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

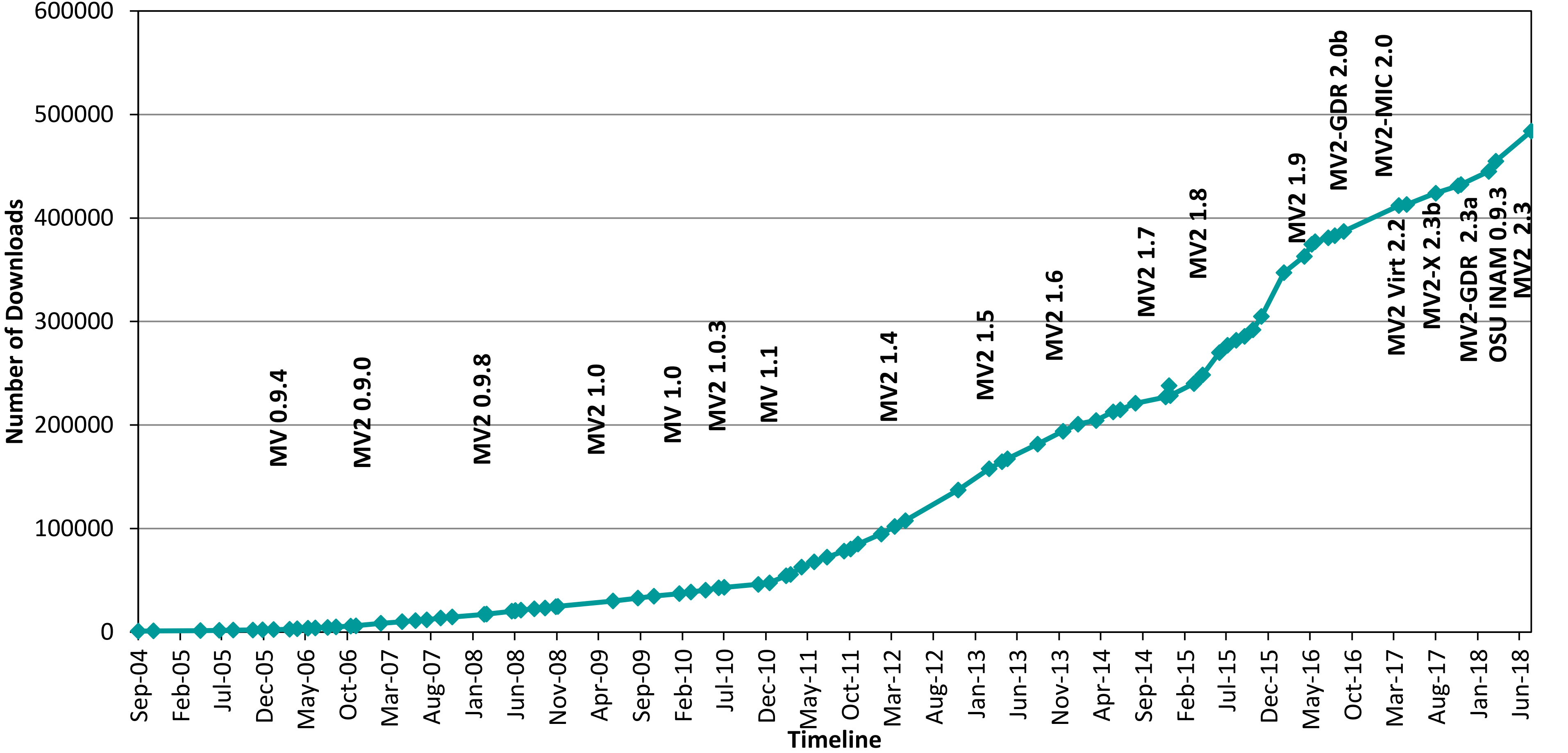
- Overview of the MVAPICH2 Project
- MVAPICH2-GPU with GPUDirect-RDMA (GDR)
- What's new with MVAPICH2-GDR
- High-Performance Deep Learning (HiDL) with MVAPICH2-GDR
- Conclusions

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**
- Empowering Top500 systems for over a decade

Partner in the upcoming TACC Frontera System

MVAPICH2 Release Timeline and Downloads



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, ARM, NVIDIA GPGPU)

Transport Mechanisms

- Shared Memory
- CMA
- IVSHMEM
- XPMEM

Modern Features

- MCDRAM*
- NVLink*
- CAPI*

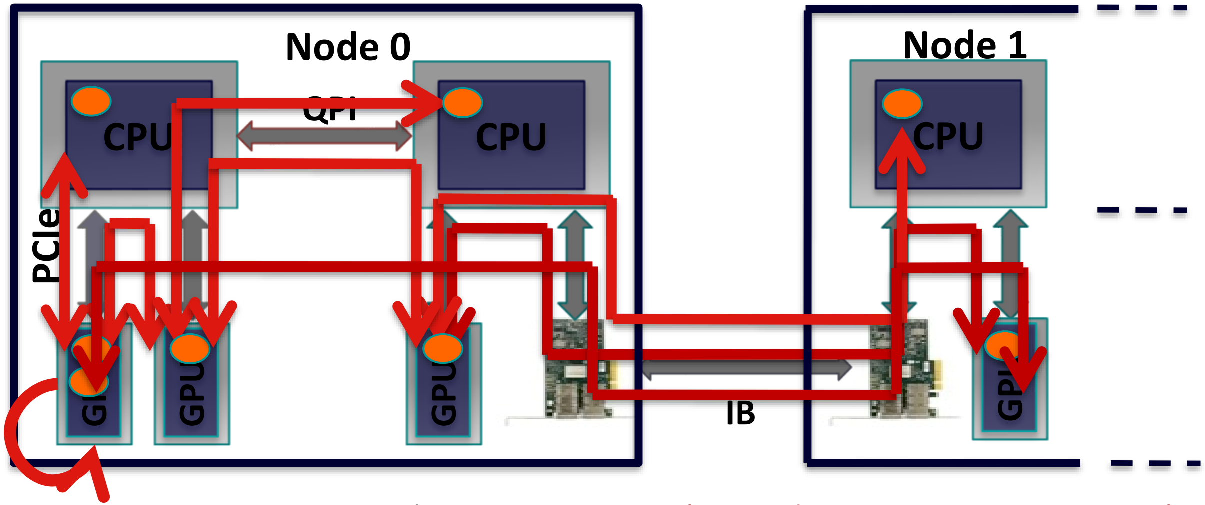
* Upcoming

MVAPICH2 Software Family

High-Performance Parallel Programming Libraries	
MVAPICH2	Support for InfiniBand, Omni-Path, Ethernet/iWARP, and RoCE
MVAPICH2-X	Advanced MPI features, OSU INAM, PGAS (OpenSHMEM, UPC, UPC++, and CAF), and MPI+PGAS programming models with unified communication runtime
MVAPICH2-GDR	Optimized MPI for clusters with NVIDIA GPUs
MVAPICH2-Virt	High-performance and scalable MPI for hypervisor and container based HPC cloud
MVAPICH2-EA	Energy aware and High-performance MPI
MVAPICH2-MIC	Optimized MPI for clusters with Intel KNC
Microbenchmarks	
OMB	Microbenchmarks suite to evaluate MPI and PGAS (OpenSHMEM, UPC, and UPC++) libraries for CPUs and GPUs
Tools	
OSU INAM	Network monitoring, profiling, and analysis for clusters with MPI and scheduler integration
OEMT	Utility to measure the energy consumption of MPI applications

MVAPICH2-GDR: Optimizing MPI Data Movement on GPU Clusters

- Connected as PCIe devices – Flexibility but Complexity



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

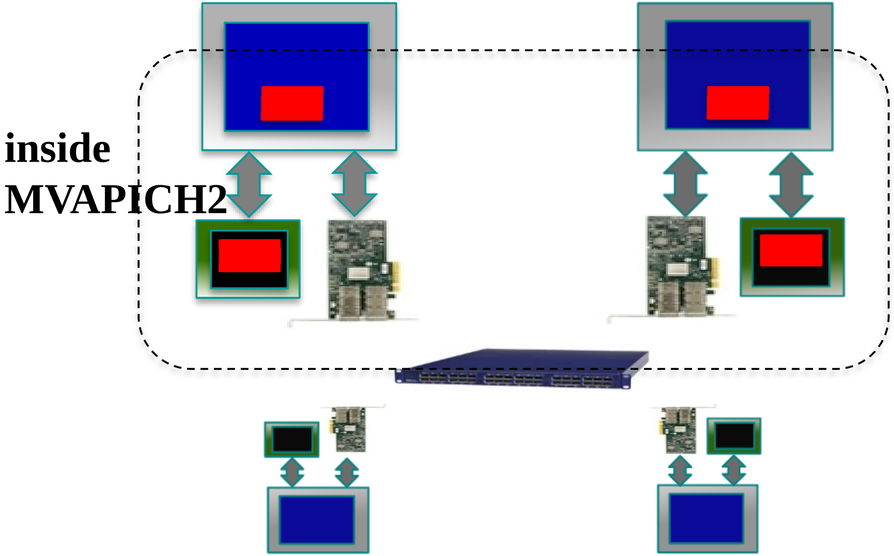
- Standard MPI interfaces used for unified data movement
- Takes advantage of Unified Virtual Addressing (>= CUDA 4.0)
- Overlaps data movement from GPU with RDMA transfers

