MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

Performance Study of CUDA-Aware MPI libraries for GPU-enabled OpenPOWER Systems

Kawthar Shafie Khorassani, Ching-Hsiang Chu, Hari Subramoni, and Dhabaleswar K (DK) Panda

{**shafiekhorassani.1**, chu.368}@osu.edu, {subramon, panda}@cse.ohio-state.edu

Network-based Computing Laboratory

Department of Computer Science and Engineering

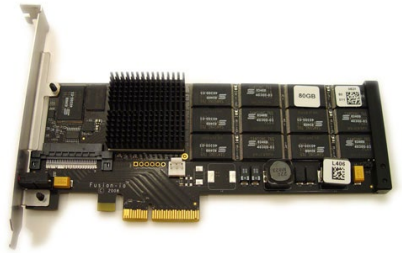
The Ohio State University

Drivers of Modern HPC Cluster Architectures

Multi-core Processors

High Performance Interconnects - InfiniBand
<1usec latency, 200Gbps Bandwidth>

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



SSD, NVMe-SSD, NVRAM

- Multi-core/many-core technologies
- Remote Direct Memory Access (RDMA)-enabled networking (InfiniBand and RoCE)
- Solid State Drives (SSDs), Non-Volatile Random-Access Memory (NVRAM), NVMe-SSD
- Accelerators (NVIDIA GPGPUs and Intel Xeon Phi)
- Available on HPC Clouds, e.g., Amazon EC2, NSF Chameleon, Microsoft Azure, etc.

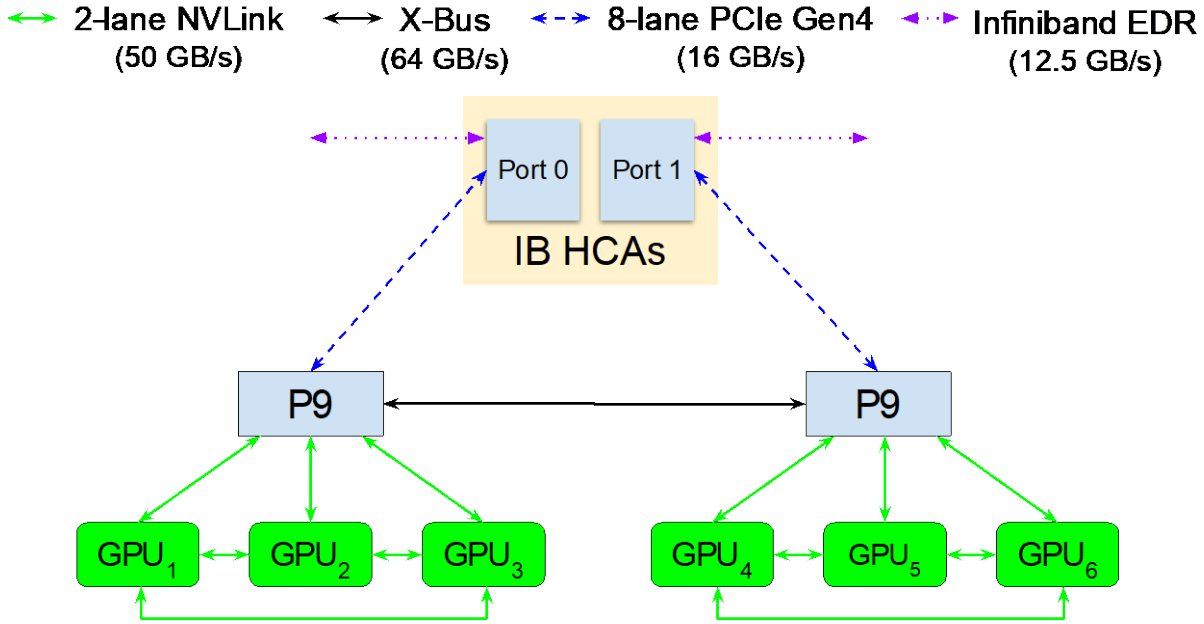
Summit

Sierra

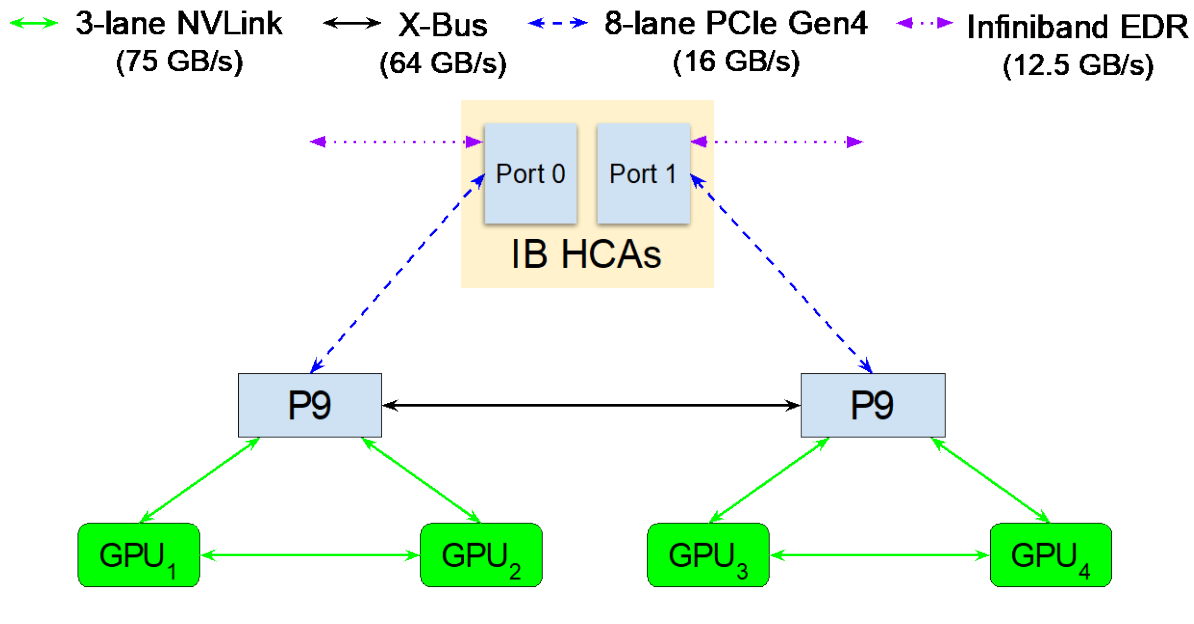
Sunway TaihuLight

K - Computer

Hardware Configuration – OpenPOWER GPU-Enabled Systems



Summit - #1 Supercomputer



Lassen - #10 Supercomputer

- **NVLink:** Between CPU and GPU and between GPUs
- **HBM2** (High Bandwidth Memory): Memory interface used in GPUs
- **IB** (InfiniBand): Between multiple OpenPOWER nodes

- **PCIe** (Peripheral Component Interconnect Express): Between CPU to Mellanox Socket-Direct InfiniBand EDR HCA
- **X-Bus:** Between two IBM POWER9 processes

Message Passing Interface (MPI)

- The defacto standard programming model in HPC
- Used in Parallel Applications to enable communication between processes
- MPI used to execute applications at scale
- CUDA-Aware MPI's for optimizing data movement on GPU clusters
  - Point-to-point Communication
    - Intra-node (GPU-GPU, GPU-Host, and Host-GPU)
    - Inter-node (GPU-GPU, GPU-Host, and Host-GPU)
  - Collective Communication
  - One-Sided Communication

