MVAPICH2-GDR: Pushing the Frontier of HPC and Deep Learning

Talk at Mellanox booth (SC 2018)

by

Dhabaleswar K. (DK) Panda
The Ohio State University
E-mail: panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda
Outline

• Overview of the MVAPICH2 Project
• MVAPICH2-GPU with GPUDirect-RDMA (GDR)
• What’s new with MVAPICH2-GDR
• High-Performance Deep Learning (HiDL) with MVAPICH2-GDR
• Conclusions
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 2,950 organizations in 86 countries
  – More than 506,000 (> 0.5 million) downloads from the OSU site directly
  – Empowering many TOP500 clusters (Nov ‘18 ranking)
    • 3rd ranked 10,649,640-core cluster (Sunway TaihuLight) at NSC, Wuxi, China
    • 14th, 556,104 cores (Oakforest-PACS) in Japan
    • 17th, 367,024 cores (Stampede2) at TACC
    • 27th, 241,108-core (Pleiades) at NASA and many others
  – Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, and OpenHPC)
  – [http://mvapich.cse.ohio-state.edu](http://mvapich.cse.ohio-state.edu) Partner in the upcoming TACC Frontera System
  – Empowering Top500 systems for over a decade
MVAPICH2 Release Timeline and Downloads

Network Based Computing Laboratory
# Architecture of MVAPICH2 Software Family

## High Performance Parallel Programming Models

<table>
<thead>
<tr>
<th>Message Passing Interface (MPI)</th>
<th>PGAS (UPC, OpenSHMEM, CAF, UPC++)</th>
<th>Hybrid --- MPI + X (MPI + PGAS + OpenMP/Cilk)</th>
</tr>
</thead>
</table>

## 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

**Transport Protocols**: RC, XRC, UD, DC

**Modern Features**: UMR, ODP, SR-IOV, Multi Rail

**Support for Modern Multi-/Many-core Architectures**

**Support for Modern Multi-/Many-core Architectures**

**Transport Protocols**: MCDRAM*, NVLink*, CAPI*

### Modern Features

* Upcoming

---

Network Based Computing Laboratory

SC ‘18
## MVAPICH2 Software Family

### High-Performance Parallel Programming Libraries

<table>
<thead>
<tr>
<th>Library</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVAPICH2</td>
<td>Support for InfiniBand, Omni-Path, Ethernet/iWARP, and RoCE</td>
</tr>
<tr>
<td>MVAPICH2-X</td>
<td>Advanced MPI features, OSU INAM, PGAS (OpenSHMEM, UPC, UPC++, and CAF), and MPI+PGAS programming models with unified communication runtime</td>
</tr>
<tr>
<td>MVAPICH2-GDR</td>
<td>Optimized MPI for clusters with NVIDIA GPUs</td>
</tr>
<tr>
<td>MVAPICH2-Virt</td>
<td>High-performance and scalable MPI for hypervisor and container based HPC cloud</td>
</tr>
<tr>
<td>MVAPICH2-EA</td>
<td>Energy aware and High-performance MPI</td>
</tr>
<tr>
<td>MVAPICH2-MIC</td>
<td>Optimized MPI for clusters with Intel KNC</td>
</tr>
</tbody>
</table>

### Microbenchmarks

<table>
<thead>
<tr>
<th>Library</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMB</td>
<td>Microbenchmarks suite to evaluate MPI and PGAS (OpenSHMEM, UPC, and UPC++) libraries for CPUs and GPUs</td>
</tr>
</tbody>
</table>

### Tools

<table>
<thead>
<tr>
<th>Library</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSU INAM</td>
<td>Network monitoring, profiling, and analysis for clusters with MPI and scheduler integration</td>
</tr>
<tr>
<td>OEMT</td>
<td>Utility to measure the energy consumption of MPI applications</td>
</tr>
</tbody>
</table>
MVAPICH2-GDR: Optimizing MPI Data Movement on GPU Clusters

- Connected as PCIe devices – **Flexibility but Complexity**

  ![Diagram showing connections between nodes and devices]

  - Memory buffers
  - **1. Intra-GPU**
  - **2. Intra-Socket GPU-GPU**
  - **3. Inter-Socket GPU-GPU**
  - **4. Inter-Node GPU-GPU**
  - **5. Intra-Socket GPU-Host**
  - **6. Inter-Socket GPU-Host**
  - **7. Inter-Node GPU-Host**

  - **8. Inter-Node GPU-GPU** with IB adapter on remote socket and more . . .