At Sender:

```
MPI_Send(s_devbuf, size, …);
```

At Receiver:

```
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

    http://www.mellanox.com/page/products_dyn?product_family=116

  - Strongly recommended
  - GDRCOPY module from NVIDIA

    https://github.com/NVIDIA/gdrcopy

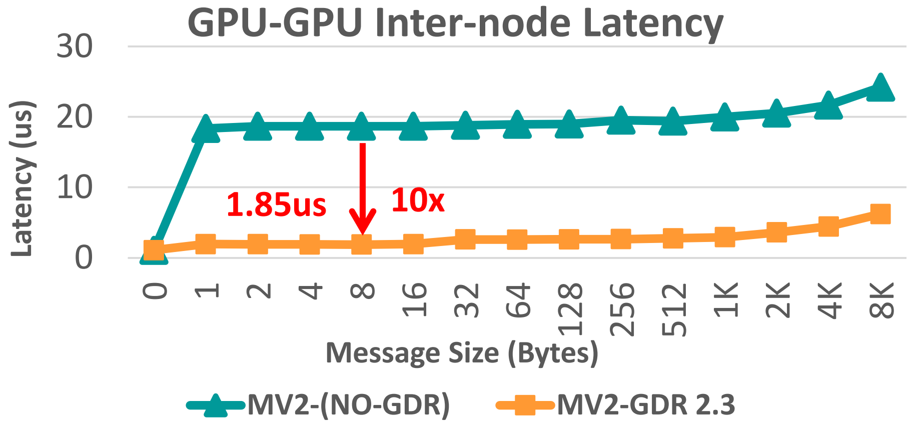
- Contact MVAPICH help list with any questions related to the package

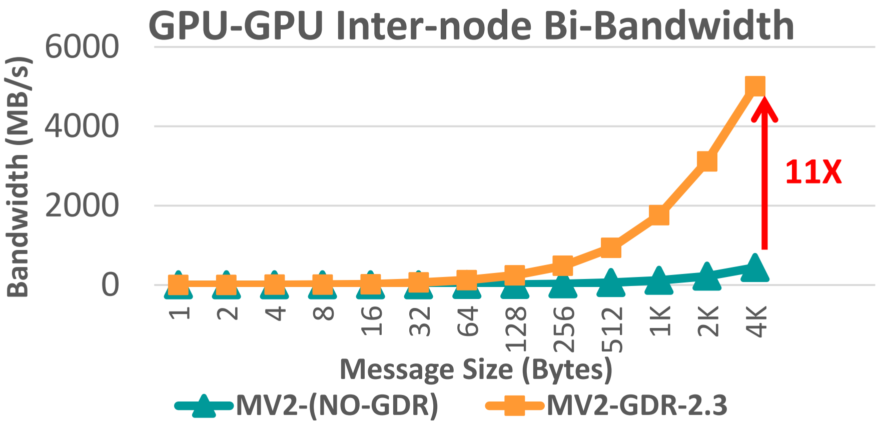
  mvapich-help@cse.ohio-state.edu

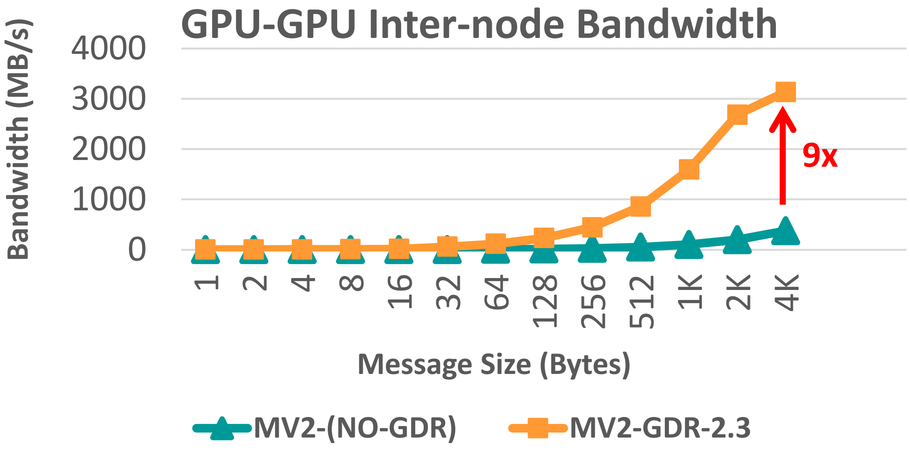
MVAPICH2-GDR 2.3 GA

- 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

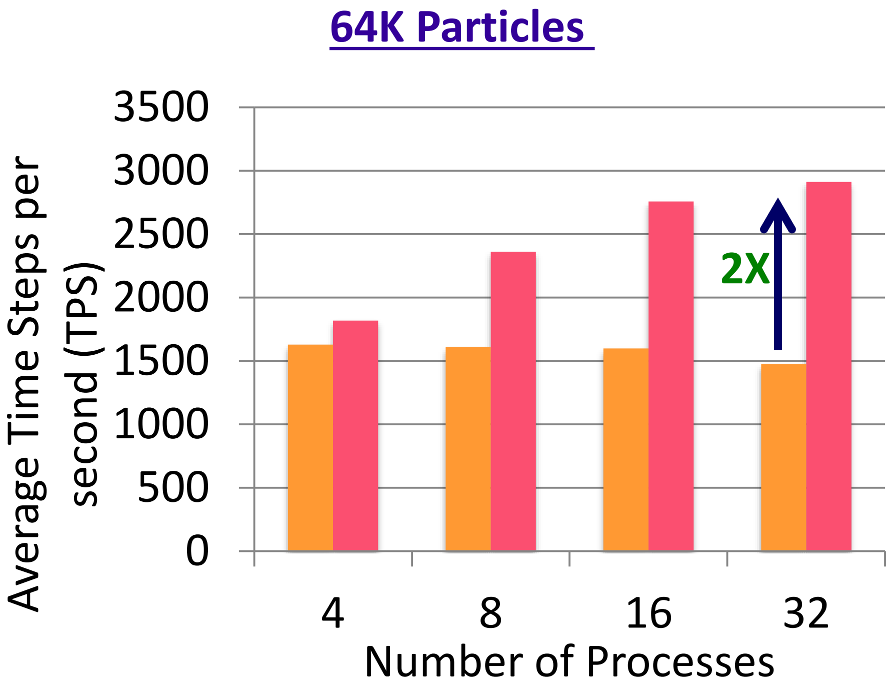


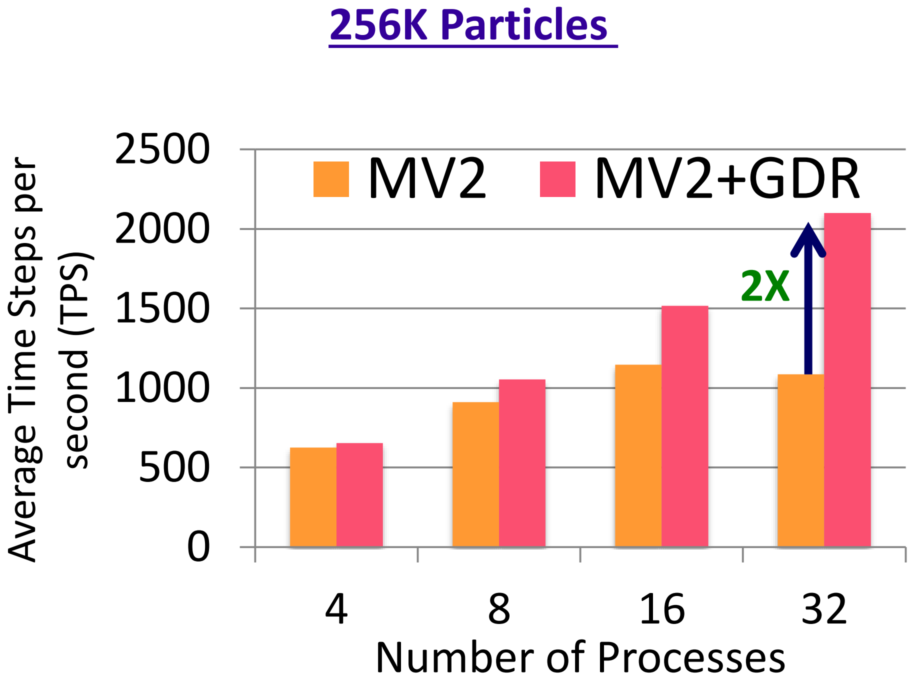




MVAPICH2-GDR-2.3
Intel Haswell (E5-2687W @ 3.10 GHz) node - 20 cores
NVIDIA Volta V100 GPU
Mellanox Connect-X4 EDR HCA
CUDA 9.0
Mellanox OFED 4.0 with GPU-Direct-RDMA

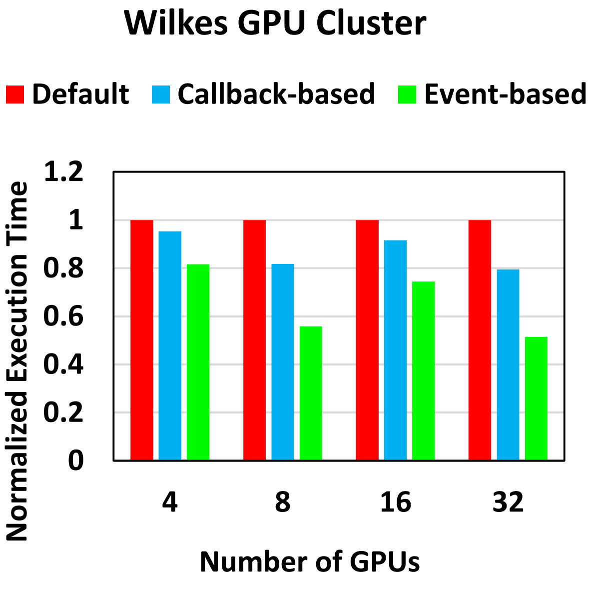
Application-Level Evaluation (HOOMD-blue)

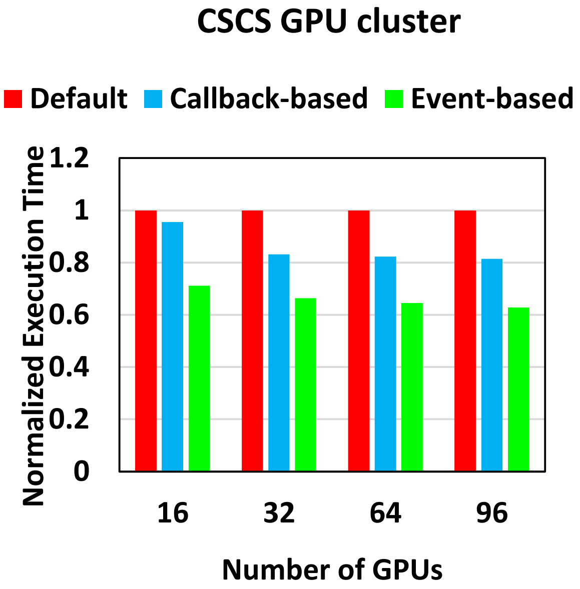


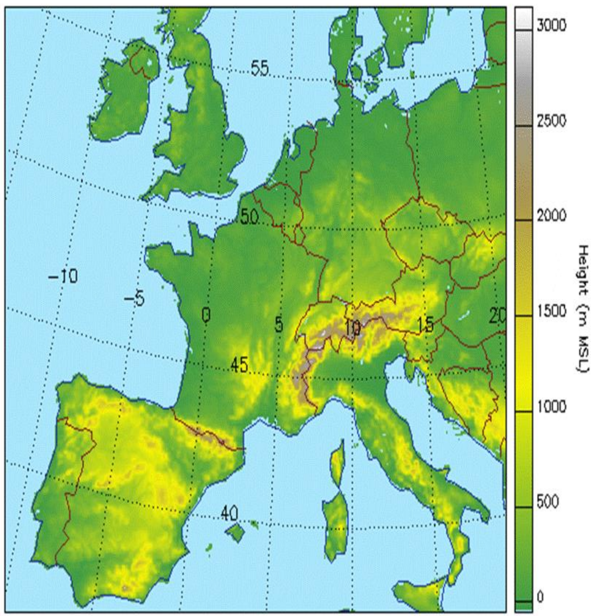


- Platform: Wilkes (Intel Ivy Bridge + NVIDIA Tesla K20c + Mellanox Connect-IB)
- HoomdBlue Version 1.0.5
    - GDRCOPY enabled: MV2_USE_CUDA=1 MV2_IBA_HCA=mlx5_0 MV2_IBA_EAGER_THRESHOLD=32768 MV2_VBUF_TOTAL_SIZE=32768 MV2_USE_GPUDIRECT_LOOPBACK_LIMIT=32768 MV2_USE_GPUDIRECT_GDRCOPY=1 MV2_USE_GPUDIRECT_GDRCOPY_LIMIT=16384

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