CUDA-aware MPI

- 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

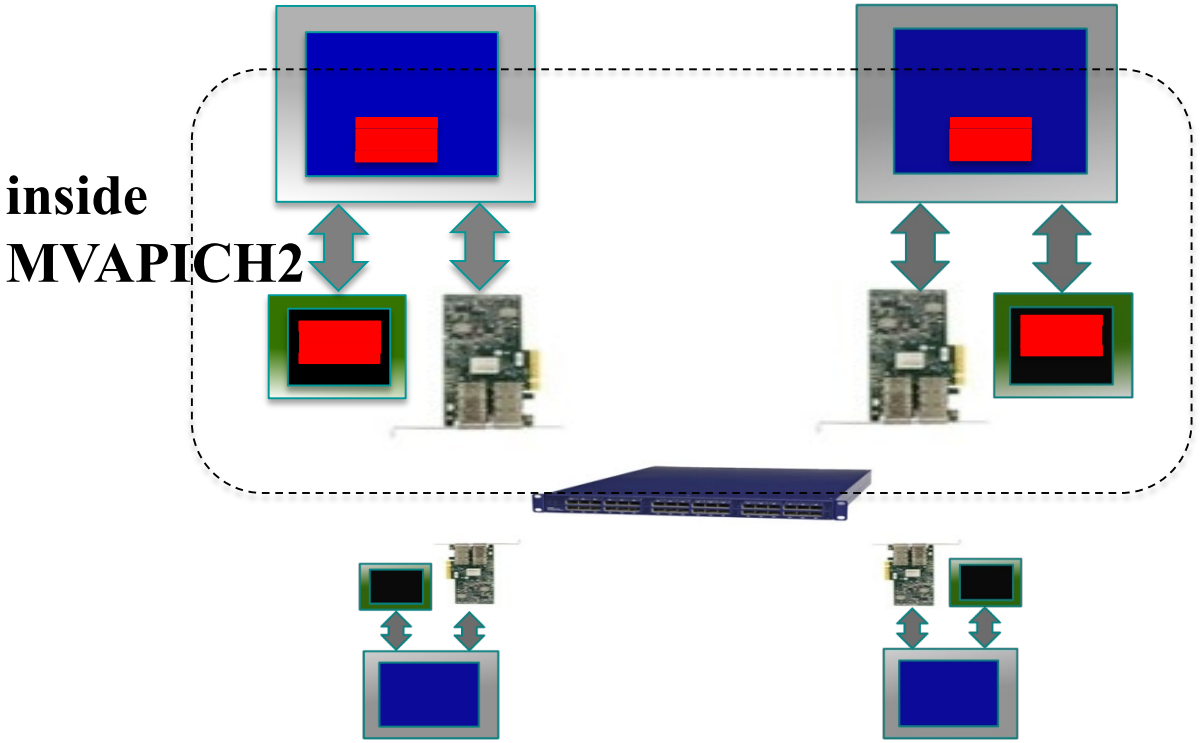
At Sender:

```
MPI_Send(s_devbuf, size, …);
```

At Receiver:

```
MPI_Recv(r_devbuf, size, …);
```

High Performance and High Productivity



CUDA-aware MPI - Communication

- MPI communication from NVIDIA GPU device memory
- High performance Remote Direct Memory Access (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)
- **CUDA IPC** (available since CUDA 4.1) in intra-node communication for multiple GPU adapters/node

Challenges

Advent of GPU-enabled OpenPOWER systems introduces new challenges:

- **Variety of interconnects**
  → Which interconnects dominate performance?
- **Lack of comprehensive evaluation of CUDA-aware MPI libraries**
  → Do MPI libraries fully leverage the interconnects?
- **Optimize communication libraries**
  → How can MPI libraries be optimized further?
- **Additional factors to consider when adjusting end applications**
  → What MPI primitives should be used to maximize performance gain?

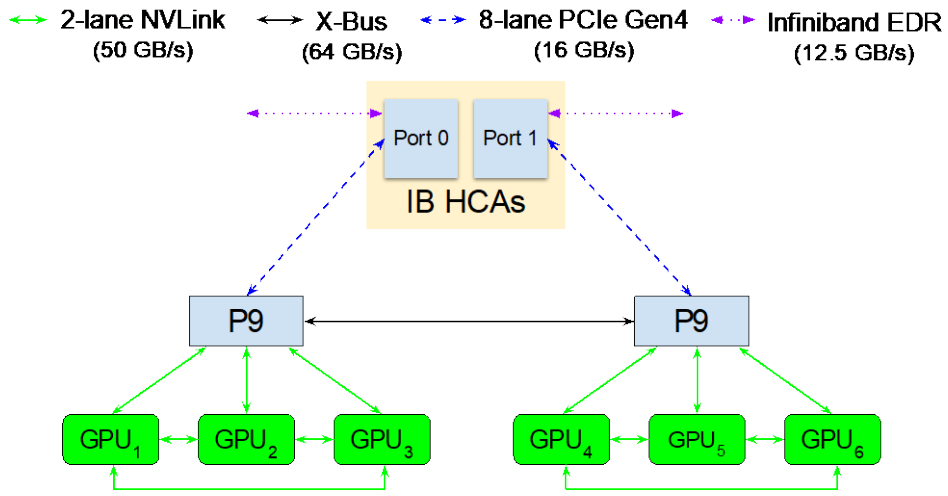
Goals of Benchmarking

Can the state-of-the-art MPI libraries fully leverage the interconnects on GPU-enabled OpenPOWER architectures?

- Evaluate CUDA-aware MPI libraries on various systems
- Determine achievable performance of MPI libraries
- Derive insights into expected performance of MPI libraries
- Evaluate bottlenecks of various configurations
- Compare performance of various MPI libraries

Native Performance of Interconnects

- NVLink between CPU and GPU
  - ***BandwidthTest*** from NVIDIA CUDA sample: Multiple cudaMemcpyAsync back-to-back between system & GPU memory
  - https://github.com/NVIDIA/cuda-samples
- GPU HBM2 and NVLink between GPUs
  - ***simpleIPC*** test from NVIDIA CUDA sample: CPU processes transfer data within a GPU and between GPUs
- X-Bus
  - **STREAM** benchmark
  - https://www.cs.virginia.edu/stream/
- InfiniBand
  - InfiniBand verbs performance test (**ib_read_bw**)



POWER9 Topology (Summit)

Native Performance of Interconnects

Theoretical and achievable peak bandwidth of interconnects on Lassen and Summit OpenPOWER Systems

		Lassen		Summit			
	GPU HBM2	3-lane NVLink2 CPU-GPU	3-lane NVLink2 GPU-GPU	2-lane NVLink2 CPU-GPU	2-lane NVLink2 GPU-GPU	X-Bus	InfiniBand EDR x 2
Theoretical Peak Bandwidth (Uni-directional)	900 GB/s	75 GB/s	75 GB/s	50 GB/s	50 GB/s	64 GB/s	12.5 GB/s
Achievable Peak Bandwidth (Uni-directional)	768.91 GB/s	68.78 GB/s	70.56 GB/s	45.9 GB/s	47 GB/s	58.01 GB/s	11.82 GB/s
Fraction of Peak	**85.43%**	**91.70%**	**91.81%**	**91.80%**	**94%**	**90.64%**	**94.56%**

