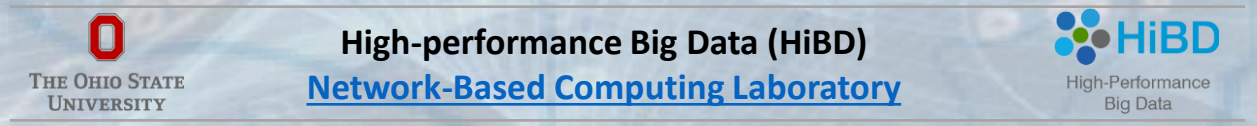
http://www.NeuroscienceNetwork.org/ACNN_Workshop_2016.html

Workshop Sponsors

The National Science Foundation
http://www.nsf.gov

Midwest Big Data Hub
http://MidwestBigDataHub.org

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



The Michigan Institute for Data Science (MIDAS)
http://midas.umich.edu

The Indiana Imaging Research Facility (IRF)
https://www.indiana.edu/~irf/home

CWRU Biomedical and Healthcare Informatics
https://goo.gl/l19s07

Michigan Nutrition Obesity Research Center (MNORC) http://mmoc.med.umich.edu