- 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/

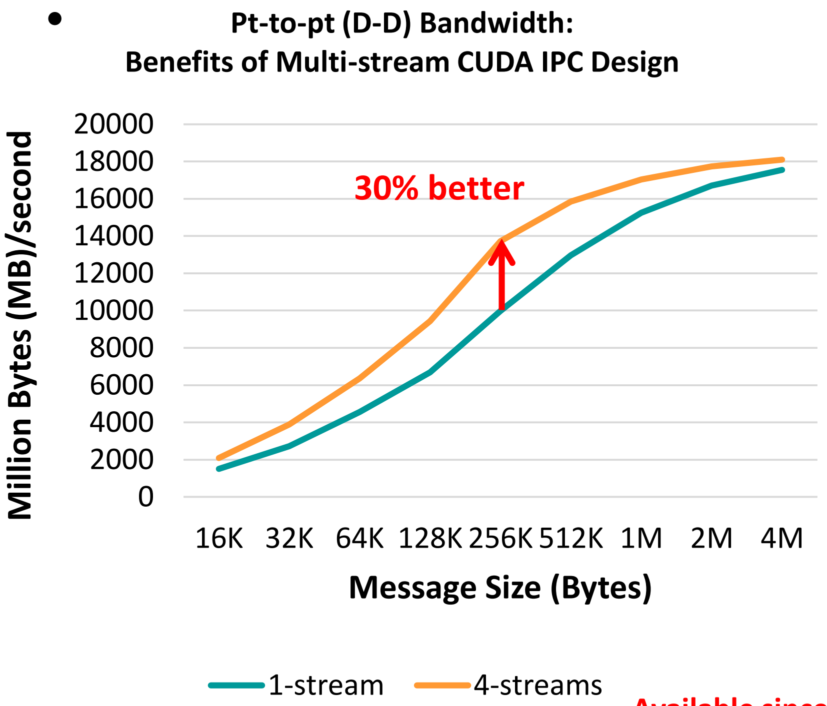
On-going collaboration with CSCS and MeteoSwiss (Switzerland) in co-designing MV2-GDR and Cosmo Application

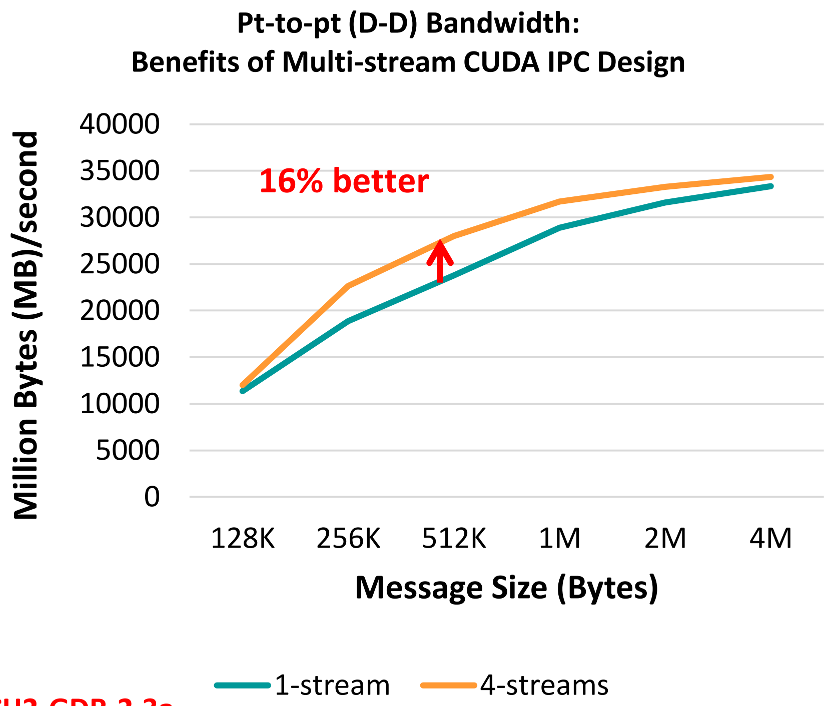
C. Chu, K. Hamidouche, A. Venkatesh, D. Banerjee , H. Subramoni, and D. K. Panda, Exploiting Maximal Overlap for Non-Contiguous Data Movement Processing on Modern GPU-enabled Systems, IPDPS'16

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

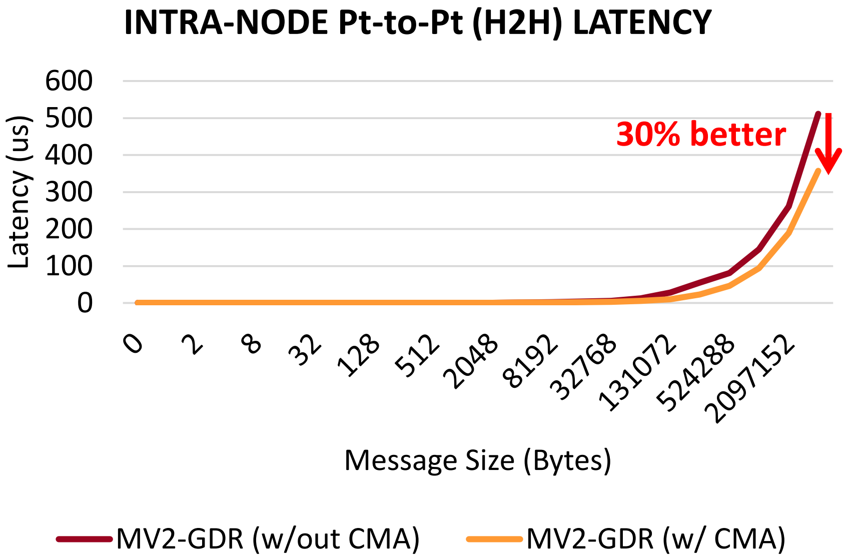
- Up to **16% higher** Device to Device (D2D) bandwidth on OpenPOWER + NVLink inter-connect
- Up to **30% higher** D2D bandwidth on DGX-1 with NVLink
- 

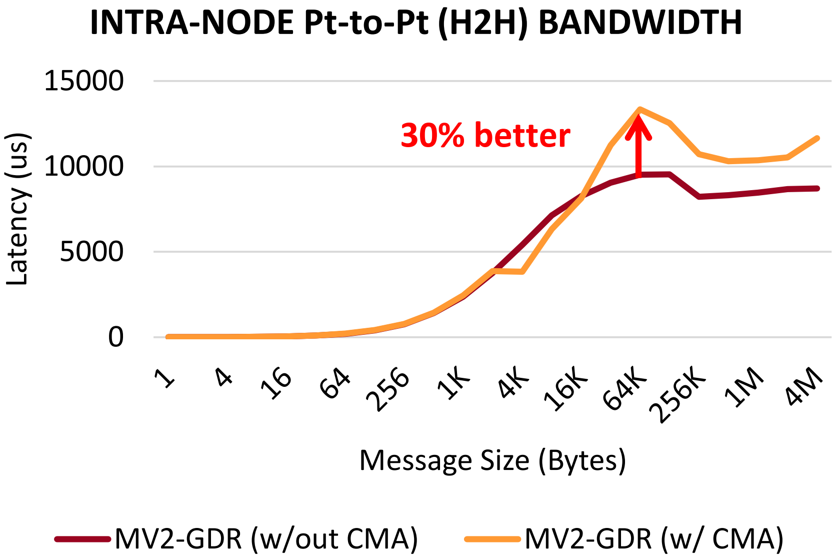


Available since MVAPICH2-GDR-2.3a

CMA-based Intra-node Communication Support

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

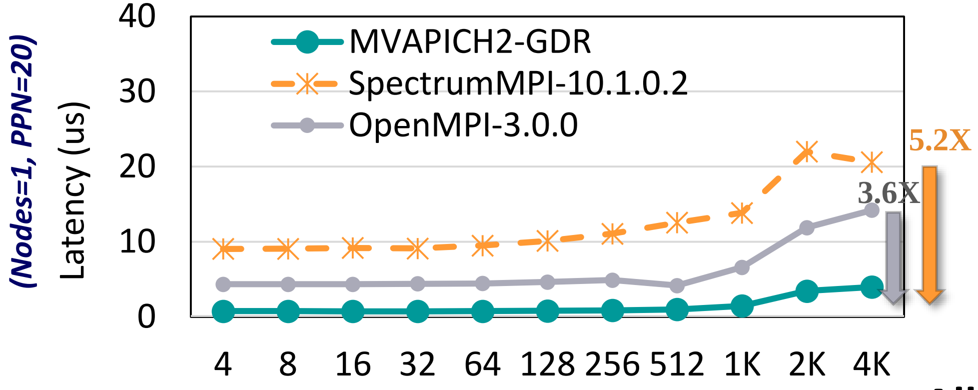


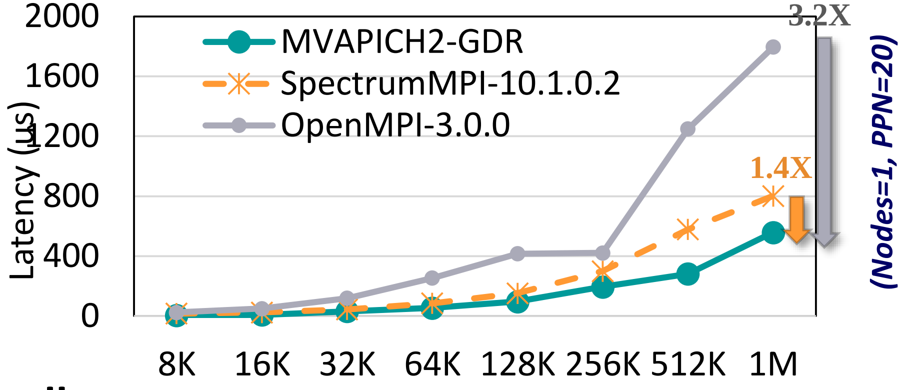


MVAPICH2-GDR-2.3
Intel Broadwell (E5-2680 v4 @ 3240 GHz) node – 28 cores
NVIDIA Tesla K-80 GPU, and Mellanox Connect-X4 EDR HCA
CUDA 8.0, Mellanox OFED 4.0 with GPU-Direct-RDMA

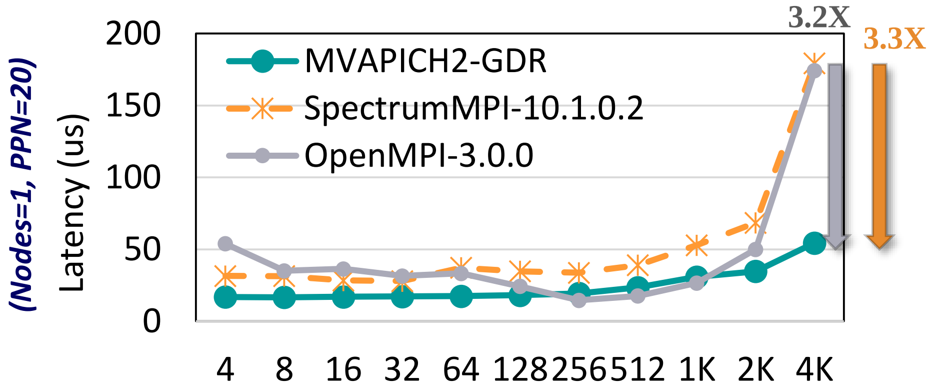
Scalable Host-based Collectives on OpenPOWER (Intra-node Reduce & AlltoAll)

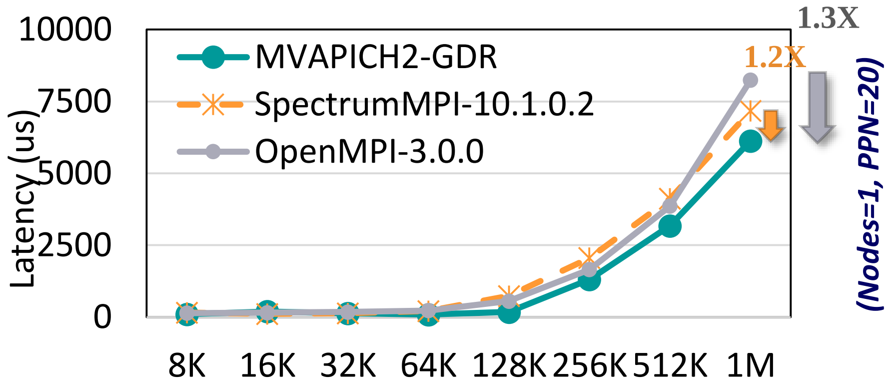
Reduce





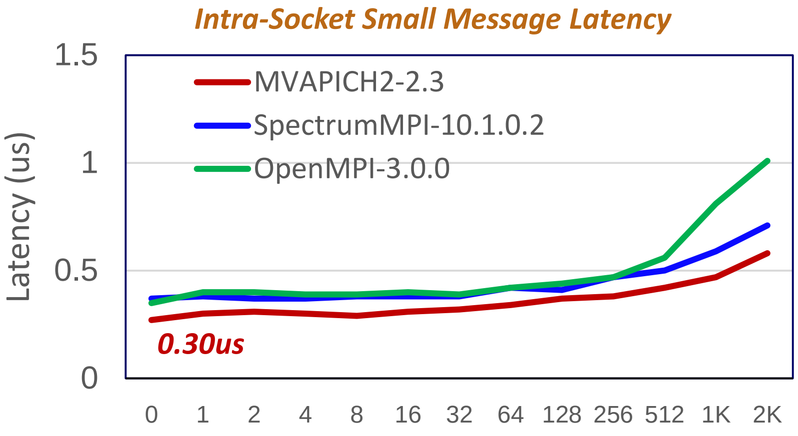
Alltoall

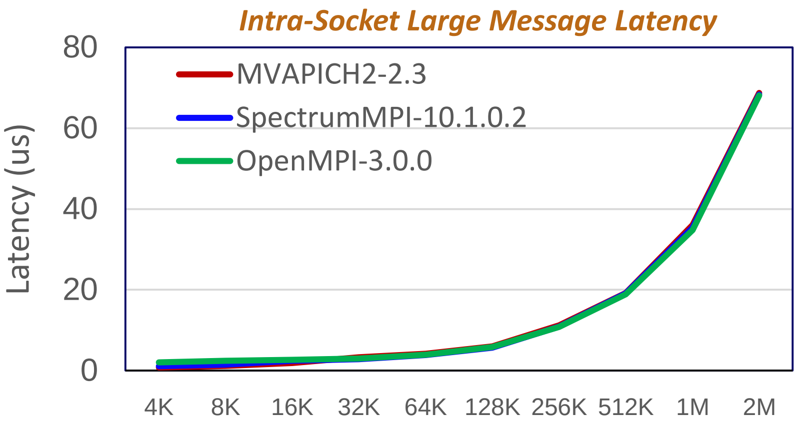


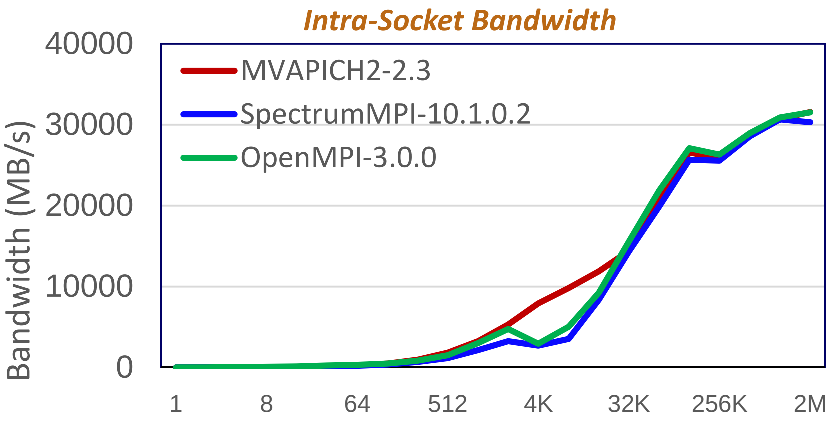


