



MVA PICH

MPI, PGAS and Hybrid MPI+PGAS Library



HiDL

*High-Performance
Deep Learning*



HiBD

High-Performance
Big Data

RDMA-based Networking Technologies and Middleware for Next-Generation Clusters and Data Centers

Keynote Talk at KBNet '18

by

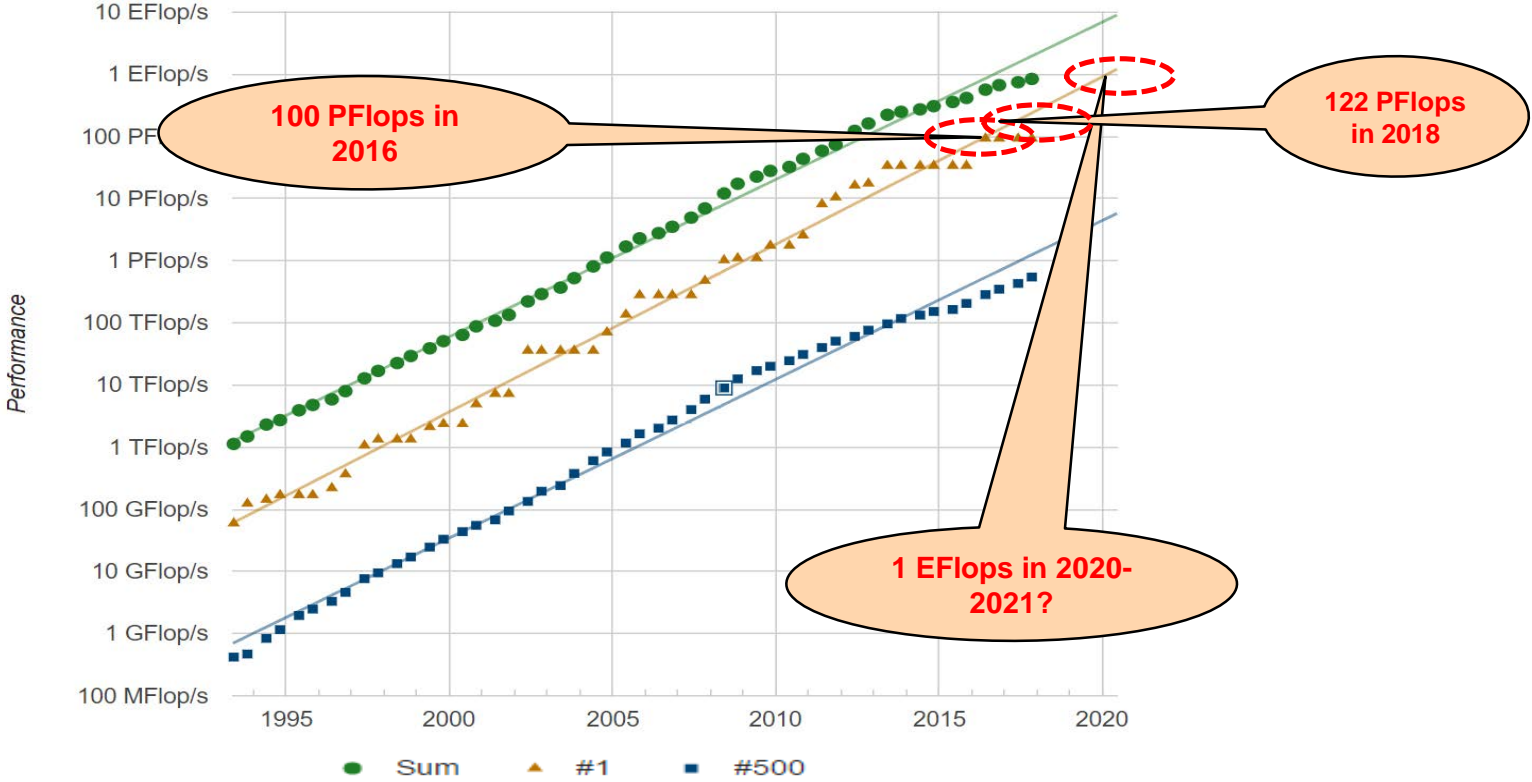
Dhabaleswar K. (DK) Panda

The Ohio State University

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

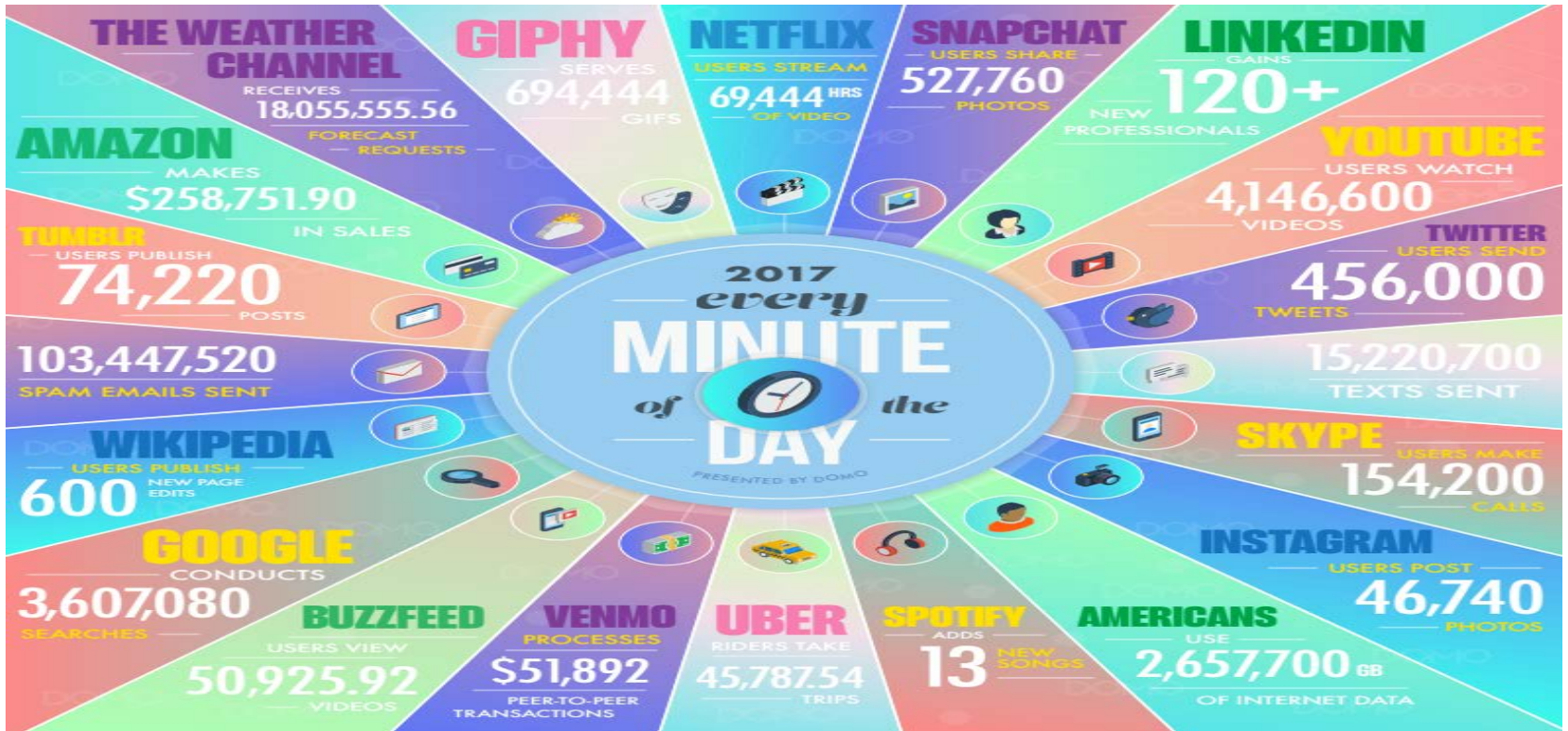
<http://www.cse.ohio-state.edu/~panda>

High-End Computing (HEC): Towards Exascale



Expected to have an ExaFlop system in 2020-2021!

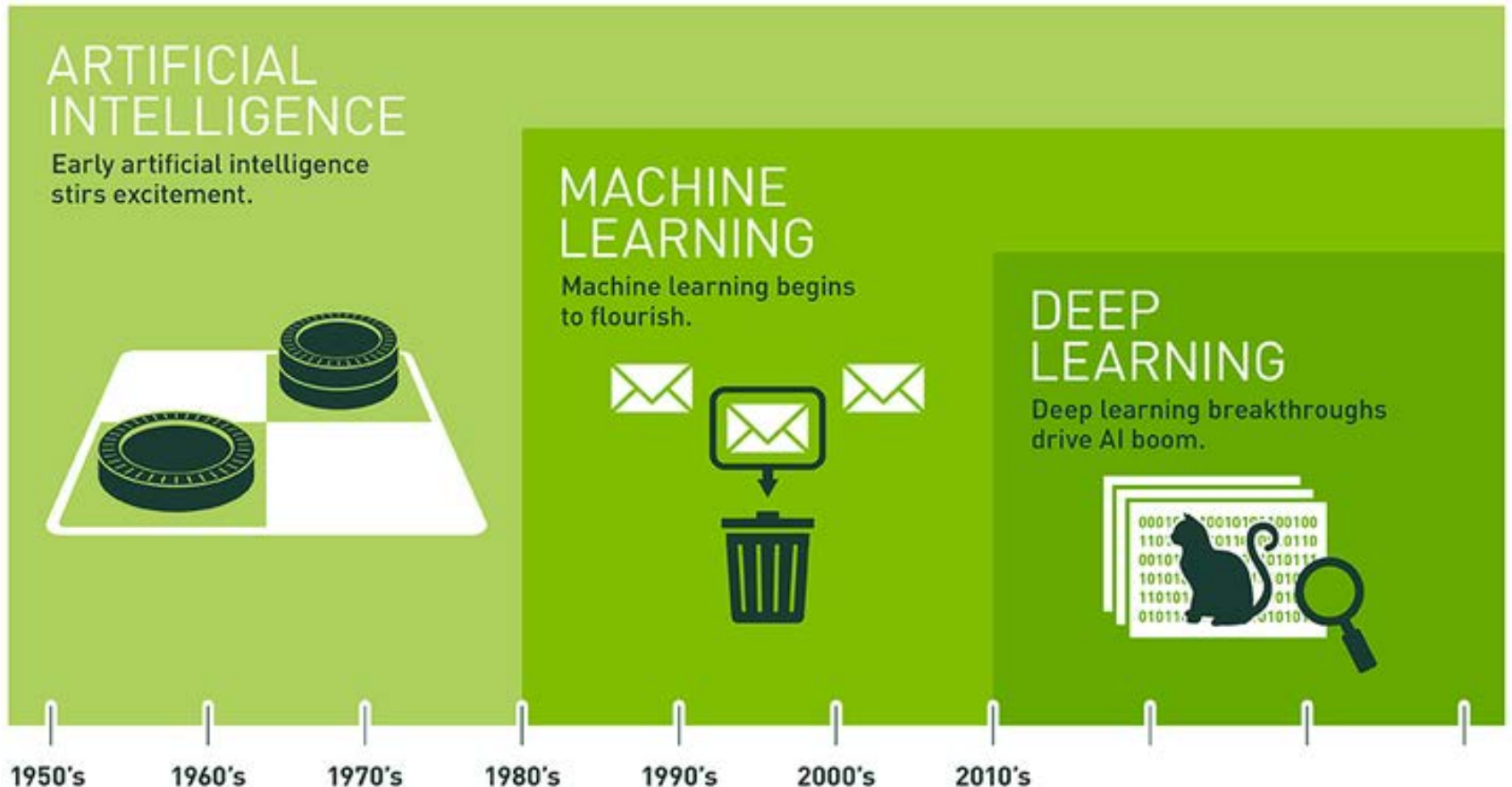
Big Data – How Much Data Is Generated Every Minute on the Internet?



