MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

HiDL
High-Performance Deep Learning

HiBD
High-Performance Big Data

Networking Technologies and Middleware for Next-Generation Clusters and Data Centers: Challenges and Opportunities

Keynote Talk at HPSR '18

by

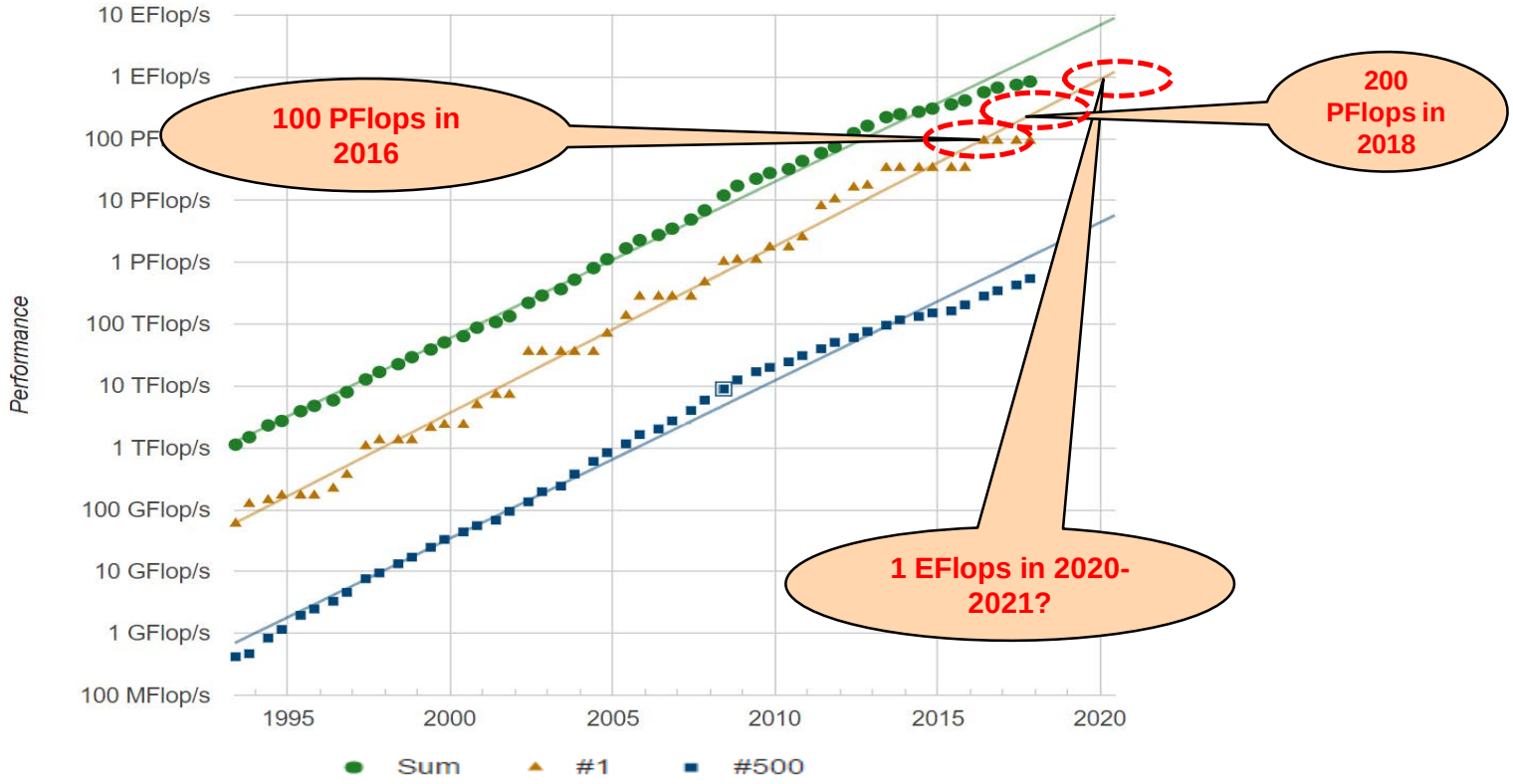
Dhabaleswar K. (DK) Panda

The Ohio State University

E-mail: panda@cse.ohio-state.edu

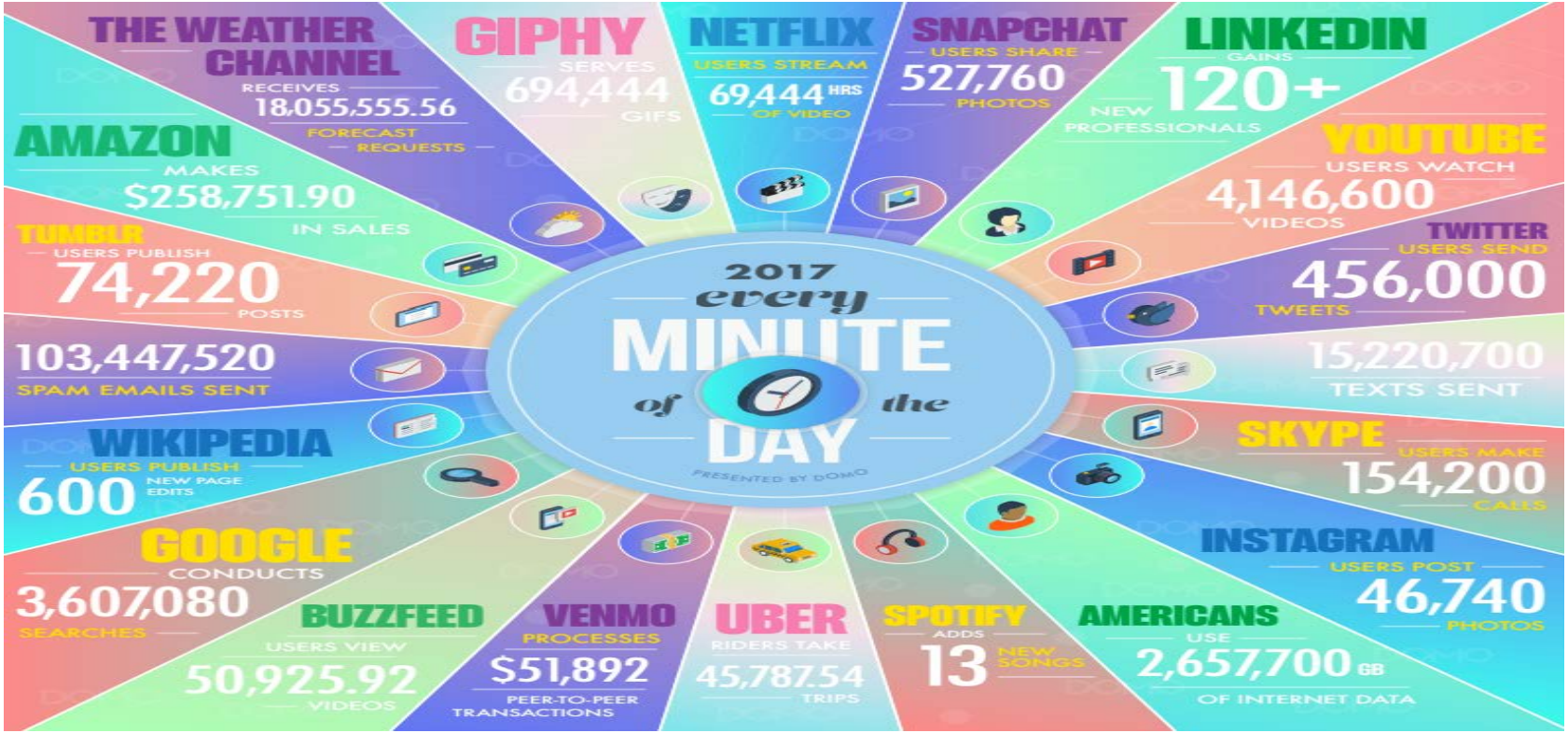
http://www.cse.ohio-state.edu/~panda

High-End Computing (HEC): Towards Exascale



Expected to have an ExaFlop system in 2020-2021!

Big Data – How Much Data Is Generated Every Minute on the Internet?

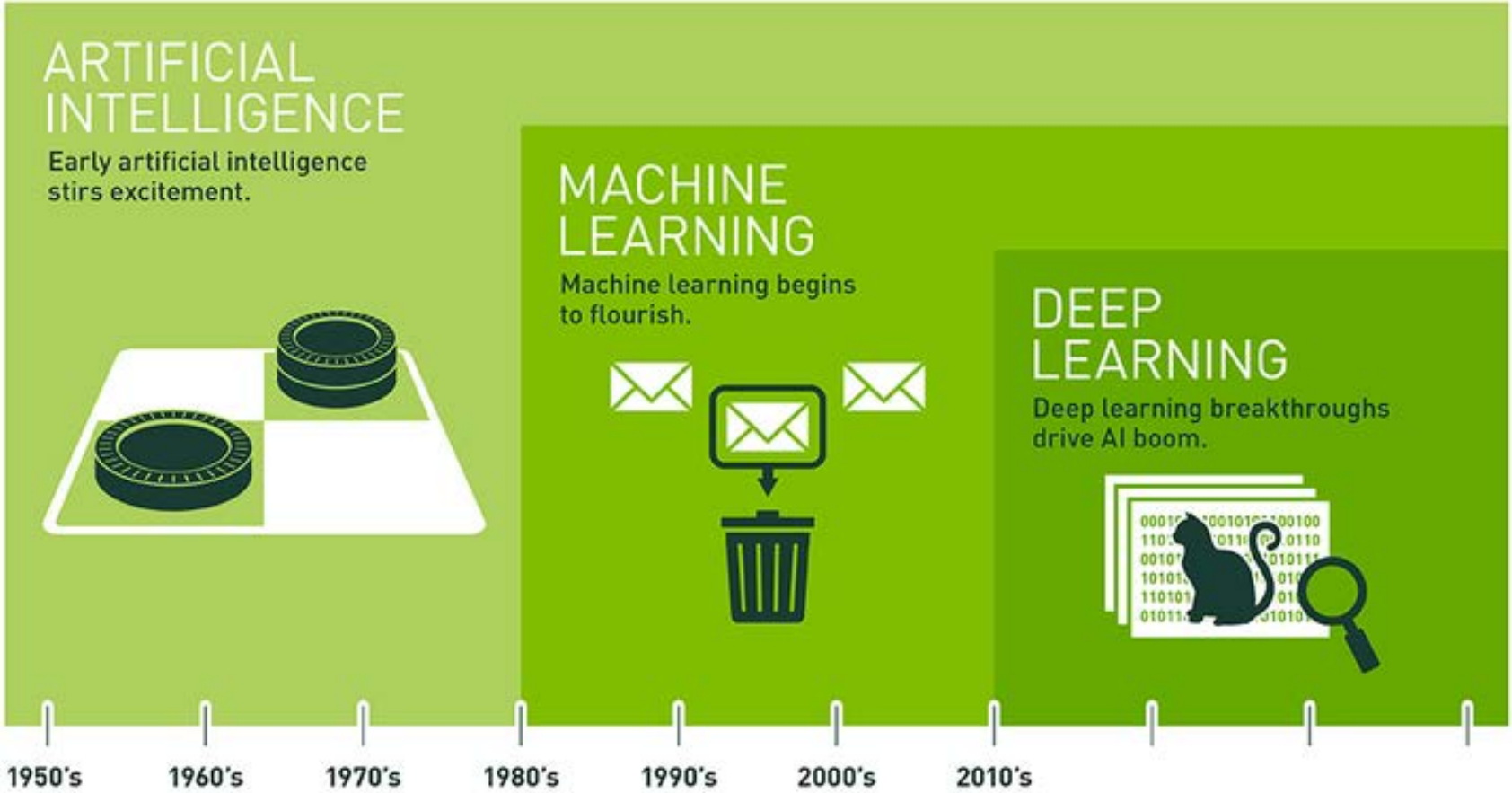


The global Internet population grew 7.5% from 2016 and now represents

3.7 Billion People.

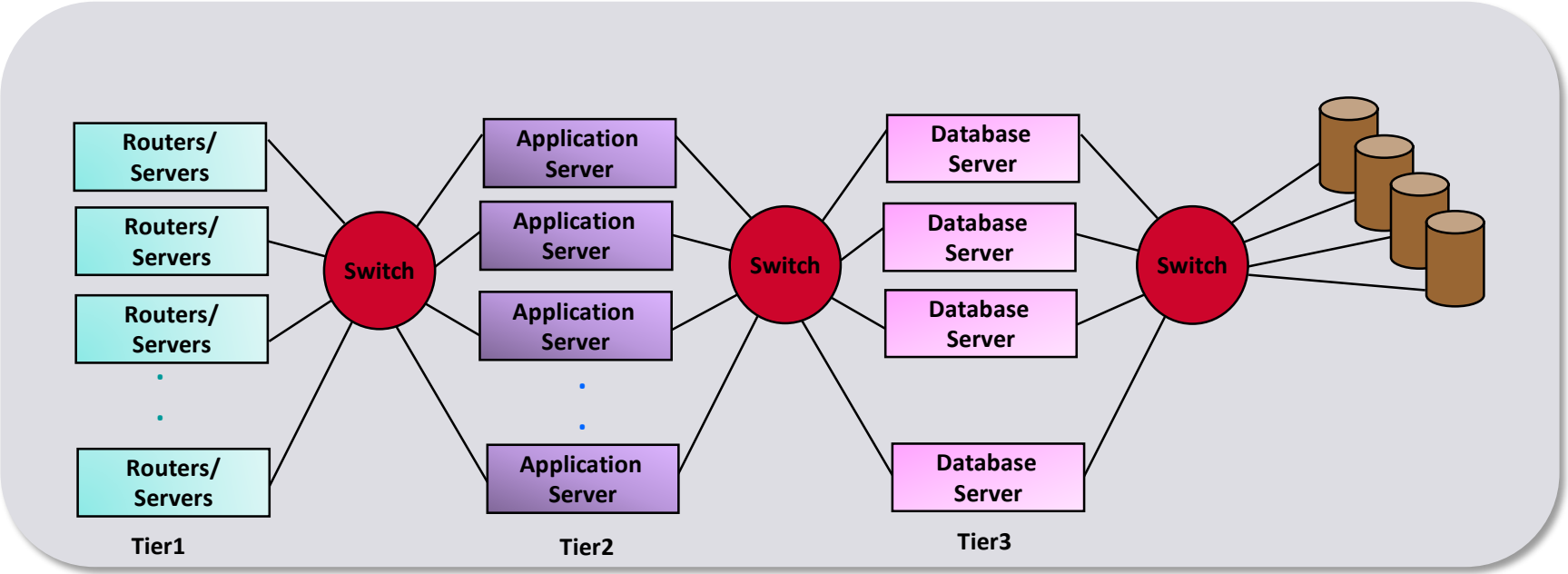
Courtesy: https://www.domo.com/blog/data-never-sleeps-5/

Resurgence of AI/Machine Learning/Deep Learning



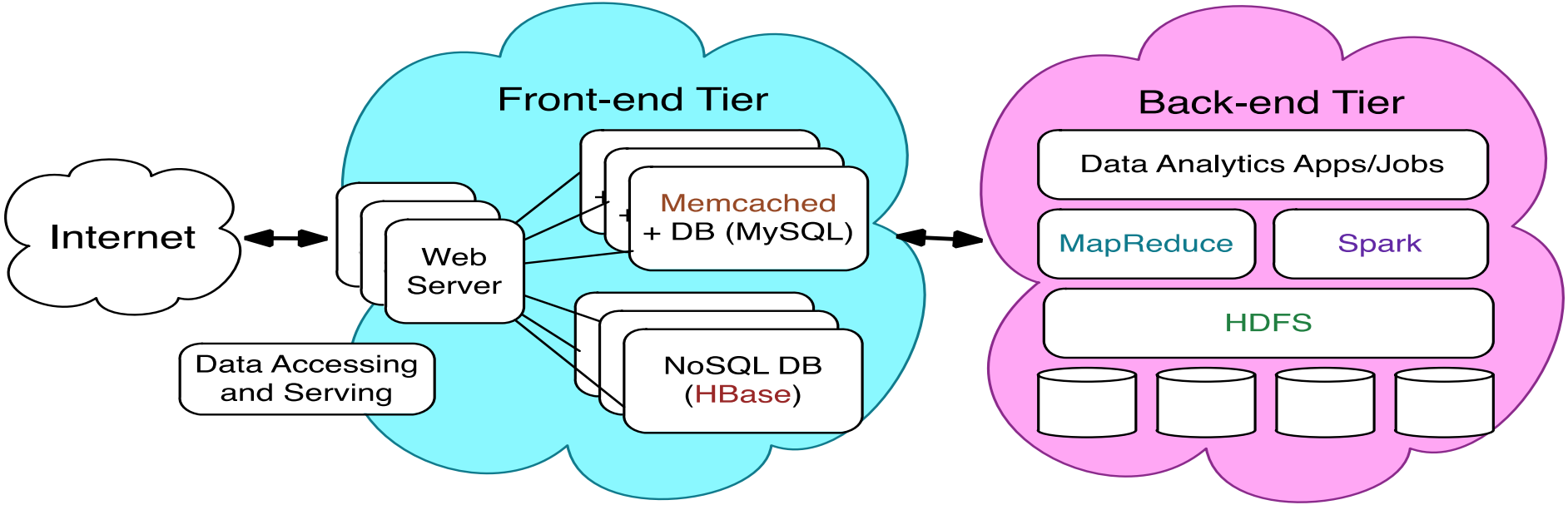
Courtesy: http://www.zdnet.com/article/caffe2-deep-learning-wide-ambitions-flexibility-scalability-and-advocacy/

A Typical Multi-Tier Data Center Architecture

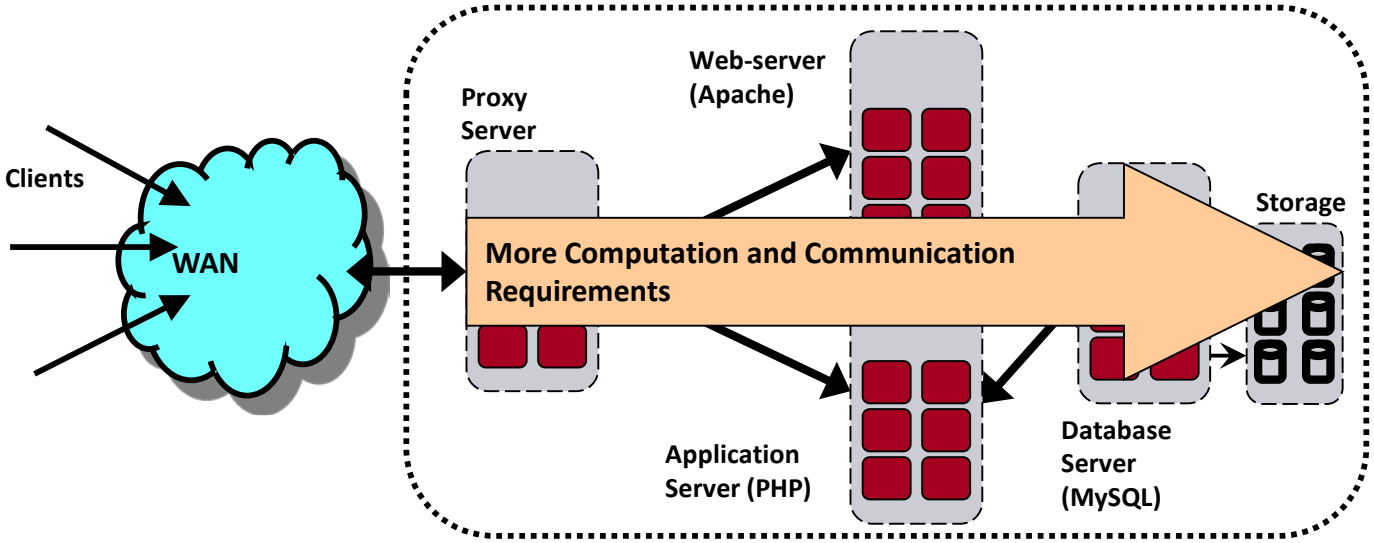


Data Management and Processing on Modern Datacenters

- Substantial impact on designing and utilizing data management and processing systems in multiple tiers
  - Front-end data accessing and serving (Online)
    - Memcached + DB (e.g. MySQL), HBase
  - Back-end data analytics (Offline)
    - HDFS, MapReduce, Spark

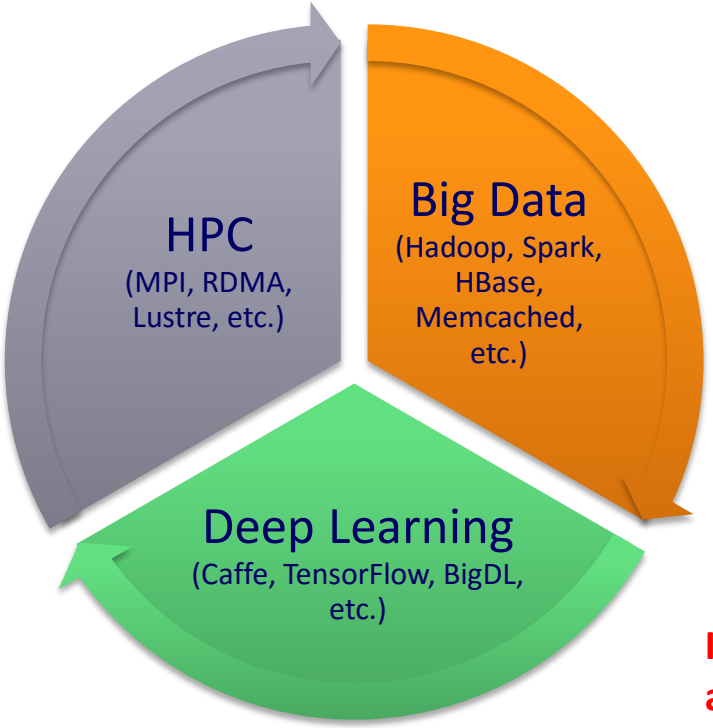


Communication and Computation Requirements



- Requests are received from clients over the WAN
- Proxy nodes perform caching, load balancing, resource monitoring, etc.
- If not cached, the request is forwarded to the next tiers → Application Server
- Application server performs the business logic (CGI, Java servlets, etc.)
  - Retrieves appropriate data from the database to process the requests

Increasing Usage of HPC, Big Data and Deep Learning on Modern Datacenters



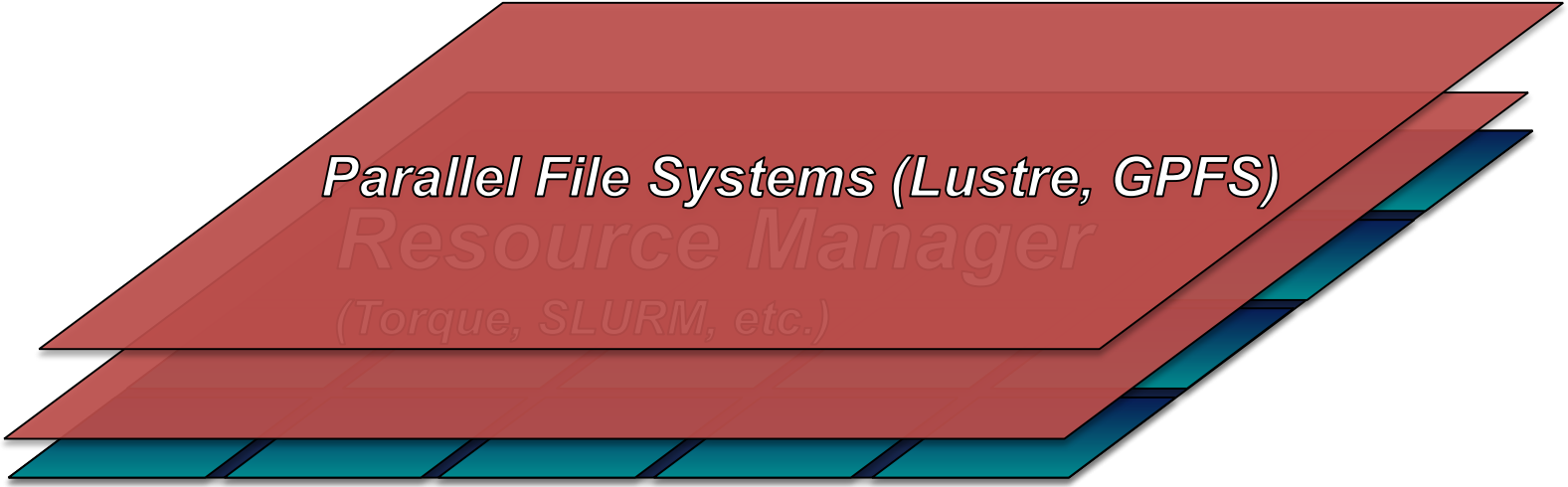
Convergence of HPC, Big Data, and Deep Learning!

