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.

Resurgence of AI/Machine Learning/Deep Learning

ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.

MACHINE LEARNING
Machine learning begins to flourish.

DEEP LEARNING
Deep learning breakthroughs drive AI boom.


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 $\rightarrow$ 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

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

100 times in the last 17 years
Available Interconnects and Protocols for Data Centers

Application / Middleware Interface

Sockets

Verbs

Protocol

Kernel Space

TCP/IP

RSockets

SDP

TCP/IP

RDMA

RDMA

User Space

User Space

User Space

User Space

Adapter

Ethernet Driver

IPoIB

Hardware Offload

User Space

RDMA

User Space

User Space

User Space

Switch

Ethernet Adapter

InfiniBand Adapter

Ethernet Adapter

InfiniBand Adapter

Ethernet Adapter

InfiniBand Adapter

Ethernet Adapter

InfiniBand Adapter

Omni-Path Adapter

Switch

Ethernet Switch

IPoIB

Ethernet Switch

IPoIB

Ethernet Switch

RSockets

SDP

iWARP

RoCE

IB Native

Omni-Path Switch

1/10/25/40/50/100 GigE-TOE

1/10/25/40/50/100 GigE-TOE

RSockets

Sockets

OFI

100 Gb/s
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)

Initiator processor is involved only to:
1. Post send WQE
2. Pull out completed CQE from the send CQ

No involvement from the target processor

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](http://www.top500.org)
- Installations in the Top 50 (19 systems):

<table>
<thead>
<tr>
<th>Large-scale InfiniBand Installations</th>
<th>#2nd system (Sunway TaihuLight) also uses InfiniBand</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2,282,544 cores (Summit) at ORNL (1st)</strong></td>
<td>155,150 cores (JURECA) at FZI/Germany (38th)</td>
</tr>
<tr>
<td>1,572,480 cores (Sierra) at LLNL (3rd)</td>
<td>72,800 cores Cray CS-Storm in US (40th)</td>
</tr>
<tr>
<td>391,680 cores (ABCI) at AIST/Japan (5th)</td>
<td>72,800 cores Cray CS-Storm in US (41st)</td>
</tr>
<tr>
<td>253,600 cores (HPC4) in Italy (13th)</td>
<td>78,336 cores (Electra) at NASA/Ames (43rd)</td>
</tr>
<tr>
<td>114,480 cores (Juwels Module 1) at FZI/Germany (23rd)</td>
<td>124,200 cores (Topaz) at ERDC DSRC/USA (44th)</td>
</tr>
<tr>
<td>241,108 cores (Pleiades) at NASA/Ames (24th)</td>
<td>60,512 cores NVIDIA DGX-1 at Facebook/USA (45th)</td>
</tr>
<tr>
<td>220,800 cores (Pangea) in France (30th)</td>
<td>60,512 cores (DGX Saturn V) at NVIDIA/USA (46th)</td>
</tr>
<tr>
<td>144,900 cores (Cheyenne) at NCAR/USA (31st)</td>
<td>113,832 cores (Damson) at AWE/UK (47th)</td>
</tr>
<tr>
<td>72,000 cores (IT0 – Subsystem A) in Japan (32nd)</td>
<td>72,000 cores (HPC2) in Italy (49th)</td>
</tr>
<tr>
<td>79,488 cores (JOLIOT-CURIE SKL) at CEA/France (34th)</td>
<td>and many more!</td>
</tr>
</tbody>
</table>
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
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

RDMA over Converged Enhanced Ethernet (RoCE)

- 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

Network Stack Comparison

Packet Header Comparison

Courtesy: OFED, Mellanox
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 support
Large-scale Omni-Path Installations

- 39 Omni-Path Clusters (7.8%) in the Jun’18 Top500 list
  - [http://www.top500.org](http://www.top500.org)