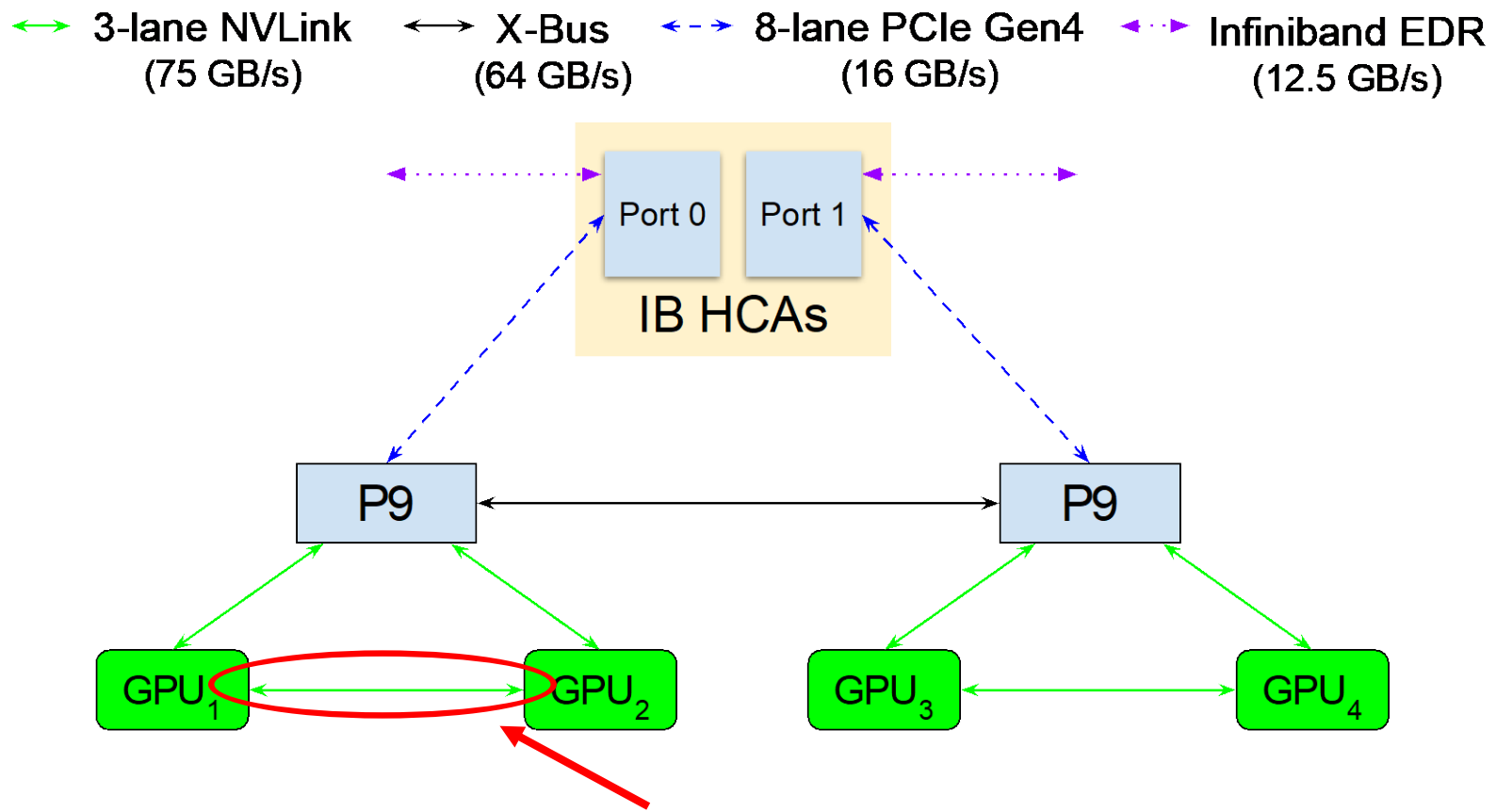
OSU Micro-Benchmarks (OMB v5.6.1)

- Benchmark to evaluate performance of traditional and CUDA-Aware MPI Libraries (point-to-point, multi-pair, and collective communication)
  - http://mvapich.cse.ohio-state.edu/benchmarks/
- Point-to-point benchmarks used:
  - Latency
  - Uni-directional Bandwidth
  - Bi-directional Bandwidth
- Communication Patterns:
  - Inter-Node, Intra-Node
  - GPU to GPU, Host to GPU

CUDA-aware MPI Libraries

- **IBM Spectrum-MPI 10.3.0.01**
  - Default CUDA-aware MPI library deployed on many OpenPOWER systems
  - https://www.ibm.com
- **OpenMPI 4.0.1 + UCX 1.6**
  - Unified Communication X (UCX): Collaboration between industry, laboratories, and academia
  - https://www.open-mpi.org
- **MVAPICH2-GDR 2.3.2**
  - Based on standard MVAPICH2 and incorporates GPUDirect RDMA technology
  - Advanced optimizations for GPU communication
  - http://mvapich.cse.ohio-state.edu

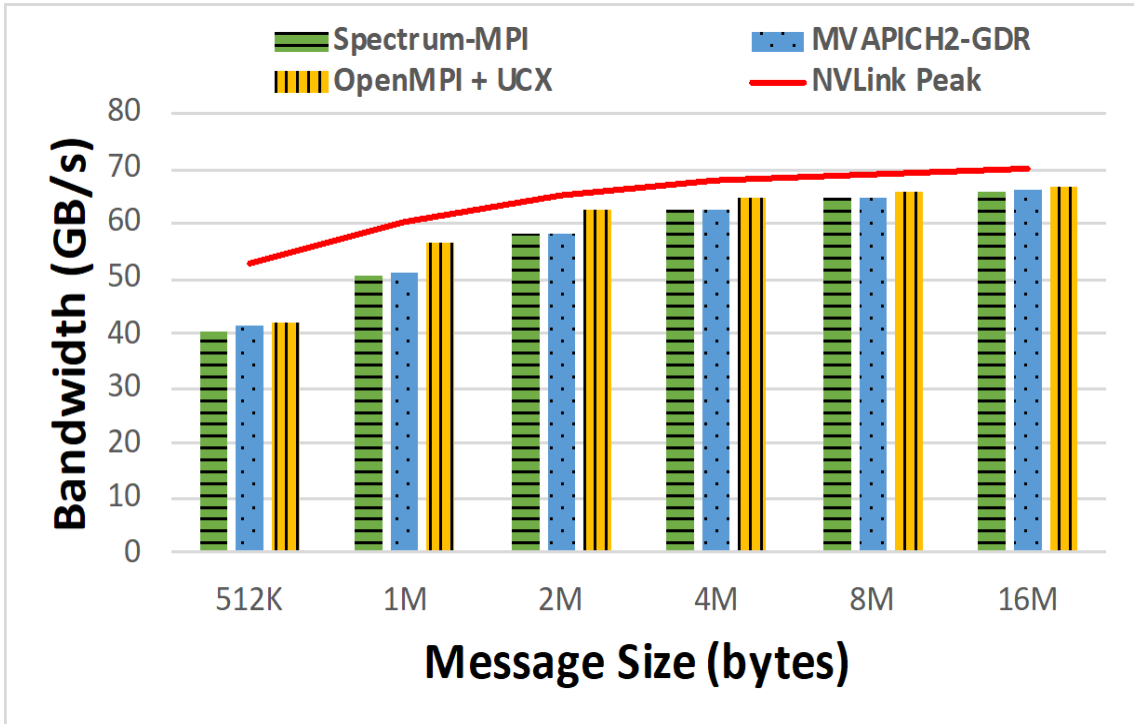
Hardware Configuration – Lassen OpenPOWER System