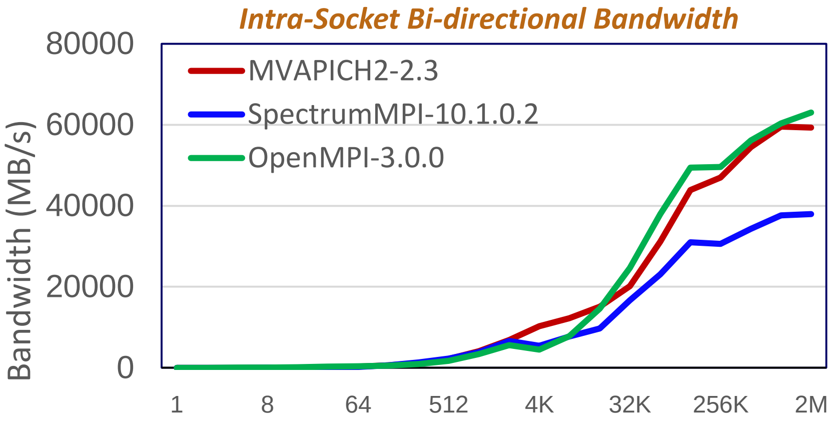
Up to 5X and 3x performance improvement by MVAPICH2 for small and large messages respectively

Intra-node Point-to-Point Performance on OpenPower



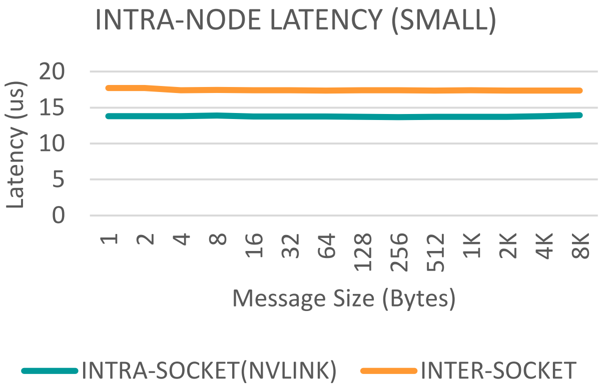


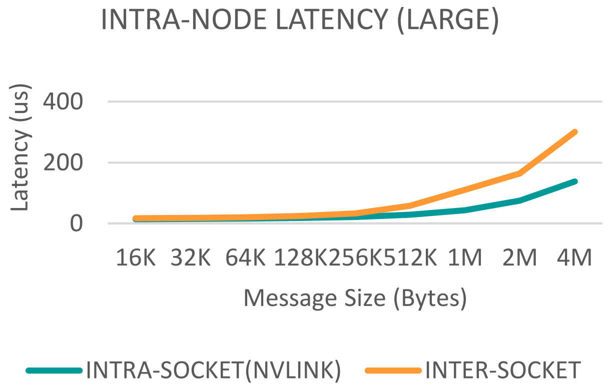


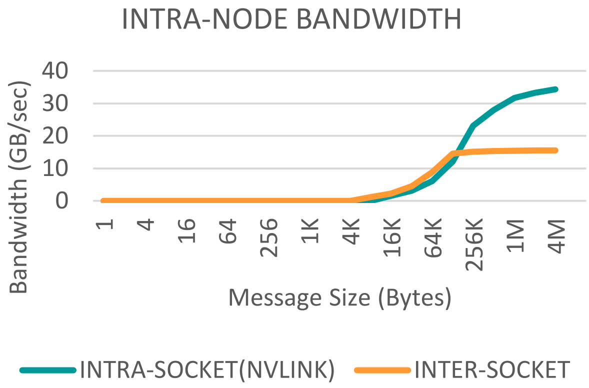


Platform: Two nodes of OpenPOWER (Power8-ppc64le) CPU using Mellanox EDR (MT4115) HCA

MVAPICH2-GDR: Performance on OpenPOWER (NVLink + Pascal)

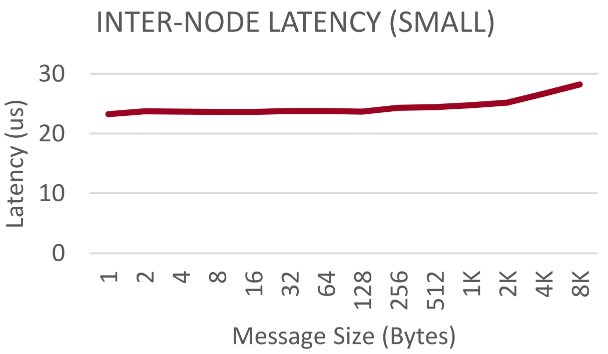


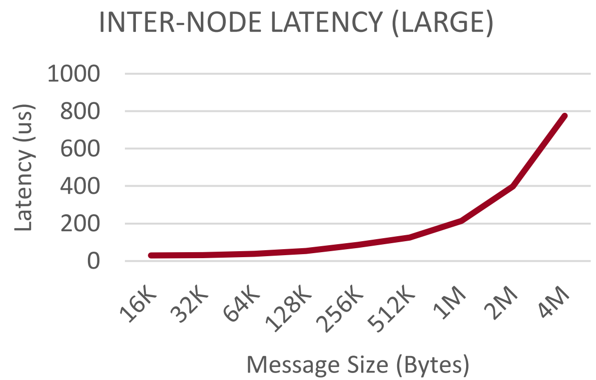


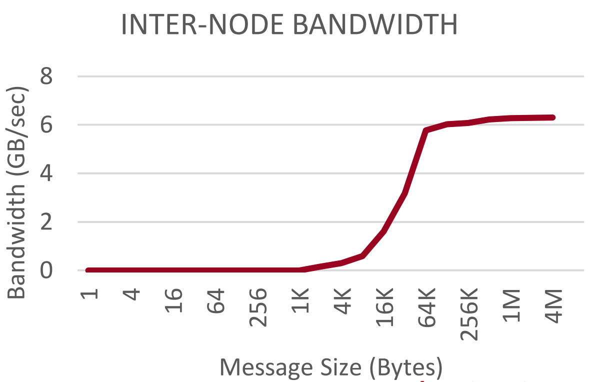


Intra-node Latency: 13.8 us (without GPUDirectRDMA)

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







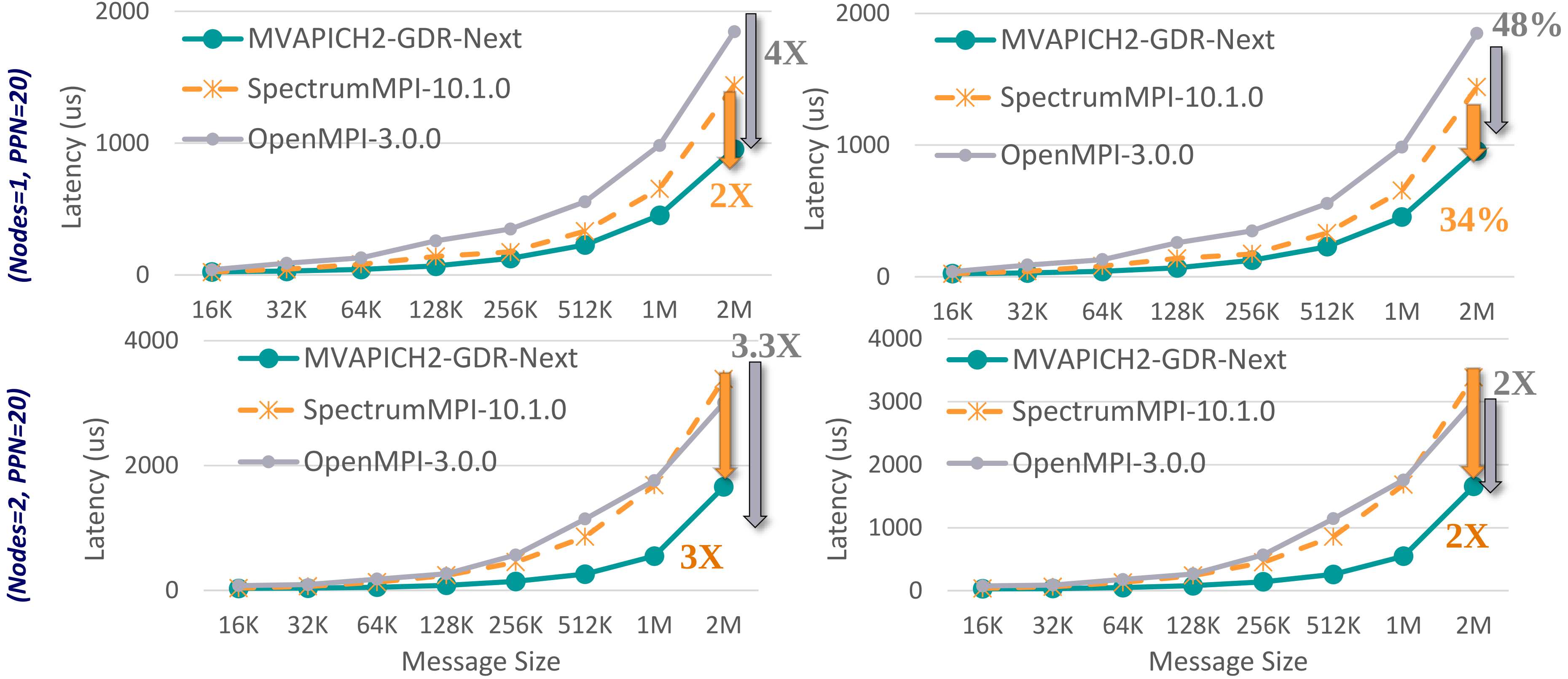
Inter-node Latency: 23 us (without GPUDirectRDMA)

Available since MVAPICH2-GDR 2.3a

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

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

Optimized All-Reduce with XPMEM on OpenPOWER

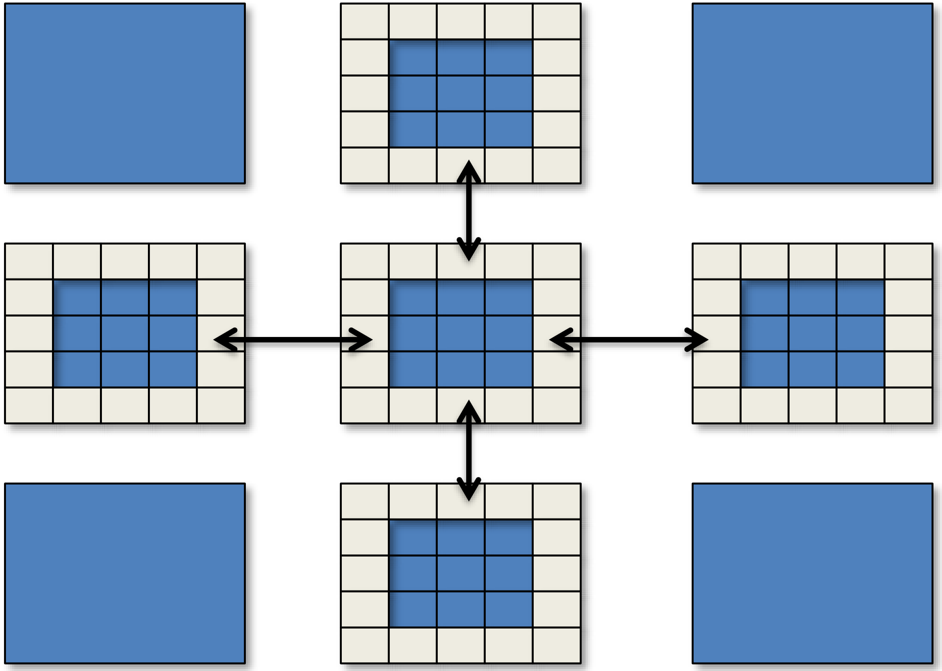


- **Optimized MPI All-Reduce Design in MVAPICH2**
  - *Up to 2X* performance improvement over Spectrum MPI and 4X over OpenMPI for intra-node

Optimized Runtime Parameters: MV2_CPU_BINDING_POLICY=hybrid MV2_HYBRID_BINDING_POLICY=bunch

Non-contiguous Data Exchange

Halo 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

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:

```
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

H. Wang, S. Potluri, D. Bureddy, C. Rosales and D. K. Panda, GPU-aware MPI on RDMA-Enabled Clusters: Design, Implementation and Evaluation, IEEE Transactions on Parallel and Distributed Systems, Vol. 25, No. 10, pp. 2595-2605 , Oct 2014.

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

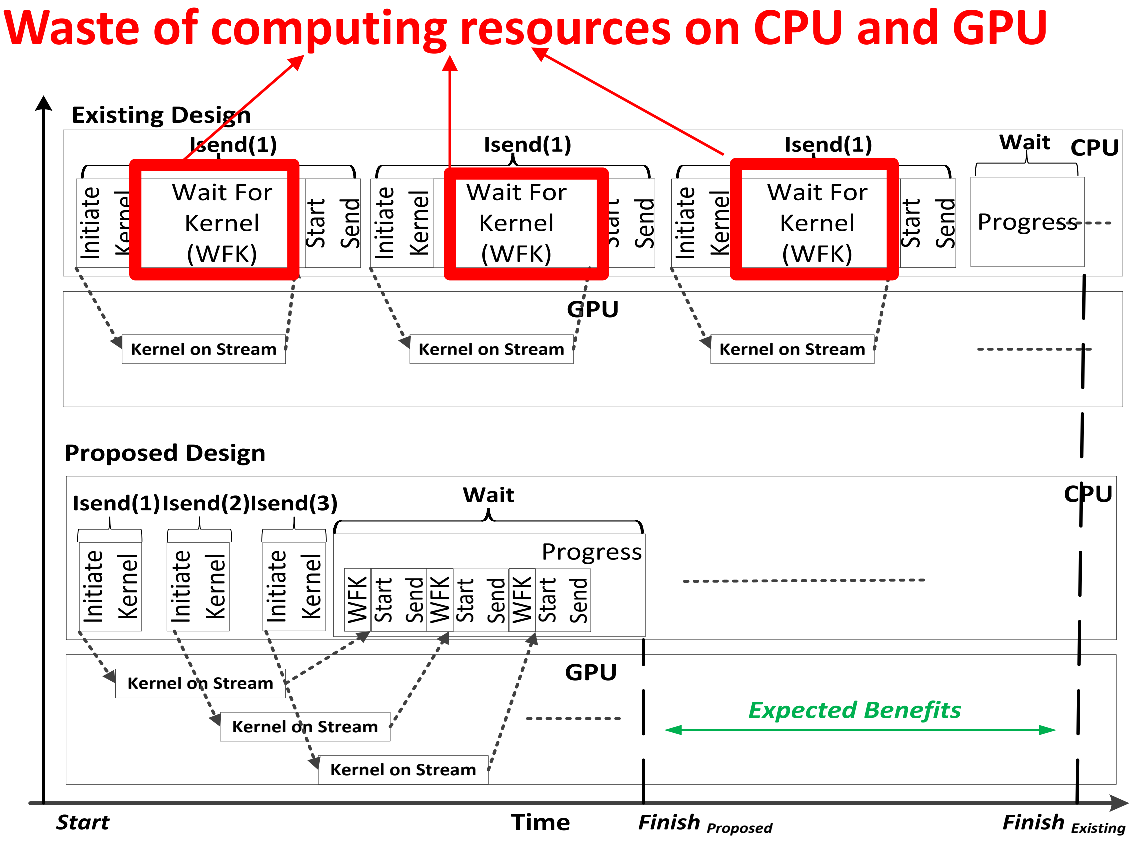
MPI Datatype Processing (Communication Optimization)

Common Scenario

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

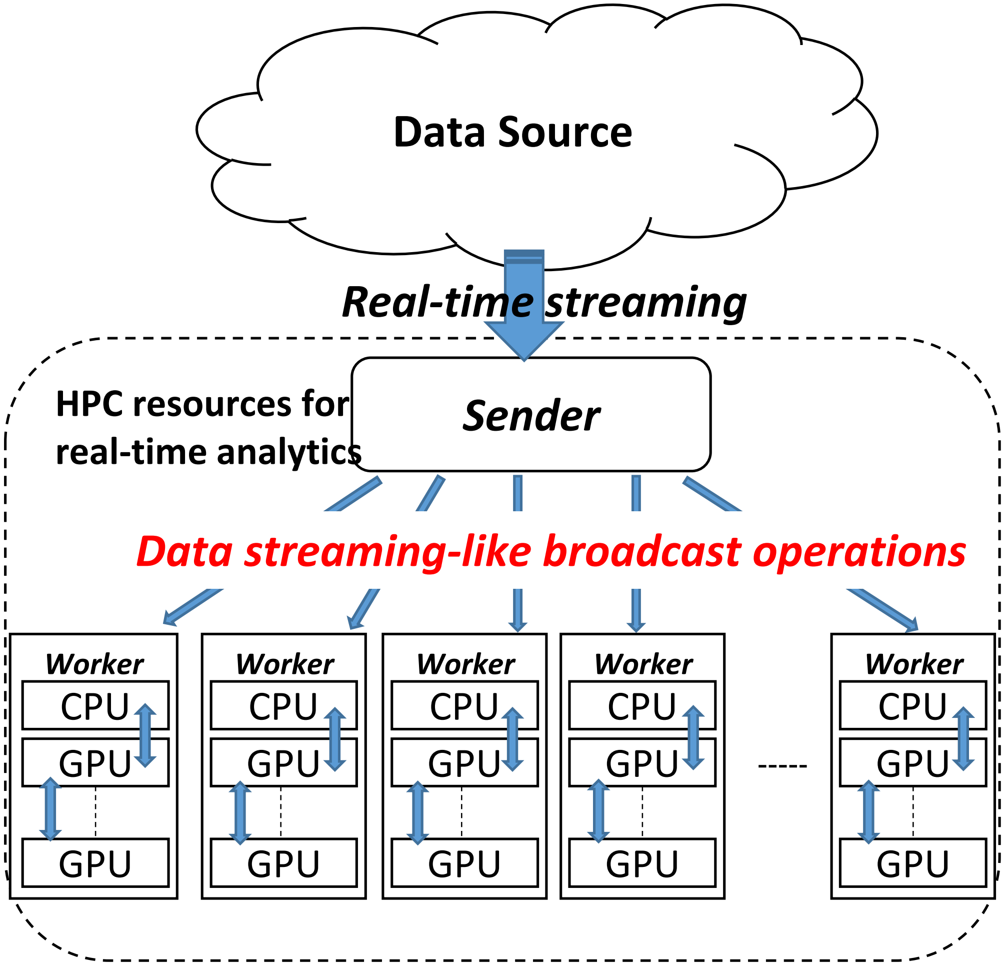
MPI_Waitall (…);

*A, B…contain non-contiguous MPI Datatype



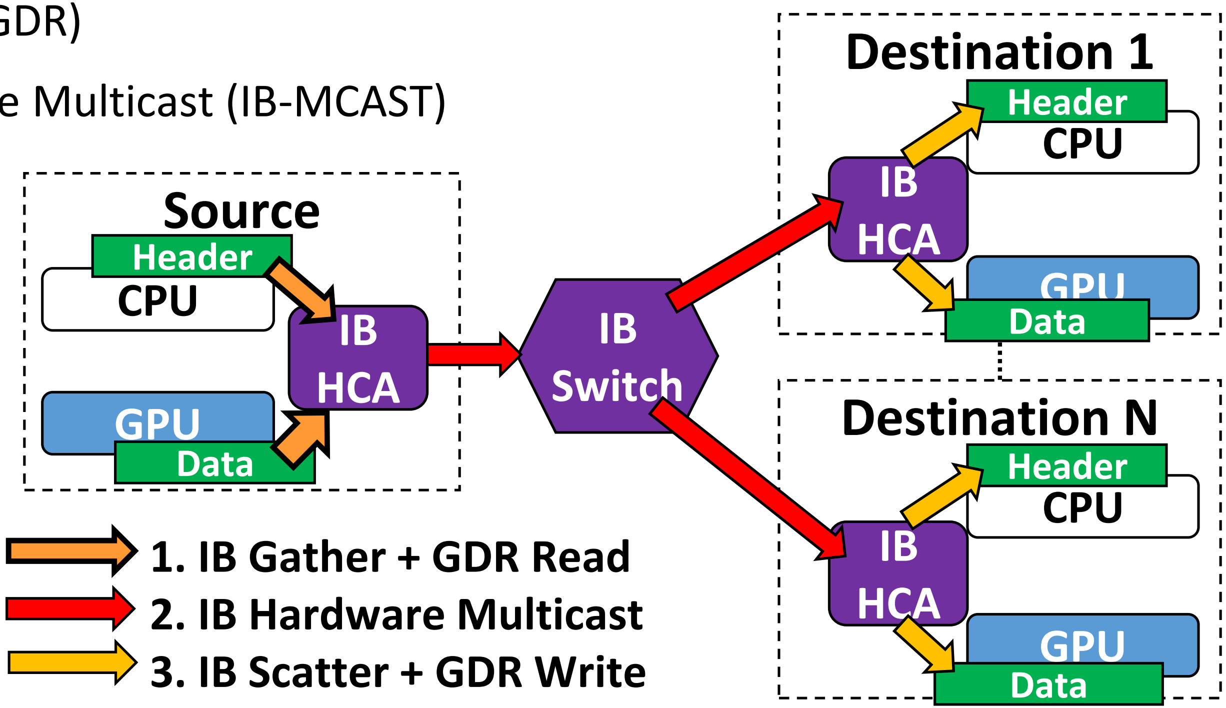
Streaming Applications

- **Streaming applications on HPC systems**
  1. **Communication (MPI)**
     - Broadcast-type operations
  2. **Computation (CUDA)**
     - Multiple GPU nodes as workers