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/>

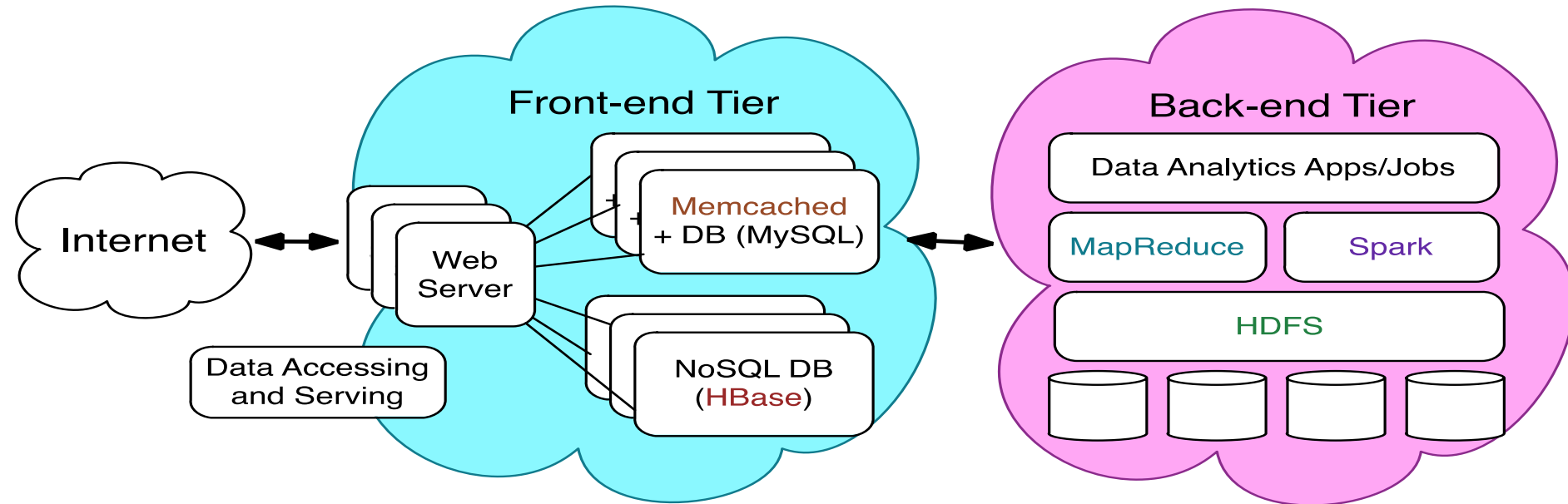
Resurgence of AI/Machine Learning/Deep Learning



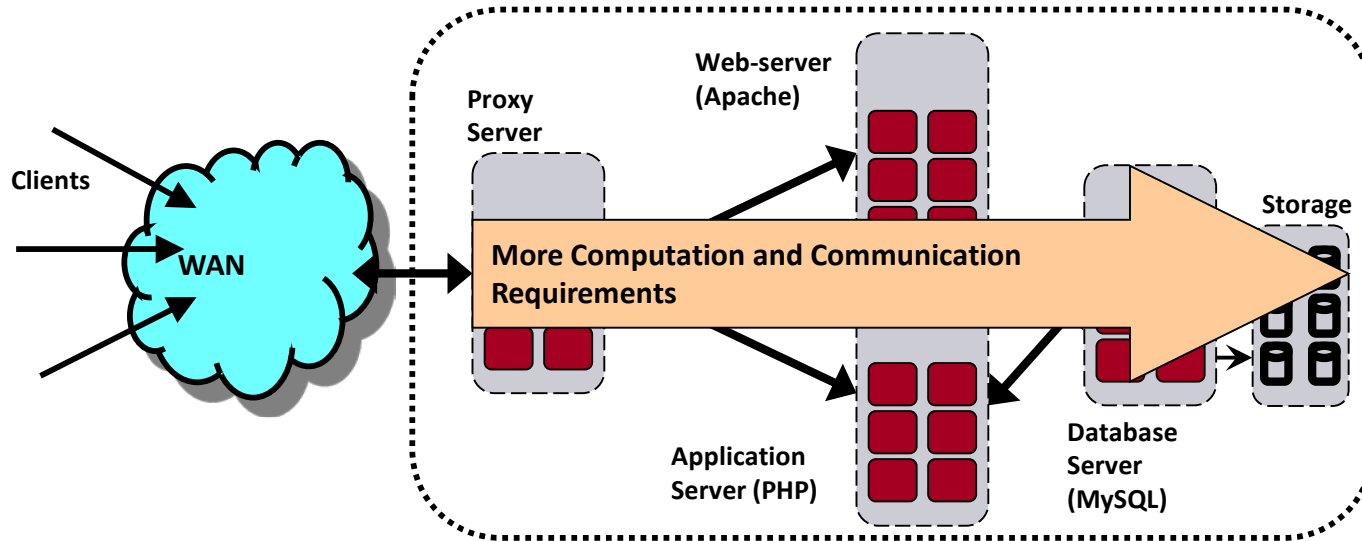
Courtesy: <http://www.zdnet.com/article/caffe2-deep-learning-wide-ambitions-flexibility-scalability-and-advocacy/>

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

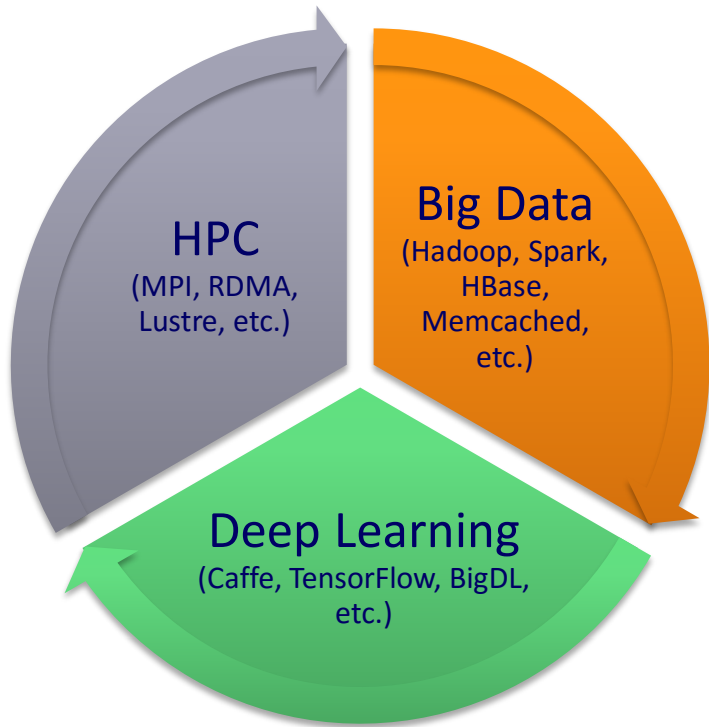


Communication and Computation Requirements



- 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?



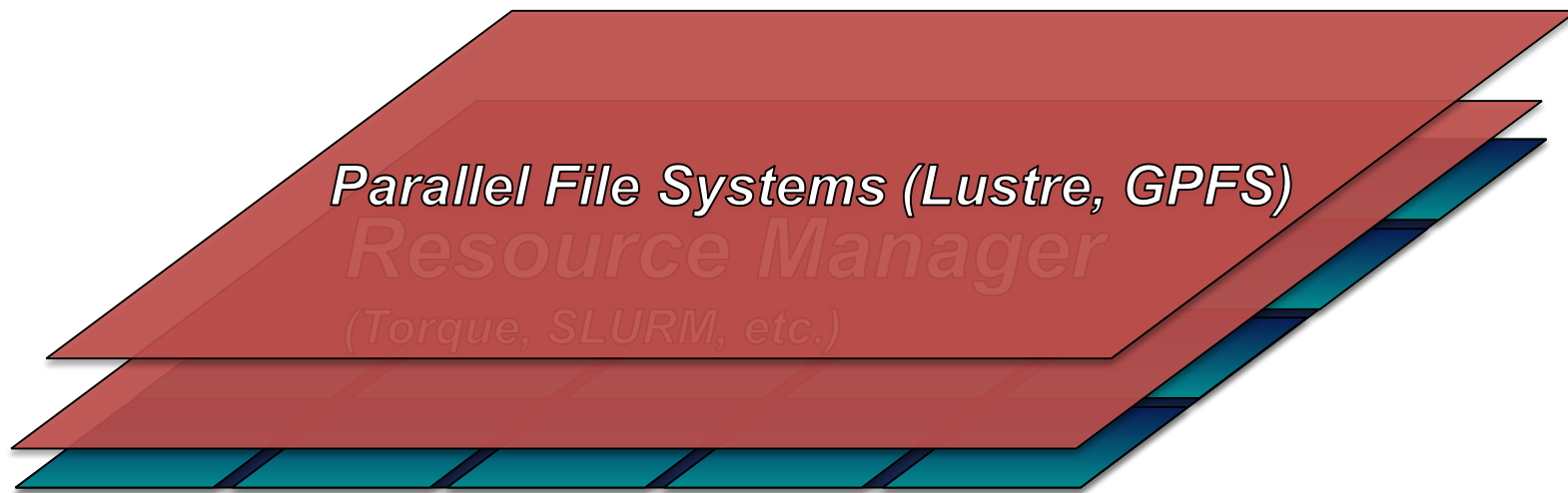
Physical Compute

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

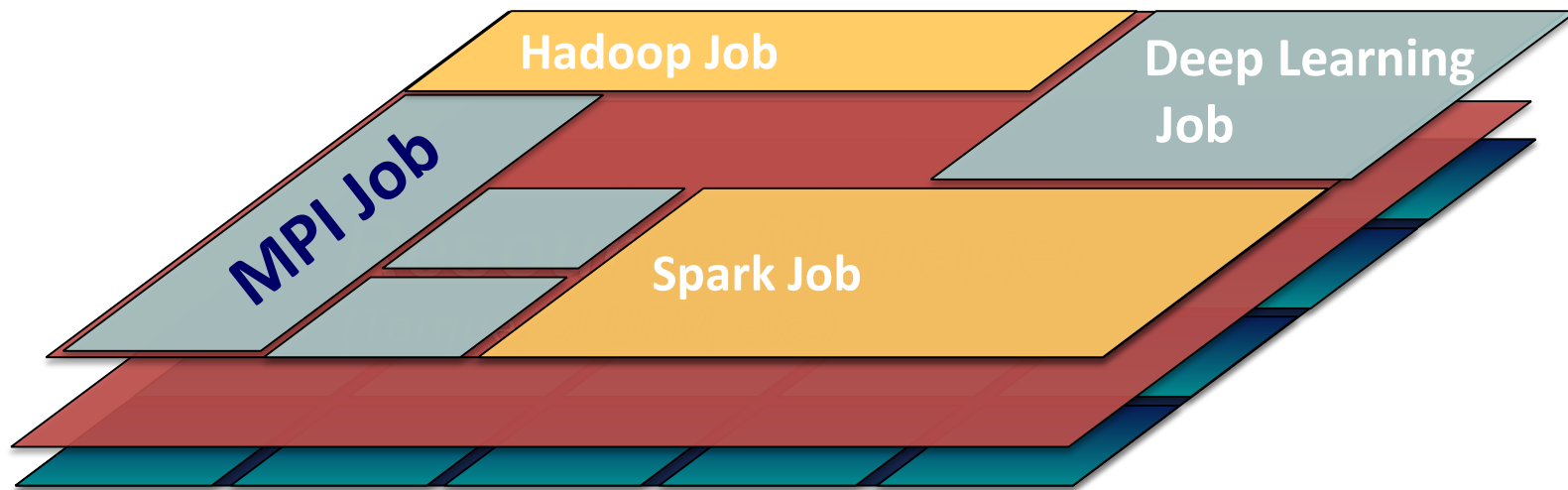


Resource Manager
(Torque, SLURM, etc.)

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?

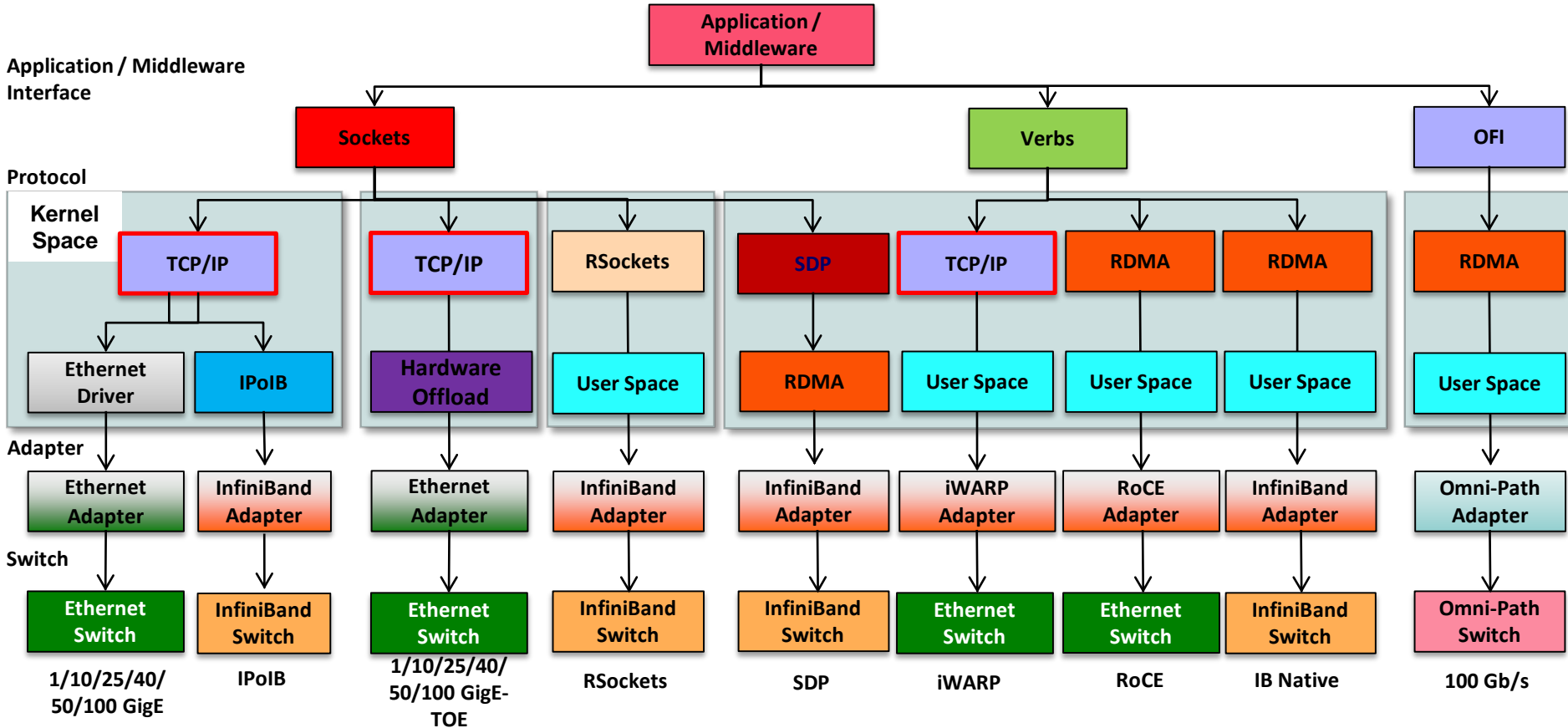


Trends in Network Speed Acceleration

| | |
|----------------------------------|---------------------------------|
| 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