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

<table>
<thead>
<tr>
<th>Features</th>
<th>IB</th>
<th>iWARP/HSE</th>
<th>RoCE</th>
<th>RoCE v2</th>
<th>Omni-Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware Acceleration</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>RDMA</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Congestion Control</td>
<td>Yes</td>
<td>Optional</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Multipathing</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Atomic Operations</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Multicast</td>
<td>Optional</td>
<td>No</td>
<td>Optional</td>
<td>Optional</td>
<td>Optional</td>
</tr>
<tr>
<td>Data Placement</td>
<td>Ordered</td>
<td>Out-of-order</td>
<td>Ordered</td>
<td>Ordered</td>
<td>Ordered</td>
</tr>
<tr>
<td>Prioritization</td>
<td>Optional</td>
<td>Optional</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Fixed BW QoS (ETS)</td>
<td>No</td>
<td>Optional</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Ethernet Compatibility</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>TCP/IP Compatibility</td>
<td>Yes (using IPoIB)</td>
<td>Yes (using IPoIB)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Designing RDMA-based Communication and I/O Libraries for Clusters and Data Center Middleware: Challenges

Cluster and Data Center Middleware (MPI, PGAS, Memcached, HDFS, MapReduce, HBase, and gRPC/TensorFlow)

Networking Technologies (InfiniBand, 1/10/40/100 GigE and Intelligent NICs)

RDMA-based Communication Substrate

Threaded Models and Synchronization

Virtualization (SR-IOV)

I/O and File Systems

QoS & Fault Tolerance

Performance Tuning

Commodity Computing System Architectures (Multi- and Many-core architectures and accelerators)

Storage Technologies (HDD, SSD, NVM, and NVMe-SSD)

Applications

Programming Models (Sockets)
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](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

**Small Message Latency**

<table>
<thead>
<tr>
<th>Message Size (bytes)</th>
<th>Latency (us)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TrueScale-QDR</td>
<td>1.19</td>
</tr>
<tr>
<td>ConnectX-3-FDR</td>
<td>1.15</td>
</tr>
<tr>
<td>ConnectIB-Dual FDR</td>
<td>1.11</td>
</tr>
<tr>
<td>ConnectX-5-EDR</td>
<td>1.04</td>
</tr>
<tr>
<td>Omni-Path</td>
<td>0.98</td>
</tr>
</tbody>
</table>

**Large Message Latency**

<table>
<thead>
<tr>
<th>Message Size (bytes)</th>
<th>Latency (us)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TrueScale-QDR</td>
<td>120</td>
</tr>
<tr>
<td>ConnectX-3-FDR</td>
<td>110</td>
</tr>
<tr>
<td>ConnectIB-Dual FDR</td>
<td>100</td>
</tr>
<tr>
<td>ConnectX-5-EDR</td>
<td>90</td>
</tr>
<tr>
<td>Omni-Path</td>
<td>80</td>
</tr>
</tbody>
</table>

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

Unidirectional Bandwidth

Bidirectional Bandwidth

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 (>= CUDA 4.0)
- Overlaps data movement from GPU with RDMA transfers

At Sender:

```c
MPI_Send(s_devbuf, size, …);
```

At Receiver:

```c
MPI_Recv(r_devbuf, size, …);
```

High Performance and High Productivity
Optimized MVAPICH2-GDR Design

**GPU-GPU Inter-node Latency**
- MV2-(NO-GDR)
- MV2-GDR-2.3a

**GPU-GPU Inter-node Bi-Bandwidth**
- MV2-(NO-GDR)
- MV2-GDR-2.3a

**GPU-GPU Inter-node Bandwidth**
- MV2-(NO-GDR)
- MV2-GDR-2.3a

**Graph Details**
- 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

**Performance Metrics**
- **Latency (us)**:
  - 1.88us
  - 10x

- **Bandwidth (MB/s)**:
  - 9x
  - 11X

**Message Size (Bytes)**:
- 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1K, 2K, 4K
Application-Level Evaluation (Cosmo) and Weather Forecasting in Switzerland

Wilkes GPU Cluster

- Default
- Callback-based
- Event-based

CSCS GPU Cluster

- 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

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

- Request
- Cache Hit
- Version Read
- Cache Miss
- Response

Front-End

Back-End
Active Caching – Performance Benefits

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

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)

- 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 (4 bytes)
  - 4080 clients OSU-IB: 556 Kops/sec, IPoIB: 233 Kops/s
  - Nearly 2X improvement in throughput

Experiments on TACC Stampede (Intel SandyBridge Cluster, IB: FDR)
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](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
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 --&gt; **scale-up** performance
  - CUDA-Aware MPI --&gt; **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**?