Increasing Need to Run these applications on the Cloud!!

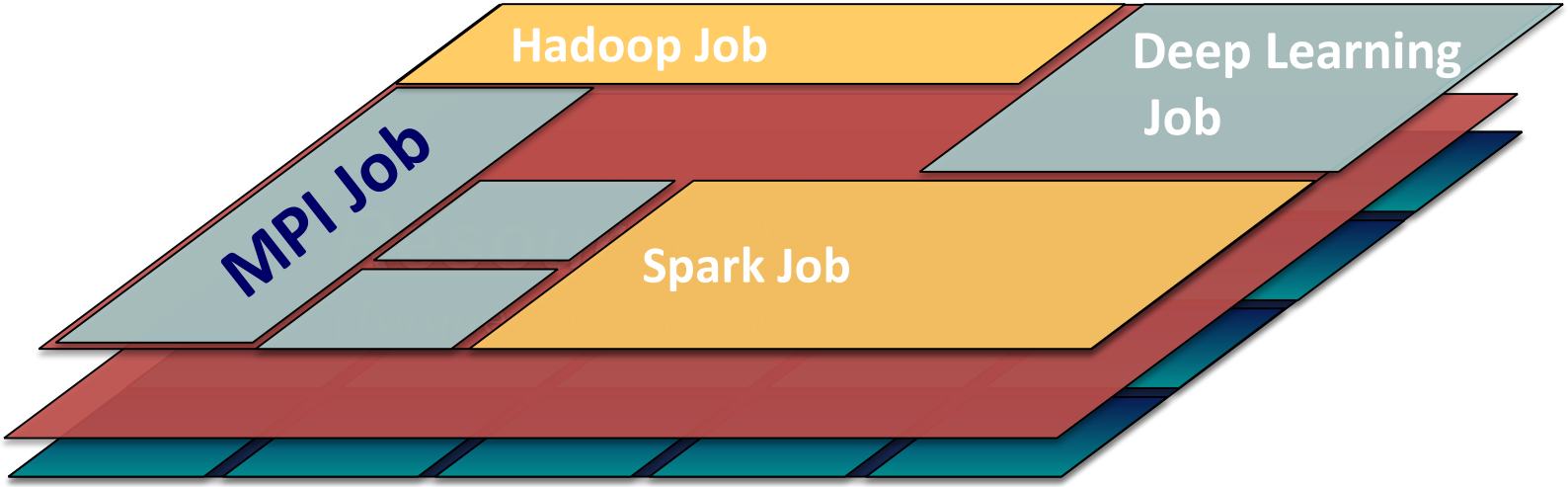
Can We Run HPC, Big Data and Deep Learning Jobs on Existing HPC Infrastructure?

Can We Run HPC, Big Data and Deep Learning Jobs on Existing HPC Infrastructure?

Can We Run HPC, Big Data and Deep Learning Jobs on Existing HPC Infrastructure?



Can We Run HPC, Big Data and Deep Learning Jobs on Existing HPC Infrastructure?



Drivers of Modern HPC Cluster and Data Center Architecture

Multi-/Many-core Processors

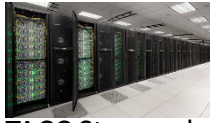
High Performance Interconnects – InfiniBand (with SR-IOV)
<1usec latency, 200Gbps Bandwidth>

Accelerators
high compute density, high performance/watt
>1 TFlop DP on a chip

SSD, NVMe-SSD, NVRAM

- Multi-core/many-core technologies
- Remote Direct Memory Access (RDMA)-enabled networking (InfiniBand, iWARP, RoCE, and Omni-Path)
- Single Root I/O Virtualization (SR-IOV)
- Solid State Drives (SSDs), NVMe/NVMf, Parallel Filesystems, Object Storage Clusters
- Accelerators (NVIDIA GPGPUs and FPGAs)

SDSC Comet



TACC Stampede

Microsoft Azure

amazon web services EC2

ORACLE Cloud

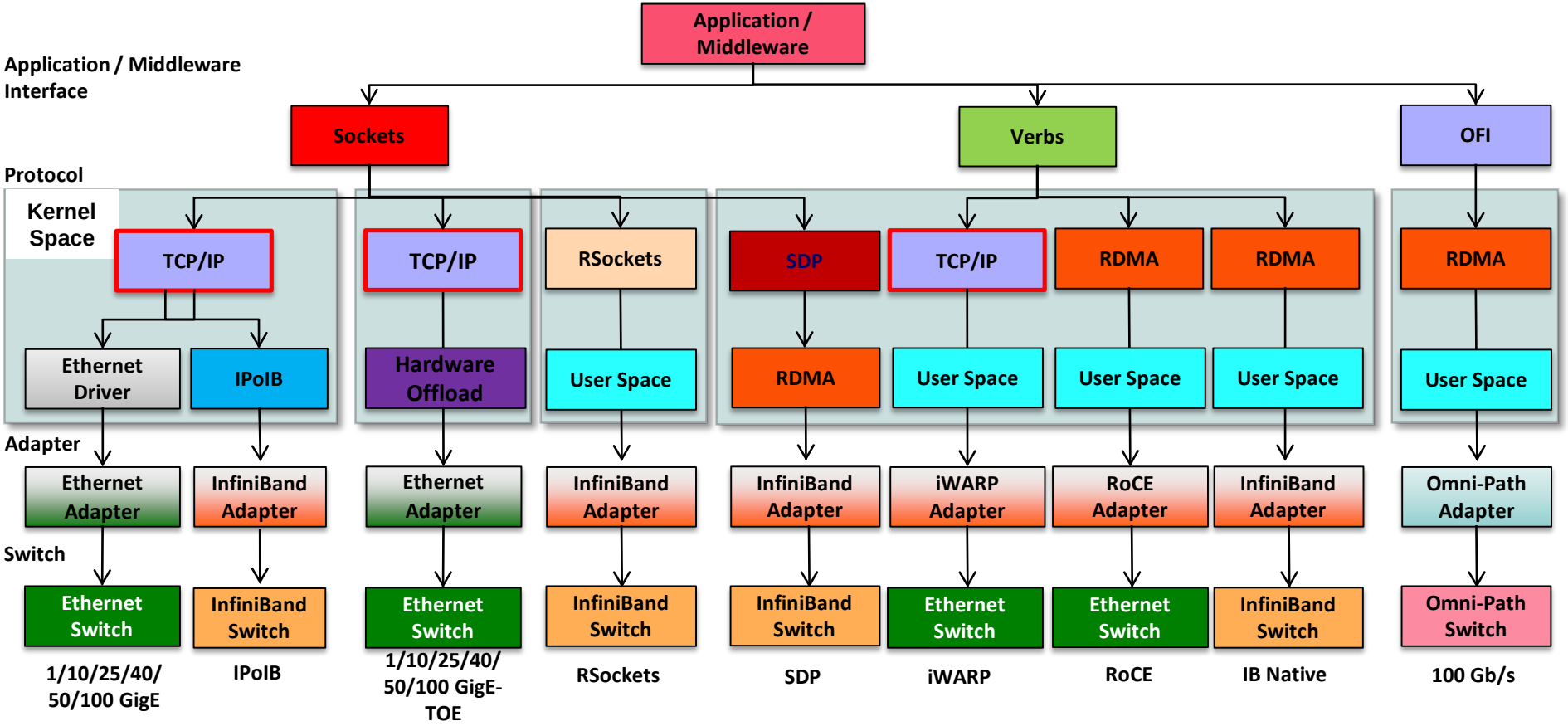
Chameleon Cloud

Trends in Network Speed Acceleration

Ethernet (1979 -)	10 Mbit/sec
Fast Ethernet (1993 -)	100 Mbit/sec
Gigabit Ethernet (1995 -)	1000 Mbit /sec
ATM (1995 -)	155/622/1024 Mbit/sec
Myrinet (1993 -)	1 Gbit/sec
Fibre Channel (1994 -)	1 Gbit/sec
InfiniBand (2001 -)	2 Gbit/sec (1X SDR)
10-Gigabit Ethernet (2001 -)	10 Gbit/sec
InfiniBand (2003 -)	8 Gbit/sec (4X SDR)
InfiniBand (2005 -)	16 Gbit/sec (4X DDR)
	24 Gbit/sec (12X SDR)
InfiniBand (2007 -)	32 Gbit/sec (4X QDR)
40-Gigabit Ethernet (2010 -)	40 Gbit/sec
InfiniBand (2011 -)	54.6 Gbit/sec (4X FDR)
InfiniBand (2012 -)	2 x 54.6 Gbit/sec (4X Dual-FDR)
25-/50-Gigabit Ethernet (2014 -)	25/50 Gbit/sec
100-Gigabit Ethernet (2015 -)	100 Gbit/sec
Omni-Path (2015 -)	100 Gbit/sec
InfiniBand (2015 -)	100 Gbit/sec (4X EDR)
InfiniBand (2016 -)	200 Gbit/sec (4X HDR)

100 times in the last 17 years

Available Interconnects and Protocols for Data Centers



Open Standard InfiniBand Networking Technology

- Introduced in Oct 2000
- High Performance Data Transfer
  - Interprocessor communication and I/O
  - Low latency (<1.0 microsec), High bandwidth (up to 25 GigaBytes/sec -> 200Gbps), and low CPU utilization (5-10%)
- Flexibility for LAN and WAN communication
- Multiple Transport Services
  - Reliable Connection (RC), Unreliable Connection (UC), Reliable Datagram (RD), Unreliable Datagram (UD), and Raw Datagram
  - Provides flexibility to develop upper layers
- Multiple Operations
  - Send/Recv
  - RDMA Read/Write
  - Atomic Operations (very unique)
    - high performance and scalable implementations of distributed locks, semaphores, collective communication operations
- Leading to big changes in designing HPC clusters, file systems, cloud computing systems, grid computing systems, ….

Large-scale InfiniBand Installations

- 163 IB Clusters (32.6%) in the Nov'17 Top500 list
  - (http://www.top500.org)
- Installations in the Top 50 (17 systems):

#1st system also uses InfiniBand

Upcoming US DOE Summit (200 PFlops, to be the new #1st) uses InfiniBand

19,860,000 core (Gyoukou) in Japan (4th)	60,512 core (DGX SATURN V) at NVIDIA/USA (36th)
241,108 core (Pleiades) at NASA/Ames (17th)	72,000 core (HPC2) in Italy (37th)
220,800 core (Pangea) in France (21st)	152,692 core (Thunder) at AFRL/USA (40th)
144,900 core (Cheyenne) at NCAR/USA (24th)	99,072 core (Mistral) at DKRZ/Germany (42nd)
155,150 core (Jureca) in Germany (29th)	147,456 core (SuperMUC) in Germany (44th)
72,800 core Cray CS-Storm in US (30th)	86,016 core (SuperMUC Phase 2) in Germany (45th)
72,800 core Cray CS-Storm in US (31st)	74,520 core (Tsubame 2.5) at Japan/GSIC (48th)
78,336 core (Electra) at NASA/USA (33rd)	66,000 core (HPC3) in Italy (51st)
124,200 core (Topaz) SGI ICE at ERDC DSRC in US (34th)	194,616 core (Cascade) at PNNL (53rd)
60,512 core (NVIDIA DGX-1/Relion) at Facebook in USA (35th)	**and many more!**

High-speed Ethernet Consortium (10GE/25GE/40GE/50GE/100GE)

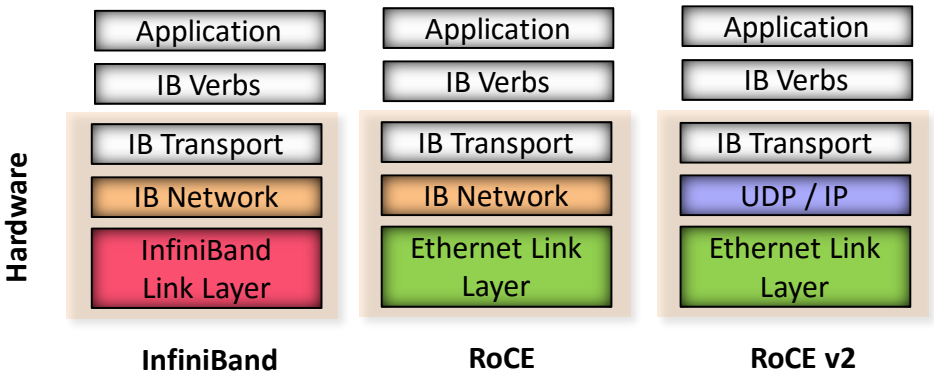
- 10GE Alliance formed by several industry leaders to take the Ethernet family to the next speed step
- Goal: To achieve a scalable and high performance communication architecture while maintaining backward compatibility with Ethernet
- http://www.ethernetalliance.org
- 40-Gbps (Servers) and 100-Gbps Ethernet (Backbones, Switches, Routers): IEEE 802.3 WG
- 25-Gbps Ethernet Consortium targeting 25/50Gbps (July 2014)
  - http://25gethernet.org
- Energy-efficient and power-conscious protocols
  - On-the-fly link speed reduction for under-utilized links
- Ethernet Alliance Technology Forum looking forward to 2026
  - http://insidehpc.com/2016/08/at-ethernet-alliance-technology-forum/

TOE and iWARP Accelerators

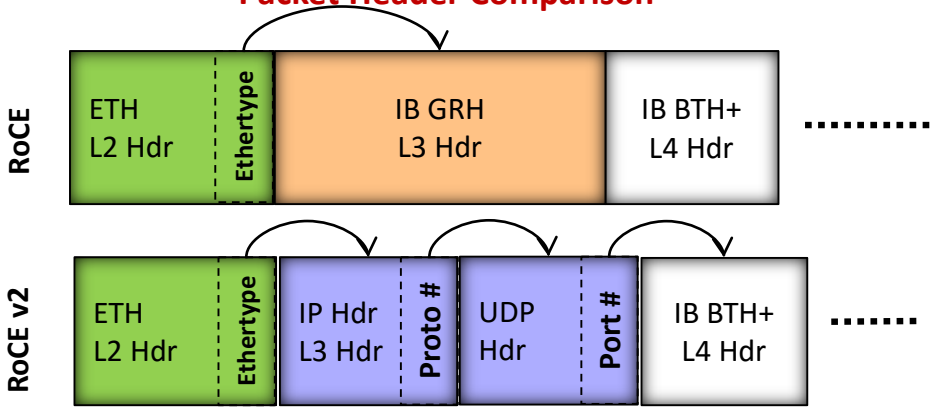
- TCP Offload Engines (TOE)
  - Hardware Acceleration for the entire TCP/IP stack
  - Initially patented by Tehuti Networks
  - Actually refers to the IC on the network adapter that implements TCP/IP
  - In practice, usually referred to as the entire network adapter
- Internet Wide-Area RDMA Protocol (iWARP)
  - Standardized by IETF and the RDMA Consortium
  - Support acceleration features (like IB) for Ethernet
- http://www.ietf.org & http://www.rdmaconsortium.org

RDMA over Converged Enhanced Ethernet (RoCE)

Network Stack Comparison



Packet Header Comparison



Courtesy: OFED, Mellanox

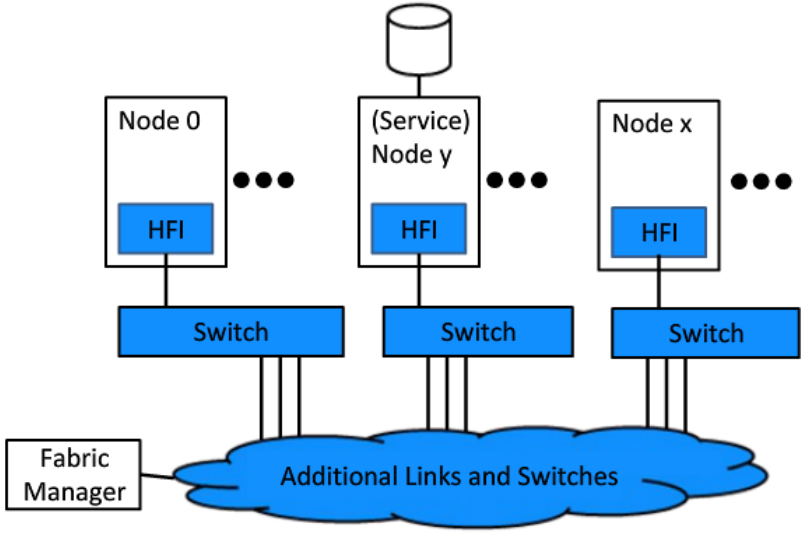
- Takes advantage of IB and Ethernet
  - Software written with IB-Verbs
  - Link layer is Converged (Enhanced) Ethernet (CE)
  - 100Gb/s support from latest EDR and ConnectX-3 Pro adapters
- Pros: IB Vs RoCE
  - Works natively in Ethernet environments
    - Entire Ethernet management ecosystem is available
  - Has all the benefits of IB verbs
  - Link layer is very similar to the link layer of native IB, so there are no missing features
- RoCE v2: Additional Benefits over RoCE
  - Traditional Network Management Tools Apply
  - ACLs (Metering, Accounting, Firewalling)
  - GMP Snooping for Optimized Multicast
  - Network Monitoring Tools

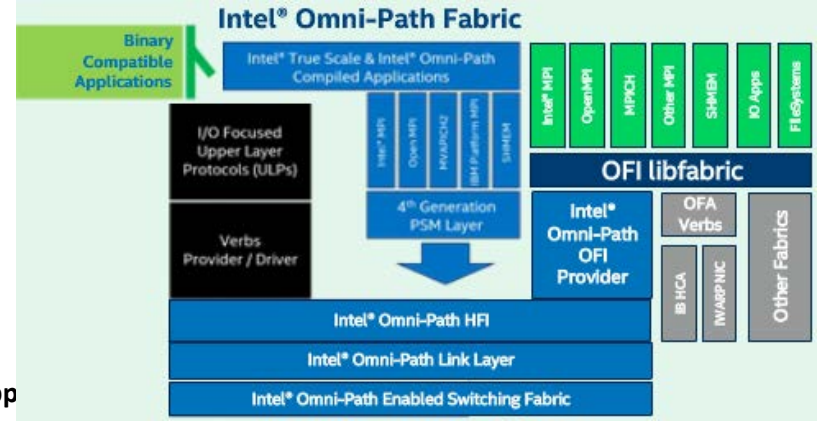
HSE Scientific Computing Installations

- 204 HSE compute systems with ranking in the Nov'17 Top500 list
  - 39,680-core installation in China (#73)
  - 66,560-core installation in China (#101) – new
  - 66,280-core installation in China (#103) – new
  - 64,000-core installation in China (#104) – new
  - 64,000-core installation in China (#105) – new
  - 72,000-core installation in China (#108) – new
  - 78,000-core installation in China (#125)
  - 59,520-core installation in China (#128) – new
  - 59,520-core installation in China (#129) – new
  - 64,800-core installation in China (#130) – new
  - 67,200-core installation in China (#134) – new
  - 57,600-core installation in China (#135) – new
  - 57,600-core installation in China (#136) – new
  - 64,000-core installation in China (#138) – new
  - 84,000-core installation in China (#139)
  - 84,000-core installation in China (#140)
  - 51,840-core installation in China (#151) – new
  - 51,200-core installation in China (#156) – new
  - **and many more!**

Omni-Path Fabric Overview

- Derived from QLogic InfiniBand
- Layer 1.5: Link Transfer Protocol
  - Features
    - Traffic Flow Optimization
    - Packet Integrity Protection
    - Dynamic Lane Switching
  - Error detection/replay occurs in Link Transfer Packet units
  - Retransmit request via NULL LTP; carries replay command flit
- Layer 2: Link Layer
  - Supports 24 bit fabric addresses
  - Allows 10KB of L4 payload; 10,368 byte max packet size
  - Congestion Management
    - Adaptive / Dispersive Routing
    - Explicit Congestion Notification
  - QoS support
    - Traffic Class, Service Level, Service Channel and Virtual Lane
- Layer 3: Data Link Layer
  - Fabric addressing, switching, resource allocation and partitioning supp





Courtesy: Intel Corporation

Large-scale Omni-Path Installations

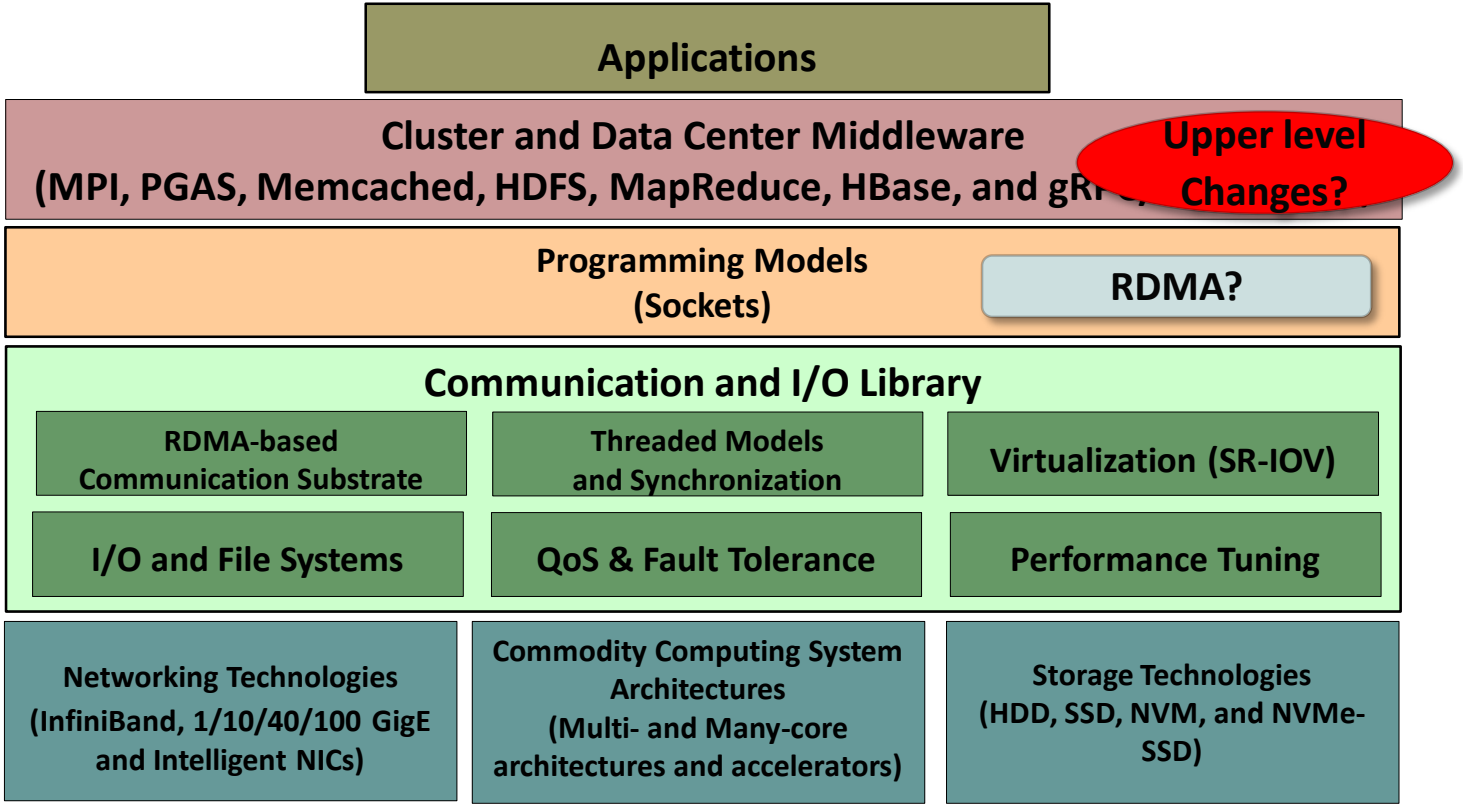
- 35 Omni-Path Clusters (7%) in the Nov'17 Top500 list
  - (http://www.top500.org)