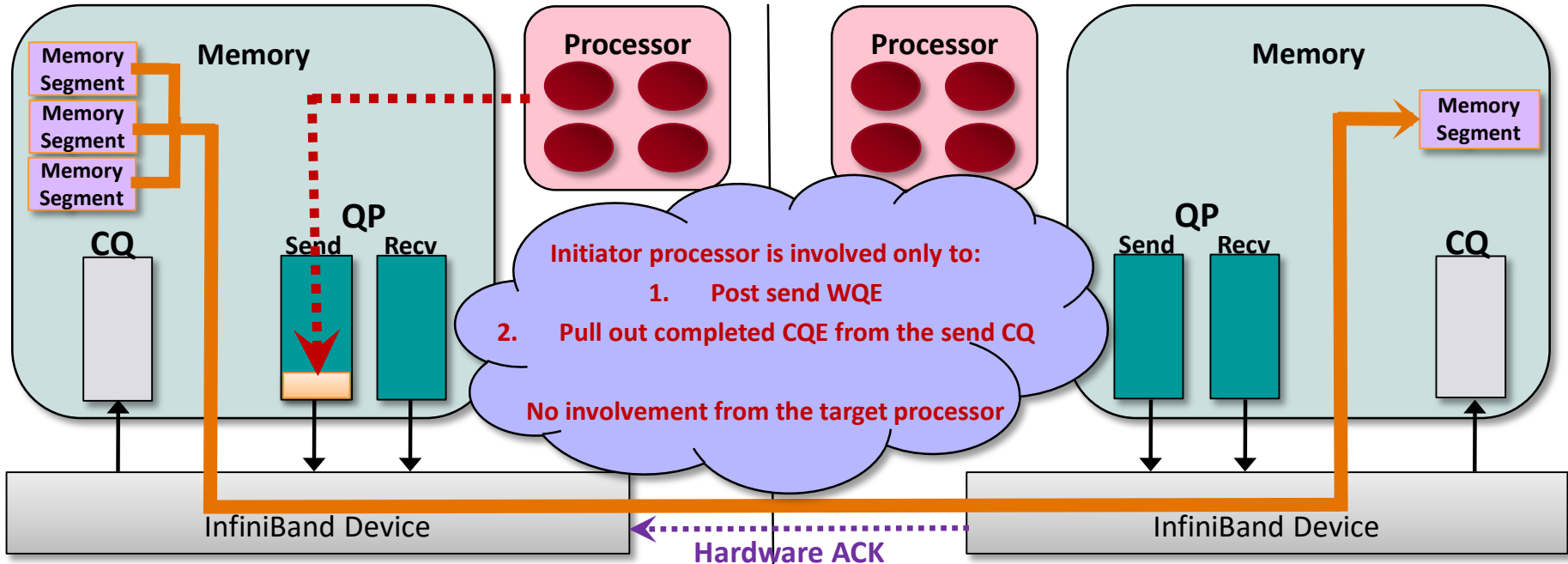
Available Interconnects and Protocols for Data Centers



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,

Communication in the Memory Semantics (RDMA Model)



Send WQE contains information about the send buffer (multiple segments) and the receive buffer (single segment)

Large-scale InfiniBand Installations

- 139 IB Clusters (27.8%) in the Jun'18 Top500 list
 - (<http://www.top500.org>)
- Installations in the Top 50 (19 systems):

**#2nd system (Sunway TaihuLight)
also uses InfiniBand**

| | |
|--|---|
| 2,282,544 cores (Summit) at ORNL (1st) | 155,150 cores (JURECA) at FZJ/Germany (38 th) |
| 1,572,480 cores (Sierra) at LLNL (3 rd) | 72,800 cores Cray CS-Storm in US (40 th) |
| 391,680 cores (ABCI) at AIST/Japan (5 th) | 72,800 cores Cray CS-Storm in US (41 st) |
| 253,600 cores (HPC4) in Italy (13 th) | 78,336 cores (Electra) at NASA/Ames (43 rd) |
| 114,480 cores (Juwels Module 1) at FZJ/Germany (23 rd) | 124,200 cores (Topaz) at ERDC DSRC/USA (44 th) |
| 241,108 cores (Pleiades) at NASA/Ames (24 th) | 60,512 cores NVIDIA DGX-1 at Facebook/USA (45 th) |
| 220,800 cores (Pangea) in France (30 th) | 60,512 cores (DGX Saturn V) at NVIDIA/USA (46 th) |
| 144,900 cores (Cheyenne) at NCAR/USA (31 st) | 113,832 cores (Damson) at AWE/UK (47 th) |
| 72,000 cores (ITO – Subsystem A) in Japan (32 nd) | 72,000 cores (HPC2) in Italy (49 th) |
| 79,488 cores (JOLIOT-CURIE SKL) at CEA/France (34 th) | and many more! |

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

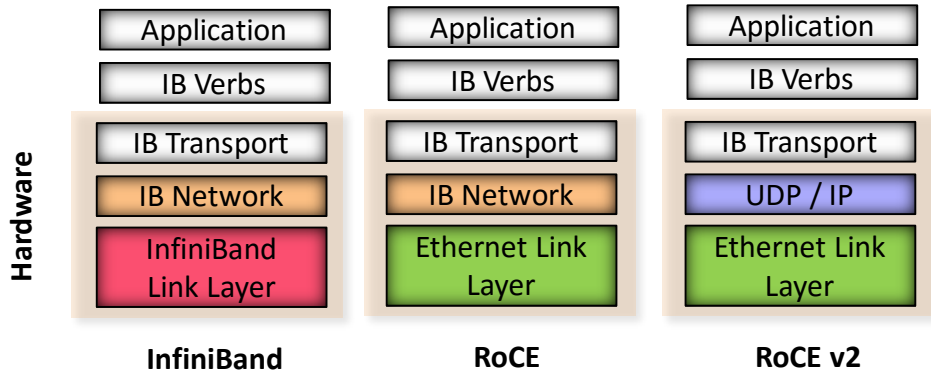
- 10GE Alliance formed by several industry leaders to take the Ethernet family to the next speed step
- Goal: To achieve a scalable and high performance communication architecture while maintaining backward compatibility with Ethernet
- <http://www.ethernetalliance.org>
- 40-Gbps (Servers) and 100-Gbps 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

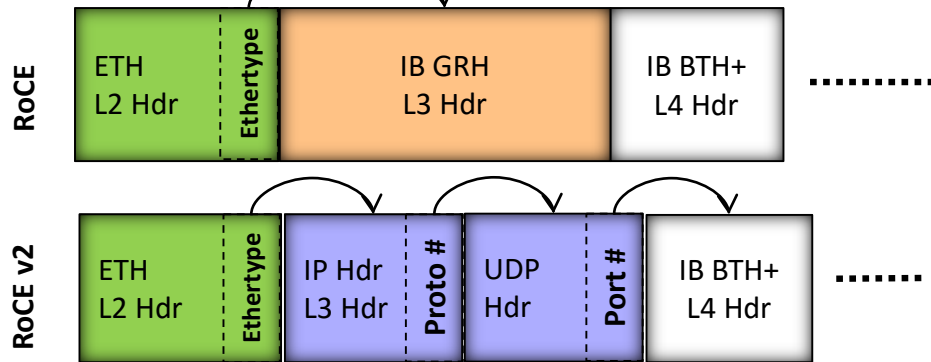
- TCP Offload Engines (TOE)
 - Hardware Acceleration for the entire TCP/IP stack
 - Initially patented by Tehuti Networks
 - Actually refers to the IC on the network adapter that implements TCP/IP
 - In practice, usually referred to as the entire network adapter
- Internet Wide-Area RDMA Protocol (iWARP)
 - 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

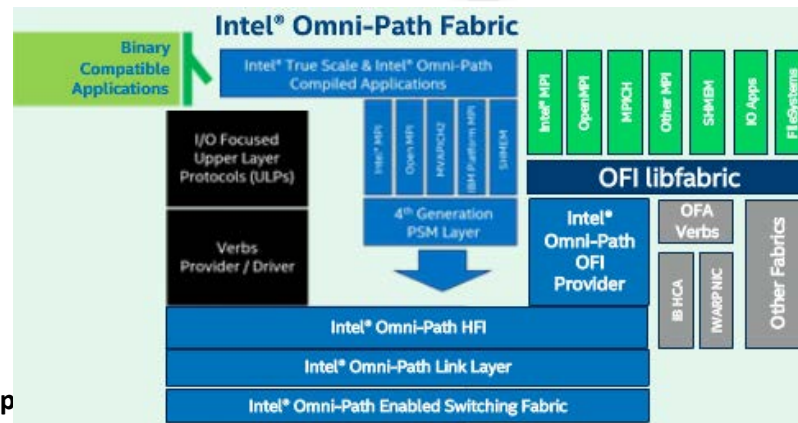
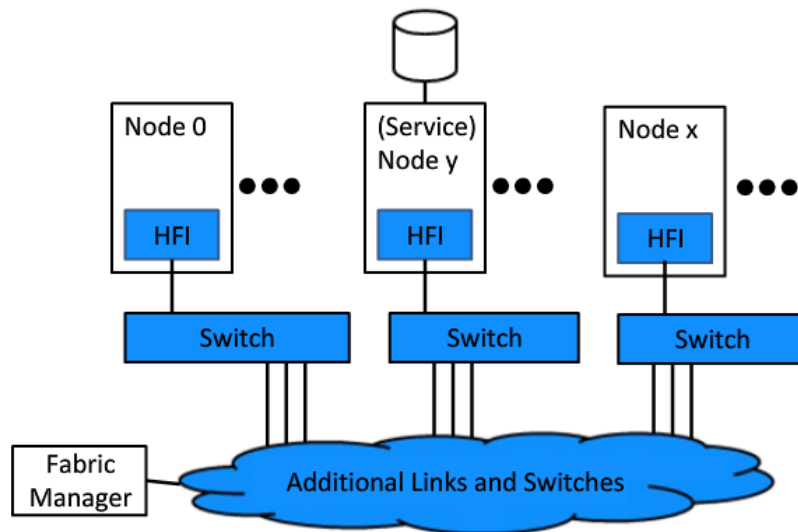
- Takes advantage of IB and Ethernet
 - Software written with IB-Verbs
 - Link layer is Converged (Enhanced) Ethernet (CE)
 - 100Gb/s support from latest EDR and ConnectX-3 Pro adapters
- Pros: IB Vs RoCE
 - Works natively in Ethernet environments
 - Entire Ethernet management ecosystem is available
 - 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

- 171 HSE compute systems with ranking in the Jun'18 Top500 list
 - 38,400-core installation in China (#95) – new
 - 38,400-core installation in China (#96) – new
 - 38,400-core installation in China (#97) – new
 - 39,680-core installation in China (#99)
 - 66,560-core installation in China (#157)
 - 66,280-core installation in China (#159)
 - 64,000-core installation in China (#160)
 - 64,000-core installation in China (#161)
 - 72,000-core installation in China (#164)
 - 64,320-core installation in China (#185) – new
 - 78,000-core installation in China (#187)
 - 75,776-core installation in China (#188) – new
 - 59,520-core installation in China (#192)
 - 59,520-core installation in China (#193)
 - 28,800-core installation in China (#195) – new
 - 62,400-core installation in China (#197) – new
 - 64,800-core installation in China (#198)
 - 66,000-core installation in China (#209) – 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](https://www.intel.com)

Large-scale Omni-Path Installations

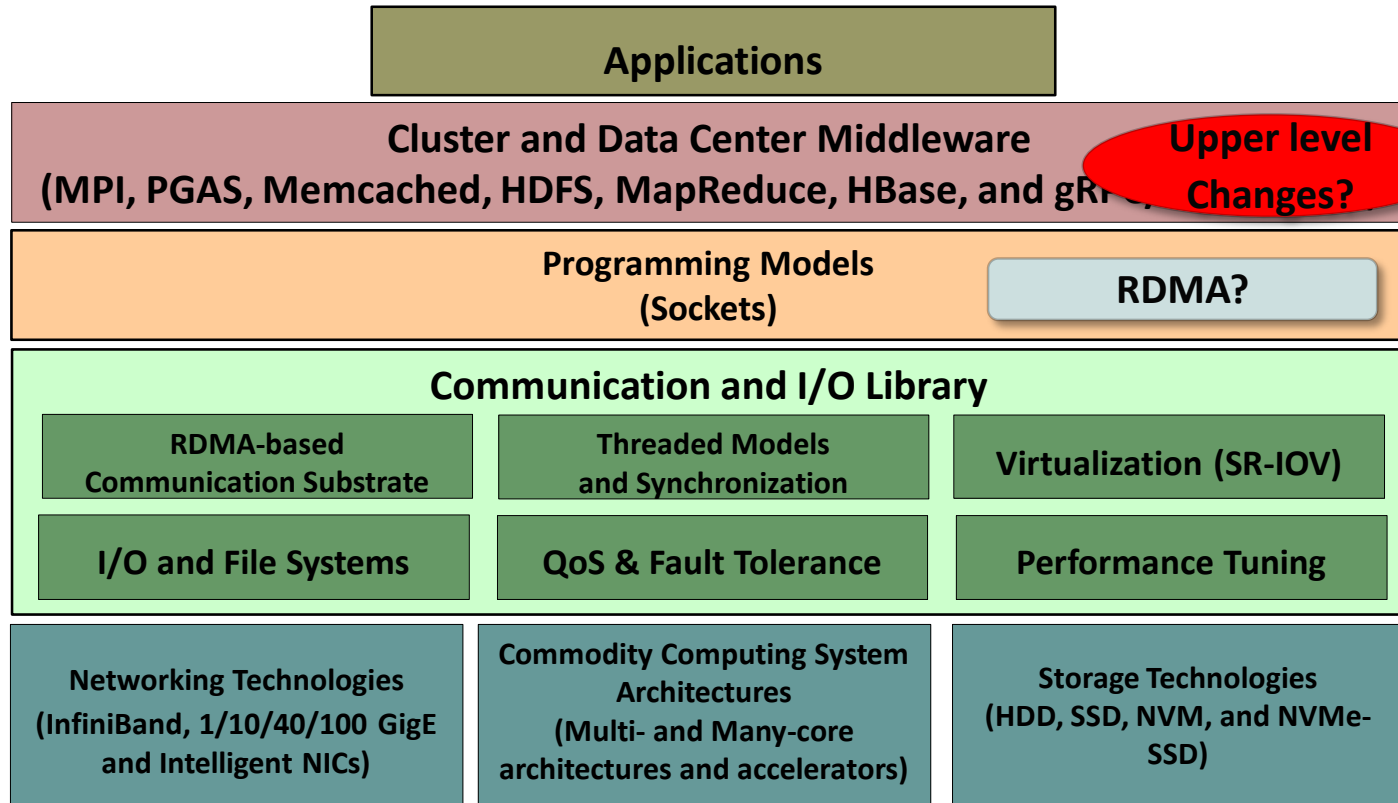
- 39 Omni-Path Clusters (7.8%) in the Jun'18 Top500 list
 - (<http://www.top500.org>)