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

- 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/
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](http://hidl.cse.ohio-state.edu)
Performance Benefit for TensorFlow (Inception3)

- 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

[Logos of various sponsors]

Equipment Support by

[Logos of various sponsors]
Personnel Acknowledgments

<table>
<thead>
<tr>
<th>Current Students (Graduate)</th>
<th>Current Students (Undergraduate)</th>
<th>Current Research Scientists</th>
<th>Current Research Specialist</th>
</tr>
</thead>
<tbody>
<tr>
<td>– A. Awan (Ph.D.)</td>
<td>– J. Hashmi (Ph.D.)</td>
<td>– X. Lu</td>
<td>– J. Smith</td>
</tr>
<tr>
<td>– S. Chakraborthy (Ph.D.)</td>
<td>– P. Kousha (Ph.D.)</td>
<td>– A. Mamidala</td>
<td></td>
</tr>
<tr>
<td>– S. Guganani (Ph.D.)</td>
<td>– H. Shi (Ph.D.)</td>
<td>– V. Meshram</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– A. Moody</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– S. Naravula</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– R. Noronha</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– X. Ouyang</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– S. Potluri</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– J. Liu</td>
<td>R. Rajachandrasekar (Ph.D.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– M. Luo</td>
<td>G. Santhanaraman (Ph.D.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– A. Mamidala</td>
<td>A. Singh (Ph.D.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– G. Marsh</td>
<td>J. Sridhar (M.S.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– V. Meshram</td>
<td>S. Sur (Ph.D.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– A. Moody</td>
<td>H. Subramoni (Ph.D.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– S. Naravula</td>
<td>K. Vaidyanathan (Ph.D.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– R. Noronha</td>
<td>A. Vishnu (Ph.D.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– X. Ouyang</td>
<td>J. Wu (Ph.D.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– S. Potluri</td>
<td>W. Yu (Ph.D.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– J. Liu</td>
<td>J. Zhang (Ph.D.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– M. Luo</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– A. Mamidala</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– G. Marsh</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– V. Meshram</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– A. Moody</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– S. Naravula</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– R. Noronha</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– X. Ouyang</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– S. Potluri</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– J. Liu</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– M. Luo</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– E. Mancini</td>
<td></td>
</tr>
<tr>
<td>Past Students</td>
<td>W. Huang (Ph.D.)</td>
<td>J. Liu (Ph.D.)</td>
<td>R. Rajachandrasekar (Ph.D.)</td>
</tr>
<tr>
<td>– A. Augustine (M.S.)</td>
<td>– W. Jhang (Ph.D.)</td>
<td>– M. Luo (Ph.D.)</td>
<td>G. Santhanaraman (Ph.D.)</td>
</tr>
<tr>
<td>– P. Balaji (Ph.D.)</td>
<td>– J. Jose (Ph.D.)</td>
<td>– A. Mamidala (Ph.D.)</td>
<td>A. Singh (Ph.D.)</td>