556,104 core (Oakforest-PACS) at JCAHPC in Japan (9th)	54,432 core (Marconi Xeon) at CINECA in Italy (72nd)
368,928 core (Stampede2) at TACC in USA (12th)	46,464 core (Peta4) at University of Cambridge in UK (75th)
135,828 core (Tsubame 3.0) at TiTech in Japan (13th)	53,352 core (Girzzly) at LANL in USA (85th)
314,384 core (Marconi XeonPhi) at CINECA in Italy (14th)	45,680 core (Endeavor) at Intel in USA (86th)
153,216 core (MareNostrum) at BSC in Spain (16th)	59,776 core (Cedar) at SFU in Canada (94th)
95,472 core (Quartz) at LLNL in USA (49th)	27,200 core (Peta HPC) in Taiwan (95th)
95,472 core (Jade) at LLNL in USA (50th)	40,392 core (Serrano) at SNL in USA (112th)
49,432 core (Mogon II) at Universitaet Mainz in Germany (65th)	40,392 core (Cayenne) at SNL in USA (113th)
38,552 core (Molecular Simulator) in Japan (70th)	39,774 core (Nel) at LLNL in USA (101st)
35,280 core (Quriosity) at BASF in Germany (71st)	**and many more!**

IB, Omni-Path, and HSE: Feature Comparison

Features	IB	iWARP/HSE	RoCE	RoCE v2	Omni-Path
Hardware Acceleration	Yes	Yes	Yes	Yes	Yes
RDMA	Yes	Yes	Yes	Yes	Yes
Congestion Control	Yes	Optional	Yes	Yes	Yes
Multipathing	Yes	Yes	Yes	Yes	Yes
Atomic Operations	Yes	No	Yes	Yes	Yes
Multicast	Optional	No	Optional	Optional	Optional
Data Placement	Ordered	Out-of-order	Ordered	Ordered	Ordered
Prioritization	Optional	Optional	Yes	Yes	Yes
Fixed BW QoS (ETS)	No	Optional	Yes	Yes	Yes
Ethernet Compatibility	No	Yes	Yes	Yes	Yes
TCP/IP Compatibility	Yes (using IPoIB)	Yes	Yes (using IPoIB)	Yes	Yes

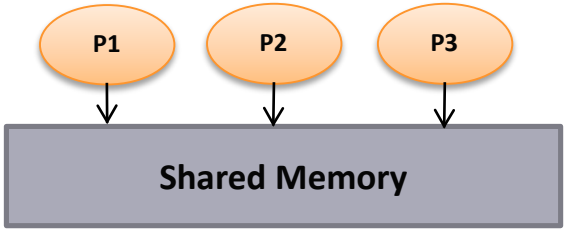
Designing Communication and I/O Libraries for Clusters and Data Center Middleware: Challenges



Designing Next-Generation Middleware for Clusters and Datacenters

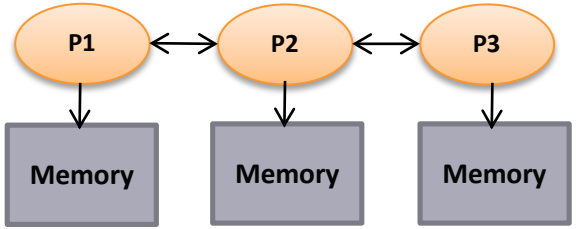
- High-Performance Programming Models Support for HPC Clusters
- RDMA-Enabled Communication Substrate for Common Services in Datacenters
- High-Performance and Scalable Memcached
- RDMA-Enabled Spark and Hadoop (HDFS, HBase, MapReduce)
- Deep Learning with Scale-Up and Scale-Out
  - Caffe, CNTK, and TensorFlow
- Virtualization Support with SR-IOV and Containers

Parallel Programming Models Overview



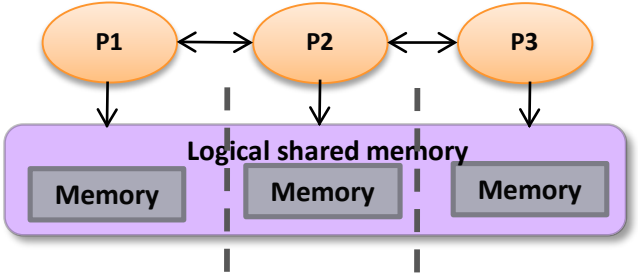
Shared Memory Model

SHMEM, DSM



Distributed Memory Model

MPI (Message Passing Interface)



Partitioned Global Address Space (PGAS)

Global Arrays, UPC, Chapel, X10, CAF, …

- Programming models provide abstract machine models
- Models can be mapped on different types of systems
  - e.g. Distributed Shared Memory (DSM), MPI within a node, etc.
- PGAS models and Hybrid MPI+PGAS models are gradually receiving importance

Supporting Programming Models for Multi-Petaflop and Exaflop Systems: Challenges

Application Kernels/Applications

Middleware

Programming Models
MPI, PGAS (UPC, Global Arrays, OpenSHMEM), CUDA, OpenMP, OpenACC, Cilk, Hadoop (MapReduce), Spark (RDD, DAG), etc.

Communication Library or Runtime for Programming Models

- Point-to-point Communication
- Collective Communication
- Energy-Awareness
- Synchronization and Locks
- I/O and File Systems
- Fault Tolerance

Networking Technologies
(InfiniBand, 40/100GigE, Aries, and Omni-Path)

Multi-/Many-core Architectures

Accelerators (GPU and FPGA)

Co-Design Opportunities and Challenges across Various Layers

Performance
Scalability
Resilience

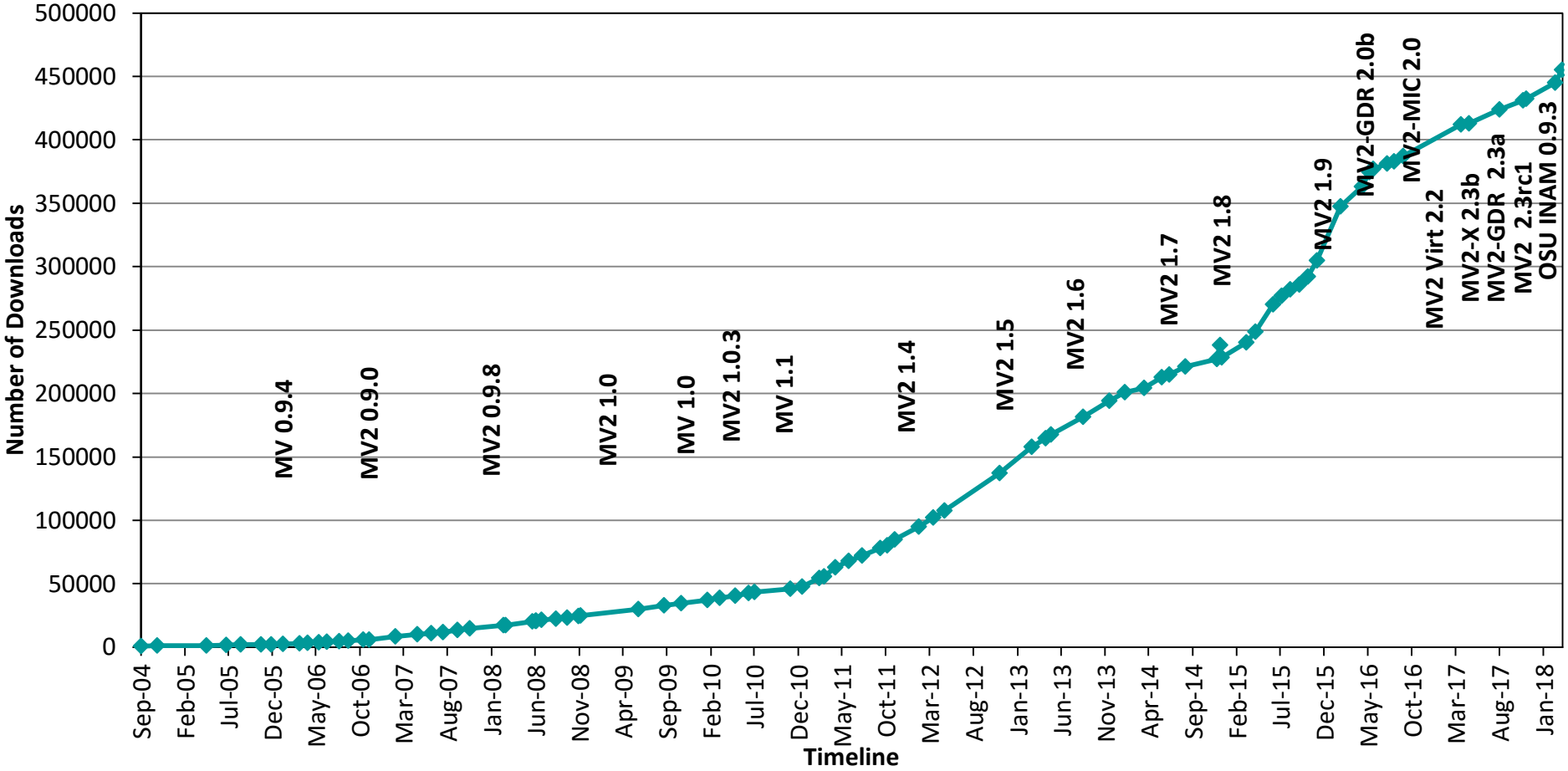
Broad Challenges in Designing Communication Middleware for (MPI+X) at Exascale

- Scalability for million to billion processors
  - Support for highly-efficient inter-node and intra-node communication (both two-sided and one-sided)
  - Scalable job start-up
- Scalable Collective communication
  - Offload
  - Non-blocking
  - Topology-aware
- Balancing intra-node and inter-node communication for next generation nodes (128-1024 cores)
  - Multiple end-points per node
- Support for efficient multi-threading
- Integrated Support for GPGPUs and Accelerators
- Fault-tolerance/resiliency
- QoS support for communication and I/O
- Support for Hybrid MPI+PGAS programming (MPI + OpenMP, MPI + UPC, MPI + OpenSHMEM, MPI+UPC++, CAF, …)
- Virtualization
- Energy-Awareness

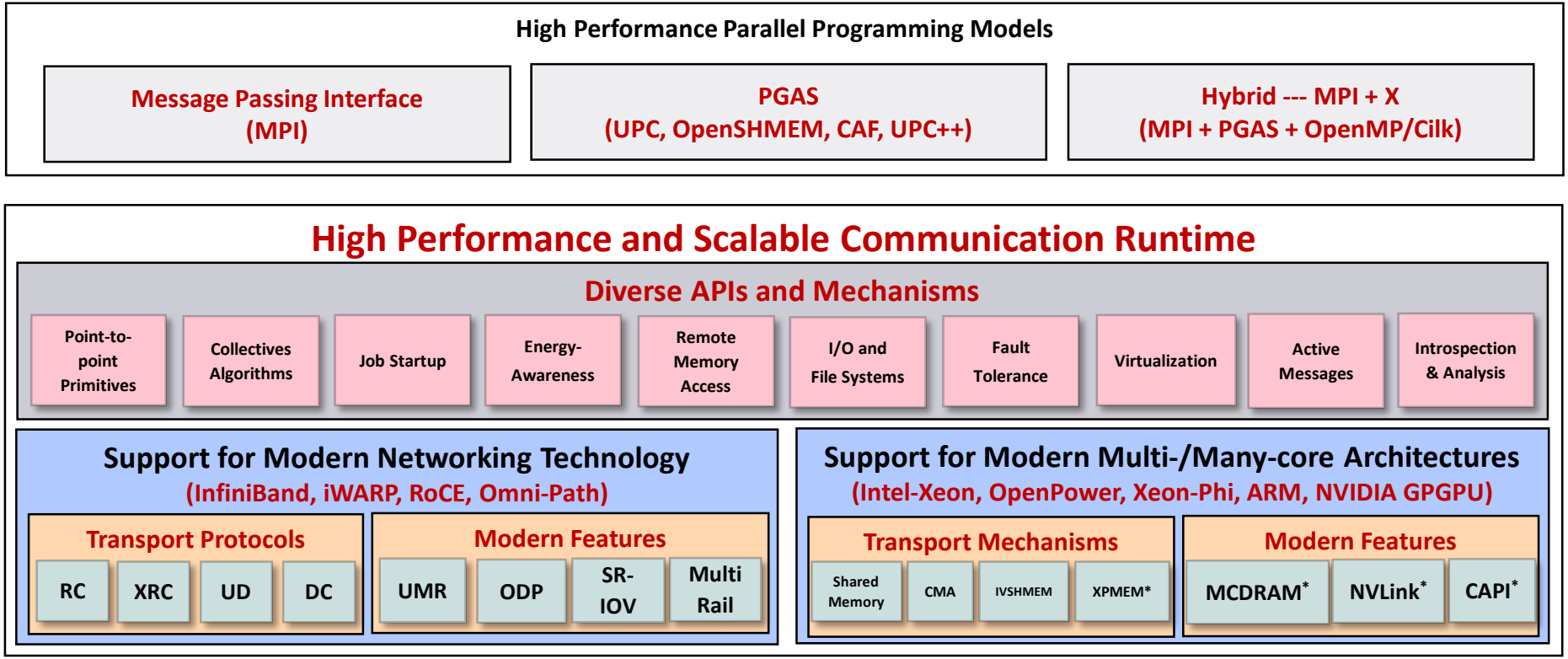
Overview of the MVAPICH2 Project

- 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-3.1), 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 2,900 organizations in 86 countries**
  - **More than 474,000 (> 0.47 million) downloads from the OSU site directl**
  - Empowering many TOP500 clusters (Nov '17 ranking)
    - **1st, 10,649,600-core (Sunway TaihuLight) at National Supercomputing Center in Wuxi, China**
    - 9th, 556,104 cores (Oakforest-PACS) in Japan
    - 12th, 368,928-core (Stampede2) at TACC
    - 17th, 241,108-core (Pleiades) at NASA
    - 48th, 76,032-core (Tsubame 2.5) at Tokyo Institute of Technology
  - 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

MVAPICH2 Release Timeline and Downloads

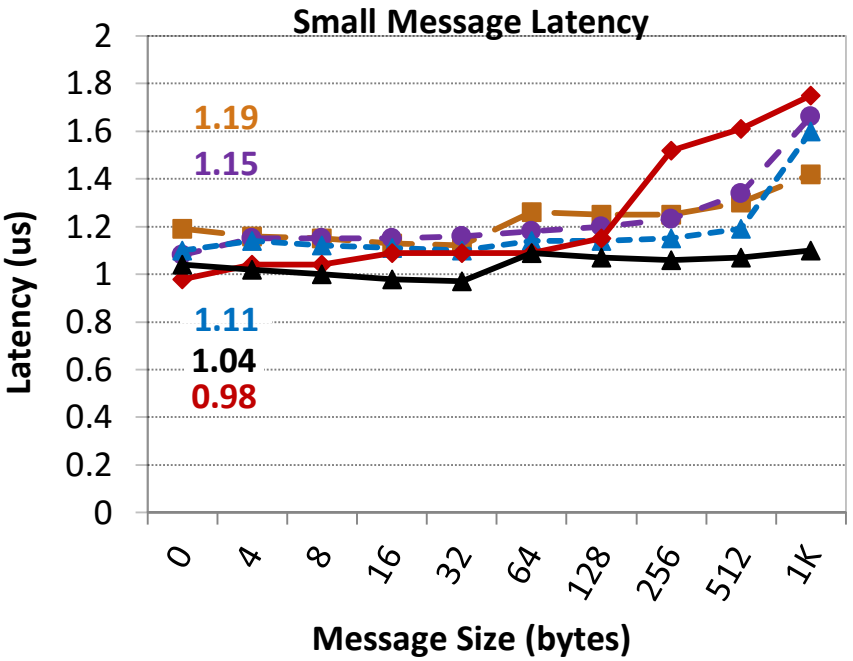


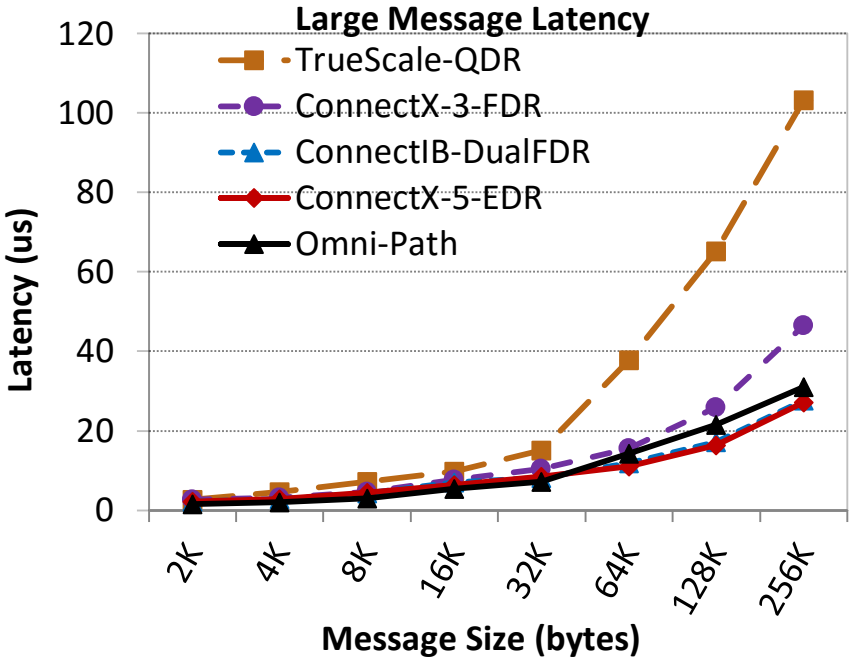
Architecture of MVAPICH2 Software Family



* Upcoming

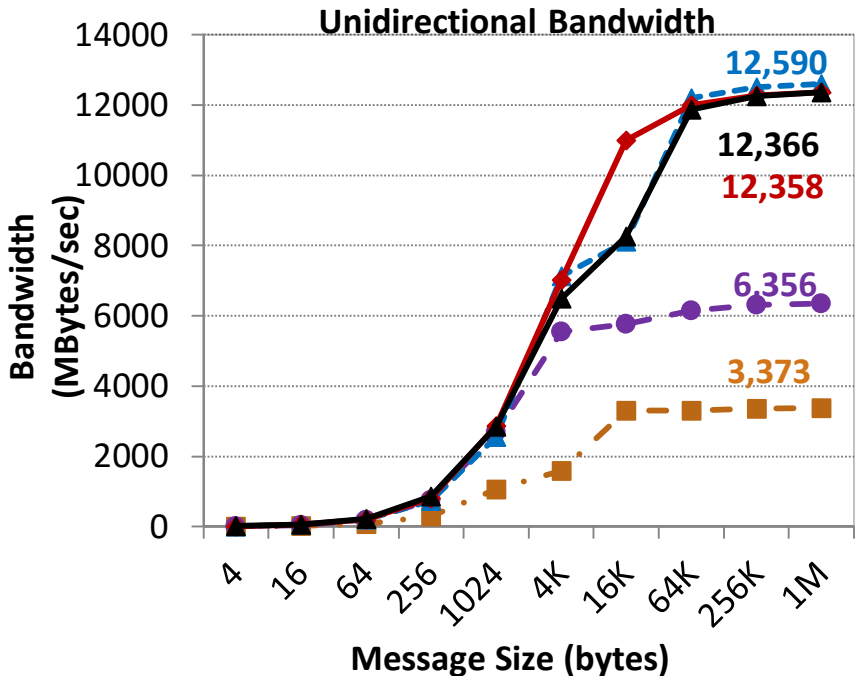
One-way Latency: MPI over IB with MVAPICH2

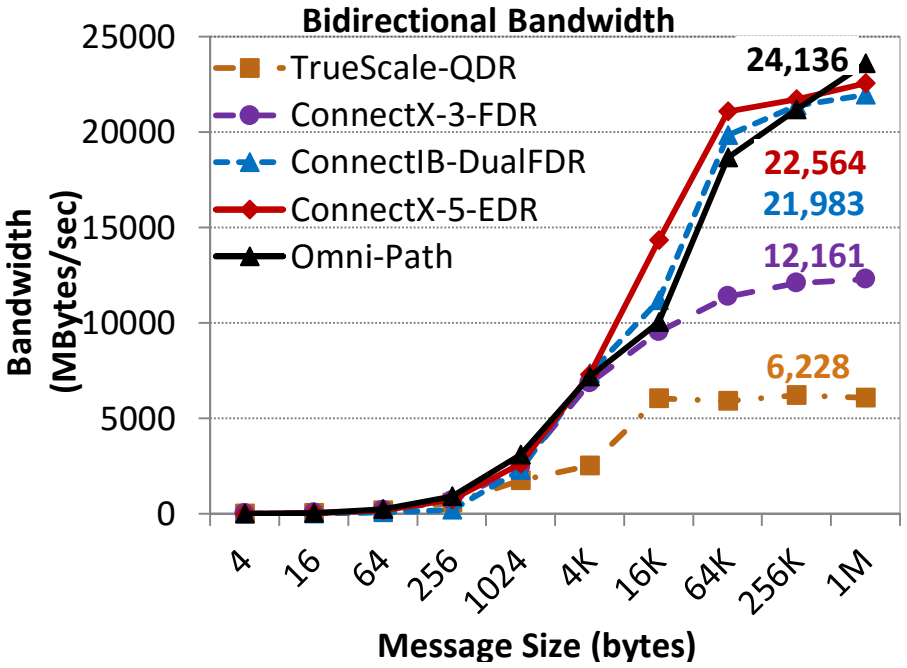




TrueScale-QDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
ConnectX-3-FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch
ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
ConnectX-5-EDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB Switch
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

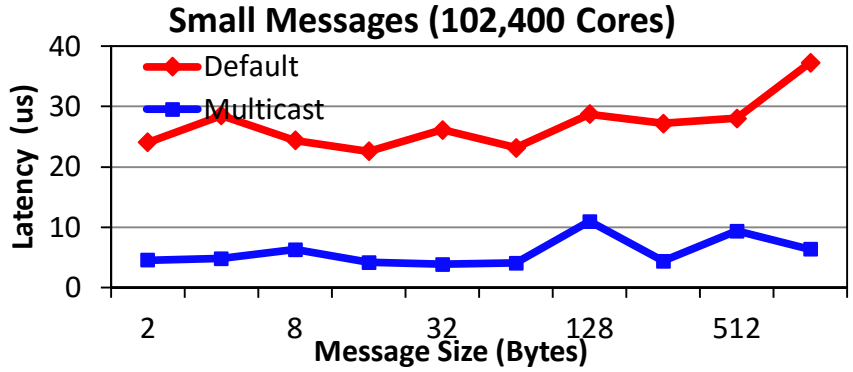
ConnectX-3-FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch

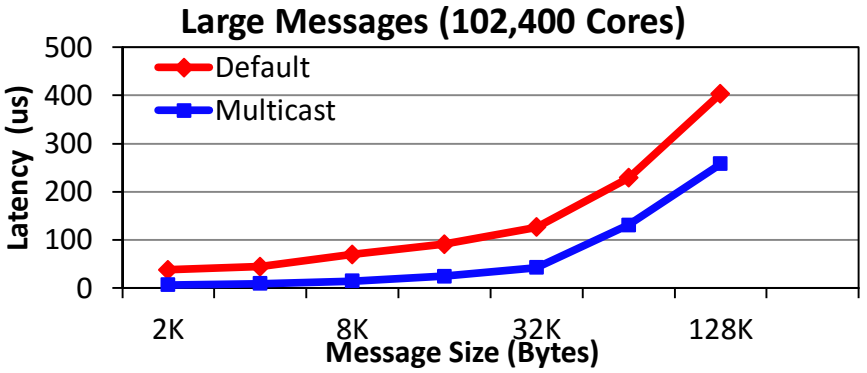
ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