| | |
|---|---|
| 570,020 core (Nurion) at KISTI/South Korea (11th) | 53,300 core (Makman-3) at Saudi Aramco/Saudi Arabia (78 th) |
| 556,104 core (Oakforest-PACS) at JCAHPC in Japan (12 th) | 34,560 core (Gaffney) at Navy DSRC/USA (85 th) |
| 367,024 core (Stampede2) at TACC in USA (15 th) | 34,560 core (Koehr) at Navy DSRC/USA (86 th) |
| 312,936 core (Marconi XeonPhi) at CINECA in Italy (18 th) | 49,432 core (Mogon II) in Germany (87 th) |
| 135,828 core (Tsubame 3.0) at TiTech in Japan (19 th) | 38,553 core (Molecular Simulator) in Japan (93 rd) |
| 153,216 core (MareNostrum) at BSC in Spain (22 nd) | 35,280 core (Curiosity) at BASF in Germany (94 th) |
| 127,520 core (Cobra) in Germany (28 th) | 54,432 core (Marconi Xeon) at CINECA in Italy (98 th) |
| 55,296 core (Mustang) at AFRL/USA (48 th) | 46,464 core (Peta4) at Cambridge/UK (101 st) |
| 95,472 core (Quartz) at LLNL in USA (63 rd) | 53,352 core (Girzzly) at LANL in USA (136 th) |
| 95,472 core (Jade) at LLNL in USA (64 th) | 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 RDMA-based Communication and I/O Libraries for Clusters and Data Center Middleware: Challenges



Designing RDMA-based 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 and TensorFlow
- Virtualization Support with SR-IOV and Containers

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

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 2,925 organizations in 86 countries**
 - **More than 487,000 (> 0.48 million) downloads from the OSU site directly**
 - Empowering many TOP500 clusters (Jul '18 ranking)
 - 2nd ranked 10,649,640-core cluster (Sunway TaihuLight) at NSC, Wuxi, China
 - 12th, 556,104 cores (Oakforest-PACS) in Japan
 - 15th, 367,024 cores (Stampede2) at TACC
 - 24th, 241,108-core (Pleiades) at NASA 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

Architecture of MVAPICH2 Software Family

High Performance Parallel Programming Models

Message Passing Interface
(MPI)

PGAS
(UPC, OpenSHMEM, CAF, UPC++)

Hybrid --- MPI + X
(MPI + PGAS + OpenMP/Cilk)

High Performance and Scalable Communication Runtime

Diverse APIs and Mechanisms

Point-to-point
Primitives

Collectives
Algorithms

Job Startup

Energy-
Awareness

Remote
Memory
Access

I/O and
File Systems

Fault
Tolerance

Virtualization

Active
Messages

Introspection
& Analysis

Support for Modern Networking Technology

(InfiniBand, iWARP, RoCE, Omni-Path)

Transport Protocols

RC

XRC

UD

DC

Modern Features

UMR

ODP

SR-
IOV

Multi
Rail

Support for Modern Multi-/Many-core Architectures

(Intel-Xeon, OpenPOWER, Xeon-Phi (MIC, KNL), NVIDIA GPGPU)

Transport Mechanisms

Shared
Memory

CMA

IVSHMEM

XPMEM*

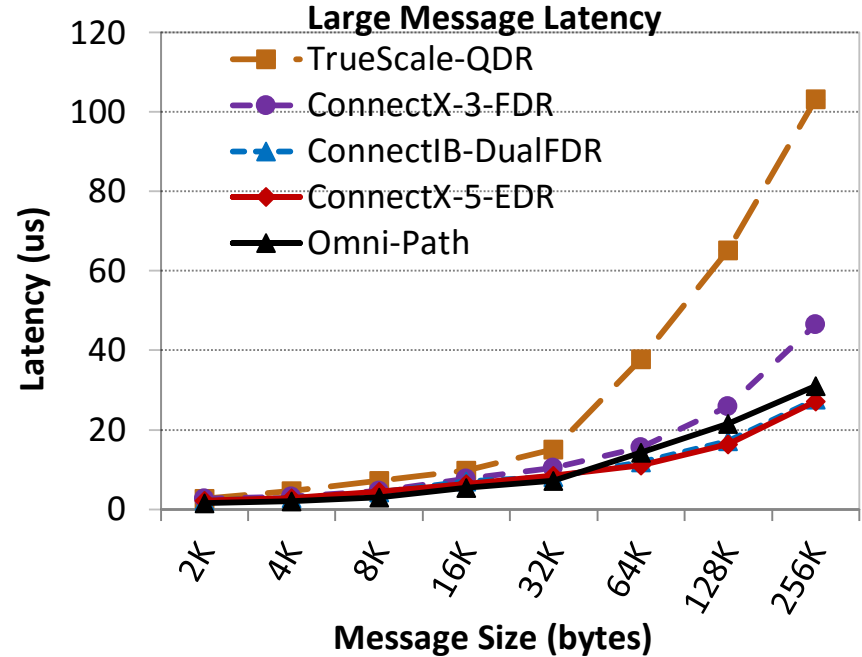
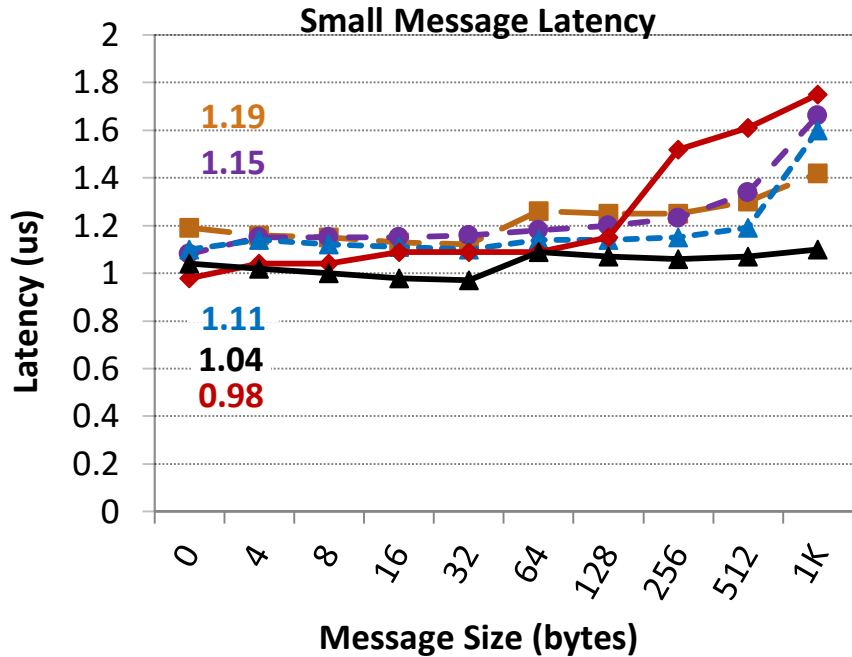
Modern Features

NVLink*

CAPI*

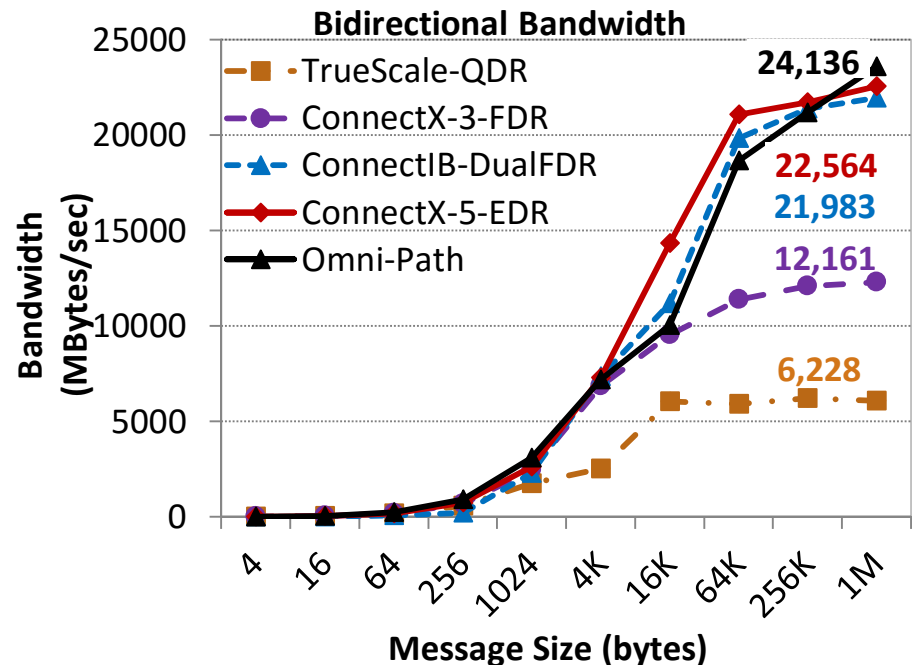
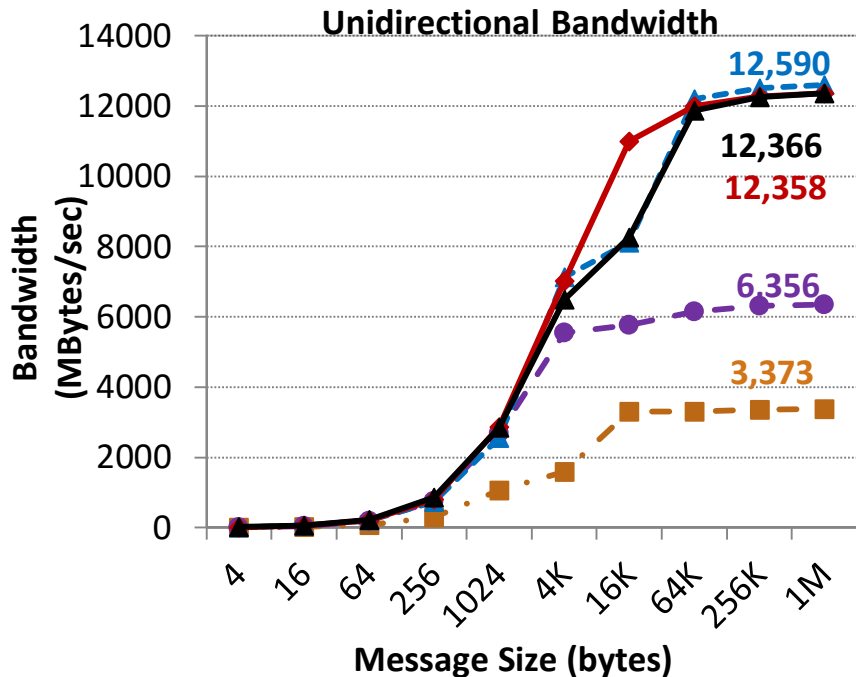
* Upcoming

One-way Latency: MPI over IB with MVAPICH2



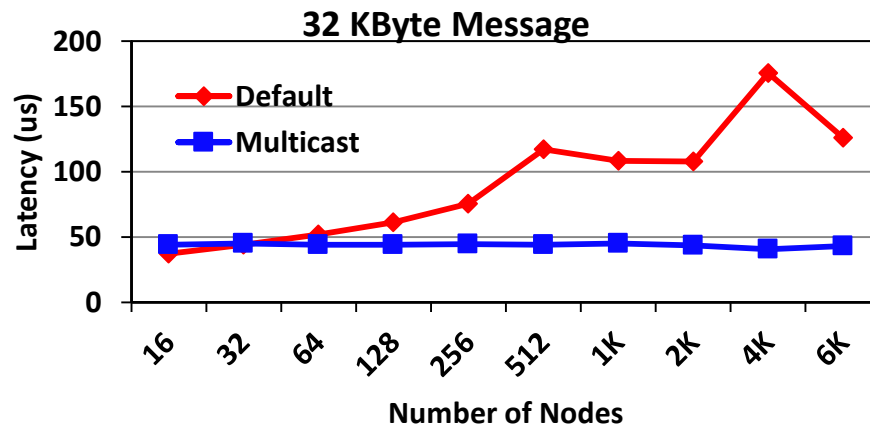
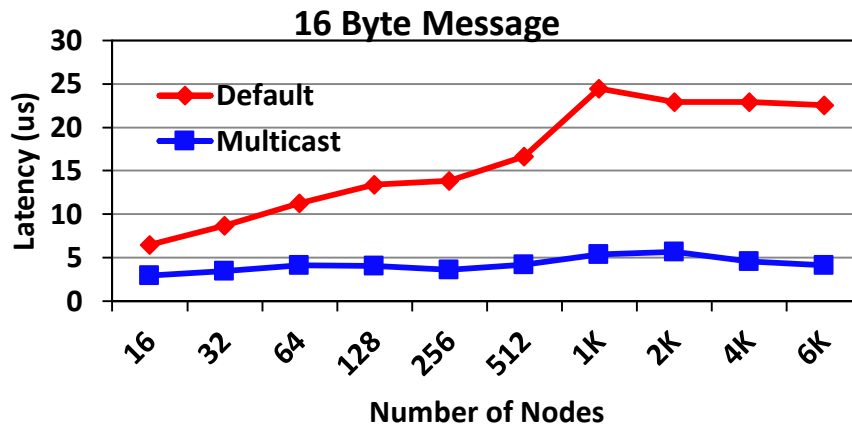
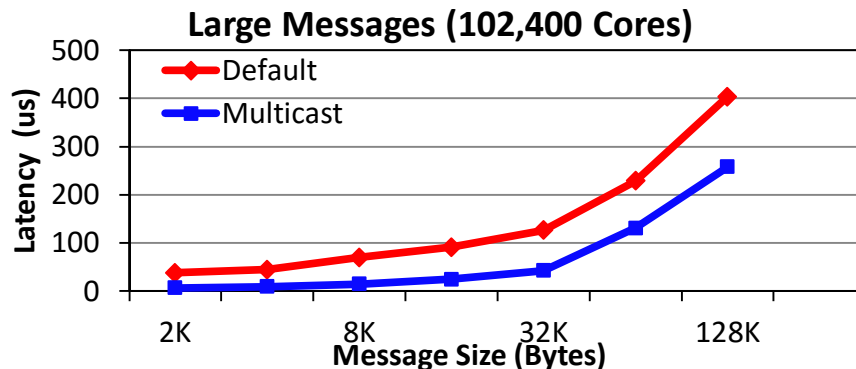
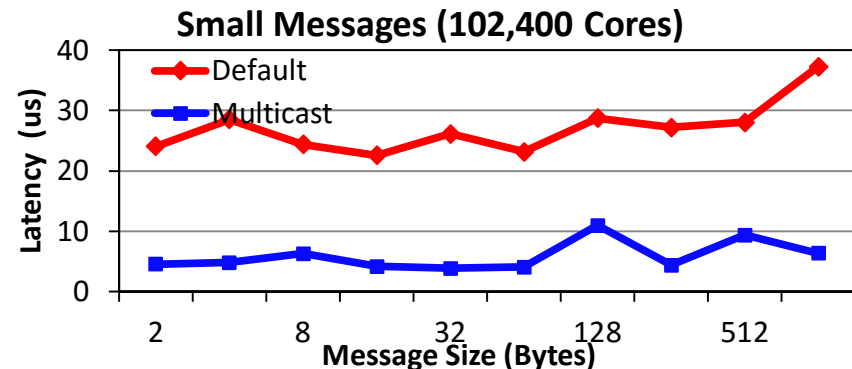
TrueScale-QDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
ConnectX-3-FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch
ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
ConnectX-5-EDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB Switch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