Hardware Multicast-based Broadcast

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

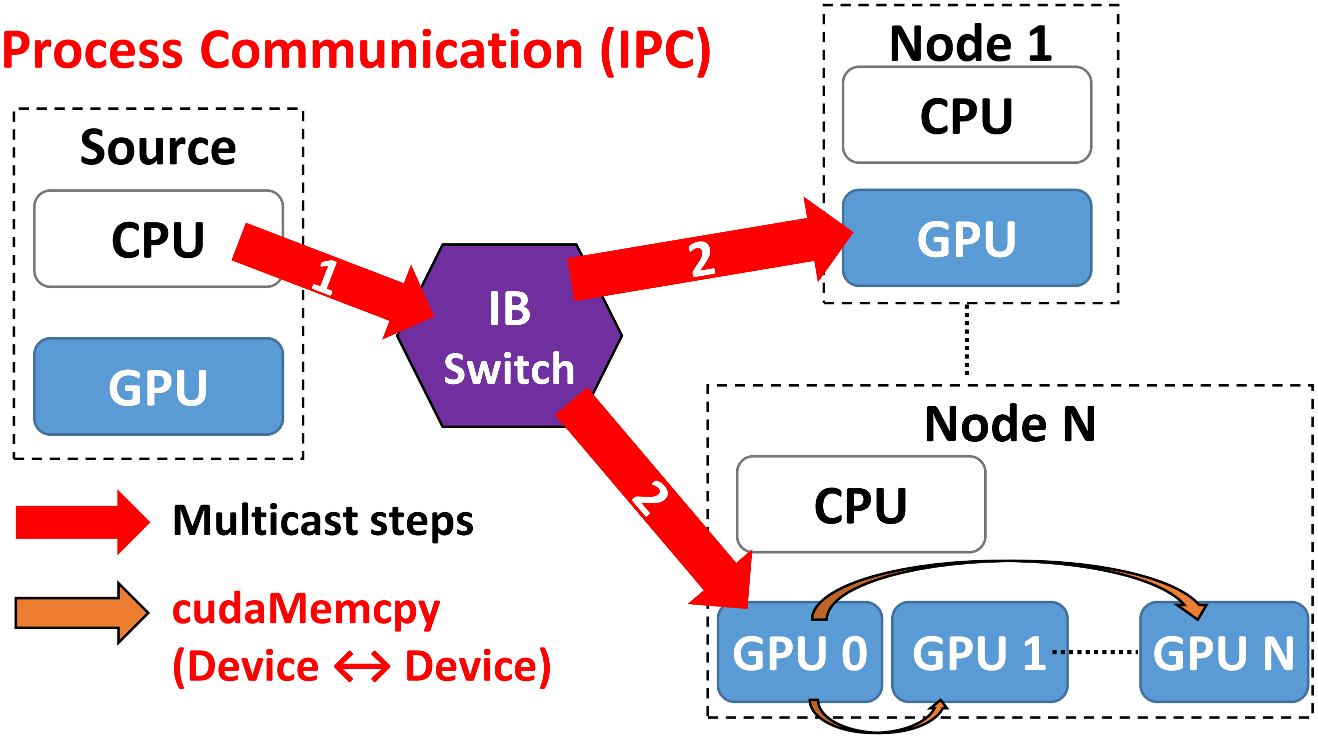


A. Venkatesh, H. Subramoni, K. Hamidouche, and D. K. Panda, "A High Performance Broadcast Design with Hardware Multicast and GPUDirect RDMA for Streaming Applications on InfiniBand Clusters," in *HiPC 2014*, Dec 2014.

Broadcast on Multi-GPU systems

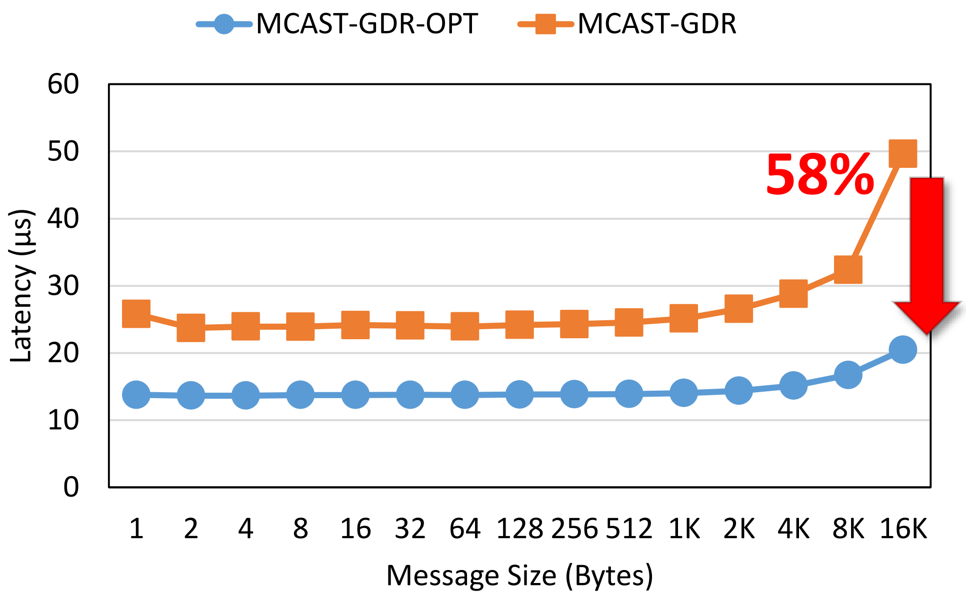
- **Proposed Intra-node Topology-Aware Broadcast**
  - CUDA InterProcess Communication (IPC)

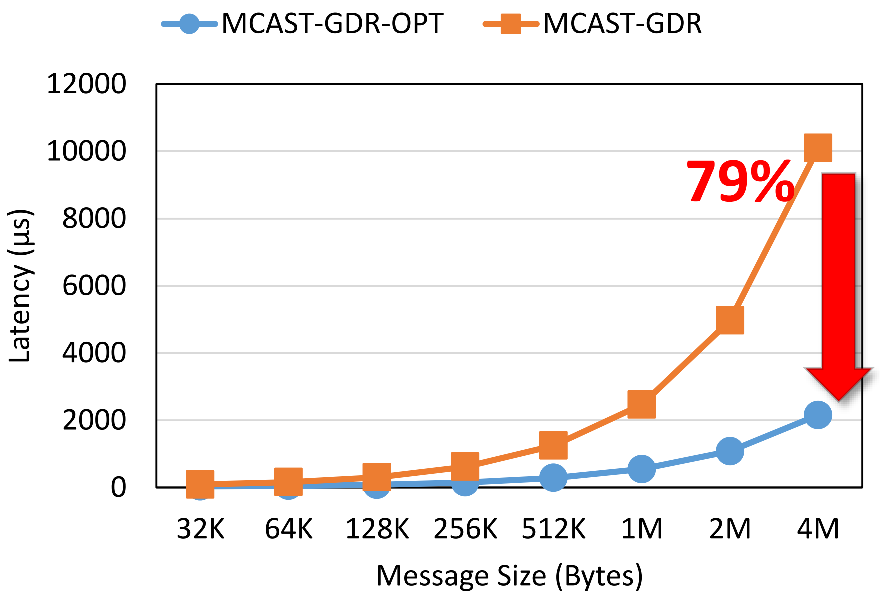
Available since MVAPICH2-GDR 2.3a



C.-H. Chu, K. Hamidouche, H. Subramoni, A. Venkatesh, B. Elton, and D. K. Panda, "Designing High Performance Heterogeneous Broadcast for Streaming Applications on GPU Clusters, " SBAC-PAD'16, Oct. 26-28, 2016.

Streaming Benchmark @ CSCS (88 GPUs)





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

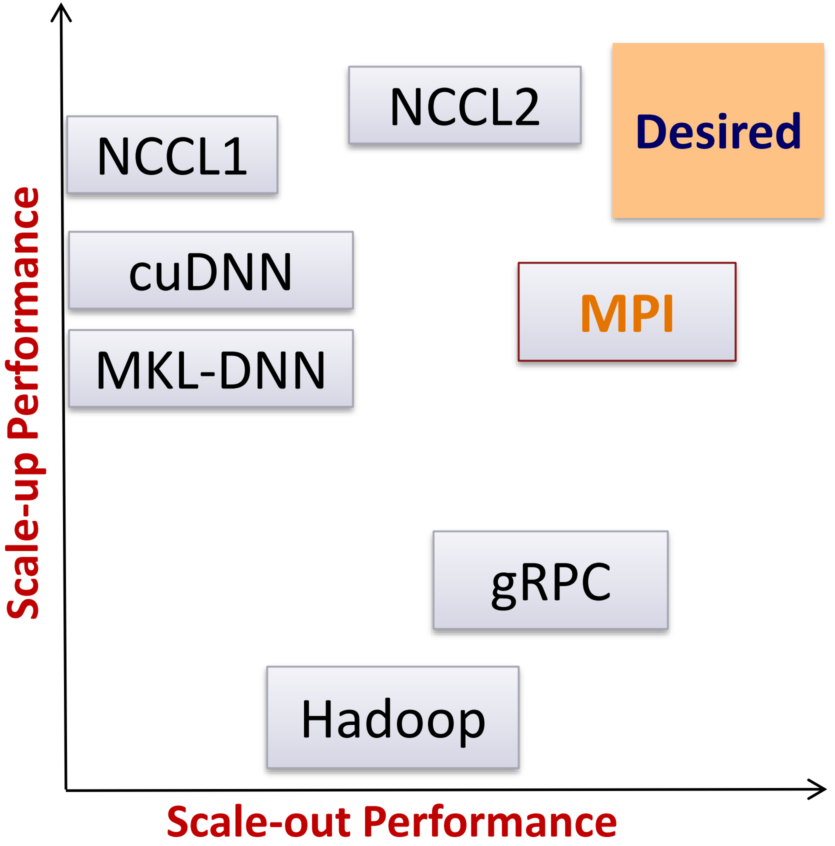
C.-H. Chu, K. Hamidouche, H. Subramoni, A. Venkatesh, B. Elton, and D. K. Panda, "Designing High Performance Heterogeneous Broadcast for Streaming Applications on GPU Clusters, " SBAC-PAD'16, Oct. 26-28, 2016.