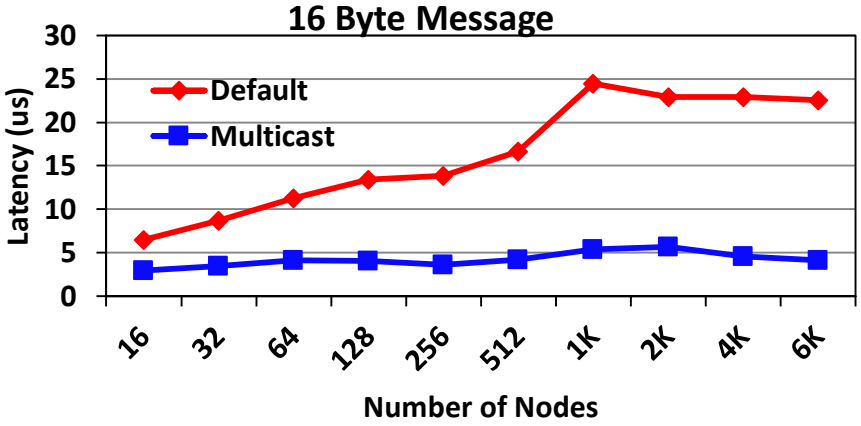
ConnectX-5-EDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 IB switch

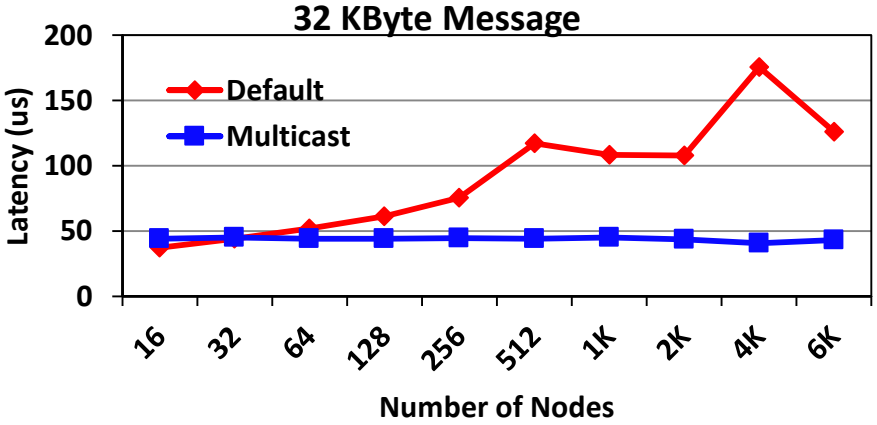
Omni-Path - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with Omni-Path switch

Hardware Multicast-aware MPI_Bcast on Stampede



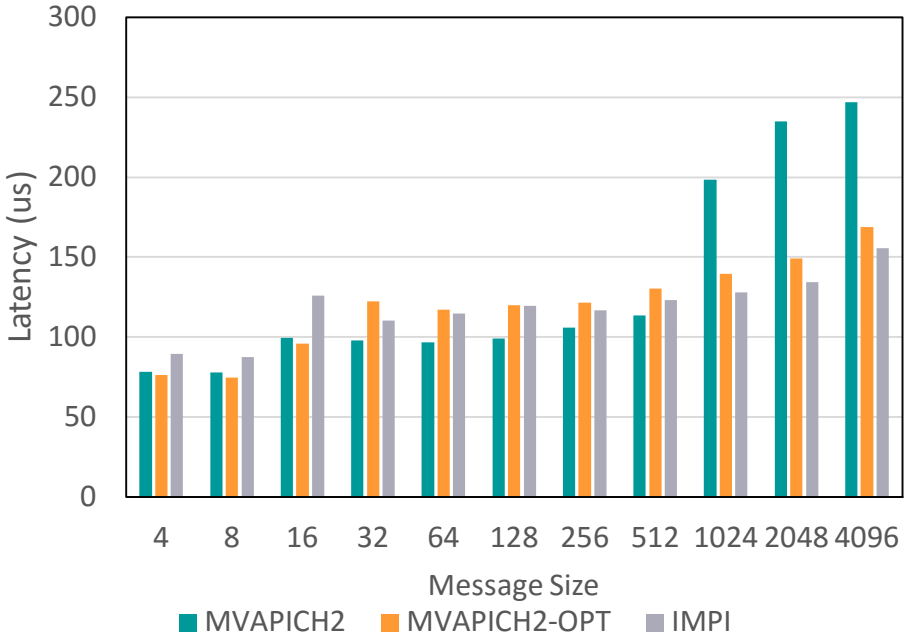


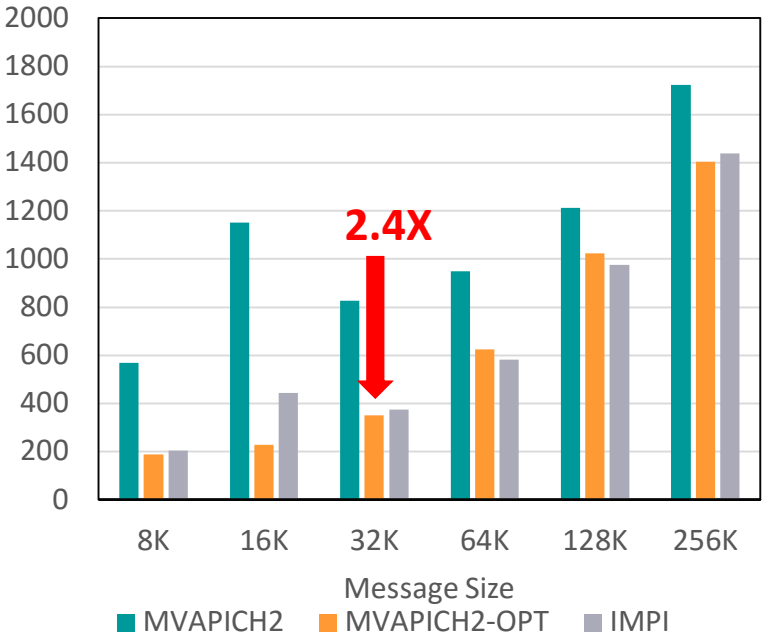




ConnectX-3-FDR (54 Gbps): 2.7 GHz Dual Octa-core (SandyBridge) Intel PCI Gen3 with Mellanox IB FDR switch

Performance of MPI_Allreduce On Stampede2 (10,240 Processes)





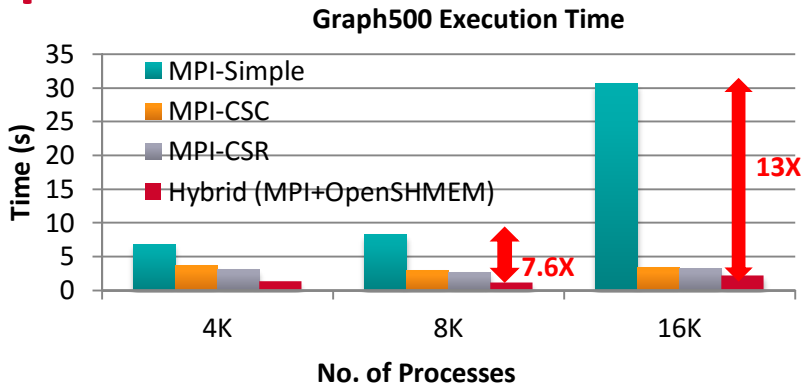
OSU Micro Benchmark 64 PPN

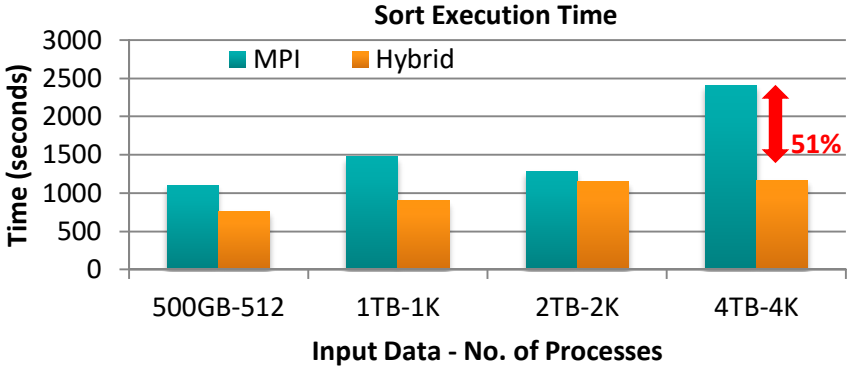
- For MPI_Allreduce latency with 32K bytes, MVAPICH2-OPT can reduce the latency by 2.4X

M. Bayatpour, S. Chakraborty, H. Subramoni, X. Lu, and D. K. Panda, Scalable Reduction Collectives with Data Partitioning-based Multi-Leader Design, SuperComputing '17.

Available in MVAPICH2-X 2.3b

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

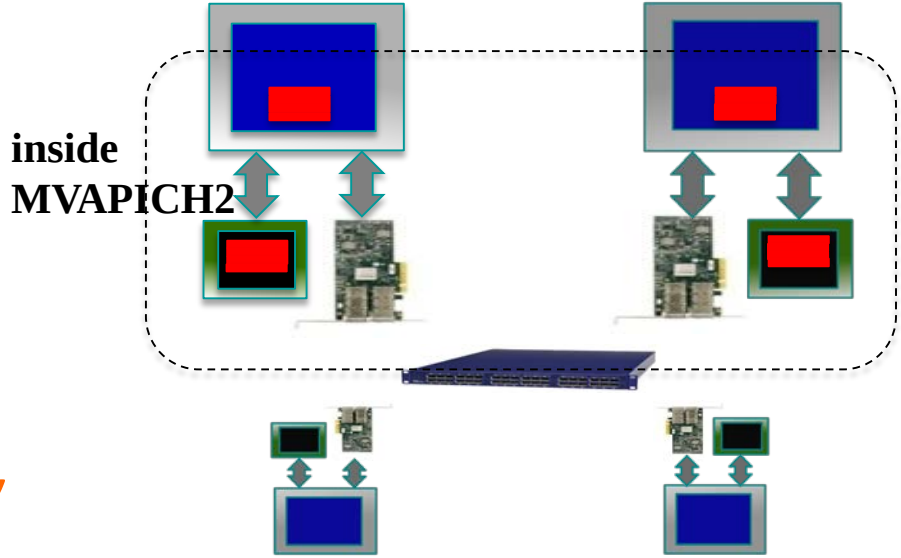
- Standard MPI interfaces used for unified data movement
- Takes advantage of Unified Virtual Addressing (>= CUDA 4.0)
- Overlaps data movement from GPU with RDMA transfers

At Sender:

```
MPI_Send(s_devbuf, size, …);
```

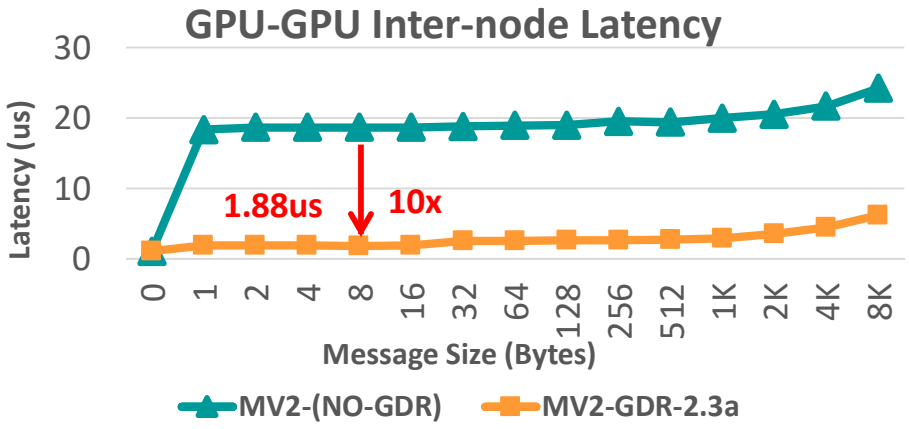
At Receiver:

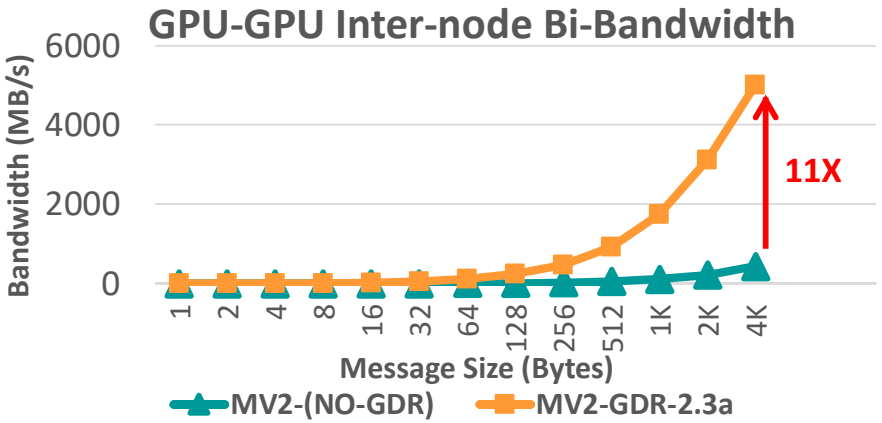
```
MPI_Recv(r_devbuf, size, …);
```

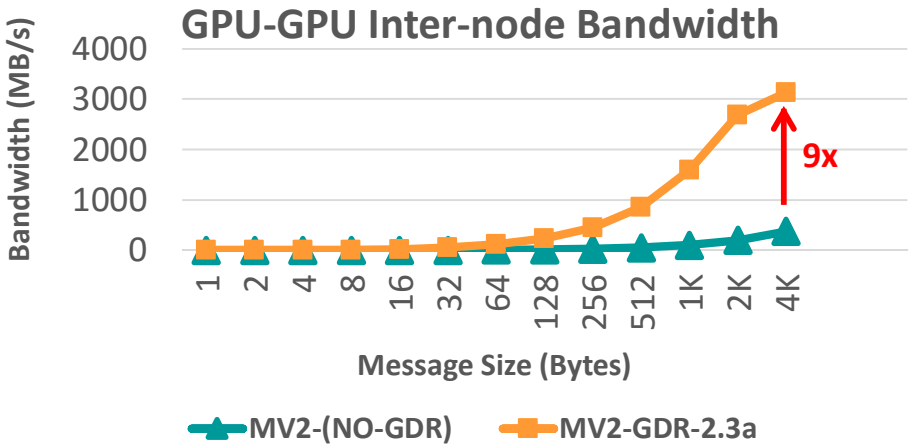


High Performance and High Productivity

Optimized MVAPICH2-GDR Design

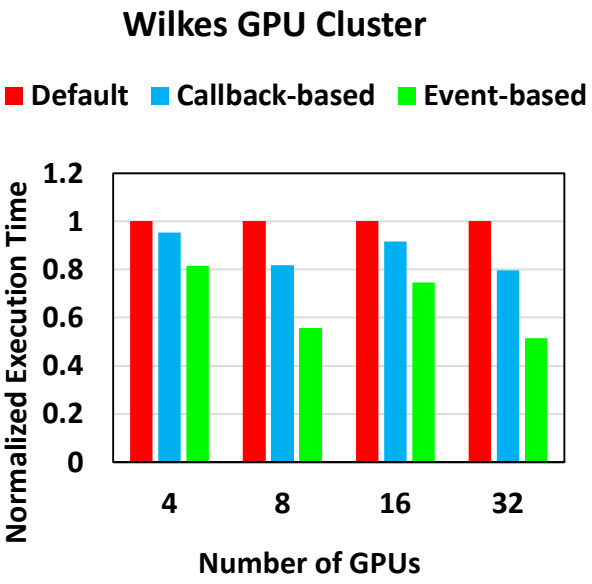


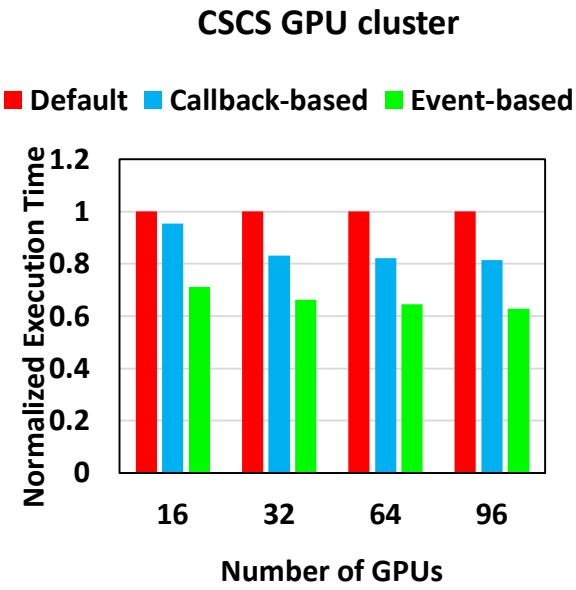


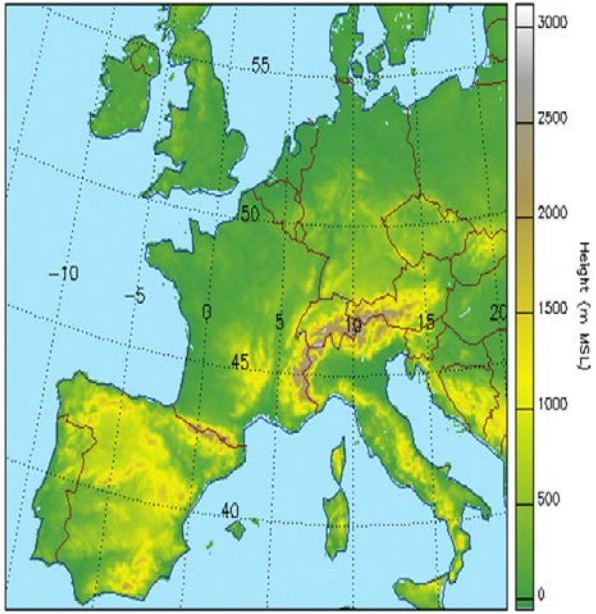


MVAPICH2-GDR-2.3a
Intel Haswell (E5-2687W @ 3.10 GHz) node - 20 cores
NVIDIA Volta V100 GPU
Mellanox Connect-X4 EDR HCA
CUDA 9.0
Mellanox OFED 4.0 with GPU-Direct-RDMA

Application-Level Evaluation (Cosmo) and Weather Forecasting in Switzerland







Cosmo model: http://www2.cosmo-model.org/content/tasks/operational/meteoSwiss/

- 2X improvement on 32 GPUs nodes
- 30% improvement on 96 GPU nodes (8 GPUs/node)

On-going collaboration with CSCS and MeteoSwiss (Switzerland) in co-designing MV2-GDR and Cosmo Application

C. Chu, K. Hamidouche, A. Venkatesh, D. Banerjee , H. Subramoni, and D. K. Panda, Exploiting Maximal Overlap for Non-Contiguous Data Movement Processing on Modern GPU-enabled Systems, IPDPS'16

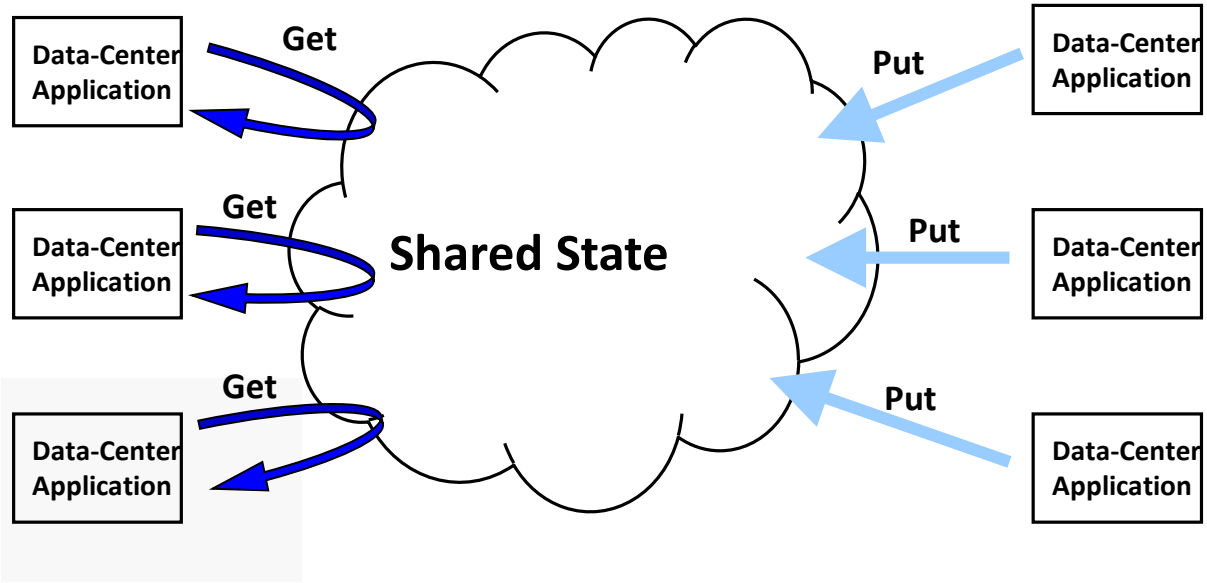
Designing Next-Generation Middleware for Clusters and Datacenters

- High-Performance Programming Models Support for HPC Clusters
- RDMA-Enabled Communication Substrate for Common Services in Datacenters
- High-Performance and Scalable Memcached
- RDMA-Enabled Spark and Hadoop (HDFS, HBase, MapReduce)
- Deep Learning with Scale-Up and Scale-Out
  - Caffe, CNTK, and TensorFlow
- Virtualization Support with SR-IOV and Containers

Data-Center Service Primitives

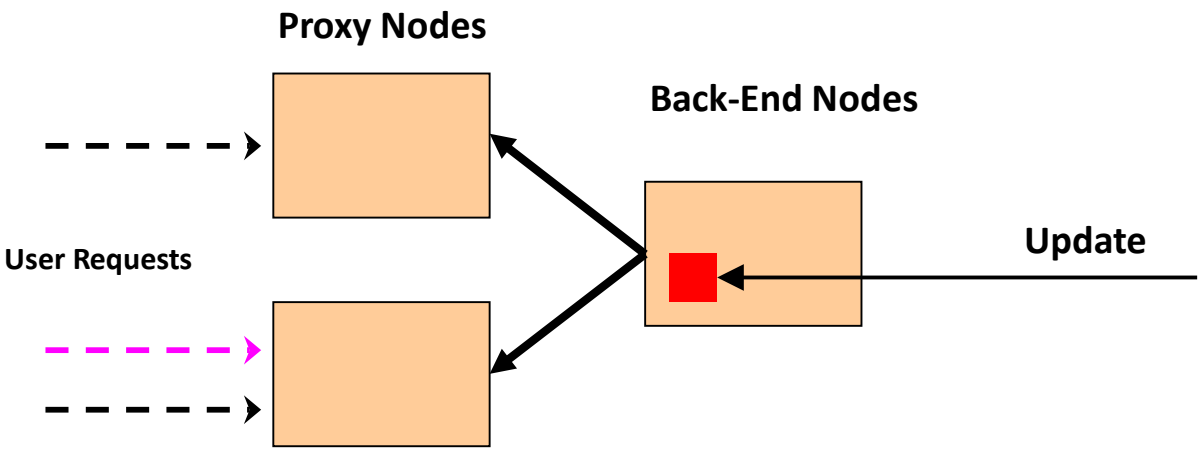
- Common Services needed by Data-Centers
  - Better resource management
  - Higher performance provided to higher layers
- Service Primitives
  - Soft Shared State
  - Distributed Lock Management
  - Global Memory Aggregator
- Network Based Designs
  - RDMA, Remote Atomic Operations