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

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

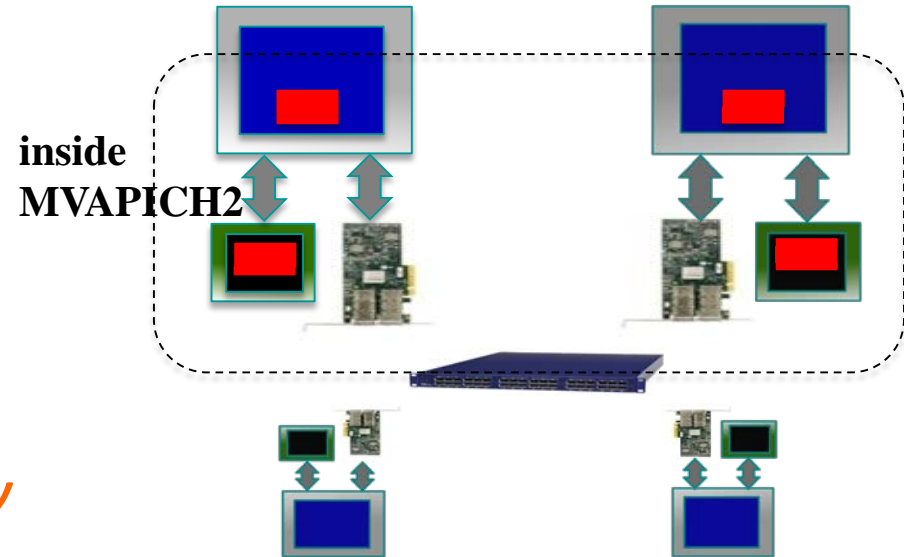
At Sender:

```
MPI_Send(s_devbuf, size, ...);
```

At Receiver:

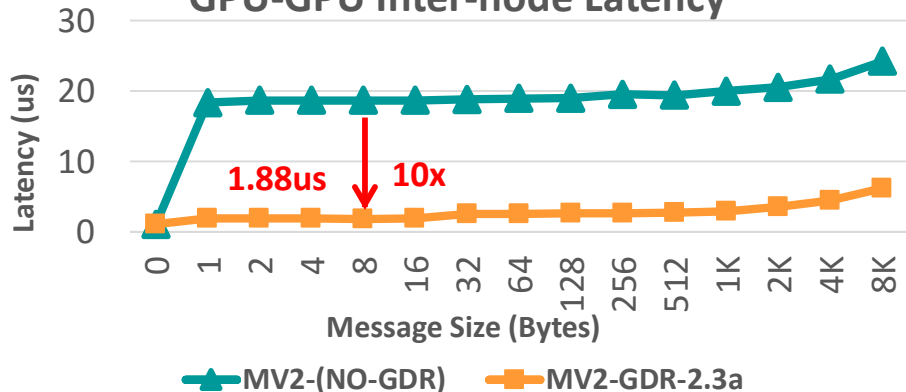
```
MPI_Recv(r_devbuf, size, ...);
```

High Performance and High Productivity

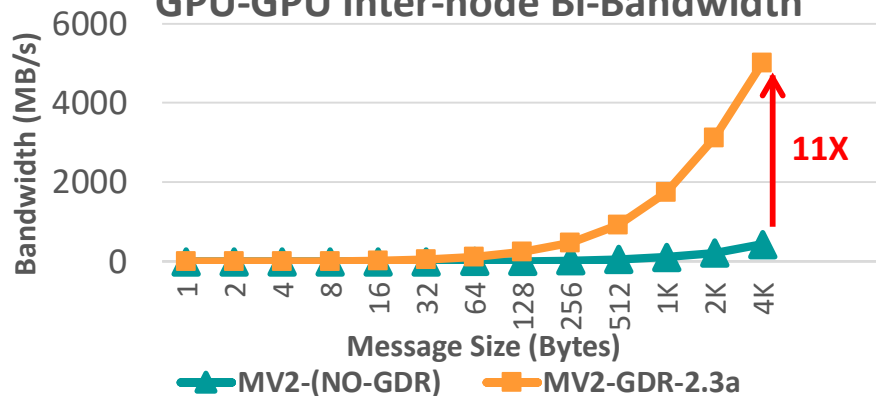


Optimized MVAPICH2-GDR Design

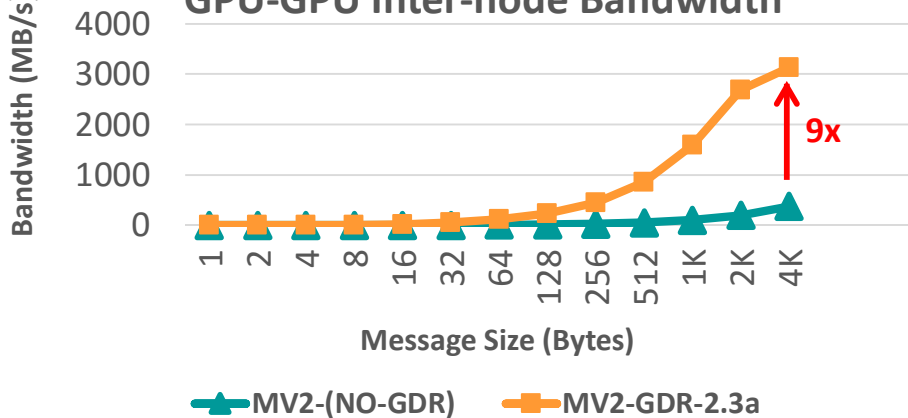
GPU-GPU Inter-node Latency



GPU-GPU Inter-node Bi-Bandwidth



GPU-GPU Inter-node Bandwidth



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

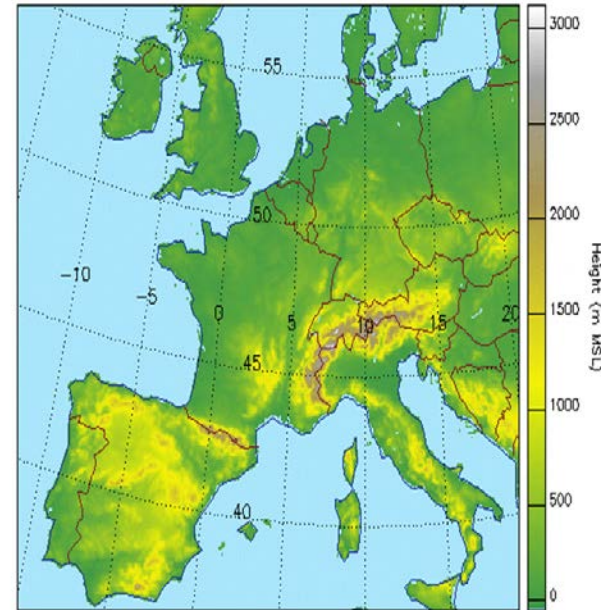
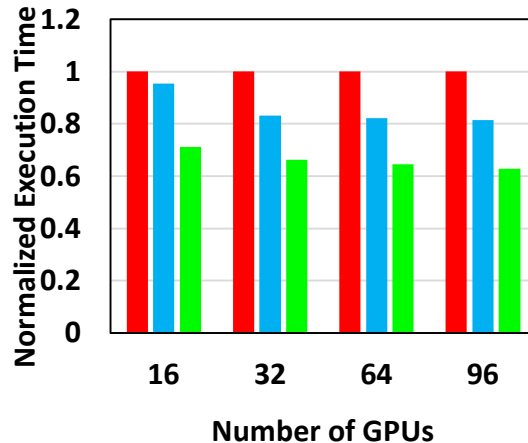
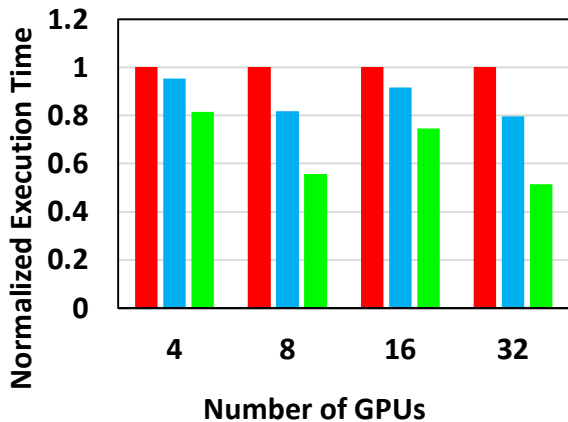
Application-Level Evaluation (Cosmo) and Weather Forecasting in Switzerland

Wilkes GPU Cluster

CSCS GPU cluster

■ Default ■ Callback-based ■ Event-based

■ Default ■ Callback-based ■ Event-based



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

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

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

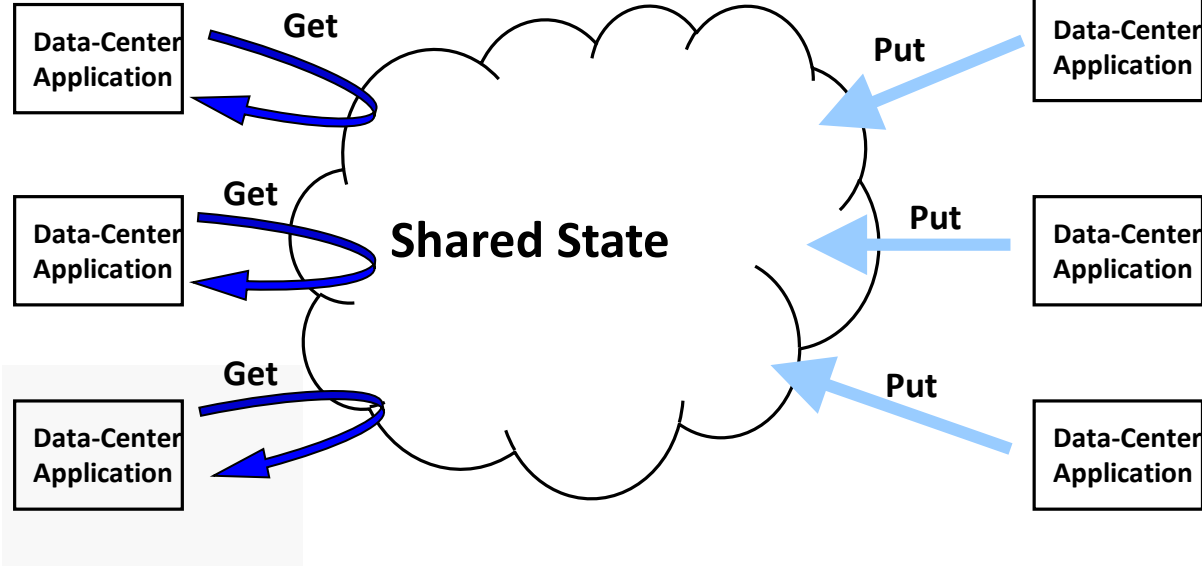
Designing RDMA-based Middleware for Clusters and Datacenters

- 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 and TensorFlow
- Virtualization Support with SR-IOV and Containers