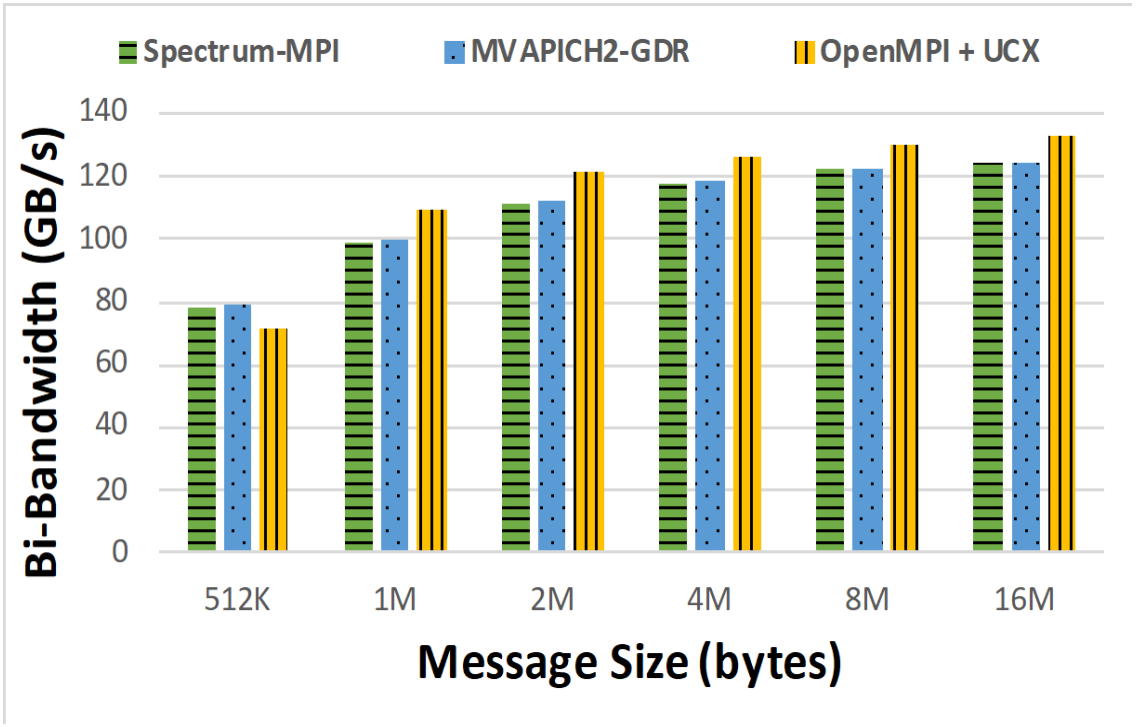


Communication through: **NVLink between GPUs**

- Theoretical Peak Bandwidth of 3-lane NVLink – 75 GB/s
- Map MPI processes to two GPUs w/ NVLink connection (CUDA_VISIBLE_DEVICES=0,1)

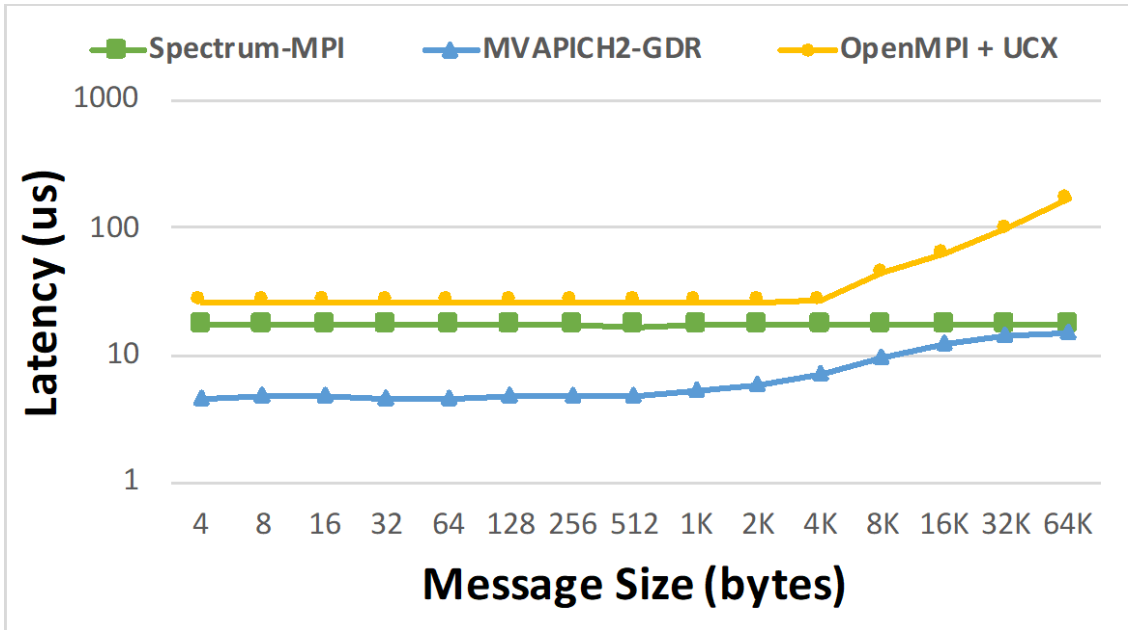
NVLink between GPUs (Lassen) - Bandwidth

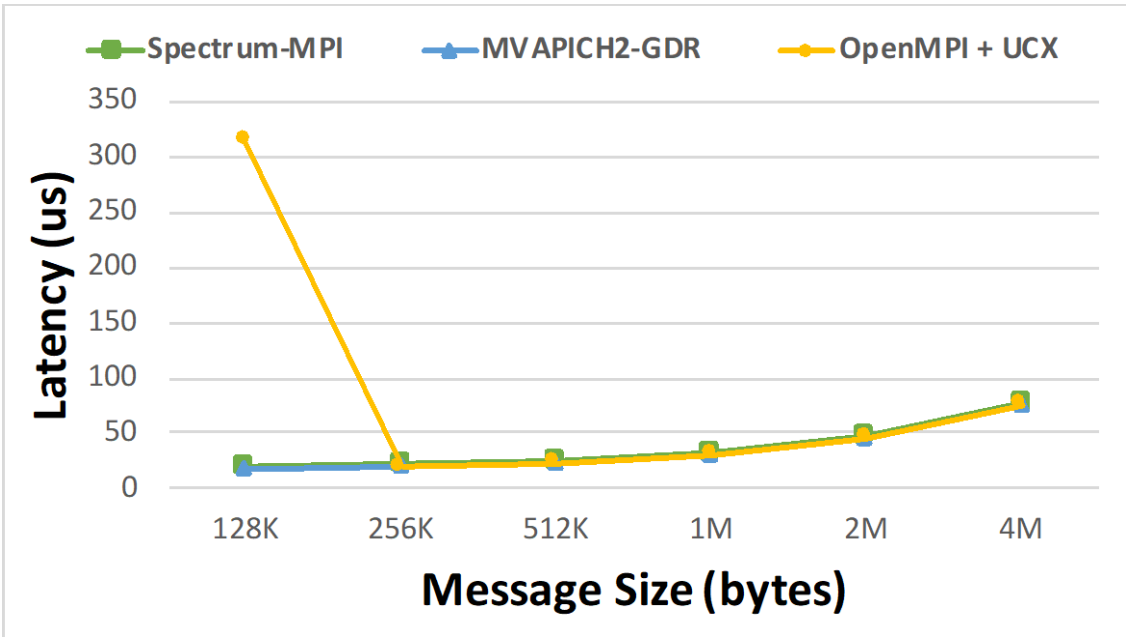




- Achievable Peak Bandwidth of 3-lane NVLink – 70.56 GB/s
- All 3 libraries deliver **~95%** of achievable bandwidth