Soft Shared State

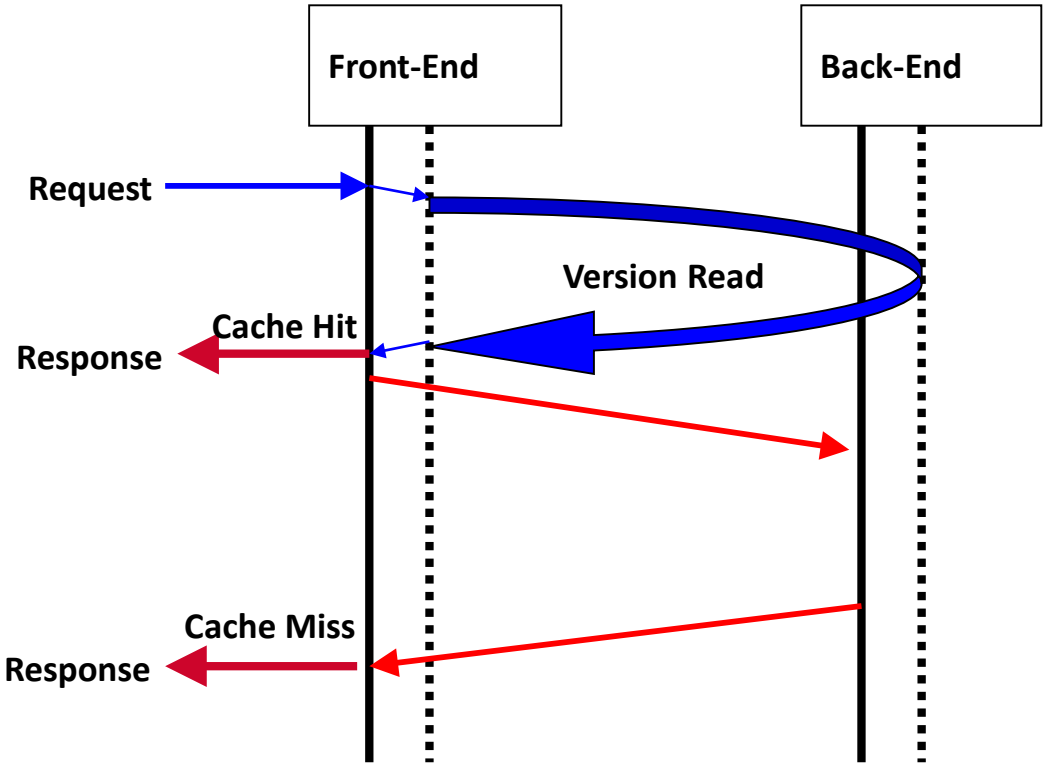


Active Caching

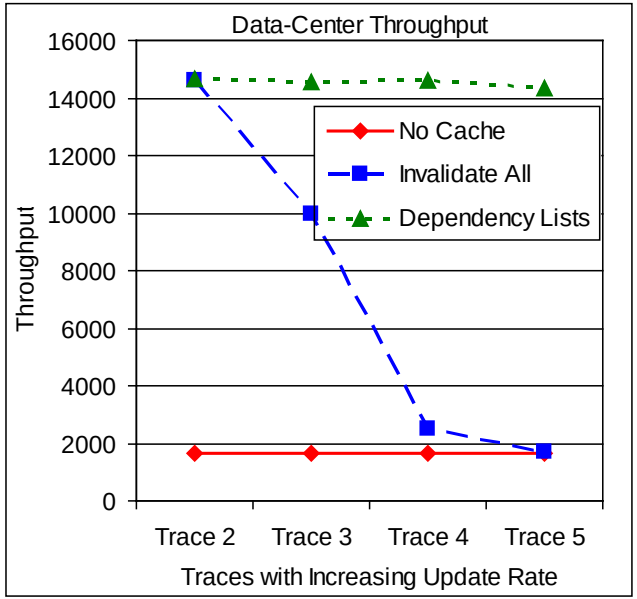
- Dynamic data caching – challenging!
- Cache Consistency and Coherence
  - Become more important than in static case

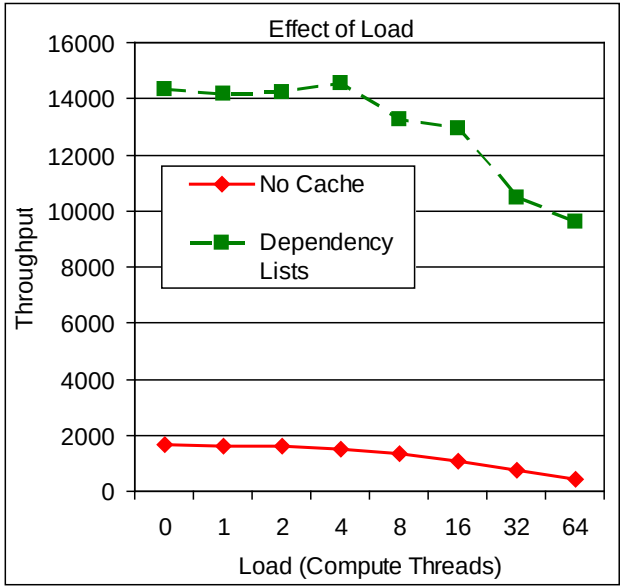


RDMA based Client Polling Design



Active Caching – Performance Benefits





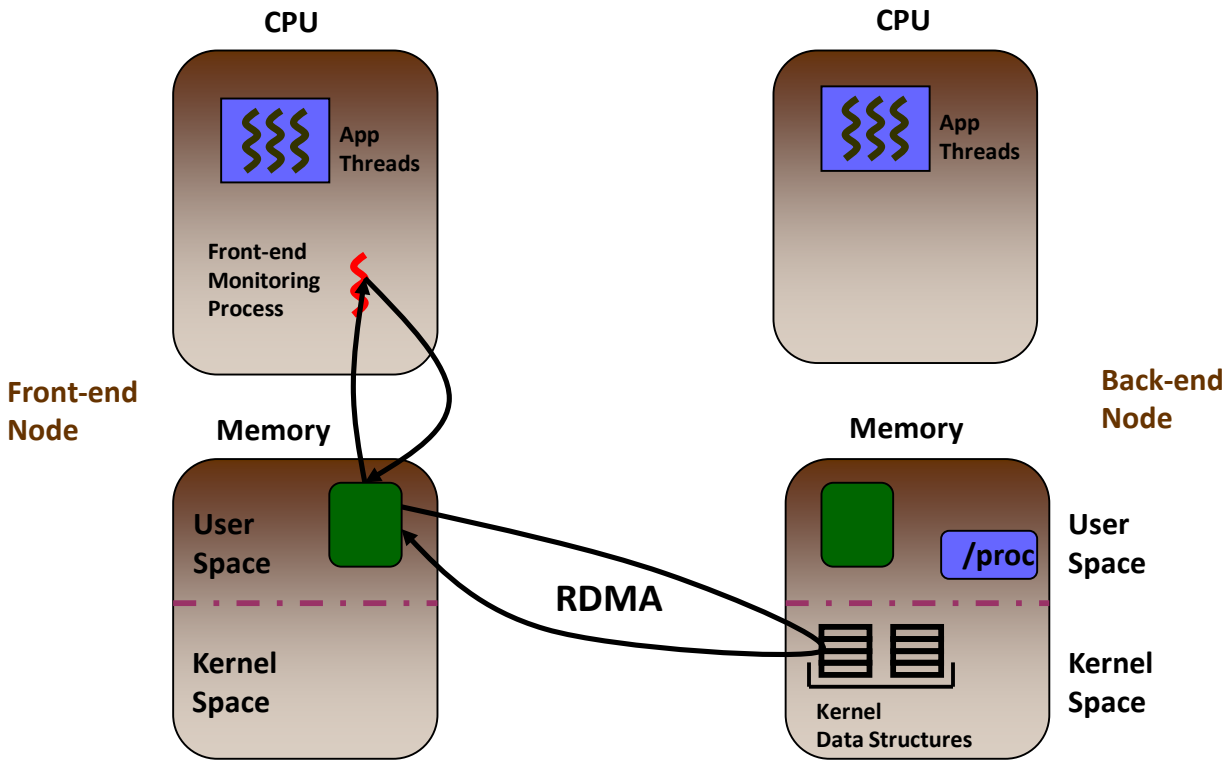
- **Higher overall performance – Up to an order of magnitude**
- **Performance is sustained under loaded conditions**

S. Narravula, P. Balaji, K. Vaidyanathan, H. -W. Jin and D. K. Panda, Architecture for Caching Responses with Multiple Dynamic Dependencies in Multi-Tier Data-Centers over InfiniBand. CCGrid-2005

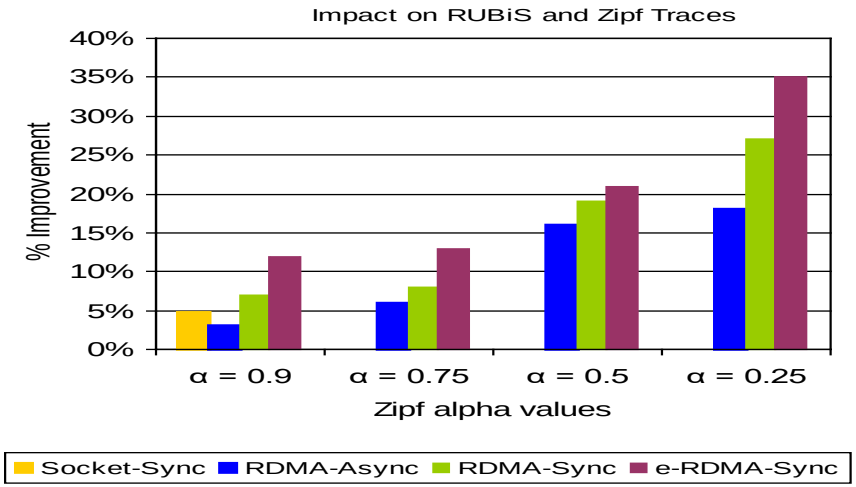
Resource Monitoring Services

- Traditional approaches
  - Coarse-grained in nature
  - Assume resource usage is consistent throughout the monitoring granularity (in the order of seconds)
- This assumption is no longer valid
  - Resource usage is becoming increasingly divergent
- Fine-grained monitoring is desired but has additional overheads
  - High overheads, less accurate, slow in response
- Can we design fine-grained resource monitoring scheme with low overhead and accurate resource usage?

Synchronous Resource Monitoring using RDMA (RDMA-Sync)



Impact of Fine-grained Monitoring with Applications



- Our schemes (RDMA-Sync and e-RDMA-Sync) achieve significant performance gain over existing schemes

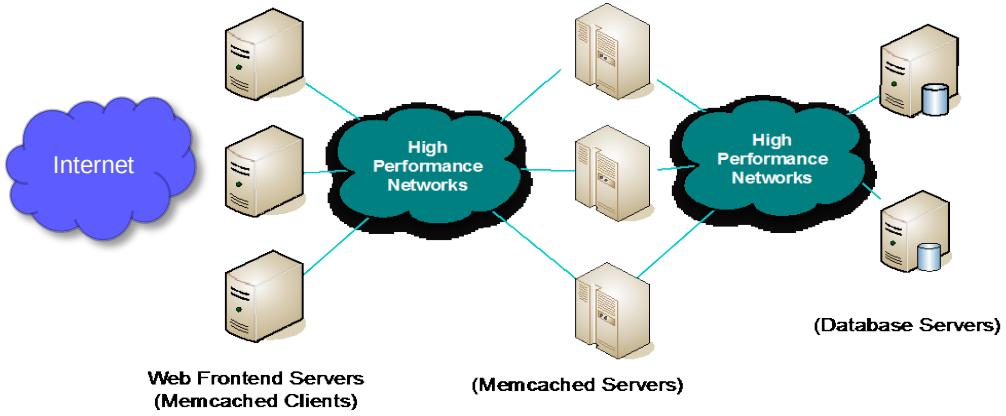
K. Vaidyanathan, H. –W. Jin and D. K. Panda. *Exploiting RDMA operations for Providing Efficient Fine-Grained Resource Monitoring in Cluster-Based Servers,* Workshop on Remote Direct Memory Access (RDMA): Applications, Implementations and Technologies, 2006

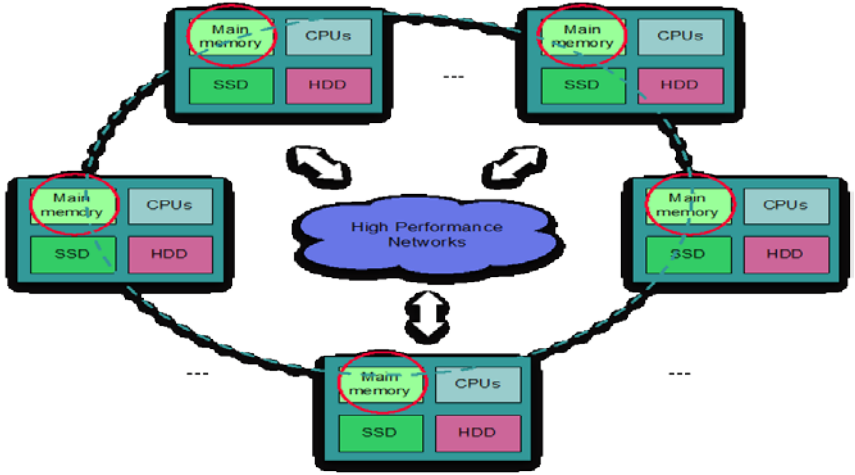
Designing Next-Generation Middleware for Clusters and Datacenters

- High-Performance Programming Models Support for HPC Clusters
- RDMA-Enabled Communication Substrate for Common Services in Datacenters
- High-Performance and Scalable Memcached
- RDMA-Enabled Spark and Hadoop (HDFS, HBase, MapReduce)
- Deep Learning with Scale-Up and Scale-Out
  - Caffe, CNTK, and TensorFlow
- Virtualization Support with SR-IOV and Containers

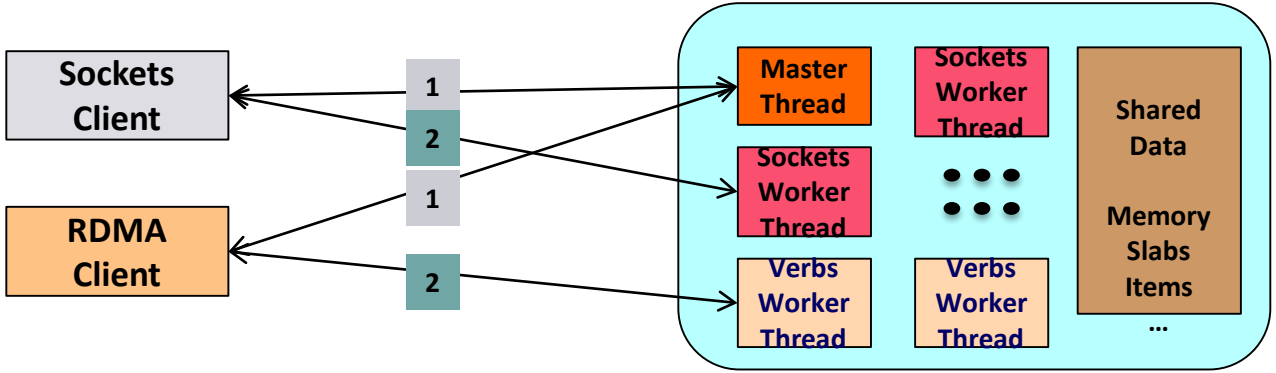
Architecture Overview of Memcached

- Three-layer architecture of Web 2.0
  - Web Servers, Memcached Servers, Database Servers
- Memcached is a core component of Web 2.0 architecture
- Distributed Caching Layer
  - Allows to aggregate spare memory from multiple nodes
  - General purpose
- Typically used to cache database queries, results of API calls
- Scalable model, but typical usage very network intensive



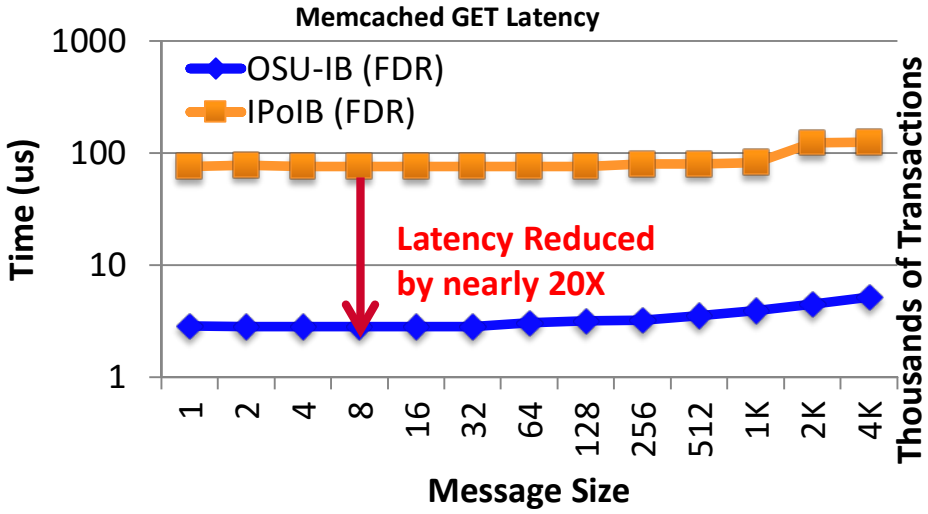


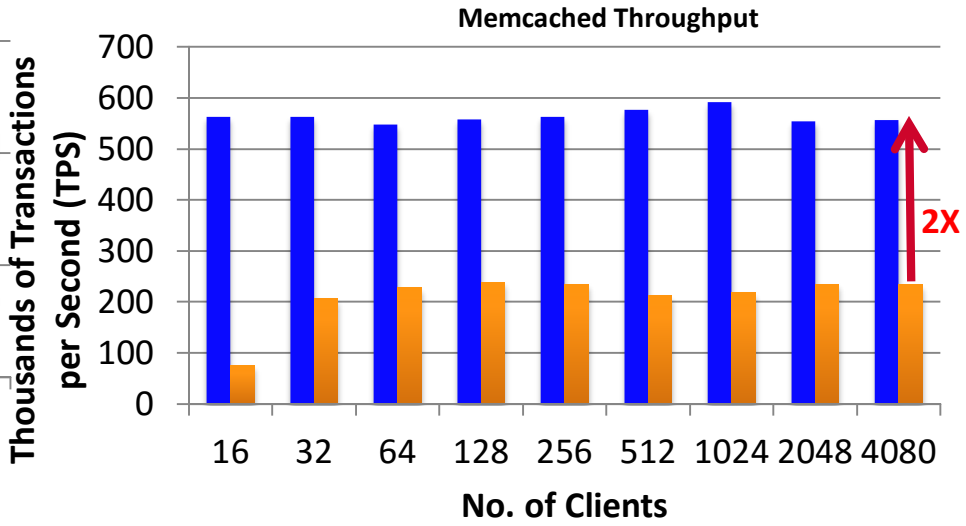
Memcached-RDMA Design



- Server and client perform a negotiation protocol
  - Master thread assigns clients to appropriate worker thread
- Once a client is assigned a verbs worker thread, it can communicate directly and is "bound" to that thread
- All other Memcached data structures are shared among RDMA and Sockets worker threads
- Memcached Server can serve both socket and verbs clients simultaneously
- Memcached applications need not be modified; uses verbs interface if available

Memcached Performance (FDR Interconnect)

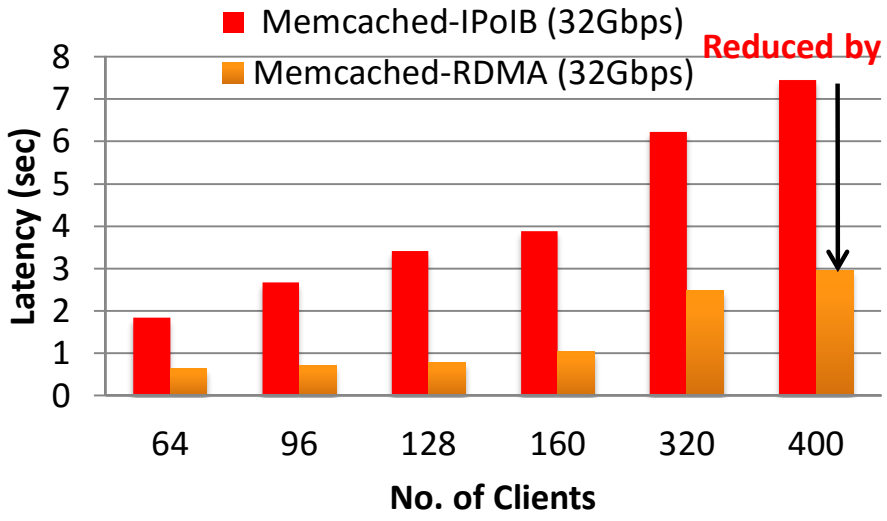


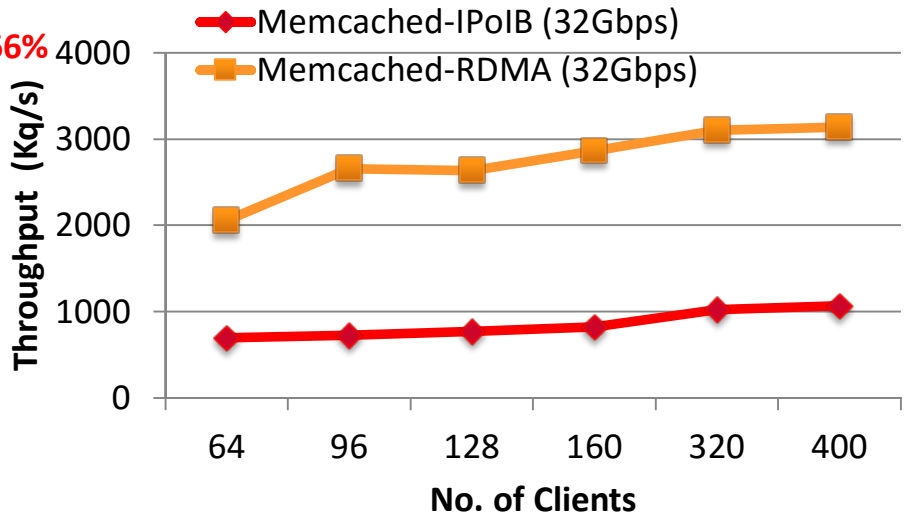


Experiments on TACC Stampede (Intel SandyBridge Cluster, IB: FDR)

- Memcached Get latency
  - 4 bytes OSU-IB: 2.84 us; IPoIB: 75.53 us
  - 2K bytes OSU-IB: 4.49 us; IPoIB: 123.42 us
- Memcached Throughput (4bytes)
  - 4080 clients OSU-IB: 556 Kops/sec, IPoIB: 233 Kops/s
  - Nearly 2X improvement in throughput