Data-Center Service Primitives

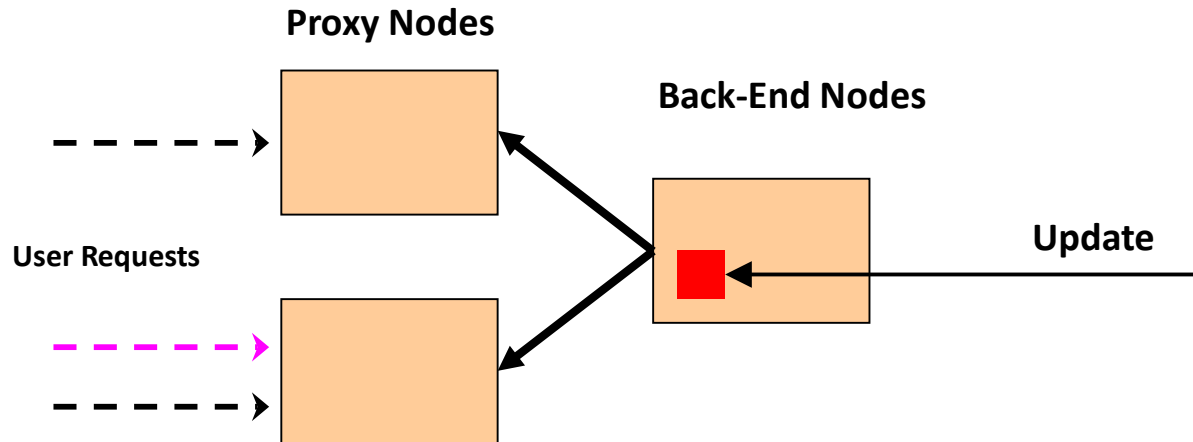
- Common Services needed by Data-Centers
 - Better resource management
 - Higher performance provided to higher layers
- Service Primitives
 - Soft Shared State
 - Distributed Lock Management
 - Global Memory Aggregator
- Network Based Designs
 - RDMA, Remote Atomic Operations

Soft Shared State

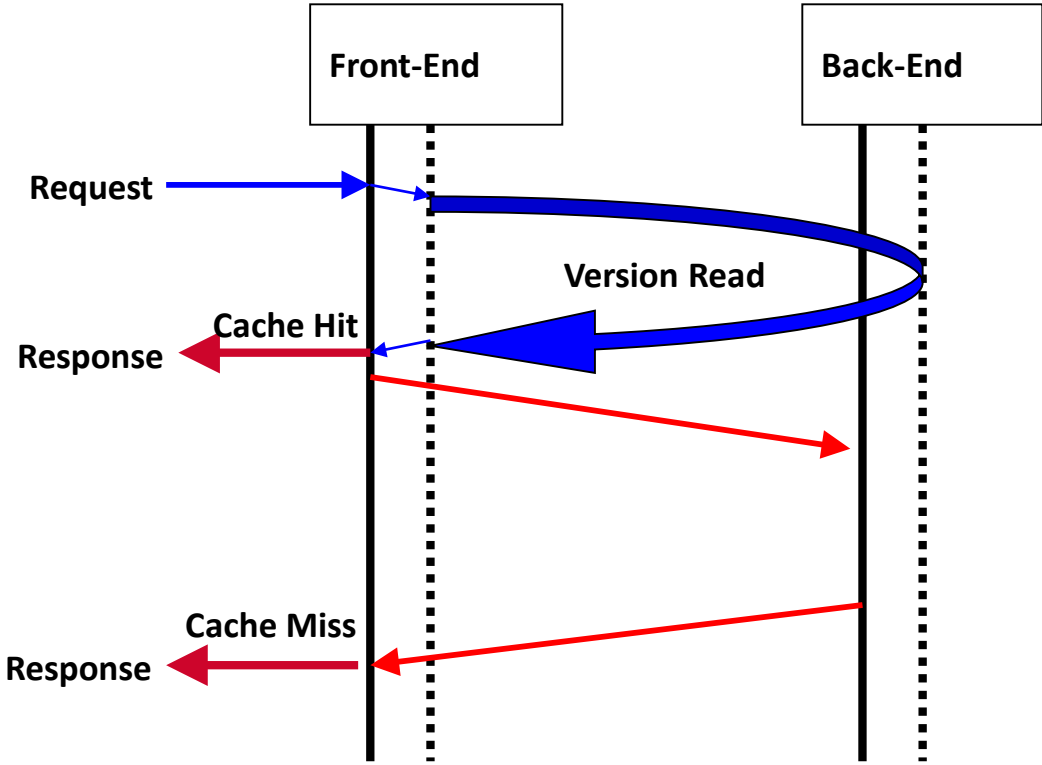


Active Caching

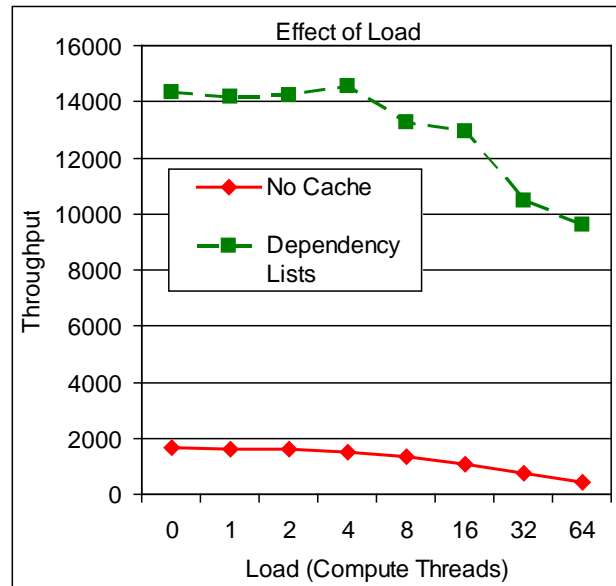
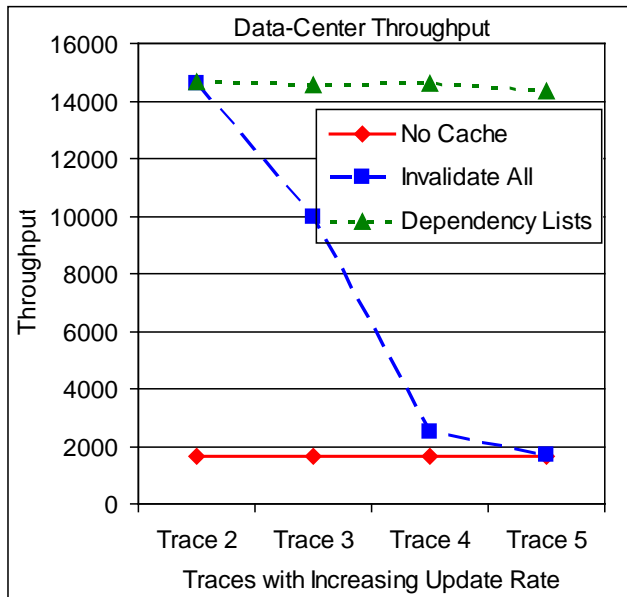
- Dynamic data caching – challenging!
- Cache Consistency and Coherence
 - Become more important than in static case



RDMA based Client Polling Design



Active Caching – Performance Benefits



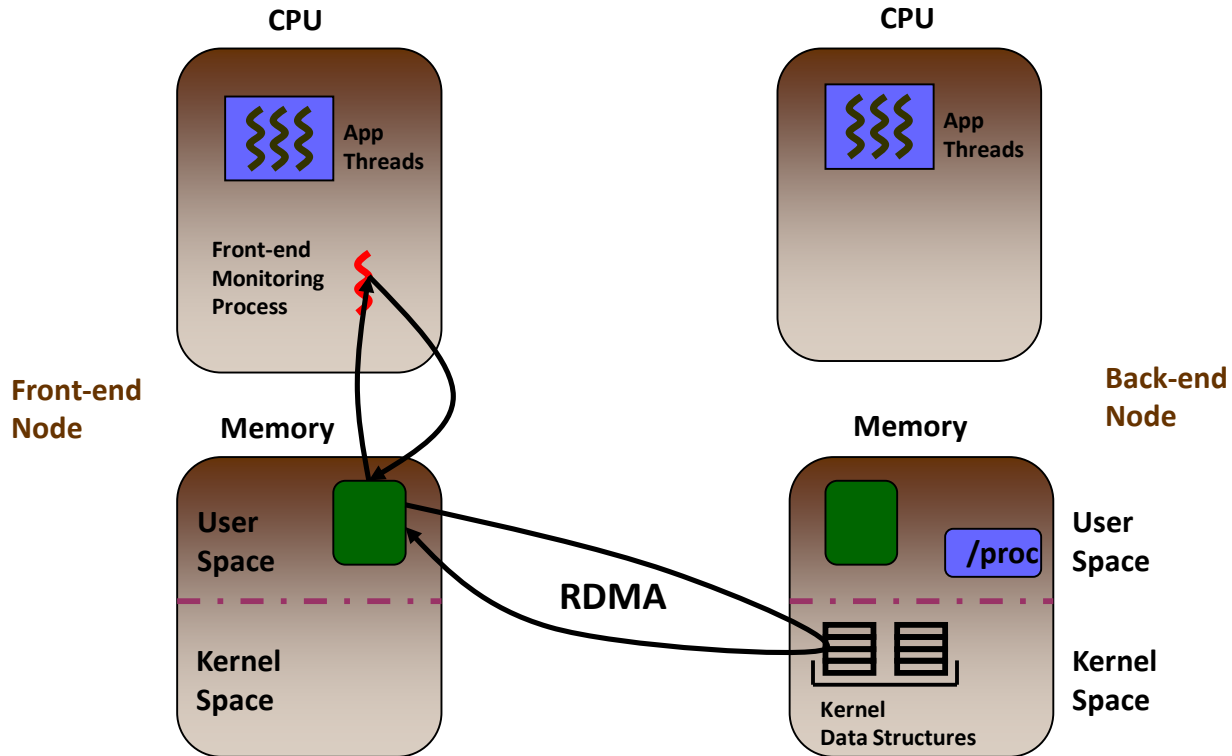
- Higher overall performance – Up to an order of magnitude
- Performance is sustained under loaded conditions

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)

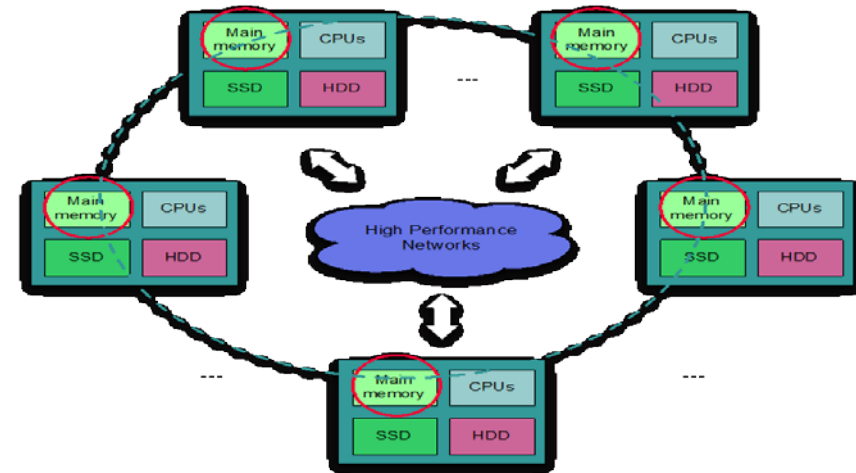
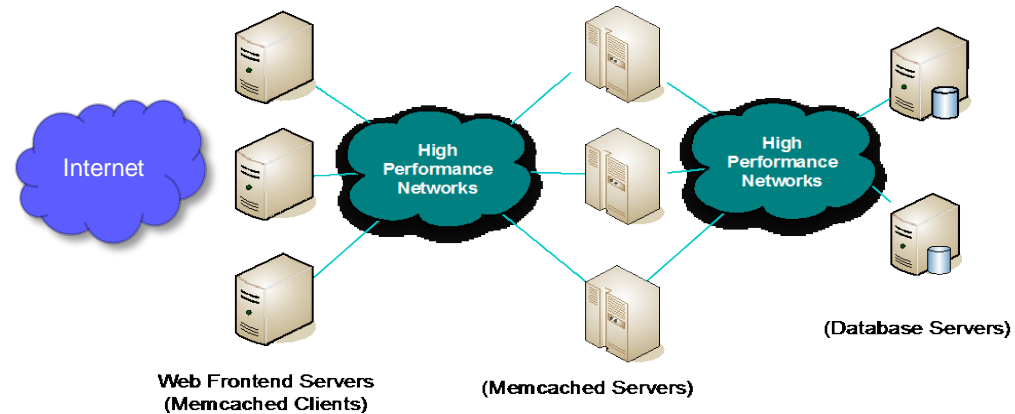


Designing RDMA-based Middleware for Clusters and Datacenters

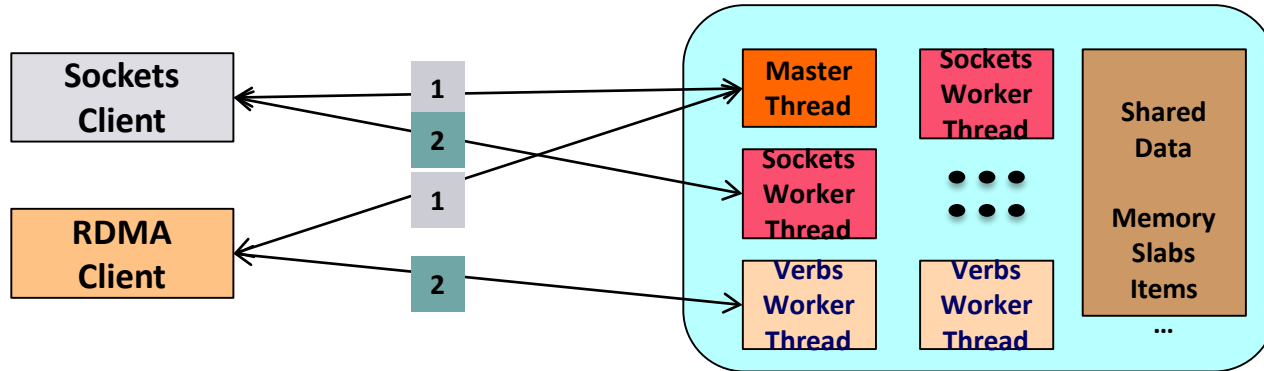
- High-Performance Programming Models Support for HPC Clusters
- RDMA-Enabled Communication Substrate for Common Services in Datacenters
- High-Performance and Scalable Memcached
- RDMA-Enabled Spark and Hadoop (HDFS, HBase, MapReduce)
- Deep Learning with Scale-Up and Scale-Out
 - Caffe and TensorFlow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