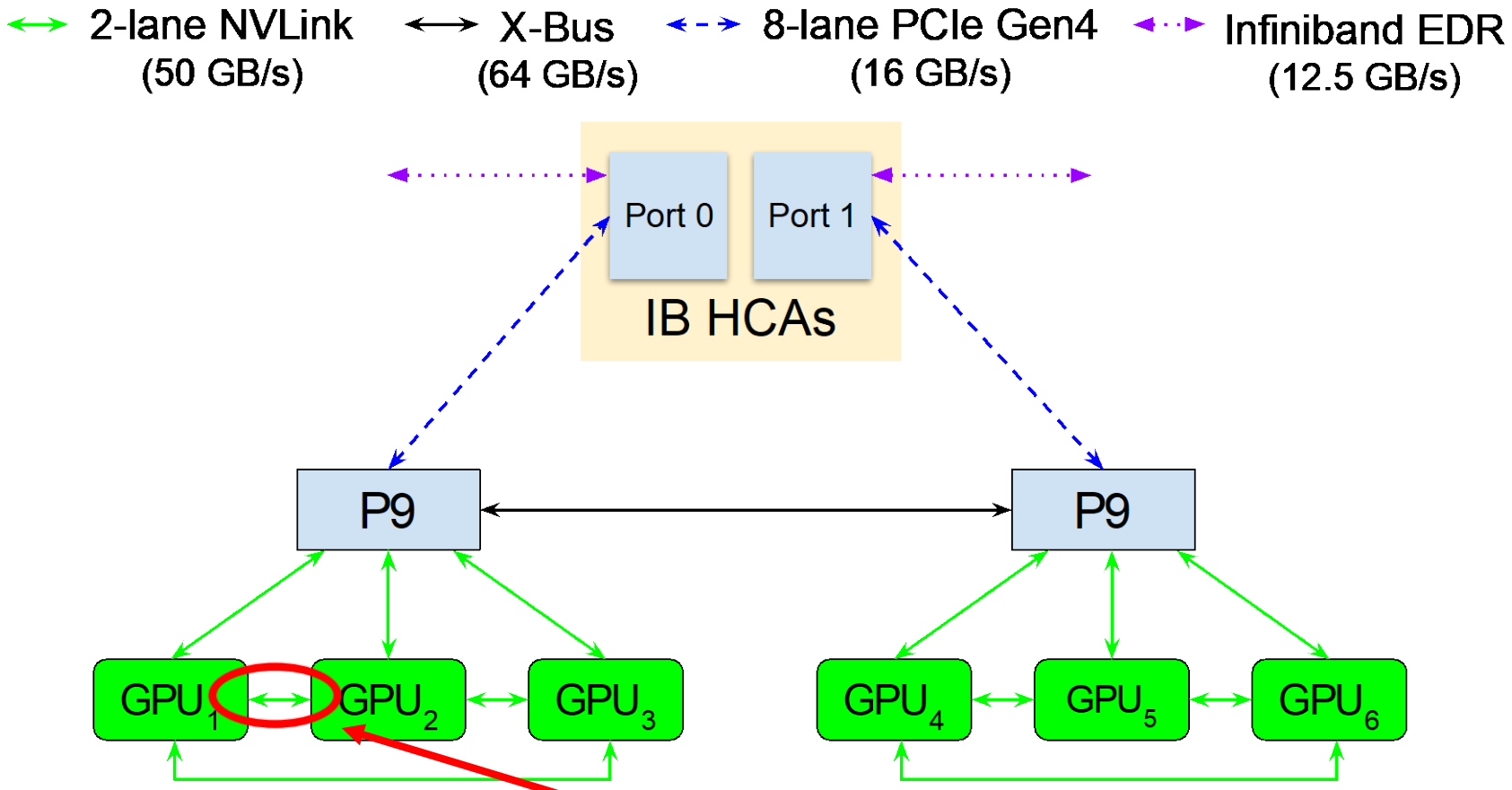
NVLink between GPUs (Lassen) - Latency





- MVAPICH2-GDR **~4x** lower than Spectrum-MPI & **~5x** lower than OpenMPI up to 4KB
- OpenMPI increases in latency from 4KB to 256KB
  - Possibly an issue with setting the thresholds for selecting communication protocols

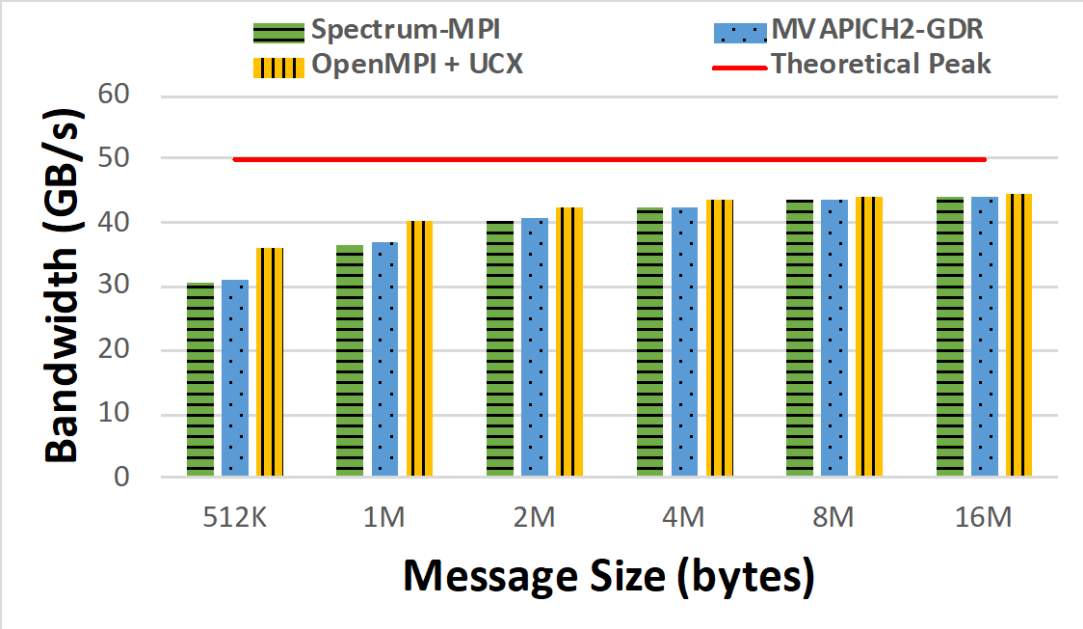
Hardware Configuration – Summit OpenPOWER System