- For each path different schemes: Shared_mem, IPC, GPUDirect RDMA, pipeline ...
- Critical for runtimes to optimize data movement while hiding the complexity
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*
CUDA-Aware MPI: MVAPICH2-GDR 1.8-2.3 Releases

- Support for MPI communication from NVIDIA GPU device memory
- High performance RDMA-based inter-node point-to-point communication (GPU-GPU, GPU-Host and Host-GPU)
- High performance intra-node point-to-point communication for multi-GPU adapters/node (GPU-GPU, GPU-Host and Host-GPU)
- Taking advantage of CUDA IPC (available since CUDA 4.1) in intra-node communication for multiple GPU adapters/node
- Optimized and tuned collectives for GPU device buffers
- MPI datatype support for point-to-point and collective communication from GPU device buffers
- Unified memory
Using MVAPICH2-GPUDirect Version

- MVAPICH2-2.3 with GDR support can be downloaded from
  
  https://mvapich.cse.ohio-state.edu/download/mvapich2gdr/

- System software requirements
  
  • Mellanox OFED 3.2 or later
  
  • NVIDIA Driver 367.48 or later
  
  • NVIDIA CUDA Toolkit 7.5/8.0/9.0 or later
  
  • Plugin for GPUDirect RDMA
    

  • Strongly recommended

  • GDRCOPY module from NVIDIA
    
    https://github.com/NVIDIA/gdrcopy

- Contact MVAPICH help list with any questions related to the package
  
  mvapich-help@cse.ohio-state.edu
MVAPICH2-GDR 2.3 GA

- Released on 11/10/2018
- Major Features and Enhancements
  - Based on MVAPICH2 2.3
  - Support for CUDA 10.0
  - Add support for Volta (V100) GPU
  - Support for OpenPOWER with NVLink
  - Efficient Multiple CUDA stream-based IPC communication for multi-GPU systems with and without NVLink
  - Enhanced performance of GPU-based point-to-point communication
  - Leverage Linux Cross Memory Attach (CMA) feature for enhanced host-based communication
  - Enhanced performance of MPI_Allreduce for GPU-resident data
  - InfiniBand Multicast (IB-MCAST) based designs for GPU-based broadcast and streaming applications
    - Basic support for IB-MCAST designs with GPUDirect RDMA
    - Advanced support for zero-copy IB-MCAST designs with GPUDirect RDMA
    - Advanced reliability support for IB-MCAST designs
  - Efficient broadcast and all-reduce designs for Deep Learning applications
  - Enhanced collective tuning on Xeon, OpenPOWER, and NVIDIA DGX-1 systems
Optimized MVAPICH2-GDR Design

GPU-GPU Inter-node Latency

- MV2-(NO-GDR)
- MV2-GDR 2.3

1.85us
10x

Message Size (Bytes)

Latency (us)

Bandwidth (MB/s)

GPU-GPU Inter-node Bandwidth

- MV2-(NO-GDR)
- MV2-GDR-2.3

9x
11X

Message Size (Bytes)

Bandwidth (MB/s)

MVAPICH2-GDR-2.3
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

Network Based Computing Laboratory
Application-Level Evaluation (HOOMD-blue)

- Platform: Wilkes (Intel Ivy Bridge + NVIDIA Tesla K20c + Mellanox Connect-IB)
- HoomdBlue Version 1.0.5
  - GDRCOPY enabled: MV2_USE_CUDA=1 MV2_IBA_HCA=mlx5_0 MV2_IBA_EAGER_THRESHOLD=32768 MV2_VBUF_TOTAL_SIZE=32768 MV2_USE_GPUDIRECT_LOOPBACK_LIMIT=32768 MV2_USE_GPUDIRECT_GDRCOPY=1 MV2_USE_GPUDIRECT_GDRCOPY_LIMIT=16384
Application-Level Evaluation (Cosmo) and Weather Forecasting in Switzerland

Wilkes GPU Cluster

CSCS GPU cluster

- 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


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

• Overview of the MVAPICH2 Project
• MVAPICH2-GPU with GPUDirect-RDMA (GDR)