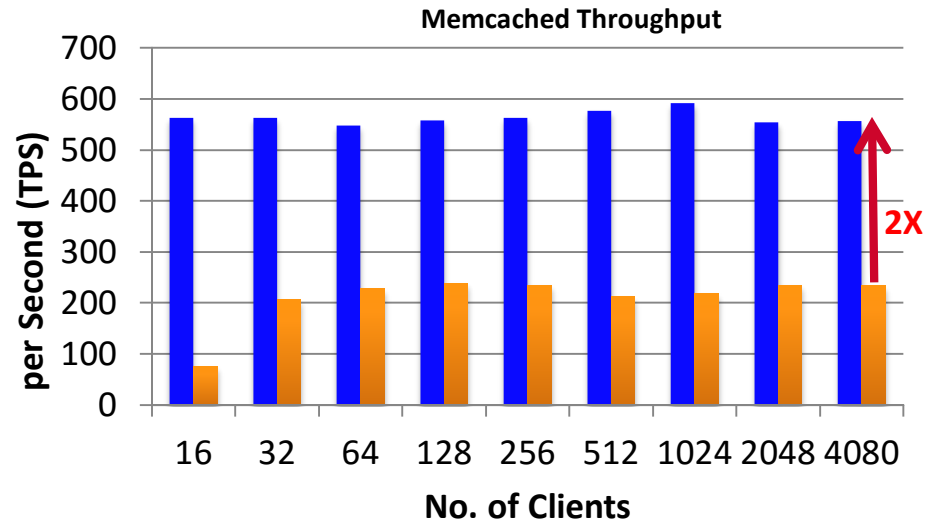
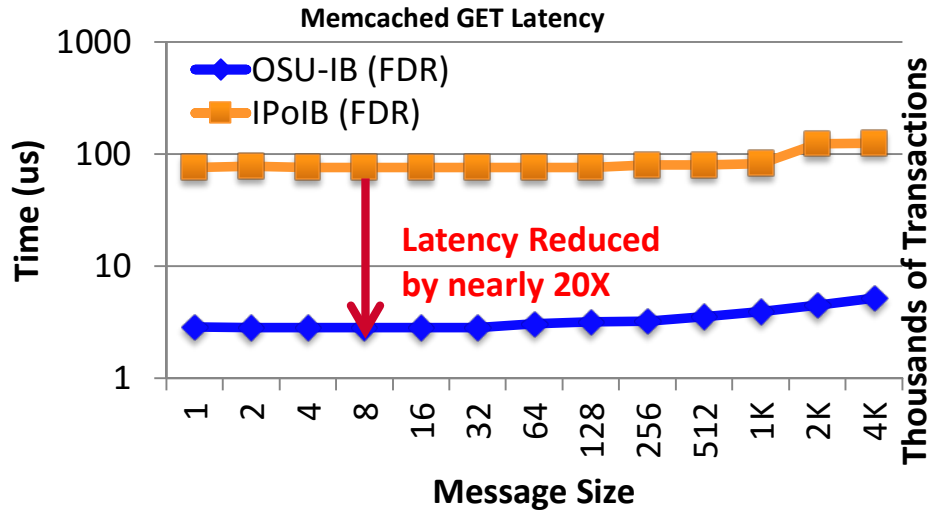


Memcached-RDMA Design



- 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

Designing RDMA-based 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 and TensorFlow
- Virtualization Support with SR-IOV and Containers

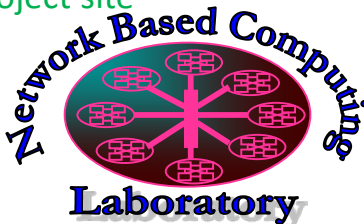
The High-Performance Big Data (HiBD) Project

- 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: 290 organizations from 34 countries
- More than 27,300 downloads from the project site

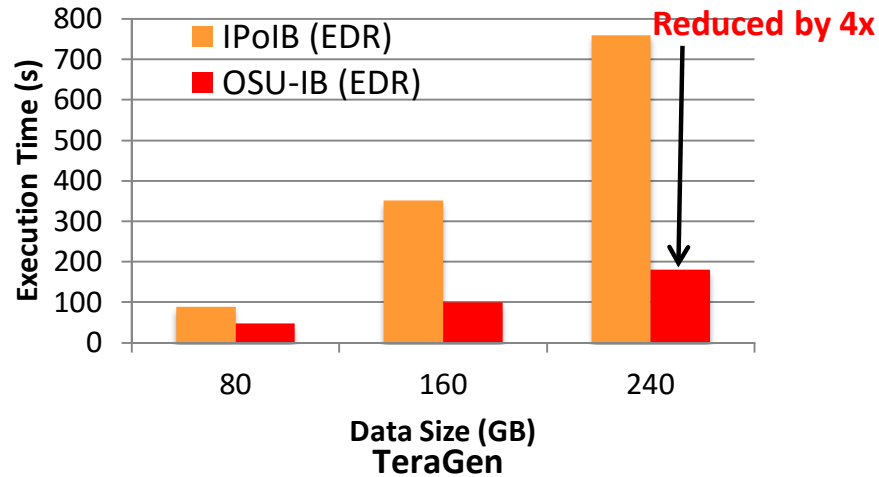
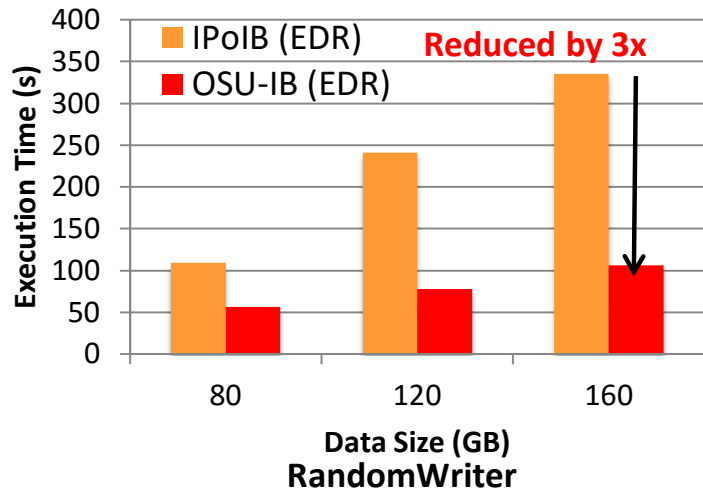
Available for InfiniBand and RoCE
Also run on Ethernet

Available for x86 and OpenPOWER

Support for Singularity and Docker



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

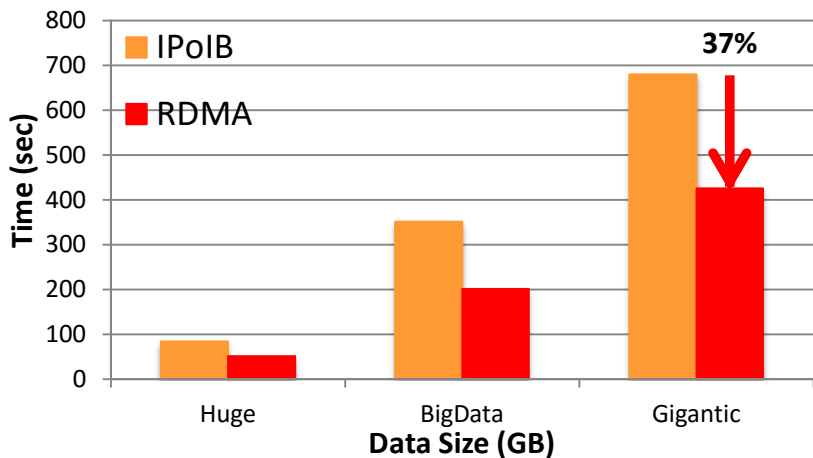
- RandomWriter

- **3x** improvement over IPoIB for 80-160 GB file size

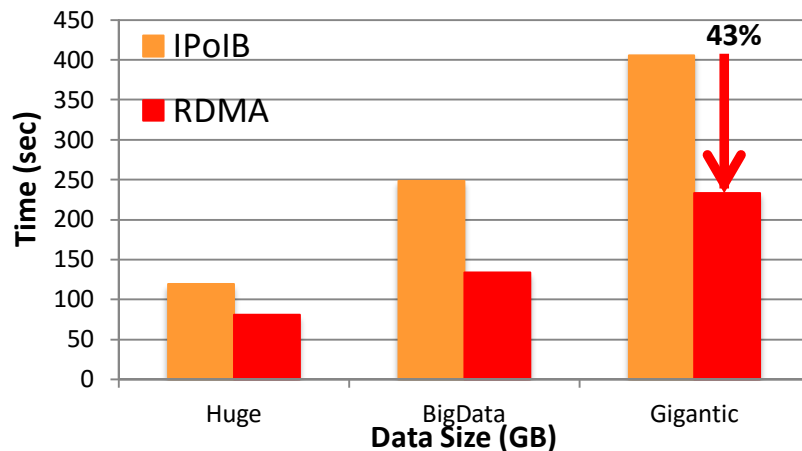
- TeraGen

- **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

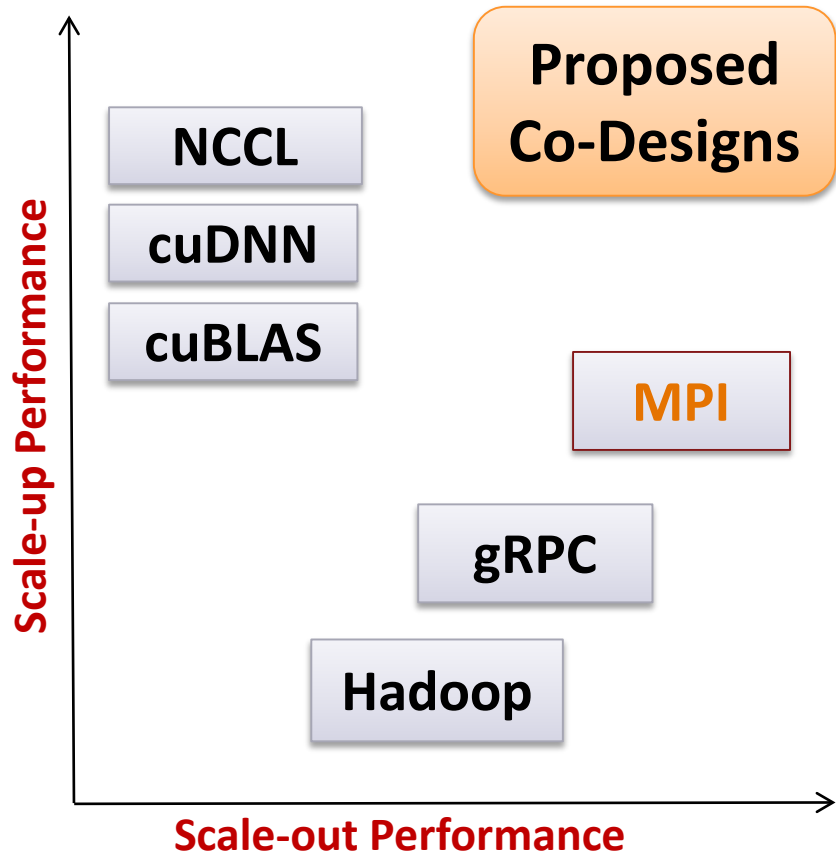
- InfiniBand FDR, SSD, 32/64 Worker Nodes, 768/1536 Cores, (768/1536M 768/1536R)
- RDMA vs. IPoIB with 768/1536 concurrent tasks, single SSD per node.
 - 32 nodes/768 cores: Total time reduced by 37% over IPoIB (56Gbps)
 - 64 nodes/1536 cores: Total time reduced by 43% over IPoIB (56Gbps)

Designing RDMA-based 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 and TensorFlow
- Virtualization Support with SR-IOV and Containers