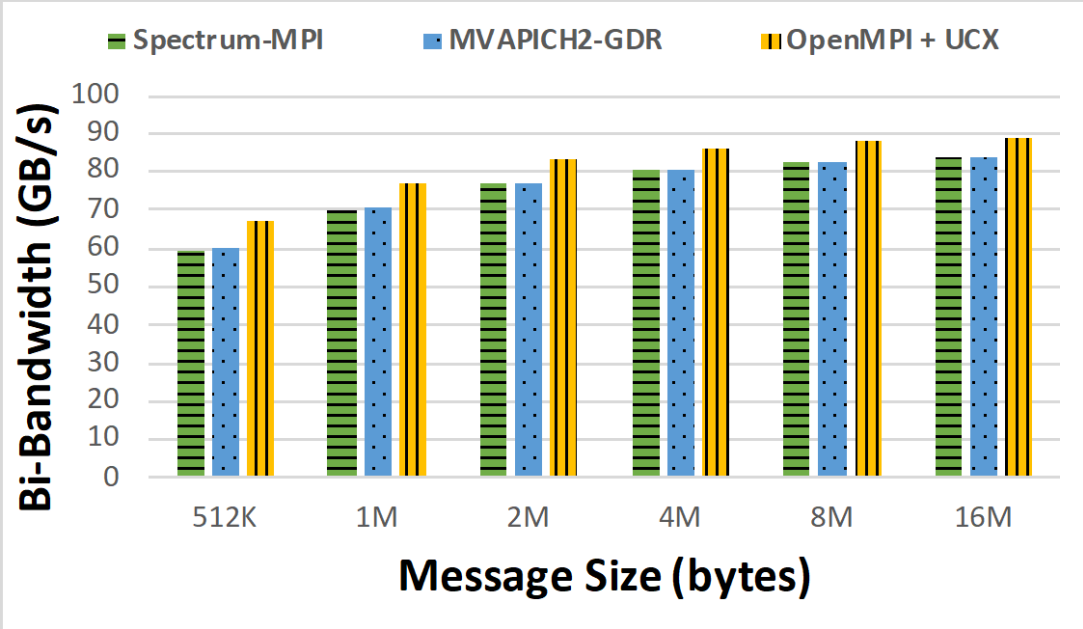


Communication through: **NVLink between GPUs**

- Theoretical Peak Bandwidth of 2-lane NVLink – 50 GB/s
- Map MPI processes to two GPUs w/ NVLink connection (CUDA_VISIBLE_DEVICES=0,1)

NVLink between GPUs (Summit) - Bandwidth

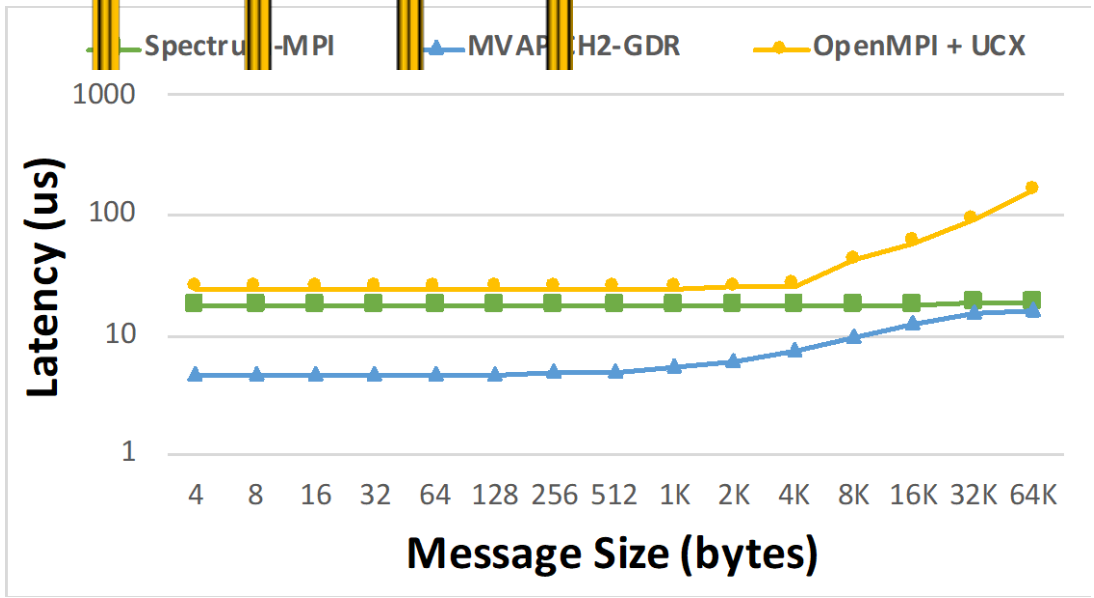


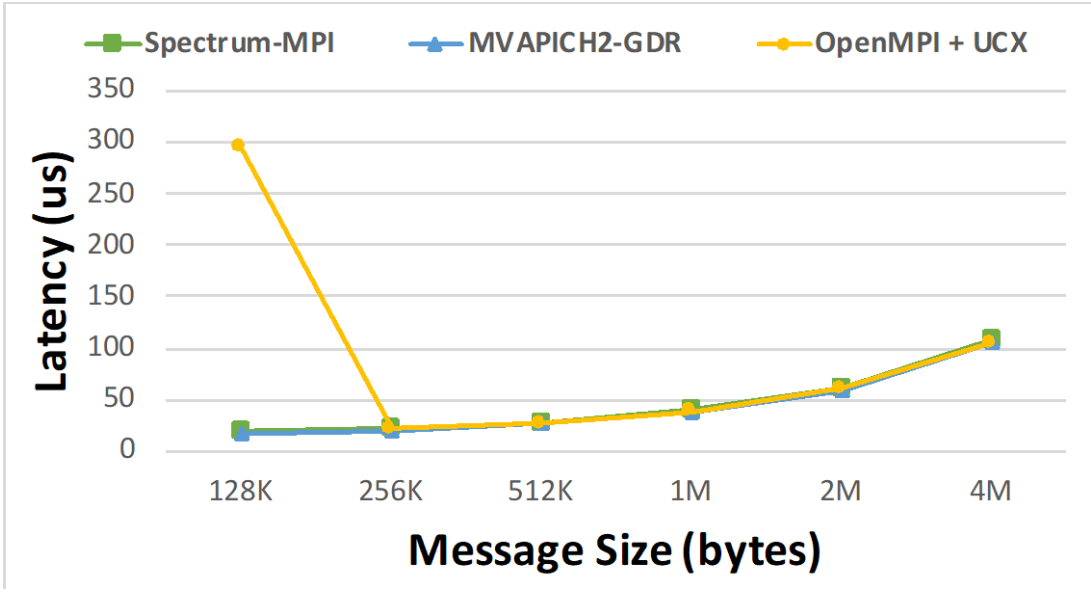


Summit **2-Lane** NVLink2 Peak Bandwidth differs from 3-Lane NVLink2 bandwidth:

- Achievable Peak Bandwidth of 2-lane NVLink – 47 GB/s
  - MVAPICH2-GDR & Spectrum-MPI Peak Bandwidth: **~44.2GB/s**
  - OpenMPI Peak Bandwidth: **~44.4GB/s**

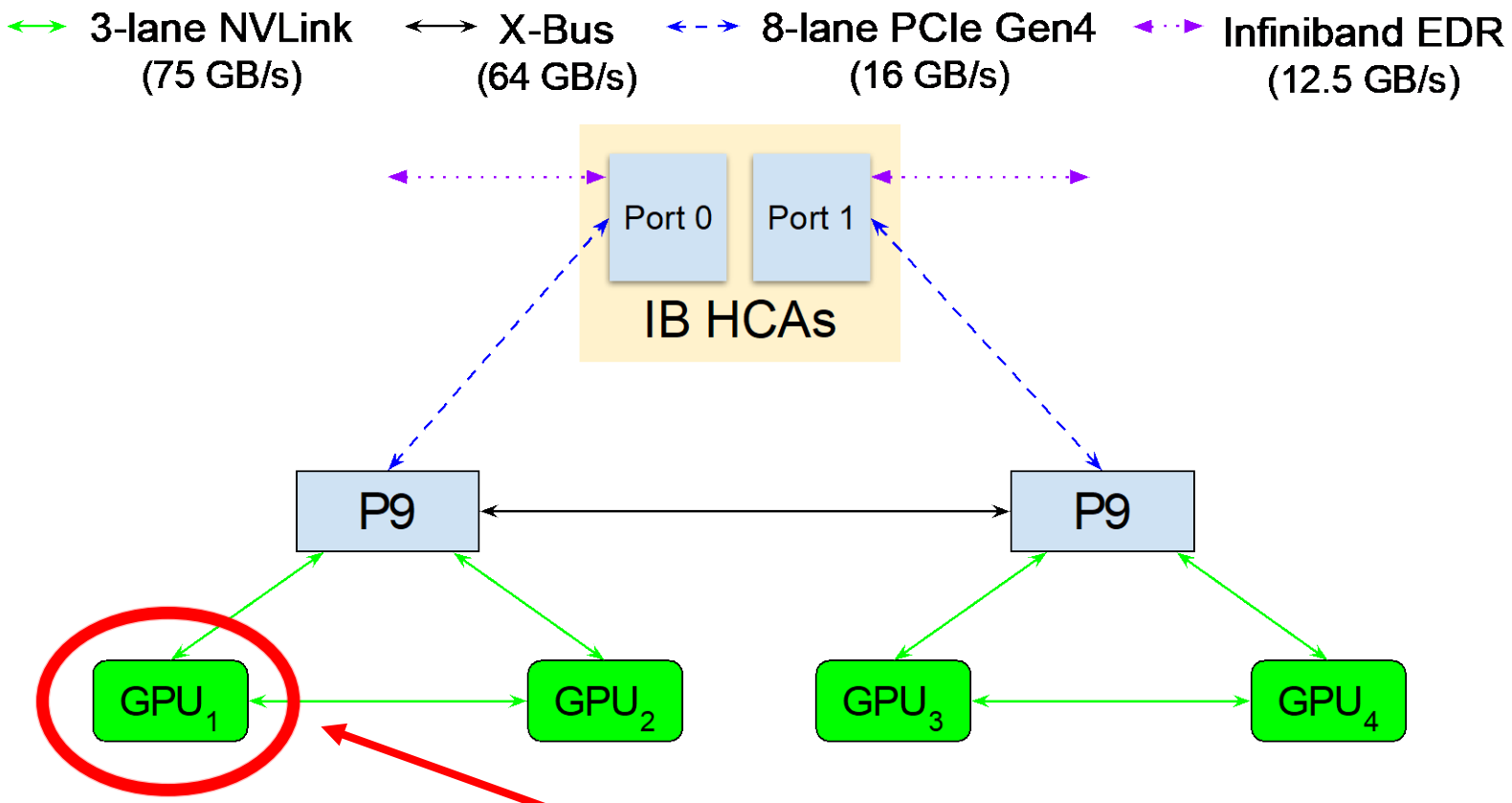
NVLink between GPUs (Summit) - Latency





- Similar trends in NVLink GPU-GPU latency on Summit as the Lassen System
- MVAPICH2-GDR outperforms Spectrum-MPI & OpenMPI
- OpenMPI has performance degradation in latency for 4KB to 256KB
  - setting the thresholds for selecting communication protocols

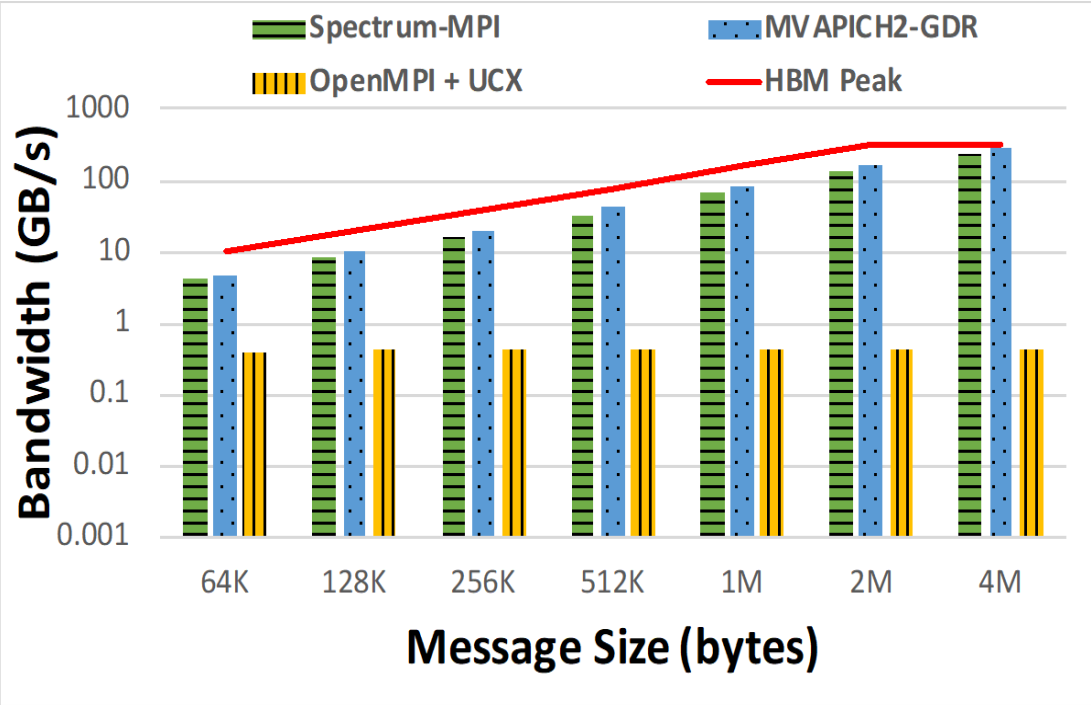
Hardware Configuration – Lassen OpenPOWER System