Micro-benchmark Evaluation for OLDP workloads

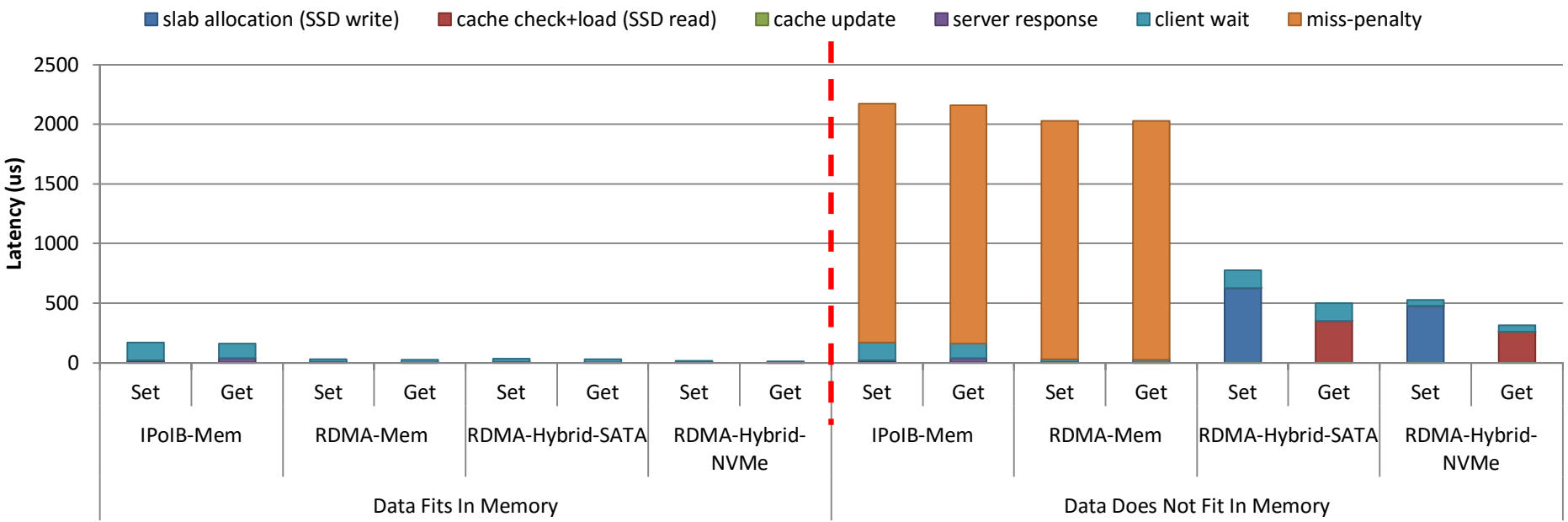




- Illustration with Read-Cache-Read access pattern using modified mysqlslap load testing tool
- Memcached-RDMA can
  - improve query latency by up to 66% over IPoIB (32Gbps)
  - throughput by up to 69% over IPoIB (32Gbps)

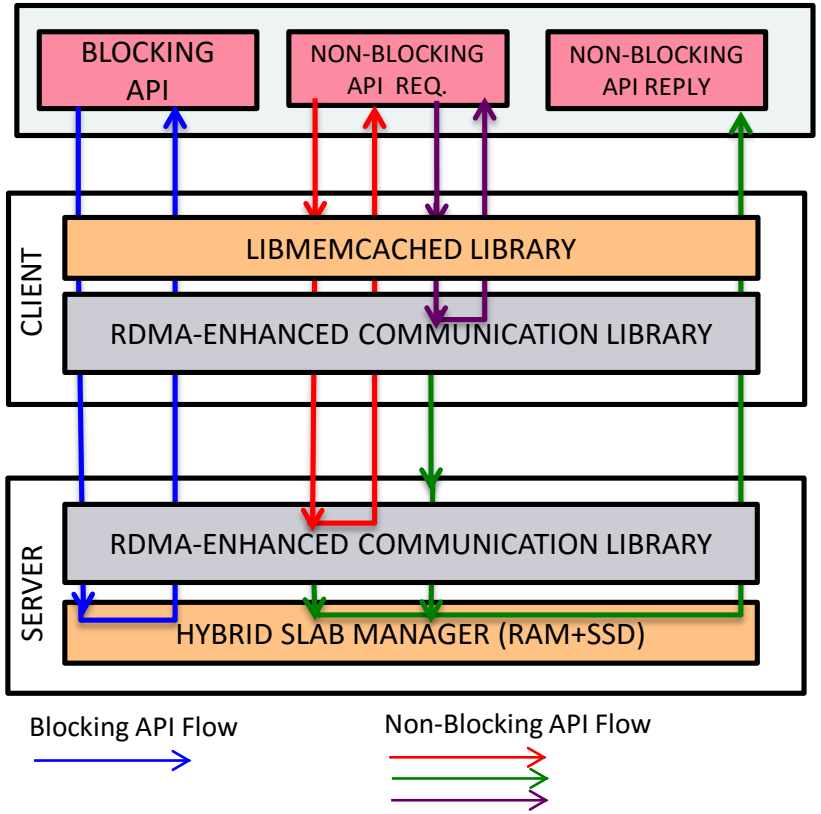
D. Shankar, X. Lu, J. Jose, M. W. Rahman, N. Islam, and D. K. Panda, Can RDMA Benefit On-Line Data Processing Workloads with Memcached and MySQL, ISPASS'15

Performance Evaluation on IB FDR + SATA/NVMe SSDs



- Memcached latency test with Zipf distribution, server with 1 GB memory, 32 KB key-value pair size, total size of data accessed is 1 GB (when data fits in memory) and 1.5 GB (when data does not fit in memory)
- When data fits in memory: RDMA-Mem/Hybrid gives 5x improvement over IPoIB-Mem
- When data does not fit in memory: RDMA-Hybrid gives 2x-2.5x over IPoIB/RDMA-Mem

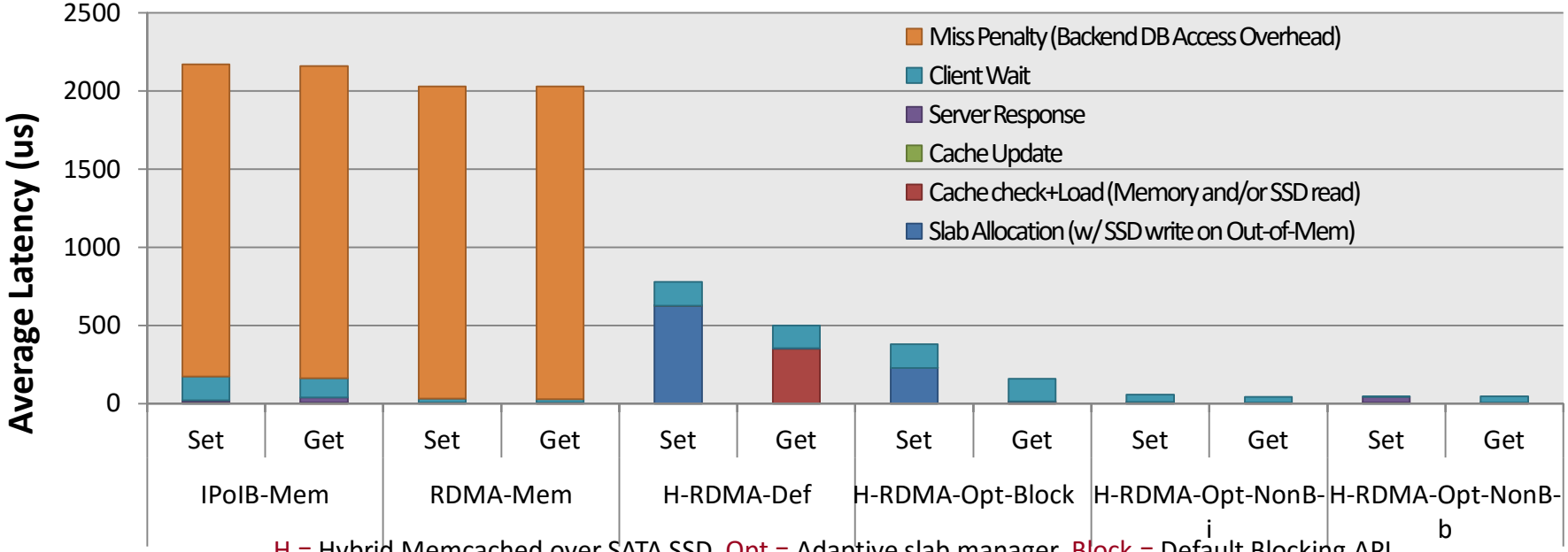
Accelerating Hybrid Memcached with RDMA, Non-blocking Extensions and SSDs



- RDMA-Accelerated Communication for Memcached Get/Set
- Hybrid 'RAM+SSD' slab management for higher data retention
- **Non-blocking API extensions**
  - **memcached_(iset/iget/bset/bget/test/wait)**
  - Achieve near in-memory speeds while hiding bottlenecks of network and SSD I/O
  - Ability to exploit communication/computation overlap
  - Optional buffer re-use guarantees
- Adaptive slab manager with different I/O schemes for higher throughput.

D. Shankar, X. Lu, N. S. Islam, M. W. Rahman, and D. K. Panda, High-Performance Hybrid Key-Value Store on Modern Clusters with RDMA Interconnects and SSDs: Non-blocking Extensions, Designs, and Benefits, IPDPS, May 2016

Performance Evaluation with Non-Blocking Memcached API



H = Hybrid Memcached over SATA SSD Opt = Adaptive slab manager Block = Default Blocking API
NonB-i = Non-blocking iset/iget API NonB-b = Non-blocking bset/bget API w/ buffer re-use guarantee

- Data does not fit in memory: Non-blocking Memcached Set/Get API Extensions can achieve
  - >16x latency improvement vs. blocking API over RDMA-Hybrid/RDMA-Mem w/ penalty
  - >2.5x throughput improvement vs. blocking API over default/optimized RDMA-Hybrid
- Data fits in memory: Non-blocking Extensions perform similar to RDMA-Mem/RDMA-Hybrid and >3.6x improvement over IPoIB-Mem

Designing Next-Generation Middleware for Clusters and Datacenters

- High-Performance Programming Models Support for HPC Clusters
- RDMA-Enabled Communication Substrate for Common Services in Datacenters
- High-Performance and Scalable Memcached
- RDMA-Enabled Spark and Hadoop (HDFS, HBase, MapReduce)
- Deep Learning with Scale-Up and Scale-Out
  - Caffe, CNTK, and TensorFlow
- Virtualization Support with SR-IOV and Containers

The High-Performance Big Data (HiBD) Project

- 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, HBase, and Spark Micro-benchmarks
- http://hibd.cse.ohio-state.edu
- Users Base: 285 organizations from 34 countries
- More than 26,700 downloads from the project site

Available for InfiniBand and RoCE

Also run on Ethernet

Available for x86 and OpenPOWER

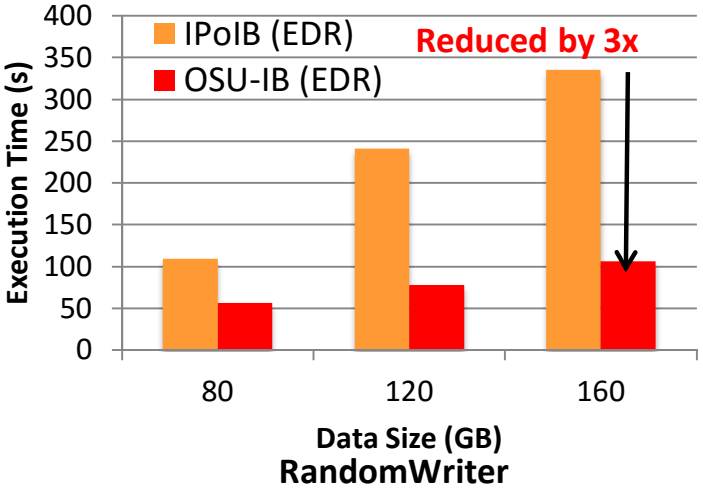
Support for Singularity and Docker

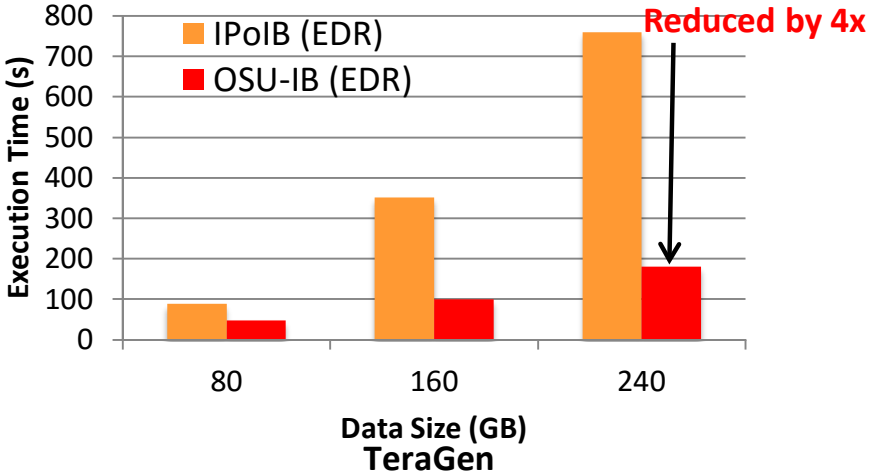
HiBD
High-Performance Big Data

Network Based Computing Laboratory

The Ohio State University

Performance Numbers of RDMA for Apache Hadoop 2.x – RandomWriter & TeraGen in OSU-RI2 (EDR)

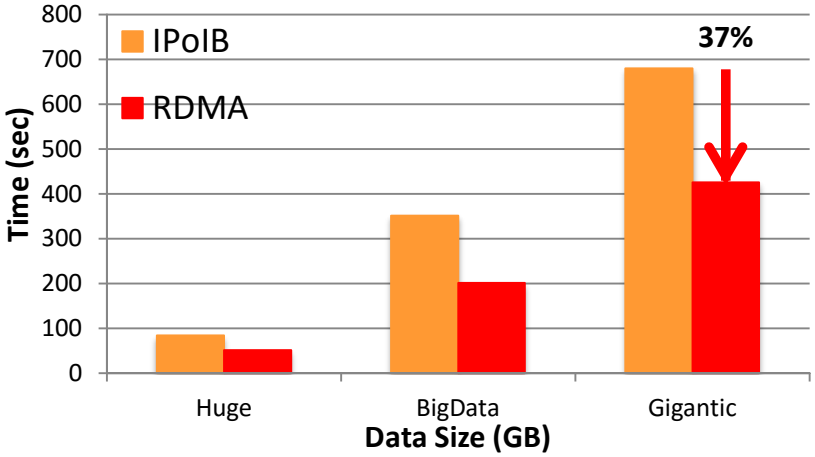




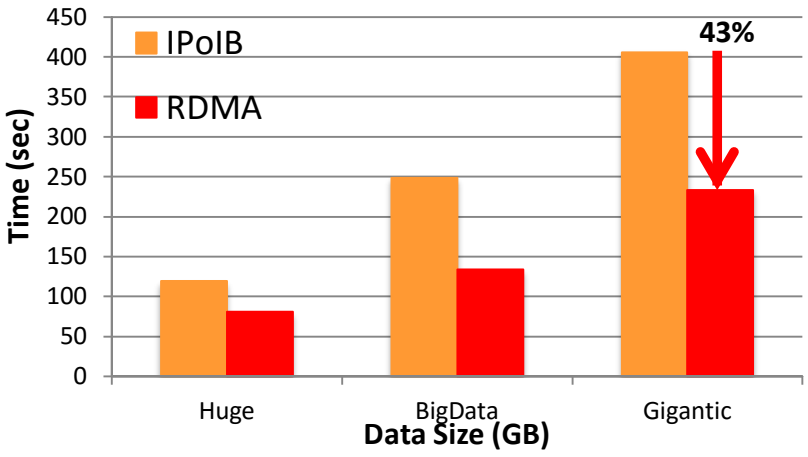
Cluster with 8 Nodes with a total of 64 maps

- RandomWriter
  - **3x** improvement over IPoIB for 80-160 GB file size
- TeraGen
  - **4x** improvement over IPoIB for 80-240 GB file size

Performance Evaluation of RDMA-Spark on SDSC Comet – HiBench PageRank



32 Worker Nodes, 768 cores, PageRank Total Time

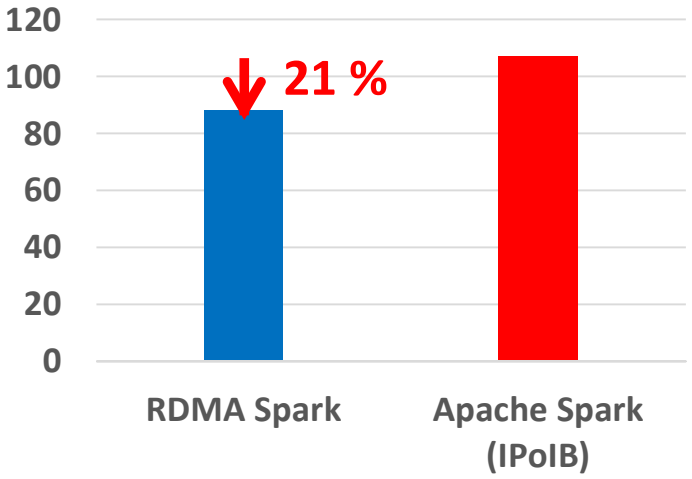


64 Worker Nodes, 1536 cores, PageRank Total Time

- InfiniBand FDR, SSD, 32/64 Worker Nodes, 768/1536 Cores, (768/1536M 768/1536R)
- RDMA 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.

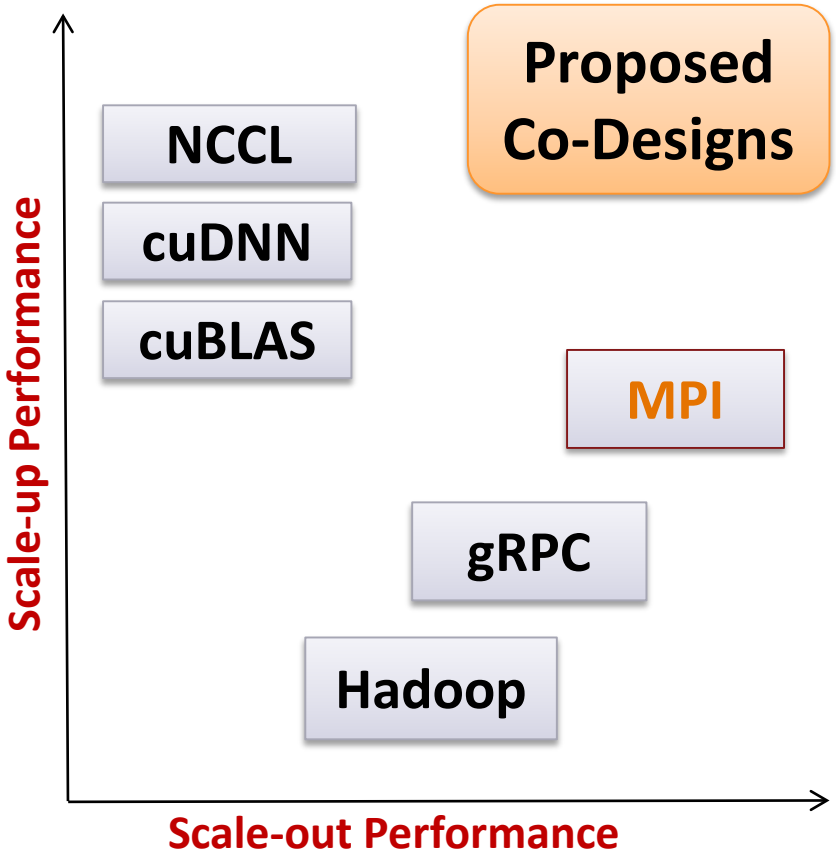
M. Tatineni, X. Lu, D. J. Choi, A. Majumdar, and D. K. Panda, Experiences and Benefits of Running RDMA Hadoop and Spark on SDSC Comet, XSEDE'16, July 2016

Designing Next-Generation Middleware for Clusters and Datacenters

- High-Performance Programming Models Support for HPC Clusters
- RDMA-Enabled Communication Substrate for Common Services in Datacenters
- High-Performance and Scalable Memcached
- RDMA-Enabled Spark and Hadoop (HDFS, HBase, MapReduce)
- Deep Learning with Scale-Up and Scale-Out
  - Caffe, CNTK, and TensorFlow
- Virtualization Support with SR-IOV and Containers

Deep Learning: New Challenges for Communication Runtimes

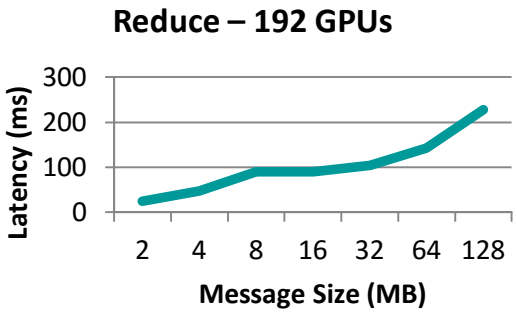
- Deep Learning frameworks are a different game altogether
  - Unusually large message sizes (order of megabytes)
  - Most communication based on GPU buffers
- Existing State-of-the-art
  - cuDNN, cuBLAS, NCCL --> **scale-up** performance
  - CUDA-Aware MPI --> **scale-out** performance
    - For small and medium message sizes only!
- Proposed: Can we **co-design** the MPI runtime (**MVAPICH2-GDR**) and the DL framework (**Caffe**) to achieve both?
  - Efficient **Overlap** of Computation and Communication
  - Efficient **Large-Message** Communication (Reductions)
  - What **application co-designs** are needed to exploit **communication-runtime co-designs**?

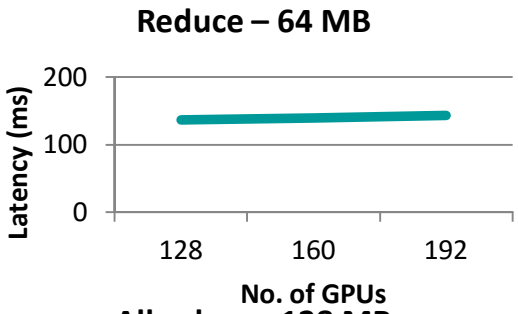


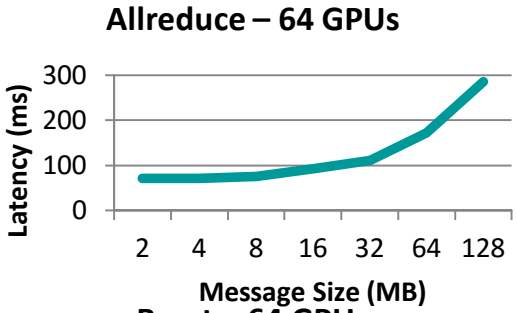
A. A. Awan, K. Hamidouche, J. M. Hashmi, and D. K. Panda, S-Caffe: Co-designing MPI Runtimes and Caffe for Scalable Deep Learning on Modern GPU Clusters. In *Proceedings of the 22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (PPoPP '17)

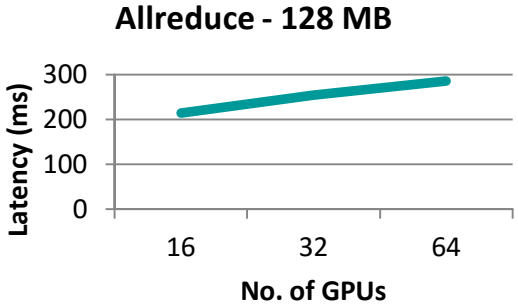
Large Message Optimized Collectives for Deep Learning

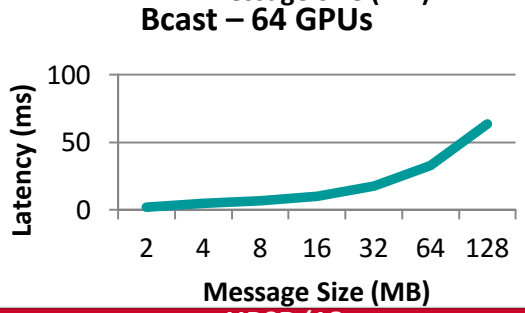
- MV2-GDR provides optimized collectives for **large message sizes**
- Optimized Reduce, Allreduce, and Bcast
- **Good scaling with large number of GPUs**
- **Available since MVAPICH2-GDR 2.2GA**