</tr>
<tr>
<td>– R. Biswas (M.S.)</td>
<td>– S. Kini (M.S.)</td>
<td>– G. Marsh (M.S.)</td>
<td>J. Sridhar (M.S.)</td>
</tr>
<tr>
<td>– S. Bhagvat (M.S.)</td>
<td>– M. Koop (Ph.D.)</td>
<td>– V. Meshram (M.S.)</td>
<td>S. Sur (Ph.D.)</td>
</tr>
<tr>
<td>– A. Bhat (M.S.)</td>
<td>– K. Kulkarni (M.S.)</td>
<td>– A. Moody (M.S.)</td>
<td>H. Subramoni (Ph.D.)</td>
</tr>
<tr>
<td>– D. Buntinas (Ph.D.)</td>
<td>– R. Kumar (M.S.)</td>
<td>– S. Naravula (Ph.D.)</td>
<td>K. Vaidyanathan (Ph.D.)</td>
</tr>
<tr>
<td>– L. Chai (Ph.D.)</td>
<td>– S. Krishnamoorthy (M.S.)</td>
<td>– R. Noronha (Ph.D.)</td>
<td>A. Vishnu (Ph.D.)</td>
</tr>
<tr>
<td>– B. Chandrasekharan (M.S.)</td>
<td>– K. Kandala (Ph.D.)</td>
<td>– X. Ouyang (Ph.D.)</td>
<td>J. Wu (Ph.D.)</td>
</tr>
<tr>
<td>– N. Dandapanthula (M.S.)</td>
<td>– M. Li (Ph.D.)</td>
<td>– S. Pai (M.S.)</td>
<td>W. Yu (Ph.D.)</td>
</tr>
<tr>
<td>– V. Dhanraj (M.S.)</td>
<td>– P. Lai (M.S.)</td>
<td>– S. Potluri (Ph.D.)</td>
<td>J. Zhang (Ph.D.)</td>
</tr>
<tr>
<td>– T. Gangadharpappa (M.S.)</td>
<td>– J. Lin</td>
<td>– S. Marcarelli</td>
<td></td>
</tr>
<tr>
<td>– K. Gopalakrishnan (M.S.)</td>
<td>– M. Luo</td>
<td>– J. Vienne</td>
<td></td>
</tr>
<tr>
<td></td>
<td>– E. Mancini</td>
<td>– H. Wang</td>
<td></td>
</tr>
<tr>
<td>Past Post-Docs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>– D. Banerjee</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>– X. Besseron</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>– H.-W. Jin</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Past Students</td>
<td>W. Huang (Ph.D.)</td>
<td>J. Liu (Ph.D.)</td>
<td>R. Rajachandrasekar (Ph.D.)</td>
</tr>
<tr>
<td>– A. Augustine (M.S.)</td>
<td>– W. Jhang (Ph.D.)</td>
<td>– M. Luo (Ph.D.)</td>
<td>G. Santhanaraman (Ph.D.)</td>
</tr>
<tr>
<td>– P. Balaji (Ph.D.)</td>
<td>– J. Jose (Ph.D.)</td>
<td>– A. Mamidala (Ph.D.)</td>
<td>A. Singh (Ph.D.)</td>
</tr>
<tr>
<td>– R. Biswas (M.S.)</td>
<td>– S. Kini (M.S.)</td>
<td>– G. Marsh (M.S.)</td>
<td>J. Sridhar (M.S.)</td>
</tr>
<tr>
<td>– S. Bhagvat (M.S.)</td>
<td>– M. Koop (Ph.D.)</td>
<td>– V. Meshram (M.S.)</td>
<td>S. Sur (Ph.D.)</td>
</tr>
<tr>
<td>– A. Bhat (M.S.)</td>
<td>– K. Kulkarni (M.S.)</td>
<td>– A. Moody (M.S.)</td>
<td>H. Subramoni (Ph.D.)</td>
</tr>
<tr>
<td>– D. Buntinas (Ph.D.)</td>
<td>– R. Kumar (M.S.)</td>
<td>– S. Naravula (Ph.D.)</td>
<td>K. Vaidyanathan (Ph.D.)</td>
</tr>
<tr>
<td>– L. Chai (Ph.D.)</td>
<td>– S. Krishnamoorthy (M.S.)</td>
<td>– R. Noronha (Ph.D.)</td>
<td>A. Vishnu (Ph.D.)</td>
</tr>
<tr>
<td>– B. Chandrasekharan (M.S.)</td>
<td>– K. Kandala (Ph.D.)</td>
<td>– X. Ouyang (Ph.D.)</td>
<td>J. Wu (Ph.D.)</td>
</tr>
<tr>
<td>– N. Dandapanthula (M.S.)</td>
<td>– M. Li (Ph.D.)</td>
<td>– S. Pai (M.S.)</td>
<td>W. Yu (Ph.D.)</td>
</tr>
<tr>
<td>– V. Dhanraj (M.S.)</td>
<td>– P. Lai (M.S.)</td>
<td>– S. Potluri (Ph.D.)</td>
<td>J. Zhang (Ph.D.)</td>
</tr>
<tr>
<td>– T. Gangadharpappa (M.S.)</td>
<td>– J. Lin</td>
<td>– S. Marcarelli</td>
<td></td>
</tr>
<tr>
<td>– K. Gopalakrishnan (M.S.)</td>
<td>– M. Luo</td>
<td>– J. Vienne</td>
<td></td>
</tr>
<tr>
<td></td>
<td>– E. Mancini</td>
<td>– H. Wang</td>
<td></td>
</tr>
</tbody>
</table>

Network Based Computing Laboratory

KBNet ‘18

60
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/

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

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