• What’s new with MVAPICH2-GDR
  • Multi-stream Communication for IPC
  • CMA-based Intra-node Communication Support
  • Support for OpenPower and NVLink
  • Maximal overlap in MPI Datatype Processing
  • Streaming Support with IB Multicast and GDR

• High-Performance Deep Learning (HiDL) with MVAPICH2-GDR
• Conclusions
Multi-stream Communication using CUDA IPC on OpenPOWER and DGX-1

- Up to **16% higher** Device to Device (D2D) bandwidth on OpenPOWER + NVLink inter-connect
- Up to **30% higher** D2D bandwidth on DGX-1 with NVLink

**Pt-to-pt (D-D) Bandwidth:**

Benefits of Multi-stream CUDA IPC Design

Available since MVAPICH2-GDR-2.3a
CMA-based Intra-node Communication Support

- Up to **30% lower** Host-to-Host (H2H) latency and **30% higher** H2H Bandwidth

### INTRA-NODE Pt-to-Pt (H2H) LATENCY

- MV2-GDR (w/out CMA)
- MV2-GDR (w/ CMA)

### INTRA-NODE Pt-to-Pt (H2H) BANDWIDTH

- MV2-GDR (w/out CMA)
- MV2-GDR (w/ CMA)

**MVAPICH2-GDR-2.3**

- Intel Broadwell (E5-2680 v4 @ 3240 GHz) node – 28 cores
- NVIDIA Tesla K-80 GPU, and Mellanox Connect-X4 EDR HCA
- CUDA 8.0, Mellanox OFED 4.0 with GPU-Direct-RDMA
Scalable Host-based Collectives on OpenPOWER (Intra-node Reduce & AlltoAll)

**Reduce**

Latency (us) vs. Message Size (Nodes=1, PPN=20)

- MVAPICH2-GDR
- SpectrumMPI-10.1.0.2
- OpenMPI-3.0.0

*Up to 5X and 3x* performance improvement by MVAPICH2 for small and large messages respectively

**AlltoAll**

Latency (us) vs. Message Size (Nodes=1, PPN=20)

- MVAPICH2-GDR
- SpectrumMPI-10.1.0.2
- OpenMPI-3.0.0

1.2X, 1.3X, 3.2X, 3.3X performance improvement
Intra-node Point-to-Point Performance on OpenPower

Platform: Two nodes of OpenPOWER (Power8-ppc64le) CPU using Mellanox EDR (MT4115) HCA
**MVAPICH2-GDR: Performance on OpenPOWER (NVLink + Pascal)**

**INTRA-NODE LATENCY (SMALL)**

Latency (us) vs Message Size (Bytes)

---

**INTRA-NODE LATENCY (LARGE)**

Latency (us) vs Message Size (Bytes)

---

**INTRA-NODE BANDWIDTH**

Bandwidth (GB/sec) vs Message Size (Bytes)

---

**Inter-node Latency: 23 us (without GPUDirectRDMA)**

**Intra-node Latency: 13.8 us (without GPUDirectRDMA)**

**Inter-node Bandwidth: 6 GB/sec (FDR)**

**Intra-node Bandwidth: 33.2 GB/sec (NVLINK)**

Platform: OpenPOWER (ppc64le) nodes equipped with a dual-socket CPU, 4 Pascal P100-SXM GPUs, and 4X-FDR InfiniBand Inter-connect

Available since MVAPICH2-GDR 2.3a
Optimized All-Reduce with XPMEM on OpenPOWER

- Optimized MPI All-Reduce Design in MVAPICH2
  - *Up to 2X* performance improvement over Spectrum MPI and *4X* over OpenMPI for intra-node

*Optimized Runtime Parameters:* MV2_CPU_BINDING_POLICY=hybrid MV2_HYBRID_BINDING_POLICY=bunch
Non-contiguous Data Exchange

- Multi-dimensional data
  - Row based organization
  - Contiguous on one dimension
  - Non-contiguous on other dimensions

- Halo data exchange
  - Duplicate the boundary
  - Exchange the boundary in each iteration

Halo data exchange
MPI Datatype support in MVAPICH2

- Datatypes support in MPI
  - Operate on customized datatypes to improve productivity
  - Enable MPI library to optimize non-contiguous data