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

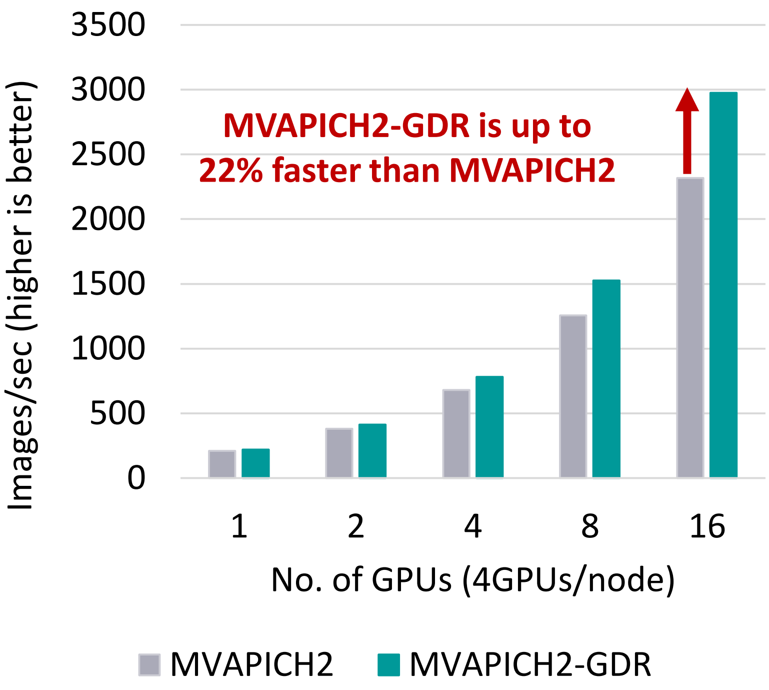
- Deep Learning frameworks are a different game altogether
  - Unusually large message sizes (order of megabytes)
  - Most communication based on GPU buffers
- Existing State-of-the-art
  - cuDNN, cuBLAS, NCCL --> **scale-up** performance
  - NCCL2, CUDA-Aware MPI --> **scale-out** performance
    - 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**?



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)

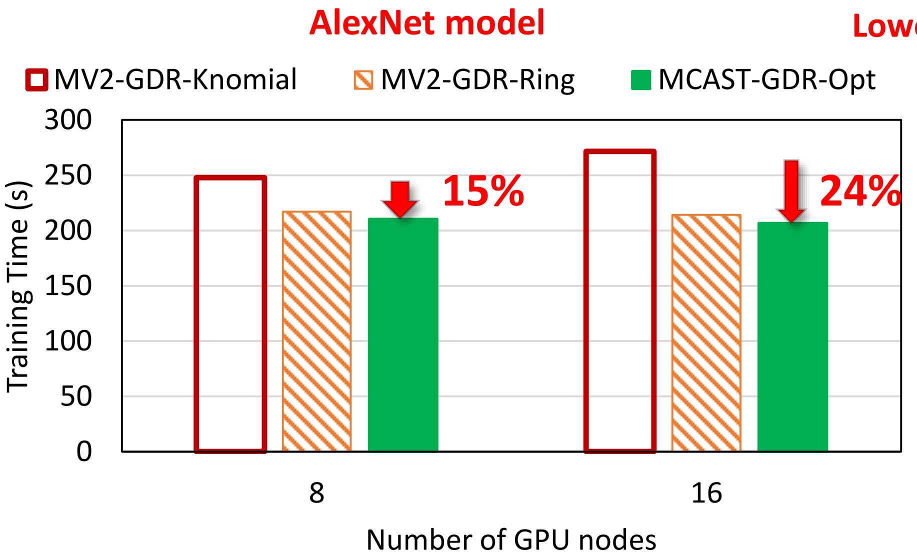
Exploiting CUDA-Aware MPI for TensorFlow (Horovod)

- MVAPICH2-GDR offers excellent performance via advanced designs for MPI_Allreduce.
- Up to **22% better** performance on Wilkes2 cluster (16 GPUs)

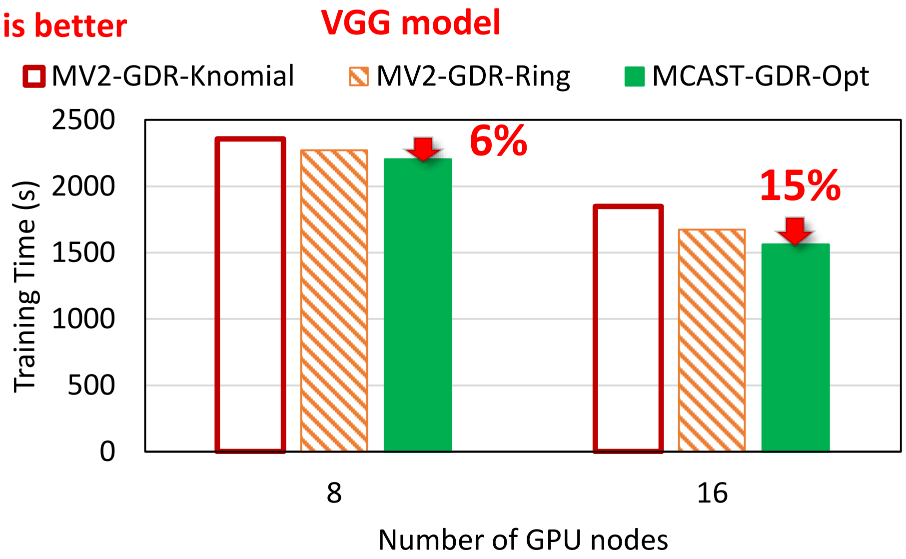


Performance Benefits with CNTK Deep Learning Framework

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



Lower is better

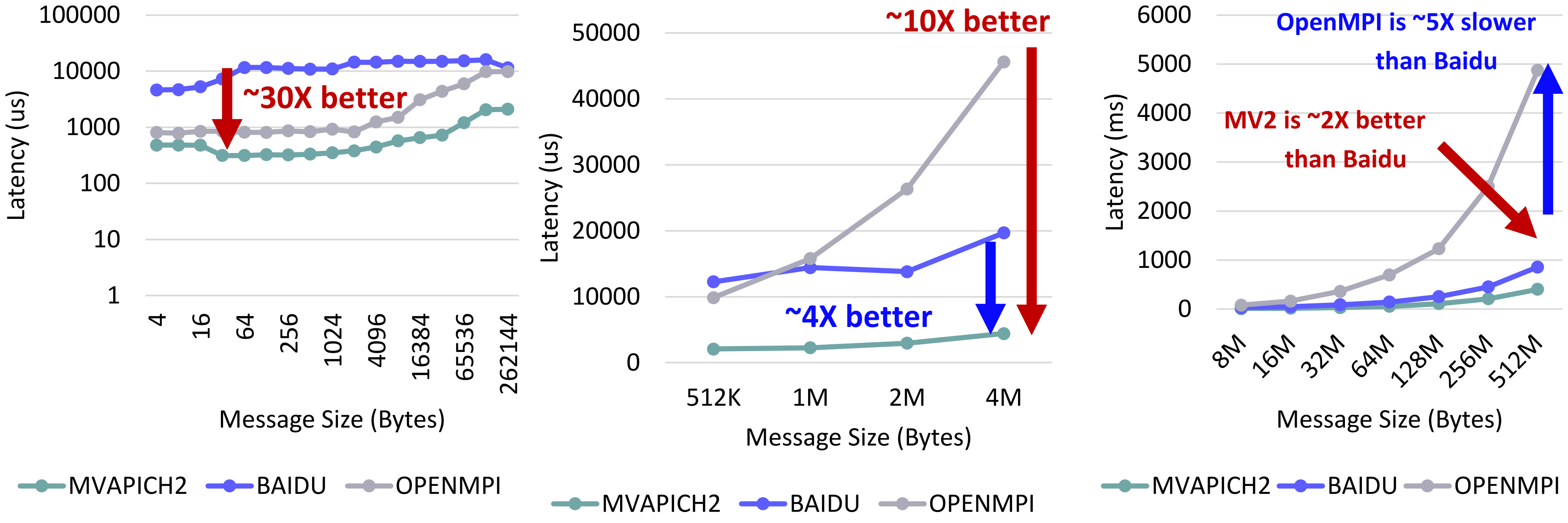


- Reduces up to 24% and 15% of latency for AlexNet and VGG models
- Higher improvement can be observed for larger system sizes

C.-H. Chu, X. Lu, A. A. Awan, H. Subramoni, J. Hashmi, B. Elton, and D. K. Panda, Efficient and Scalable Multi-Source Streaming Broadcast on GPU Clusters for Deep Learning, ICPP'17.

MVAPICH2-GDR: Allreduce Comparison with Baidu and OpenMPI

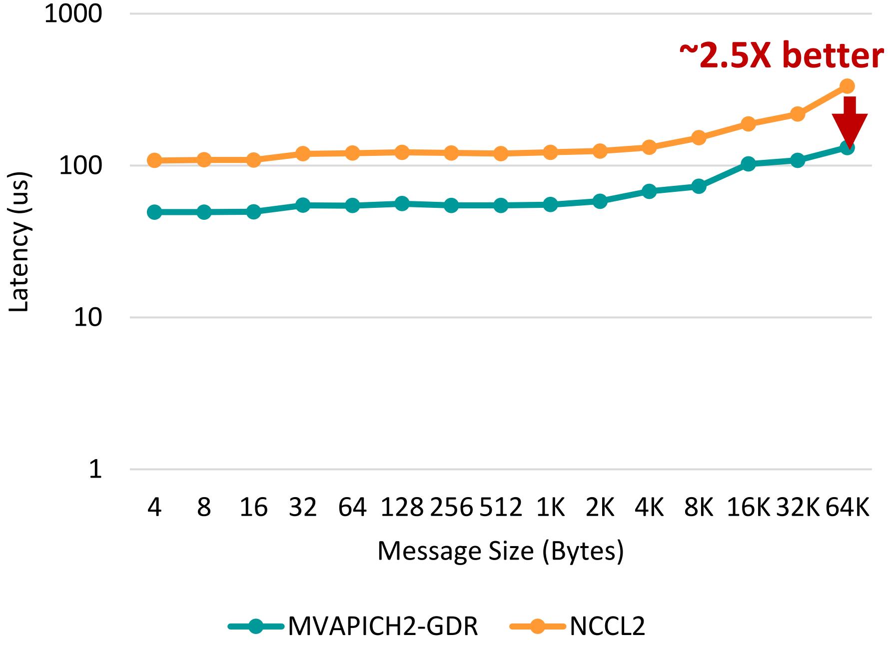
- 16 GPUs (4 nodes) MVAPICH2-GDR vs. Baidu-Allreduce and OpenMPI 3.0