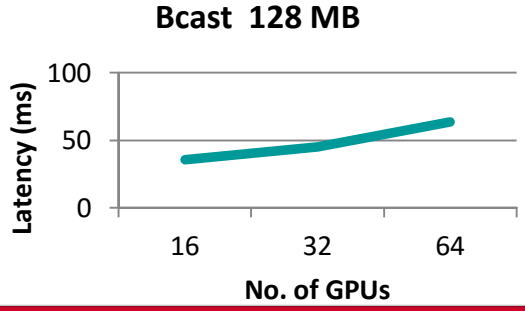






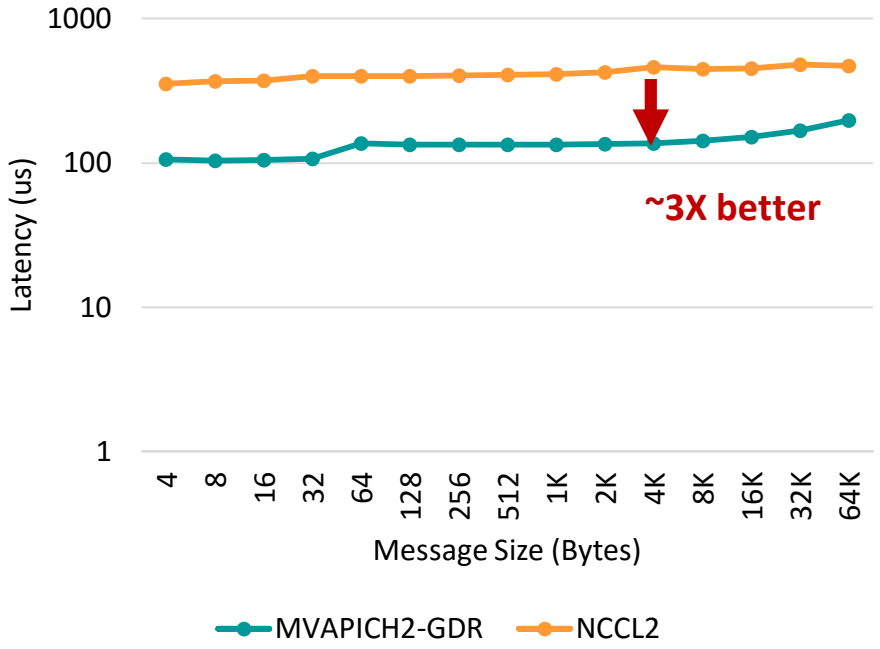


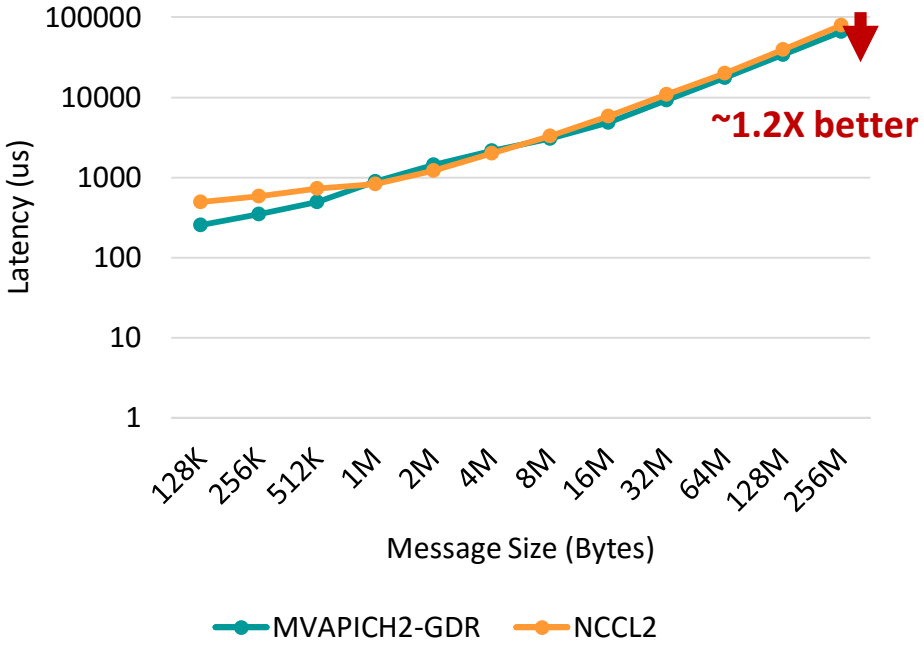




MVAPICH2-GDR vs. NCCL2 – Allreduce Operation

- **Optimized designs in MVAPICH2-GDR 2.3b* offer better/comparable performance for most cases**
- **MPI_Allreduce (MVAPICH2-GDR) vs. ncclAllreduce (NCCL2) on 16 GPUs**



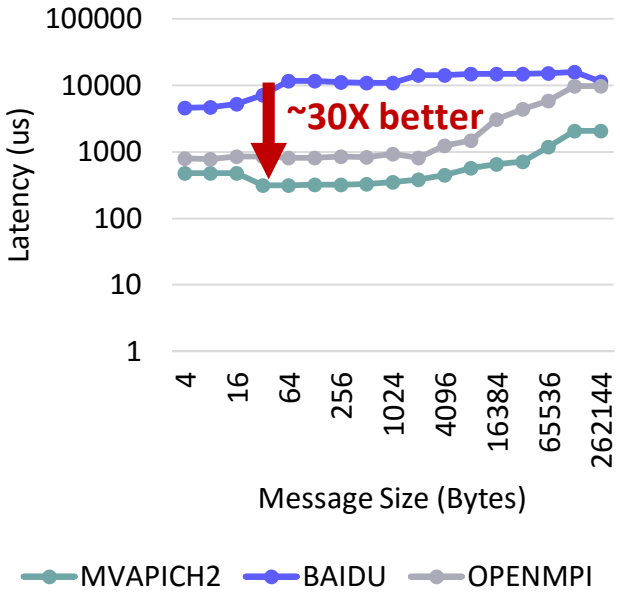


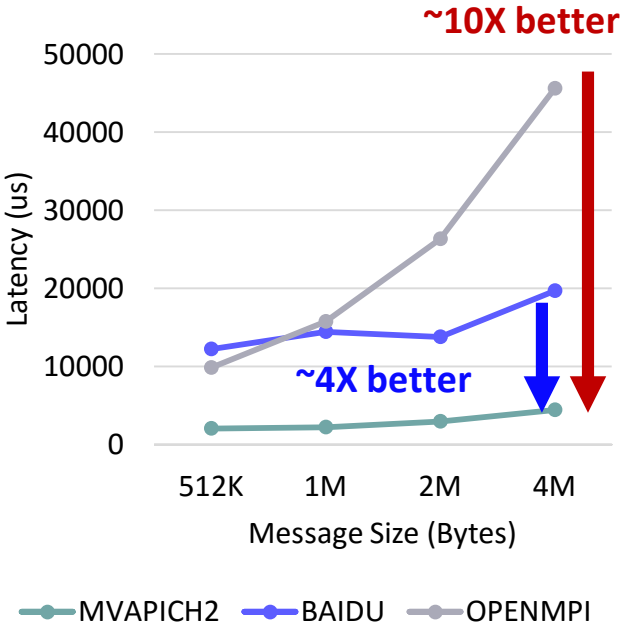
***Will be available with upcoming MVAPICH2-GDR 2.3b**

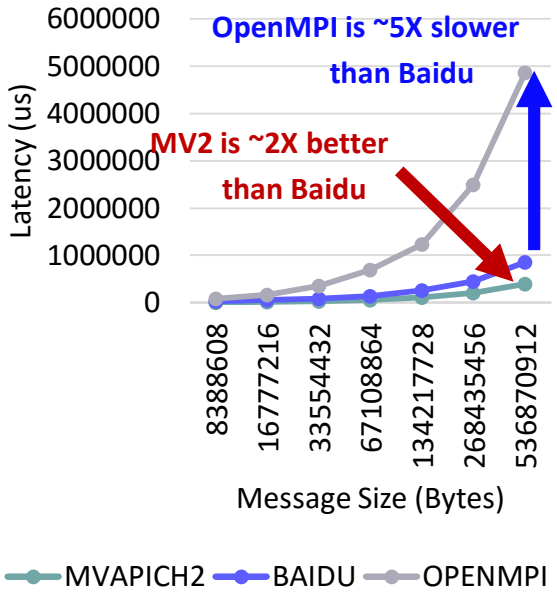
Platform: Intel Xeon (Broadwell) nodes equipped with a dual-socket CPU, 1 K-80 GPUs, and EDR InfiniBand Inter-connect

MVAPICH2: Allreduce Comparison with Baidu and OpenMPI

- 16 GPUs (4 nodes) MVAPICH2-GDR(*) vs. Baidu-Allreduce and OpenMPI 3.0







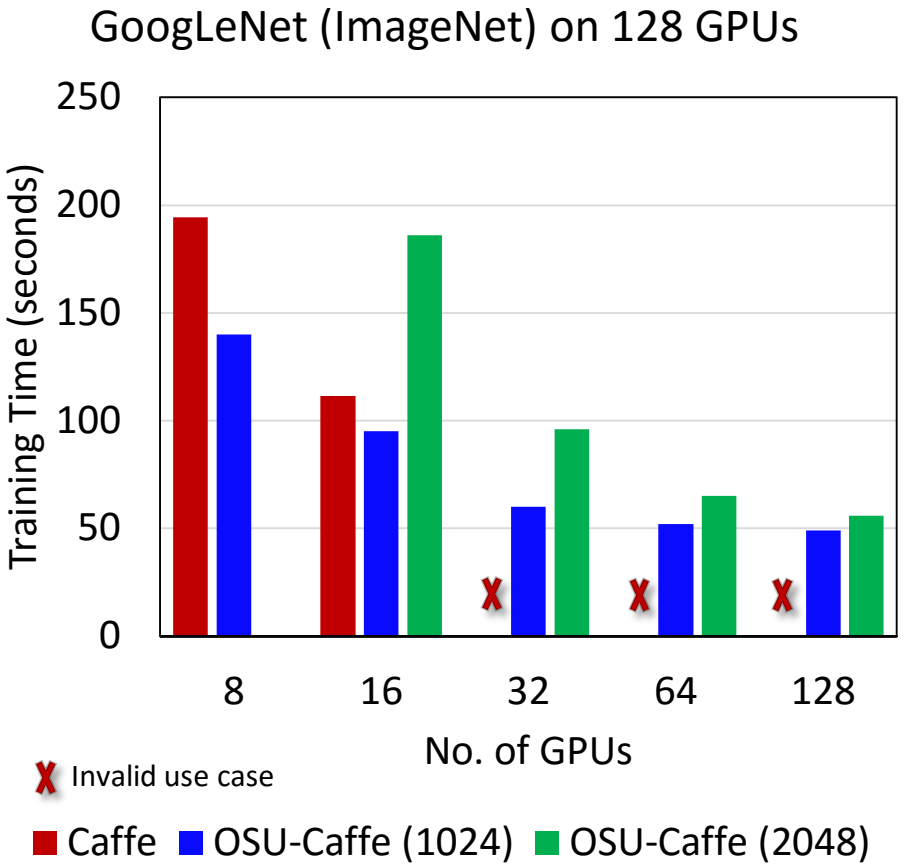
**Available with MVAPICH2-GDR 2.3a*

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 publicly available from

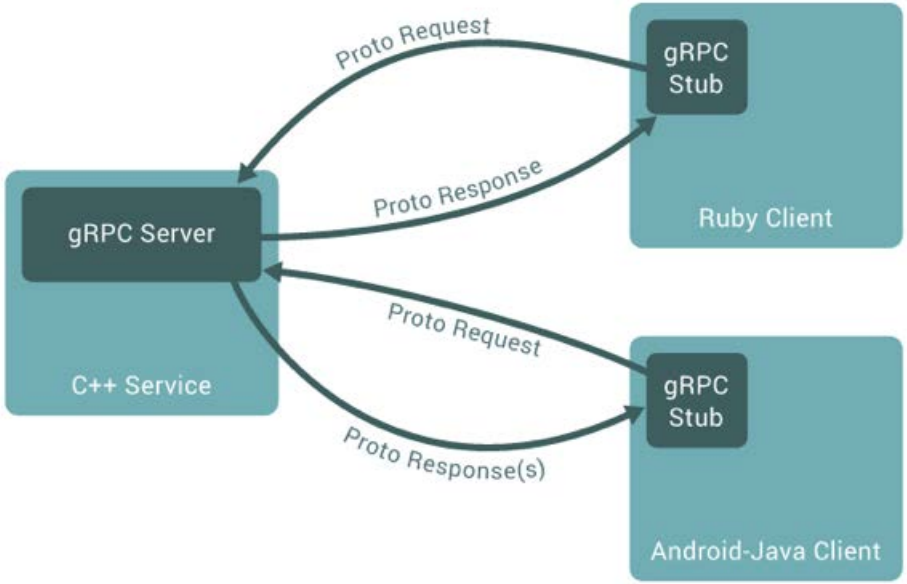
http://hidl.cse.ohio-state.edu/



Architecture Overview of gRPC

Key Features:

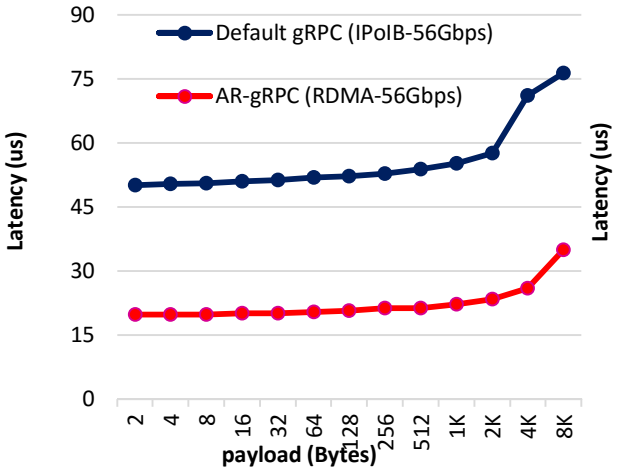
- Simple service definition
- Works across languages and platforms
  - C++, Java, Python, Android Java etc
  - Linux, Mac, Windows.
- Start quickly and scale
- Bi-directional streaming and integrated authentication
- Used by Google (several of Google's cloud products and Google externally facing APIs, TensorFlow), Netflix, Docker, Cisco, Juniper Networks etc.
- Uses sockets for communication!

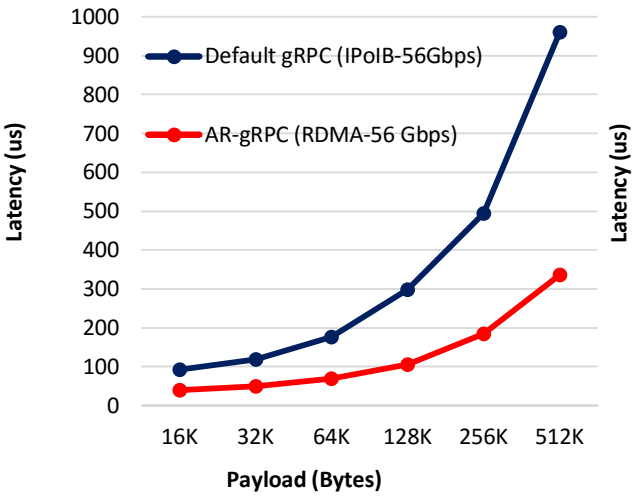


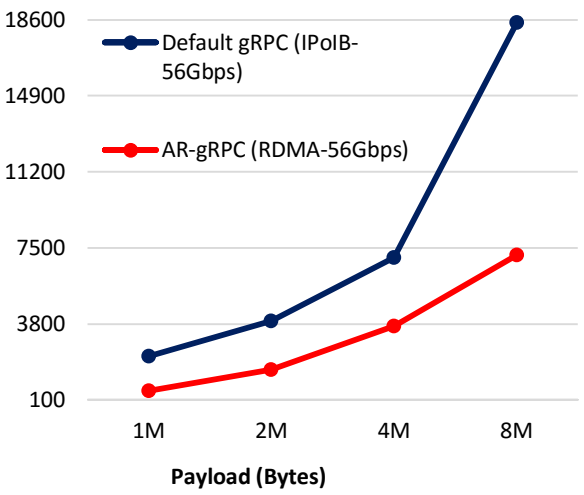
Large-scale distributed systems composed of micro services

Source: http://www.grpc.io/

Performance Benefits for AR-gRPC with Micro-Benchmark





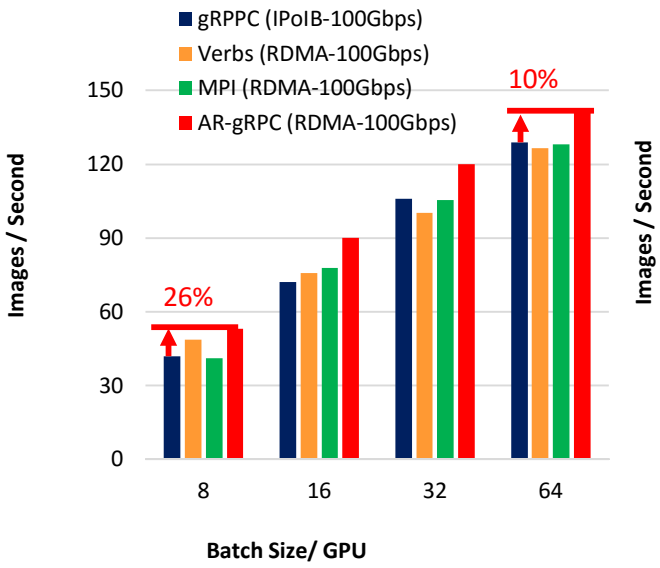


Point-to-Point Latency

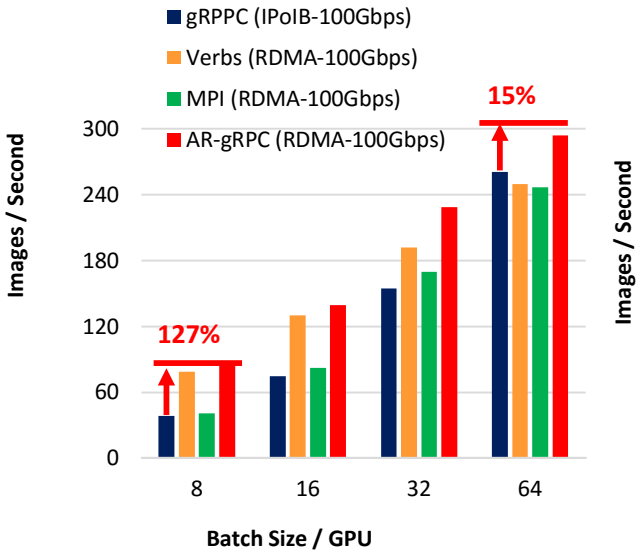
- **AR-gRPC (OSU design) Latency on SDSC-Comet-FDR**
  - **Up to 2.7x performance speedup over Default gRPC (IPoIB) for Latency for small messages.**
  - **Up to 2.8x performance speedup over Default gRPC (IPoIB) for Latency for medium messages.**
  - **Up to 2.5x performance speedup over Default gRPC (IPoIB) for Latency for large messages.**

R. Biswas, X. Lu, and D. K. Panda, Accelerating TensorFlow with Adaptive RDMA-based gRPC. (Under Review)

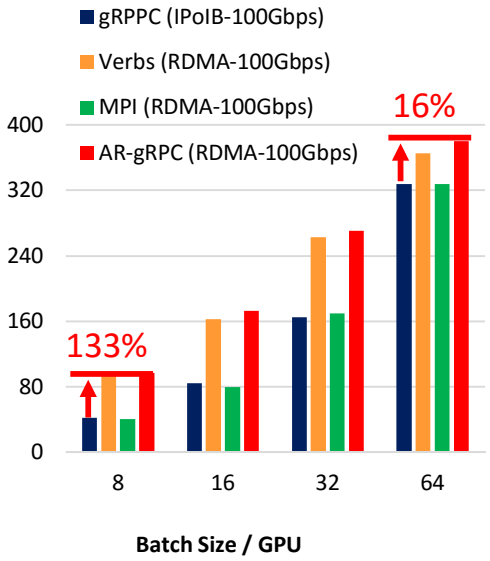
Performance Benefit for TensorFlow (Resnet50)



4 Nodes



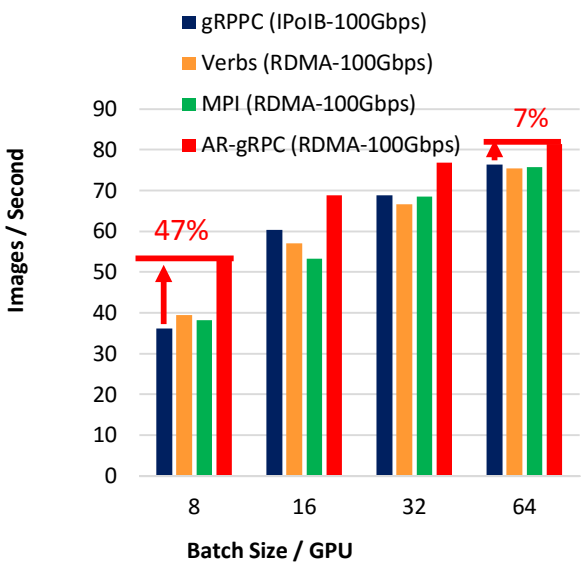
8 Nodes



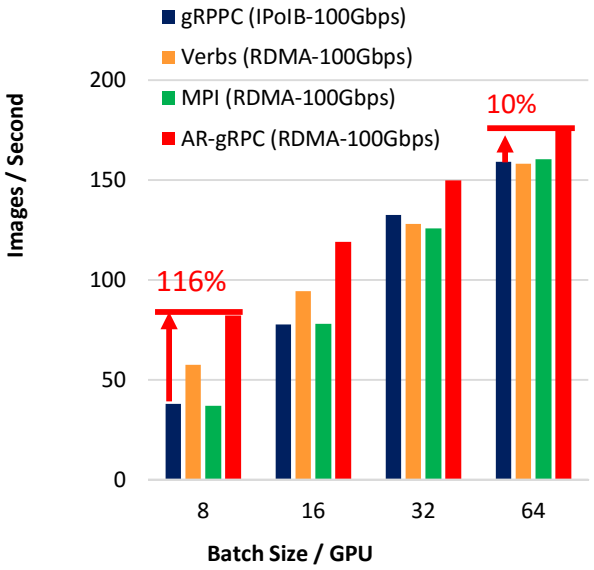
12 Nodes

- TensorFlow Resnet50 performance evaluation on an IB EDR cluster
  - Up to 26% performance speedup over Default gRPC (IPoIB) for 4 nodes
  - Up to 127% performance speedup over Default gRPC (IPoIB) for 8 nodes
  - Up to 133% performance speedup over Default gRPC (IPoIB) for 12 nodes

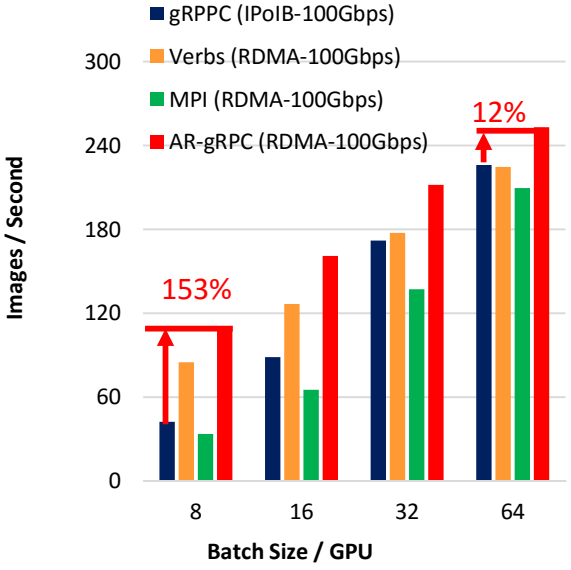
Performance Benefit for TensorFlow (Inception3)



4 Nodes



8 Nodes

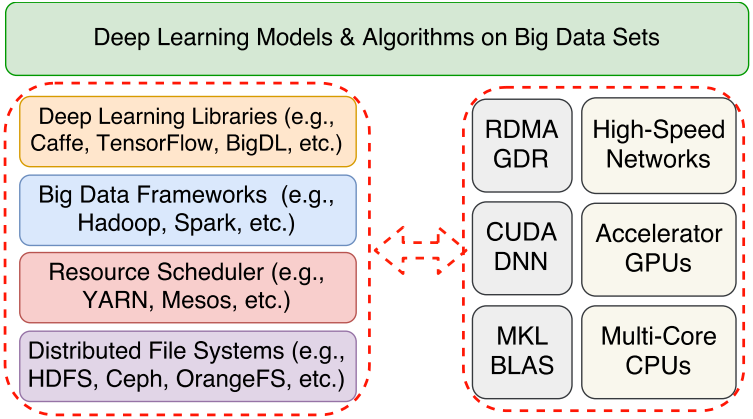


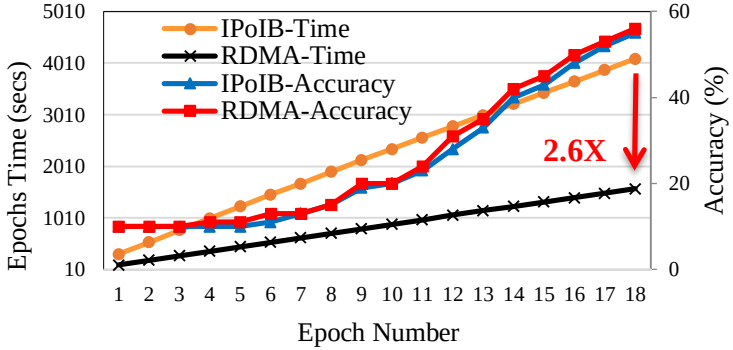
12 Nodes

- TensorFlow Inception3 performance evaluation on an IB EDR cluster
  - Up to 47% performance speedup over Default gRPC (IPoIB) for 4 nodes
  - Up to 116% performance speedup over Default gRPC (IPoIB) for 8 nodes
  - Up to 153% performance speedup over Default gRPC (IPoIB) for 12 nodes

High-Performance Deep Learning over Big Data (DLoBD) Stacks

- Benefits of Deep Learning over Big Data (DLoBD)
  - Easily integrate deep learning components into Big Data processing workflow
  - Easily access the stored data in Big Data systems
  - No need to set up new dedicated deep learning clusters; Reuse existing big data analytics clusters
- Challenges
  - Can RDMA-based designs in DLoBD stacks improve performance, scalability, and resource utilization on high-performance interconnects, GPUs, and multi-core CPUs?
  - What are the performance characteristics of representative DLoBD stacks on RDMA networks?
- Characterization on DLoBD Stacks
  - CaffeOnSpark, TensorFlowOnSpark, and BigDL
  - IPoIB vs. RDMA; In-band communication vs. Out-of-band communication; CPU vs. GPU; etc.
  - Performance, accuracy, scalability, and resource utilization
  - RDMA-based DLoBD stacks (e.g., BigDL over RDMA-Spark) can achieve 2.6x speedup compared to the IPoIB based scheme, while maintain similar accuracy





X. Lu, H. Shi, M. H. Javed, R. Biswas, and D. K. Panda, Characterizing Deep Learning over Big Data (DLoBD) Stacks on RDMA-capable Networks, HotI 2017.