**At Sender:**

```c
MPI_Type_vector (n_blocks, n_elements, stride, old_type, &new_type);
MPI_Type_commit(&new_type);
...
MPI_Send(s_buf, size, new_type, dest, tag, MPI_COMM_WORLD);
```

- Inside MVAPICH2
  - Use datatype specific CUDA Kernels to pack data in chunks
  - Efficiently move data between nodes using RDMA
  - In progress - currently optimizes vector and hindexed datatypes
  - Transparent to the user

MPI Datatype Processing (Computation Optimization)

- Comprehensive support
  - Targeted kernels for regular datatypes - vector, subarray, indexed_block
  - Generic kernels for all other irregular datatypes
- Separate non-blocking stream for kernels launched by MPI library
  - Avoids stream conflicts with application kernels
- Flexible set of parameters for users to tune kernels
  - Vector
    - MV2_CUDA_KERNEL VECTOR_TIDBLK_SIZE
    - MV2_CUDA_KERNEL VECTOR_YSIZE
  - Subarray
    - MV2_CUDA_KERNEL SUBARR TIDBLK_SIZE
    - MV2_CUDA_KERNEL_SUBARR_XDIM
    - MV2_CUDA_KERNEL_SUBARR_YDIM
    - MV2_CUDA_KERNEL_SUBARR_ZDIM
  - Indexed_block
    - MV2_CUDA_KERNEL_IDXBLK_XDIM
Common Scenario

MPI_Isend (A,.. Datatype,...)
MPI_Isend (B,.. Datatype,...)
MPI_Isend (C,.. Datatype,...)
MPI_Isend (D,.. Datatype,...)
...

MPI_Waitall (...);

*A, B...contain non-contiguous MPI Datatype
Streaming Applications

- Streaming applications on HPC systems

1. Communication (MPI)
   - Broadcast-type operations

2. Computation (CUDA)
   - Multiple GPU nodes as workers

Data Source

Real-time streaming

HPC resources for real-time analytics

Sender

Data streaming-like broadcast operations

Worker
- CPU
- GPU

Worker
- CPU
- GPU

Worker
- CPU
- GPU

Worker
- CPU
- GPU
Hardware Multicast-based Broadcast

- For GPU-resident data, using
  - GPUDirect RDMA (GDR)
  - InfiniBand Hardware Multicast (IB-MCAST)
- Overhead
  - IB UD limit
  - GDR limit

Broadcast on Multi-GPU systems

- Proposed Intra-node Topology-Aware Broadcast
  - CUDA InterProcess Communication (IPC)


Available since MVAPICH2-GDR 2.3a
Streaming Benchmark @ CSCS (88 GPUs)

- IB-MCAST + GDR + Topology-aware IPC-based schemes
  - Up to 58% and 79% reduction for small and large messages

Outline

• Overview of the MVAPICH2 Project
• MVAPICH2-GPU with GPUDirect-RDMA (GDR)
• What’s new with MVAPICH2-GDR

• High-Performance Deep Learning (HiDL) with MVAPICH2-GDR
  • Benefits of CUDA-Aware MPI with TensorFlow
  • Benefits with CNTK
  • Optimized Collectives for Deep Learning
  • Co-designed OSU-Caffe
  • Out-of-core DNN Training

• Conclusions
Deep Learning: New Challenges for MPI 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
  - NCCL2, 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?

Exploiting CUDA-Aware MPI for TensorFlow (Horovod)

- MVAPICH2-GDR offers excellent performance via advanced designs for MPI_Allreduce.
- Up to **22% better** performance on Wilkes2 cluster (16 GPUs)
Performance Benefits with CNTK Deep Learning Framework

- @ RI2 cluster, 16 GPUs, 1 GPU/node
  - Microsoft Cognitive Toolkit (CNTK) [https://github.com/Microsoft/CNTK]

![Performance Comparison Graphs]

- Reduces up to 24% and 15% of latency for AlexNet and VGG models
- Higher improvement can be observed for larger system sizes

**MVAPICH2-GDR: Allreduce Comparison with Baidu and OpenMPI**

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

*Available since MVAPICH2-GDR 2.3a*
MVAPICH2-GDR vs. NCCL2 – Reduce Operation

- Optimized designs offer better/comparable performance for most cases
- MPI_Reduce (MVAPICH2-GDR) vs. ncclReduce (NCCL2) on 16 GPUs

Platform: Intel Xeon (Broadwell) nodes equipped with a dual-socket CPU, 1 K-80 GPUs, and EDR InfiniBand Inter-connect
Optimized designs in MVAPICH2-GDR 2.3 offer better/comparable performance for most cases.