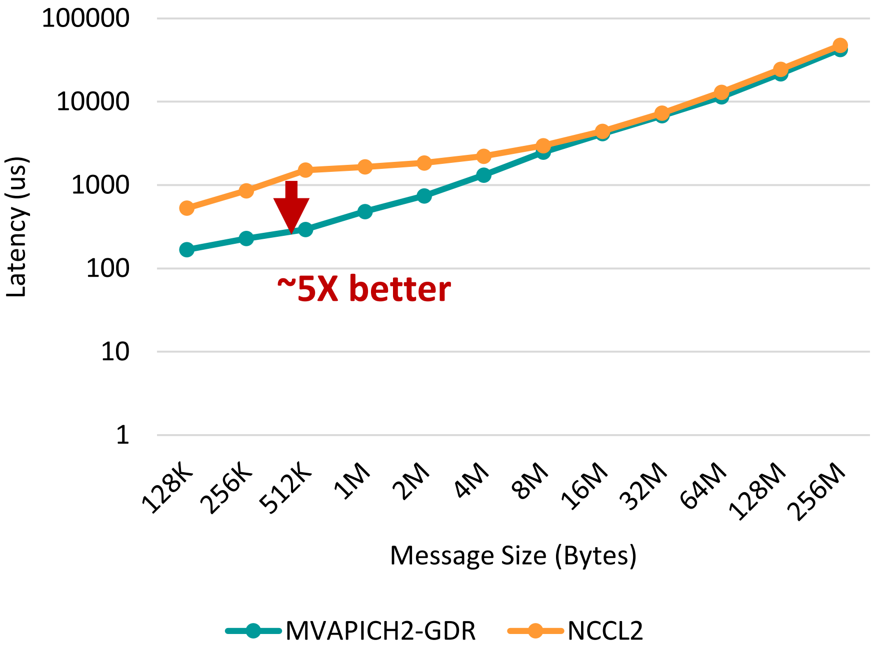


**Available since MVAPICH2-GDR 2.3a*

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**

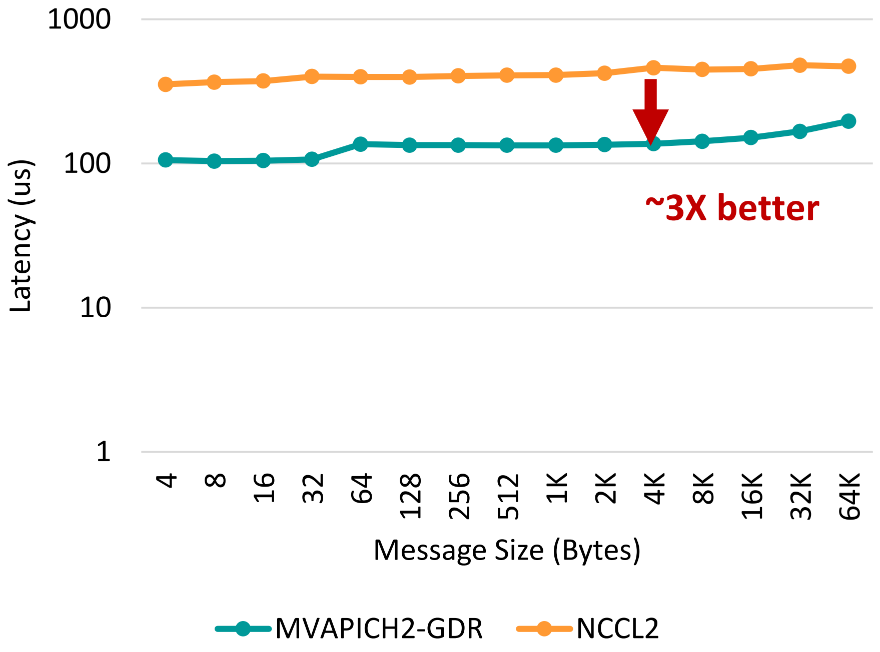


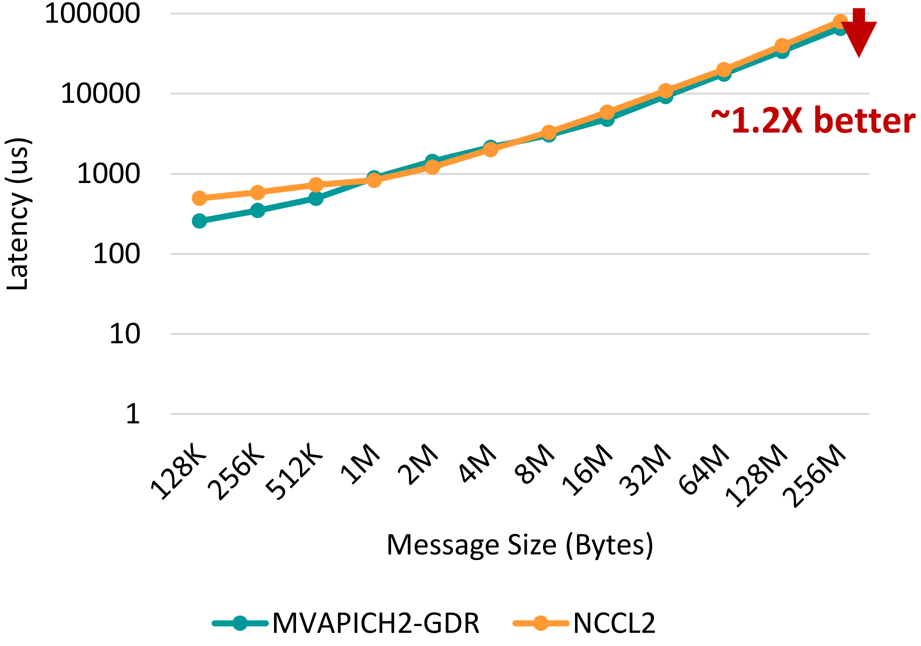


Platform: Intel Xeon (Broadwell) nodes equipped with a dual-socket CPU, 1 K-80 GPUs, and EDR InfiniBand Inter-connect

MVAPICH2-GDR vs. NCCL2 – Allreduce Operation

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





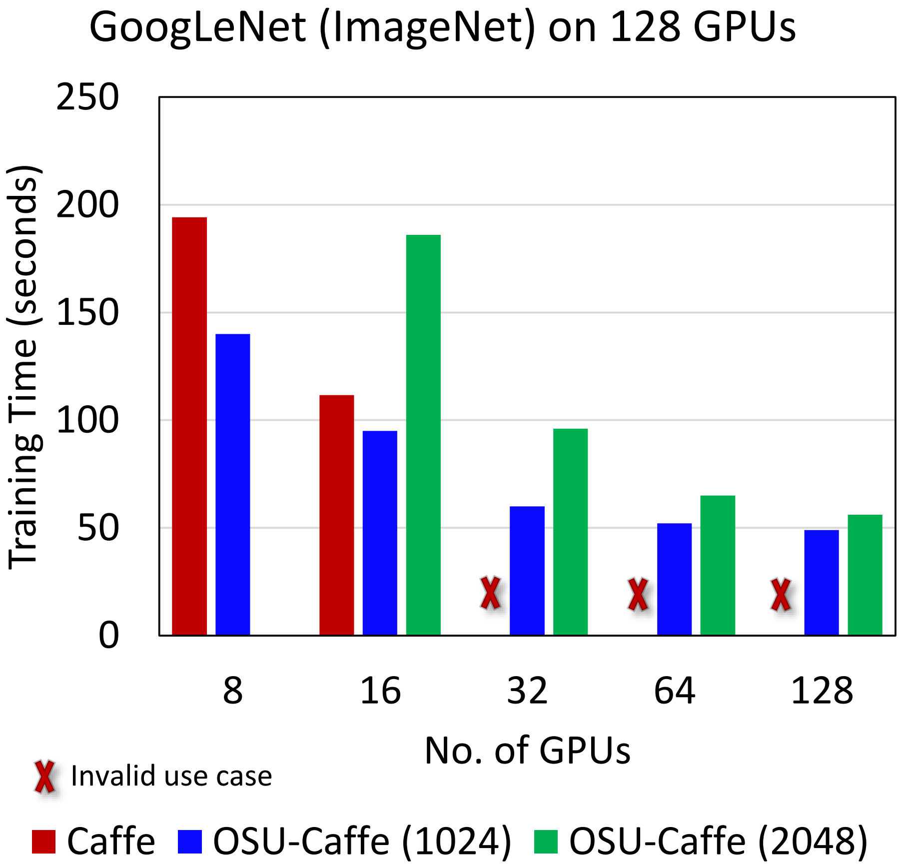
Platform: Intel Xeon (Broadwell) nodes equipped with a dual-socket CPU, 1 K-80 GPUs, and EDR InfiniBand Inter-connect

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

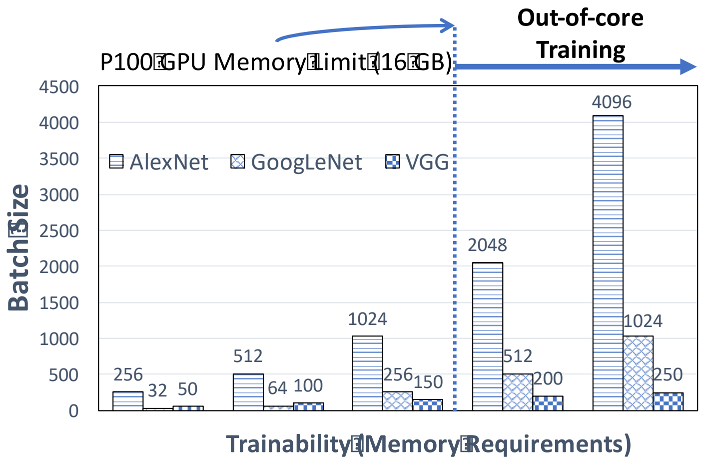
OSU-Caffe publicly available from

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



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.

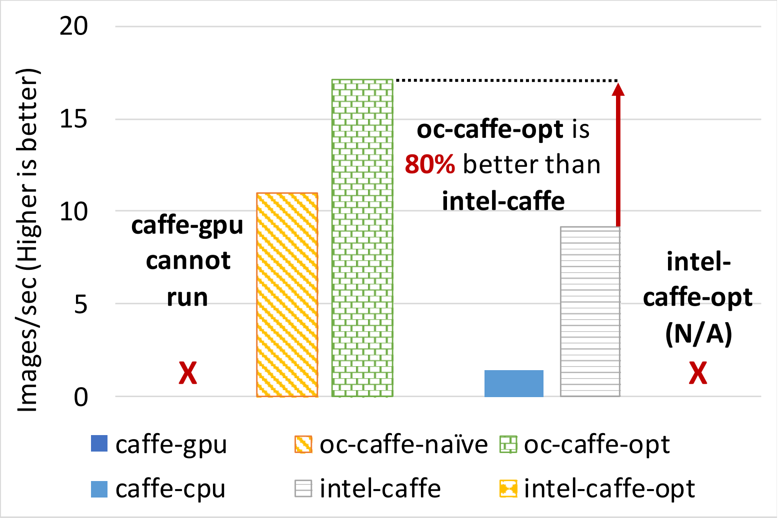


A. Awan, C. Chu, H. Subramoni, X. Lu, and D. K. Panda, OC-DNN: Exploiting Advanced Unified Memory Capabilities in CUDA 9 and Volta GPUs for Out-of-Core DNN Training, HiPC '18

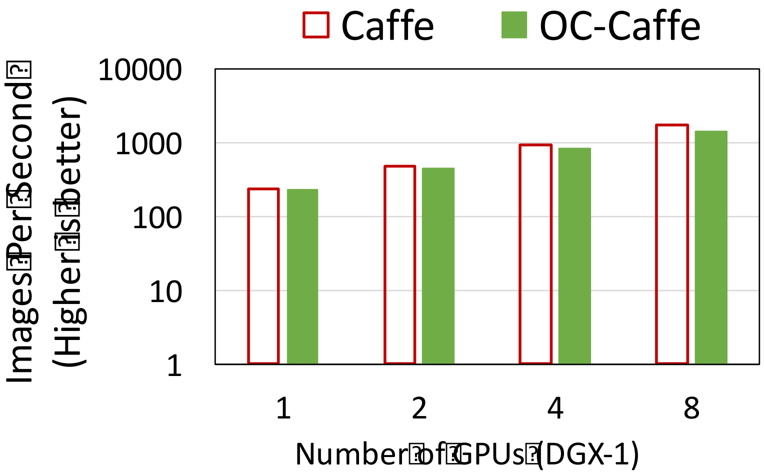
Performance Trends for OC-Caffe

- 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)



Scale-up on DGX-1



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/