Deep Learning: New Challenges for Communication Runtimes

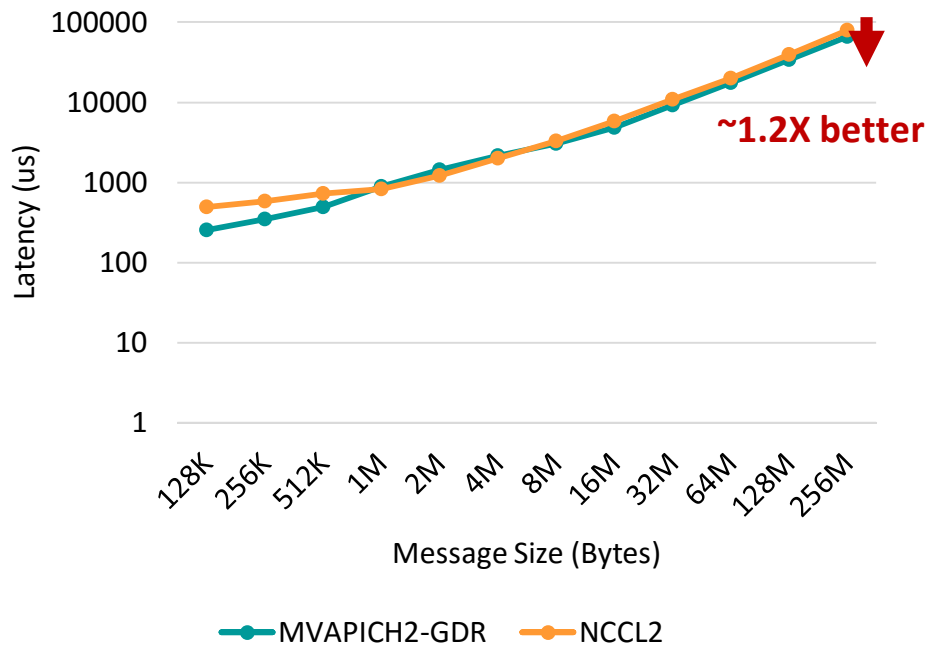
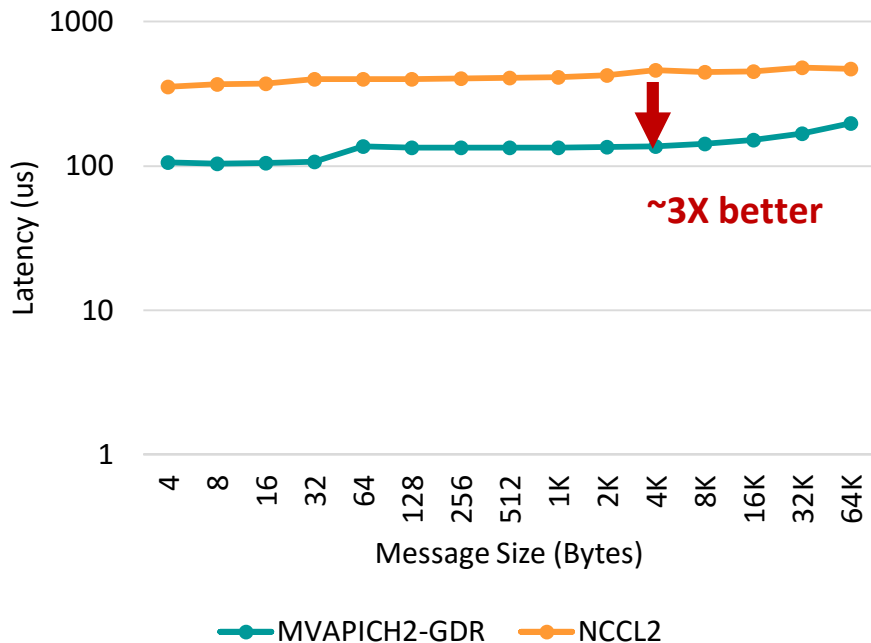
- Deep Learning frameworks are a different game altogether
 - Unusually large message sizes (order of megabytes)
 - Most communication based on GPU buffers
- Existing State-of-the-art
 - cuDNN, cuBLAS, NCCL --> **scale-up** performance
 - CUDA-Aware MPI --> **scale-out** performance
 - For small and 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)*

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

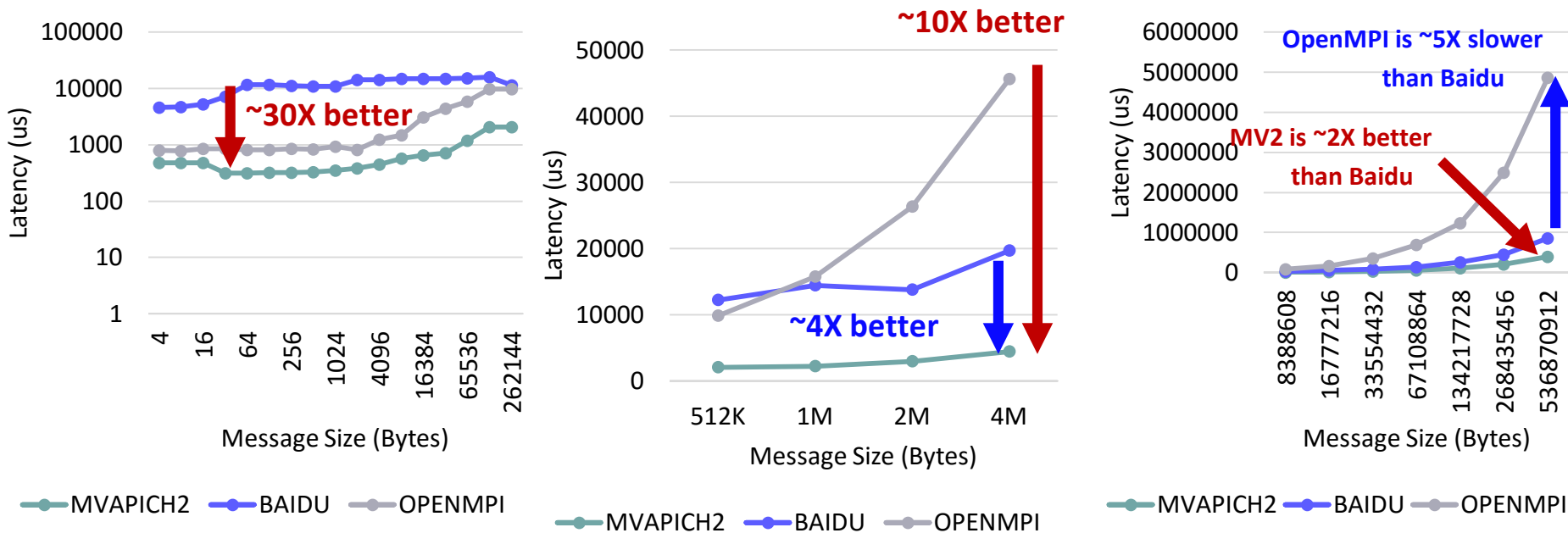


***Will be available with upcoming MVAPICH2-GDR 2.3b**

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

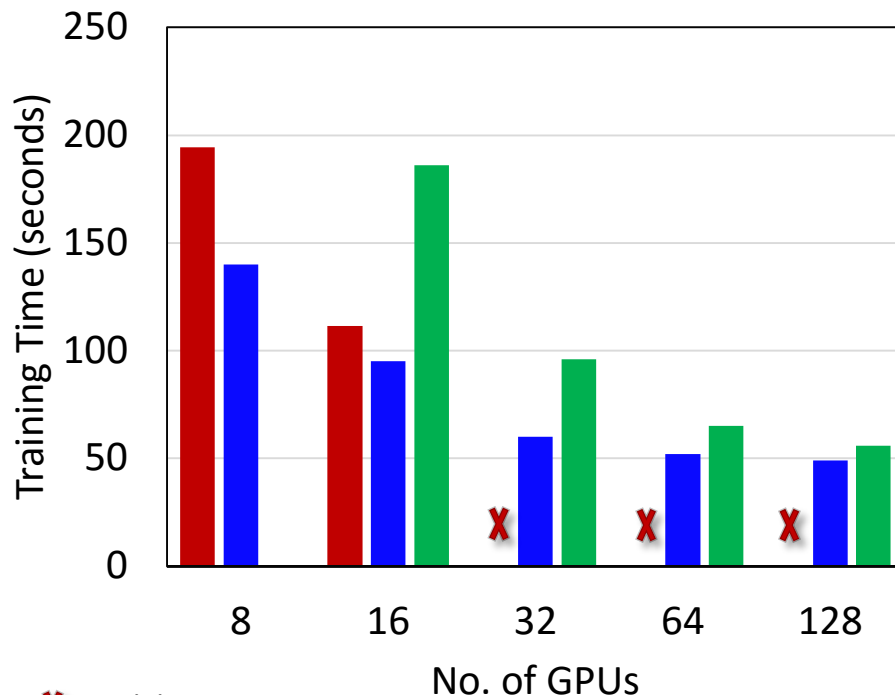
OSU-Caffe: Scalable Deep Learning

- 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/>

GoogLeNet (ImageNet) on 128 GPUs



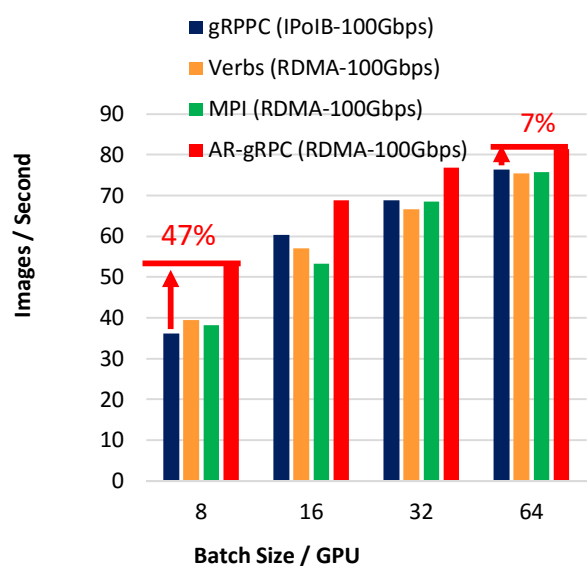
X Invalid use case

■ Caffe ■ OSU-Caffe (1024) ■ OSU-Caffe (2048)

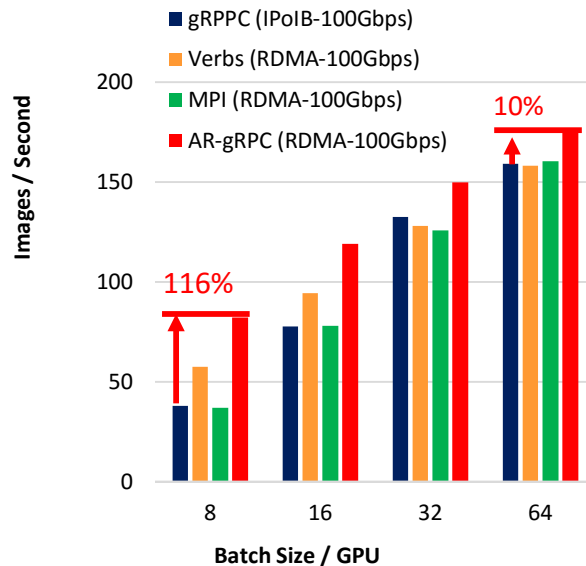
RDMA-TensorFlow Distribution

- High-Performance Design of TensorFlow over RDMA-enabled Interconnects
 - High performance RDMA-enhanced design with native InfiniBand support at the verbs-level for gRPC and TensorFlow
 - RDMA-based data communication
 - Adaptive communication protocols
 - Dynamic message chunking and accumulation
 - Support for RDMA device selection
 - Easily configurable for different protocols (native InfiniBand and IPoIB)
- Current release: **0.9.1**
 - Based on Google TensorFlow **1.3.0**
 - Tested with
 - Mellanox InfiniBand adapters (e.g., EDR)
 - NVIDIA GPGPU K80
 - Tested with CUDA 8.0 and CUDNN 5.0
 - <http://hidl.cse.ohio-state.edu>

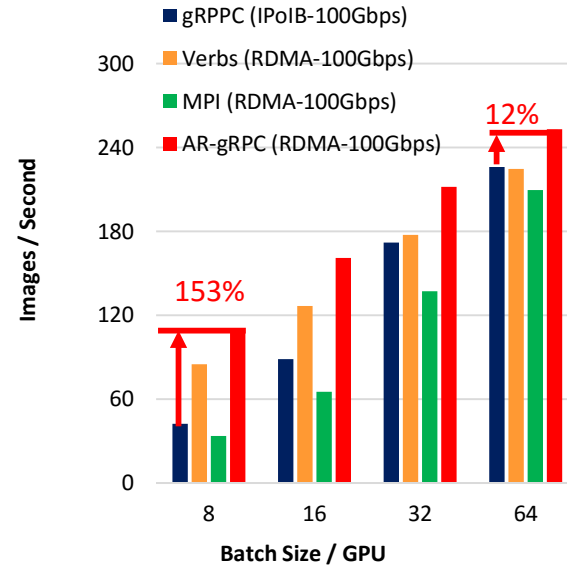
Performance Benefit for TensorFlow (Inception3)



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

Concluding Remarks

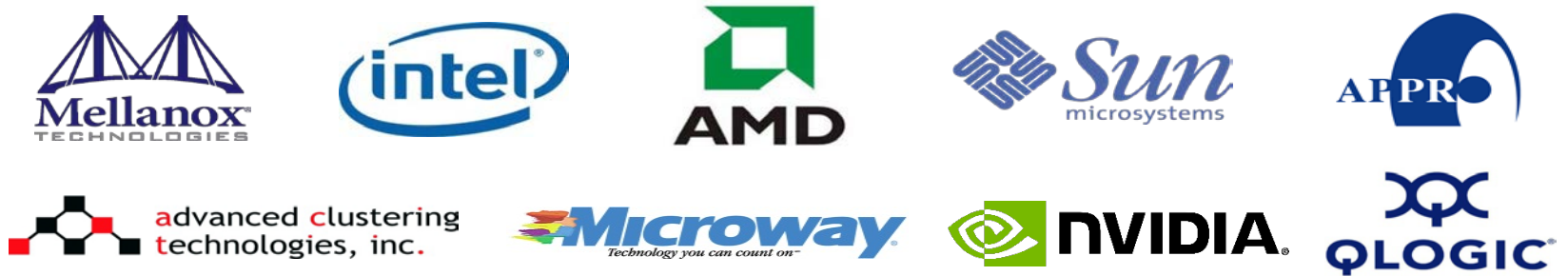
- Next generation Clusters and Data Centers need to be designed with a holistic view of HPC, Big Data, Deep Learning, and Cloud
- Presented an overview of the networking technology trends exploiting RDMA
- Presented some of the RDMA-based approaches and results along these directions
- Enable HPC, Big Data, Deep Learning and Cloud community to take advantage of modern RDMA-based 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.)
- M. Bayatpour (Ph.D.)
- S. Chakraborty (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.)

Current Students (Undergraduate)

- N. Sarkauskas (B.S.)
- V. Gangal (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.)
- R. Biswas (M.S.)
- 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.)

Past Post-Docs

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

Past Research Scientist

- K. Hamidouche
- S. Sur

Past Programmers

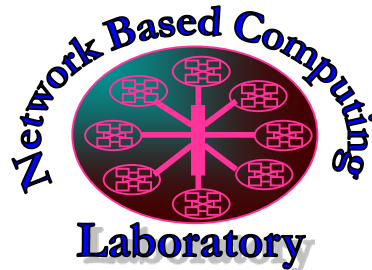
- D. Bureddy
- J. Perkins

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/Spark/Big Data Programmer/Software Engineer
 - Deep Learning 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/>



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



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



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