MPI_Allreduce (MVAPICH2-GDR) vs. ncclAllreduce (NCCL2) on 16 GPUs

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

~1.2X better

~3X better
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/

GoogLeNet (ImageNet) on 128 GPUs

<table>
<thead>
<tr>
<th>No. of GPUs</th>
<th>Training Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>250</td>
</tr>
<tr>
<td>16</td>
<td>200</td>
</tr>
<tr>
<td>32</td>
<td>200</td>
</tr>
<tr>
<td>64</td>
<td>200</td>
</tr>
<tr>
<td>128</td>
<td>200</td>
</tr>
</tbody>
</table>

- Invalid use case

Legend:
- Caffe
- OSU-Caffe (1024)
- OSU-Caffe (2048)
Out-of-Core Deep Neural Network Training with Caffe

• Large DNNs cannot be trained on GPUs due to memory limitation!
  – ResNet-50 is the state-of-the-art DNN architecture for Image Recognition but current frameworks can only go up to a small batch size of 45
  – Next generation models like Neural Machine Translation (NMT) are ridiculously large, consists of billions of parameters, and require even more memory
  – Can we design Out-of-core DNN training support using new software features in CUDA 8/9 and hardware mechanisms in Pascal/Volta GPUs?

• General intuition is that managed allocations “will be” slow!
  – The proposed framework called OC-Caffe (Out-of-Core Caffe) shows the potential of managed memory designs that can provide performance with negligible/no overhead.
  – In addition to Out-of-core Training support, productivity can be greatly enhanced in terms of DL framework design by using the new Unified Memory features.

A. Awan, C. Chu, H. Subramoni, X. Lu, and D. K. Panda, OC-DNN: Exploiting Advanced Unified Memory Capabilities in CUDA 9 and Volta GPUs for Out-of-Core DNN Training, HiPC ’18
Performance Trends for OC-Caffe

- OC-Caffe-Opt: up to **80% better** than Intel-optimized CPU Caffe for ResNet-50 training on the Volta V100 GPU with CUDA9 and CUDNN7
- OC-Caffe allows efficient scale-up on DGX-1 system with Pascal P100 GPUs with CUDA9 and CUDNN7

**Out-of-core (over-subscription)**

<table>
<thead>
<tr>
<th>Images/sec (Higher is better)</th>
<th>caffe-gpu</th>
<th>oc-caffe-naïve</th>
<th>oc-caffe-opt</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

oc-caffe-opt is 80% better than intel-caffe

*intel-caffe-opt (N/A)*

**Scale-up on DGX-1**

Images Per Second (Higher is better)

- Caffe
- OC-Caffe

<table>
<thead>
<tr>
<th>Number of GPUs (DGX-1)</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffe</td>
<td>1000</td>
<td>1000</td>
<td>10000</td>
<td>10000</td>
</tr>
<tr>
<td>OC-Caffe</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
</tr>
</tbody>
</table>

A. Awan, C. Chu, H. Subramoni, X. Lu, and D. K. Panda, OC-DNN: Exploiting Advanced Unified Memory Capabilities in CUDA 9 and Volta GPUs for Out-of-Core DNN Training, HiPC ’18
Outline

• Overview of the MVAPICH2 Project
• MVAPICH2-GPU with GPUDirect-RDMA (GDR)
• What’s new with MVAPICH2-GDR
• High-Performance Deep Learning (HiDL) with MVAPICH2-GDR
• Conclusions
Conclusions

• MVAPICH2-GDR Library provides optimized MPI communication on InfiniBand and RoCE clusters with GPUs
• Supports both X86 and OpenPower with NVLink
• Takes advantage of CUDA features like IPC and GPUDirect RDMA families
• Allows flexible solutions for streaming applications with GPUs
• Provides optimized solutions (scale-up and scale-out) for High-Performance Deep Learning
Join us at the OSU Booth

• Join the OSU Team at SC’18 at Booth #4404 and grab a free T-Shirt!
• More details about various events available at the following link
• http://mvapich.cse.ohio-state.edu/conference/735/talks/
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 Deep Learning Project
http://hidl.cse.ohio-state.edu/