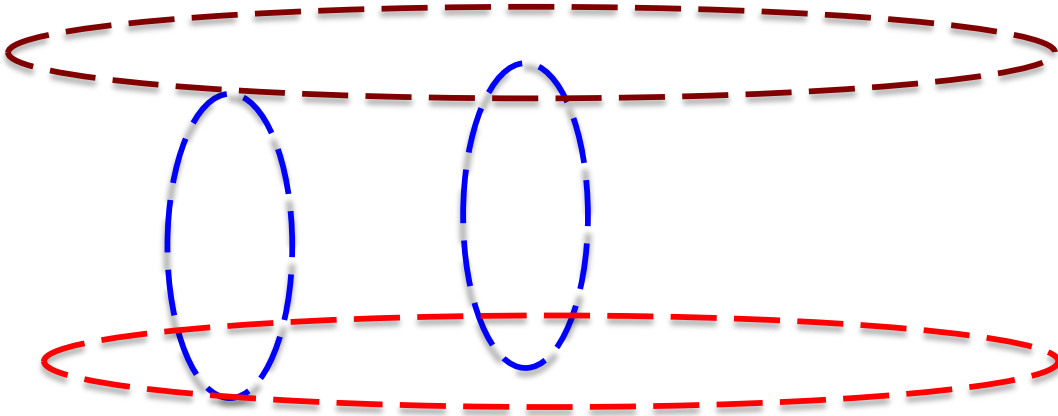
Designing Next-Generation Middleware for Clusters and Datacenters

- High-Performance Programming Models Support for HPC Clusters
- RDMA-Enabled Communication Substrate for Common Services in Datacenters
- High-Performance and Scalable Memcached
- RDMA-Enabled Spark and Hadoop (HDFS, HBase, MapReduce)
- Deep Learning with Scale-Up and Scale-Out
  - Caffe, CNTK, and TensorFlow
- Virtualization Support with SR-IOV and Containers

Single Root I/O Virtualization (SR-IOV)

- Single Root I/O Virtualization (SR-IOV) is providing new opportunities to design cloud-based datacenters with very little low overhead

- Allows a single physical device, or a Physical Function (PF), to present itself as multiple virtual devices, or Virtual Functions (VFs)
- VFs are designed based on the existing non-virtualized PFs, no need for driver change
- Each VF can be dedicated to a single VM through PCI pass-through
- Work with 10/40/100 GigE and InfiniBand



Can High-Performance and Virtualization be Combined?

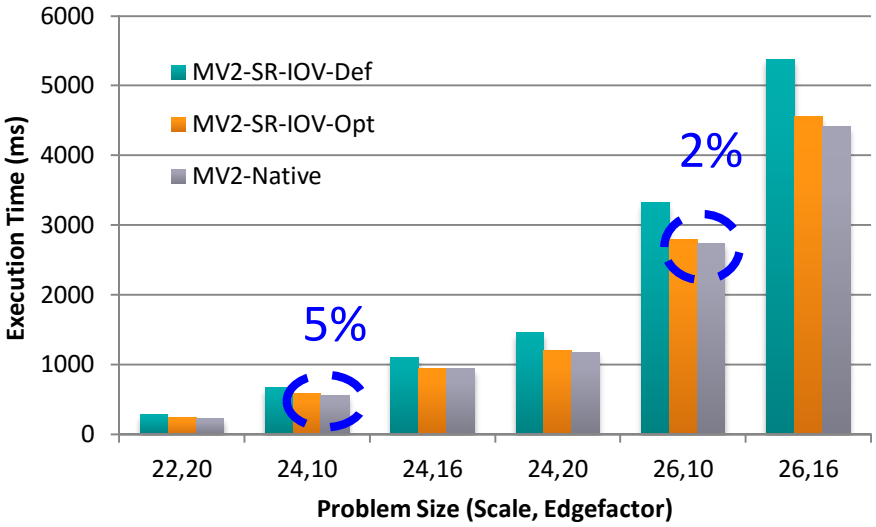
- Virtualization has many benefits
  - Fault-tolerance
  - Job migration
  - Compaction
- Have not been very popular in HPC due to overhead associated with Virtualization
- New SR-IOV (Single Root – IO Virtualization) support available with Mellanox InfiniBand adapters changes the field
- Enhanced MVAPICH2 support for SR-IOV
- MVAPICH2-Virt 2.2 supports:
  - OpenStack, Docker, and singularity

J. Zhang, X. Lu, J. Jose, R. Shi and D. K. Panda, Can Inter-VM Shmem Benefit MPI Applications on SR-IOV based Virtualized InfiniBand Clusters? EuroPar'14

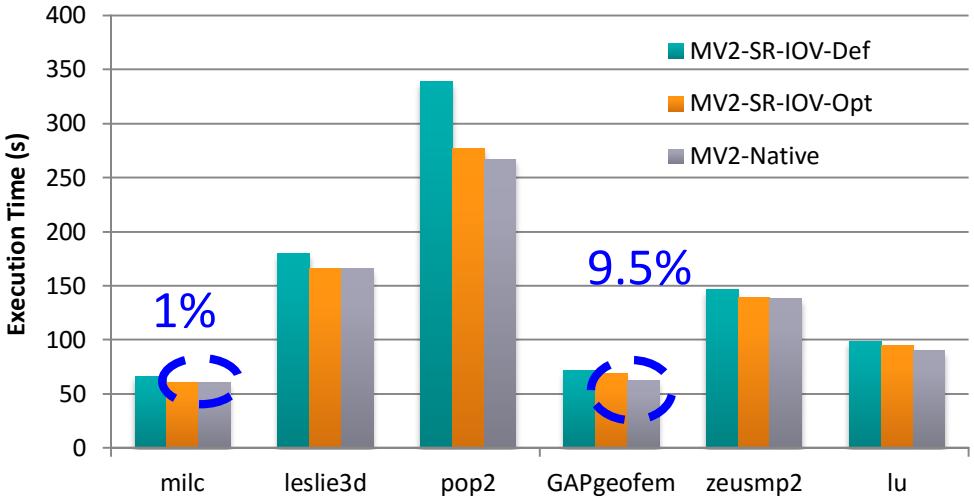
J. Zhang, X. Lu, J. Jose, M. Li, R. Shi and D.K. Panda, High Performance MPI Libray over SR-IOV enabled InfiniBand Clusters, HiPC'14

J. Zhang, X .Lu, M. Arnold and D. K. Panda, MVAPICH2 Over OpenStack with SR-IOV: an Efficient Approach to build HPC Clouds, CCGrid'15

Application-Level Performance on Chameleon



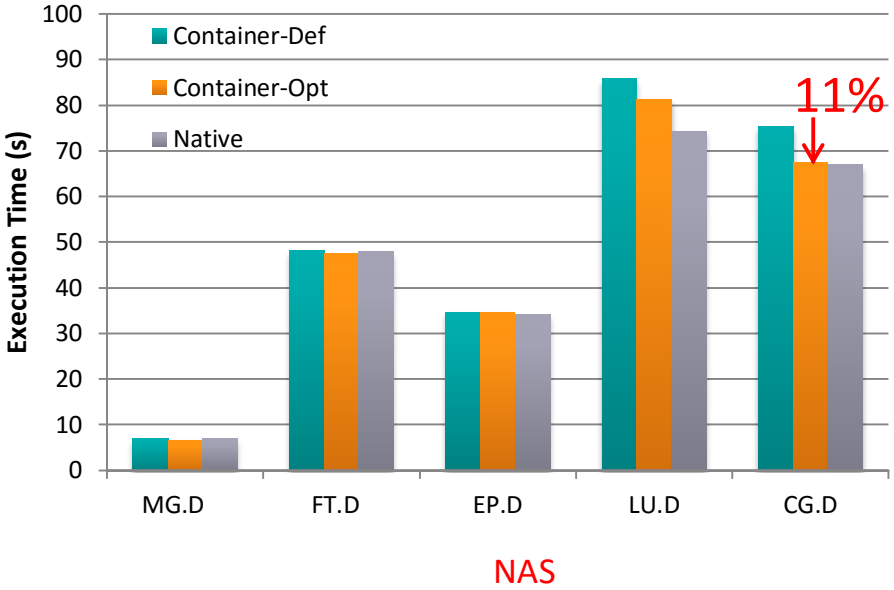
Graph500

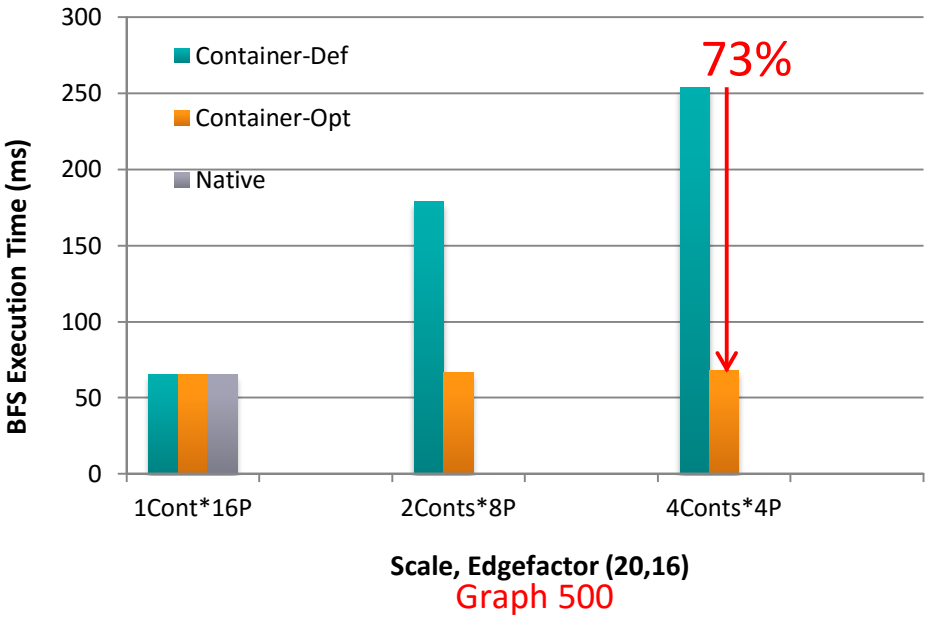


SPEC MPI2007

- 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

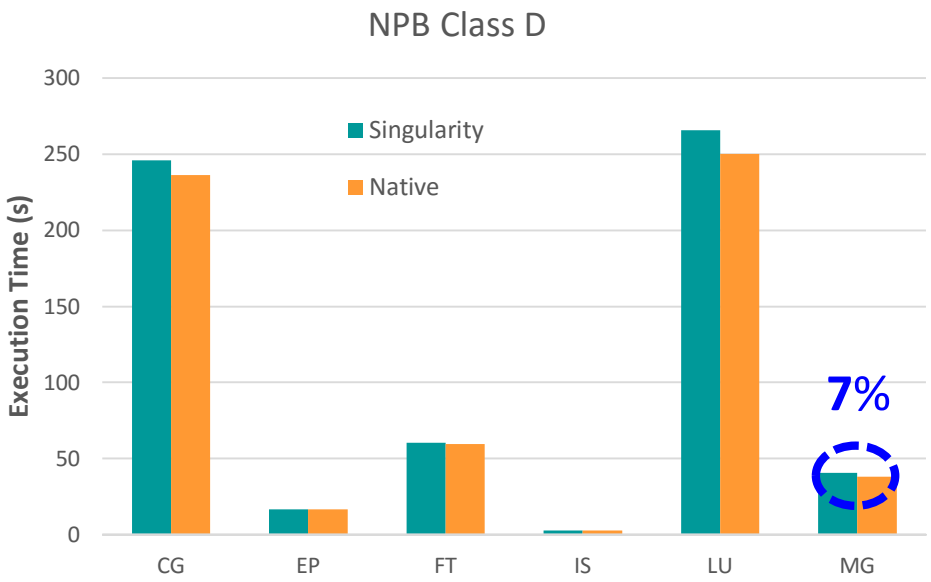
Application-Level Performance on Docker with MVAPICH2

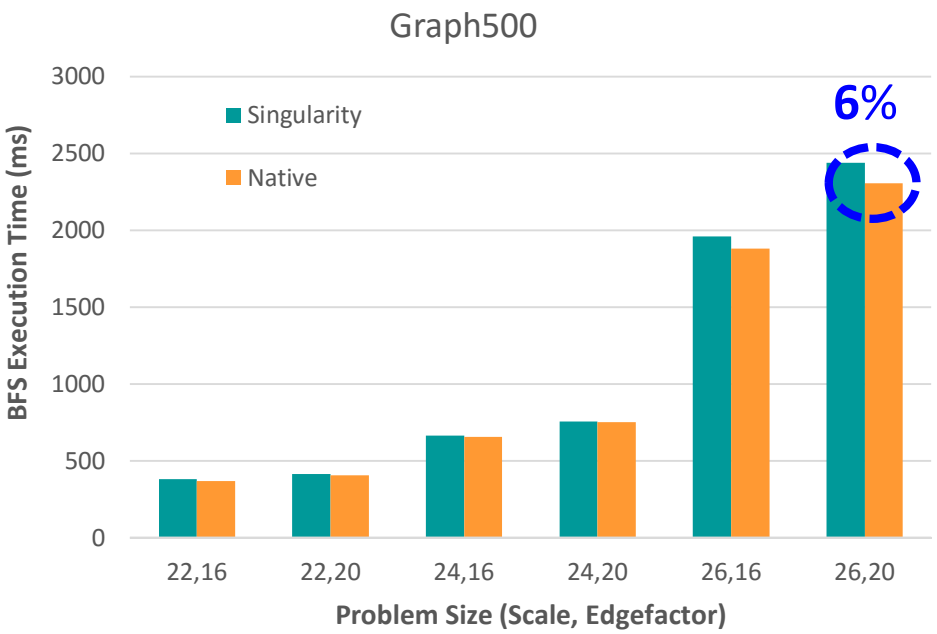




- 64 Containers across 16 nodes, pining 4 Cores per Container
- Compared to Container-Def, up to 11% and 73% of execution time reduction for NAS and Graph 500
- Compared to Native, less than 9 % and 5% overhead for NAS and Graph 500

Application-Level Performance on Singularity with MVAPICH2

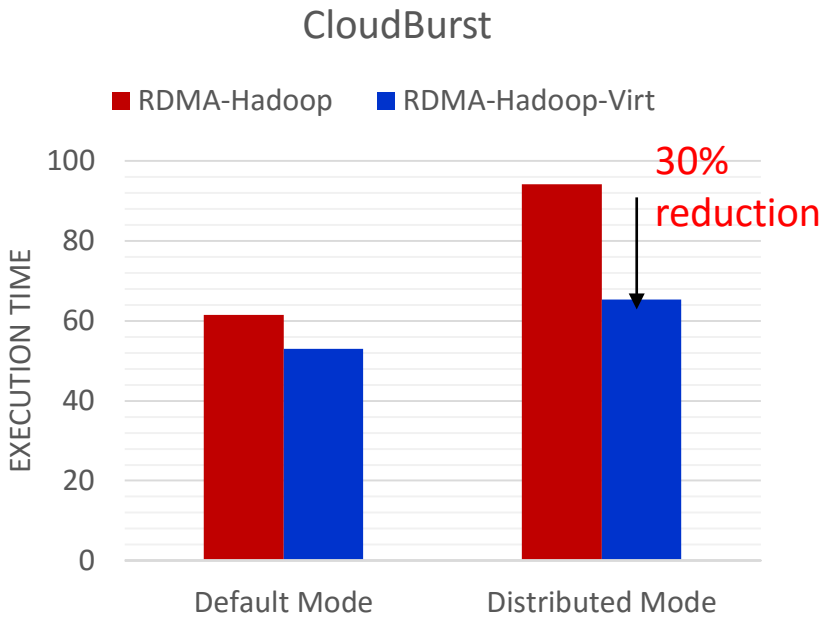


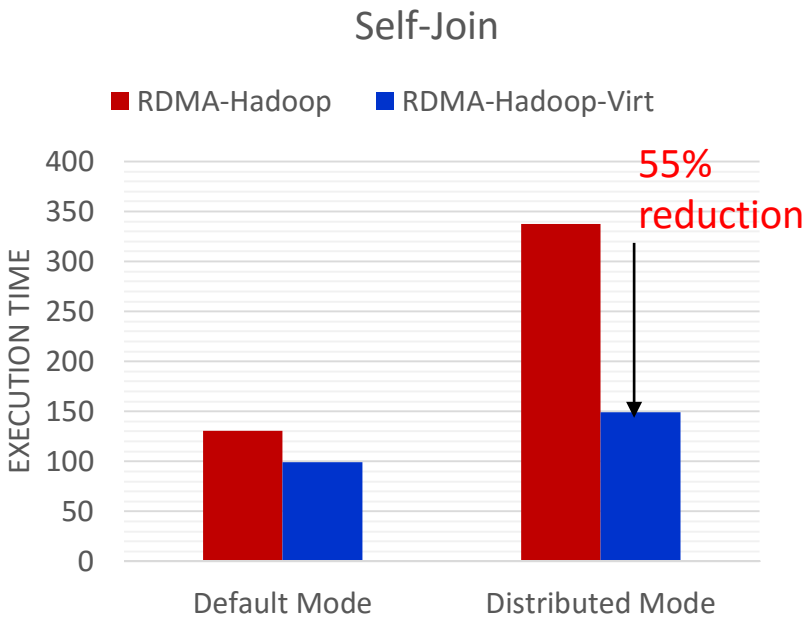


- 512 Processes across 32 nodes
- Less than 7% and 6% overhead for NPB and Graph500, respectively

J. Zhang, X .Lu and D. K. Panda, Is Singularity-based Container Technology Ready for Running MPI Applications on HPC Clouds?, UCC '17

RDMA-Hadoop with Virtualization





- 14% and 24% improvement with Default Mode for CloudBurst and Self-Join
- 30% and 55% improvement with Distributed Mode for CloudBurst and Self-Join

S. Gugnani, X. Lu, D. K. Panda. Designing Virtualization-aware and Automatic Topology Detection Schemes for Accelerating Hadoop on SR-IOV-enabled Clouds. CloudCom, 2016.

Concluding Remarks

- Next generation Clusters and Data Centers need to be designed with a holistic view of HPC, Big Data, Deep Learning, and Cloud
- Presented an overview of the networking technology trends
- Presented some of the approaches and results along these directions
- Enable HPC, Big Data, Deep Learning and Cloud community to take advantage of modern networking technologies
- Many other open issues need to be solved

Funding Acknowledgments

Funding Support by

Equipment Support by

Personnel Acknowledgments

Current Students (Graduate)

- A. Awan (Ph.D.)
- R. Biswas (M.S.)
- M. Bayatpour (Ph.D.)
- S. Chakraborthy (Ph.D.)
- C.-H. Chu (Ph.D.)
- S. Guganani (Ph.D.)
- J. Hashmi (Ph.D.)
- H. Javed (Ph.D.)
- P. Kousha (Ph.D.)
- D. Shankar (Ph.D.)
- H. Shi (Ph.D.)
- J. Zhang (Ph.D.)

Current Students (Undergraduate)

- N. Sarkauskas (B.S.)

Current Research Scientists

- X. Lu
- H. Subramoni

Current Research Specialist

- J. Smith
- M. Arnold

Current Post-doc

- A. Ruhela
- K. Manian

Past Students

- A. Augustine (M.S.)
- P. Balaji (Ph.D.)
- S. Bhagvat (M.S.)
- A. Bhat (M.S.)
- D. Buntinas (Ph.D.)
- L. Chai (Ph.D.)
- B. Chandrasekharan (M.S.)
- N. Dandapanthula (M.S.)
- V. Dhanraj (M.S.)
- T. Gangadharappa (M.S.)
- K. Gopalakrishnan (M.S.)
- W. Huang (Ph.D.)
- W. Jiang (M.S.)
- J. Jose (Ph.D.)
- S. Kini (M.S.)
- M. Koop (Ph.D.)
- K. Kulkarni (M.S.)
- R. Kumar (M.S.)
- S. Krishnamoorthy (M.S.)
- K. Kandalla (Ph.D.)
- M. Li (Ph.D.)
- P. Lai (M.S.)
- J. Liu (Ph.D.)
- M. Luo (Ph.D.)
- A. Mamidala (Ph.D.)
- G. Marsh (M.S.)
- V. Meshram (M.S.)
- A. Moody (M.S.)
- S. Naravula (Ph.D.)
- R. Noronha (Ph.D.)
- X. Ouyang (Ph.D.)
- S. Pai (M.S.)
- S. Potluri (Ph.D.)
- R. Rajachandrasekar (Ph.D.)
- G. Santhanaraman (Ph.D.)
- A. Singh (Ph.D.)
- J. Sridhar (M.S.)
- S. Sur (Ph.D.)
- H. Subramoni (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- A. Vishnu (Ph.D.)
- J. Wu (Ph.D.)
- W. Yu (Ph.D.)

Past Research Scientist

- K. Hamidouche
- S. Sur

Past Programmers

- D. Bureddy
- J. Perkins

Past Post-Docs

- D. Banerjee
- X. Besseron
- H.-W. Jin
- J. Lin
- M. Luo
- E. Mancini
- S. Marcarelli
- J. Vienne
- H. Wang

Multiple Positions Available in My Group

- Looking for Bright and Enthusiastic Personnel to join as
  - Post-Doctoral Researchers
  - PhD Students
  - MPI Programmer/Software Engineer
  - Hadoop/Big Data Programmer/Software Engineer
- If interested, please contact me at this conference and/or send an e-mail to panda@cse.ohio-state.edu

Thank You!

panda@cse.ohio-state.edu

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

MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

The High-Performance MPI/PGAS Project
http://mvapich.cse.ohio-state.edu/

HiBD
High-Performance Big Data

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

HiDL
High-Performance Deep Learning

The High-Performance Deep Learning Project
http://hidl.cse.ohio-state.edu/