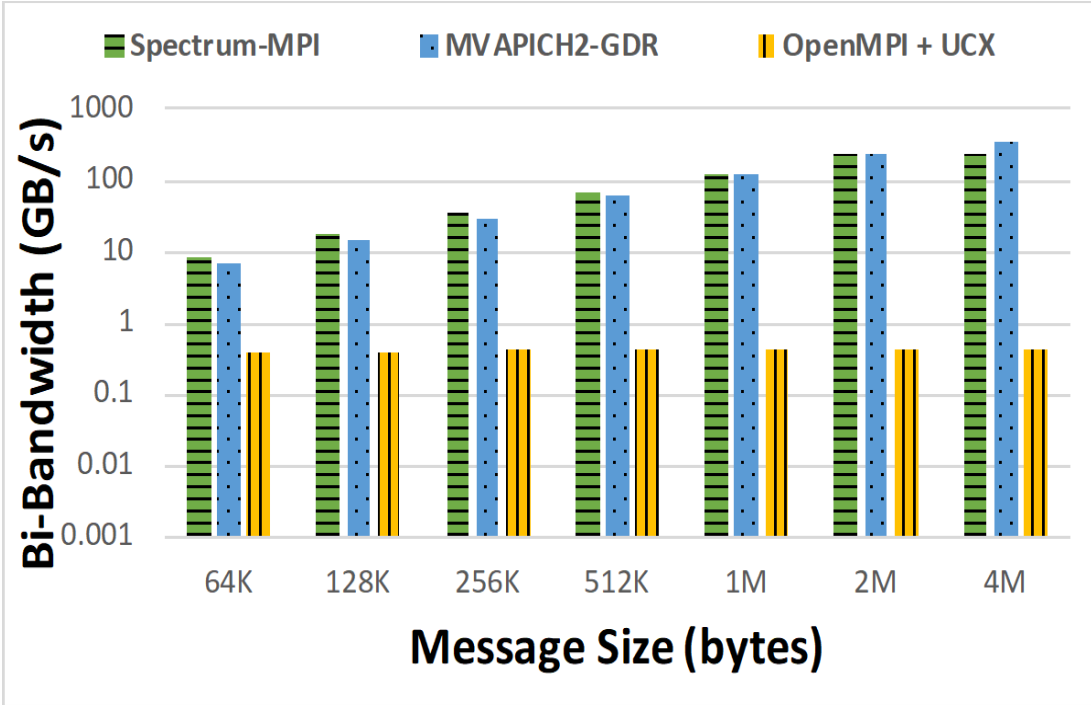


Communication through: **HBM2**

- Theoretical Peak Bandwidth – 900 GB/s
- Map two processes to the same GPU (CUDA_VISIBLE_DEVICES=0)

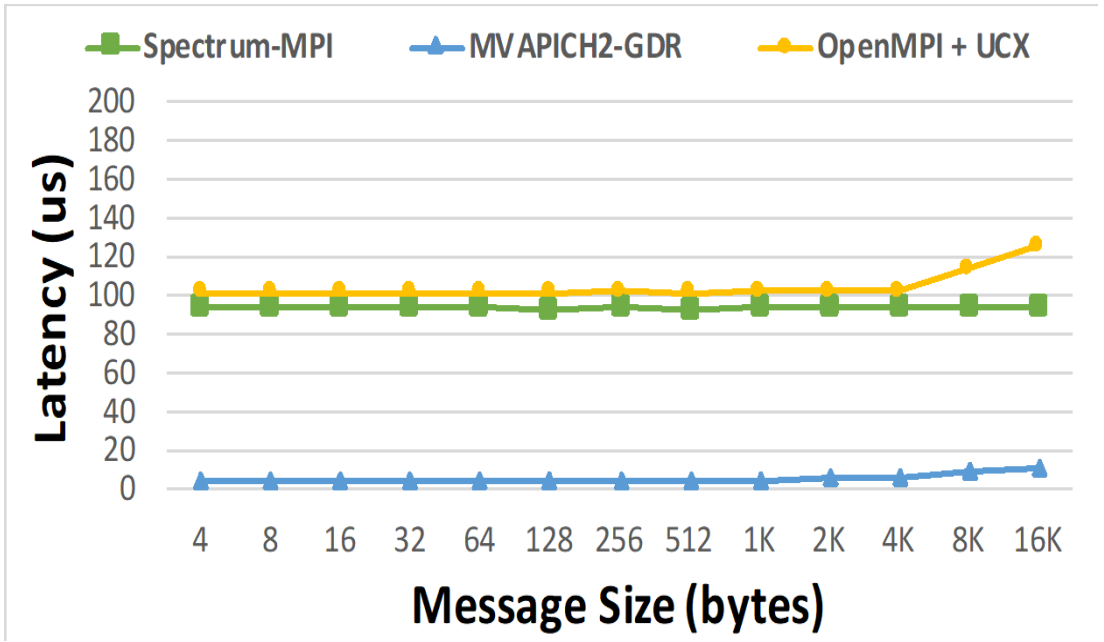
GPU HBM2 (Lassen) - Bandwidth

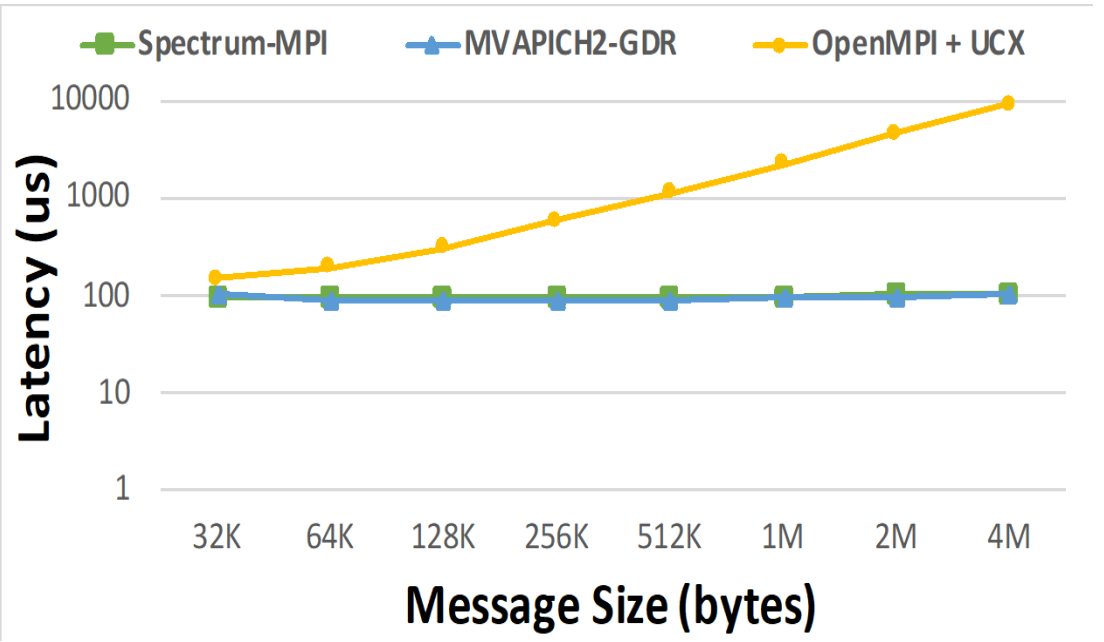




- Achievable Peak Bandwidth – 768.91 GB/s
  - MPI libraries achieve about half of the peak bandwidth for HBM2
  - Limitation of GPUs MPS feature when sharing single GPU
- OpenMPI **not** using IPC for intra-node, intra-GPU communication

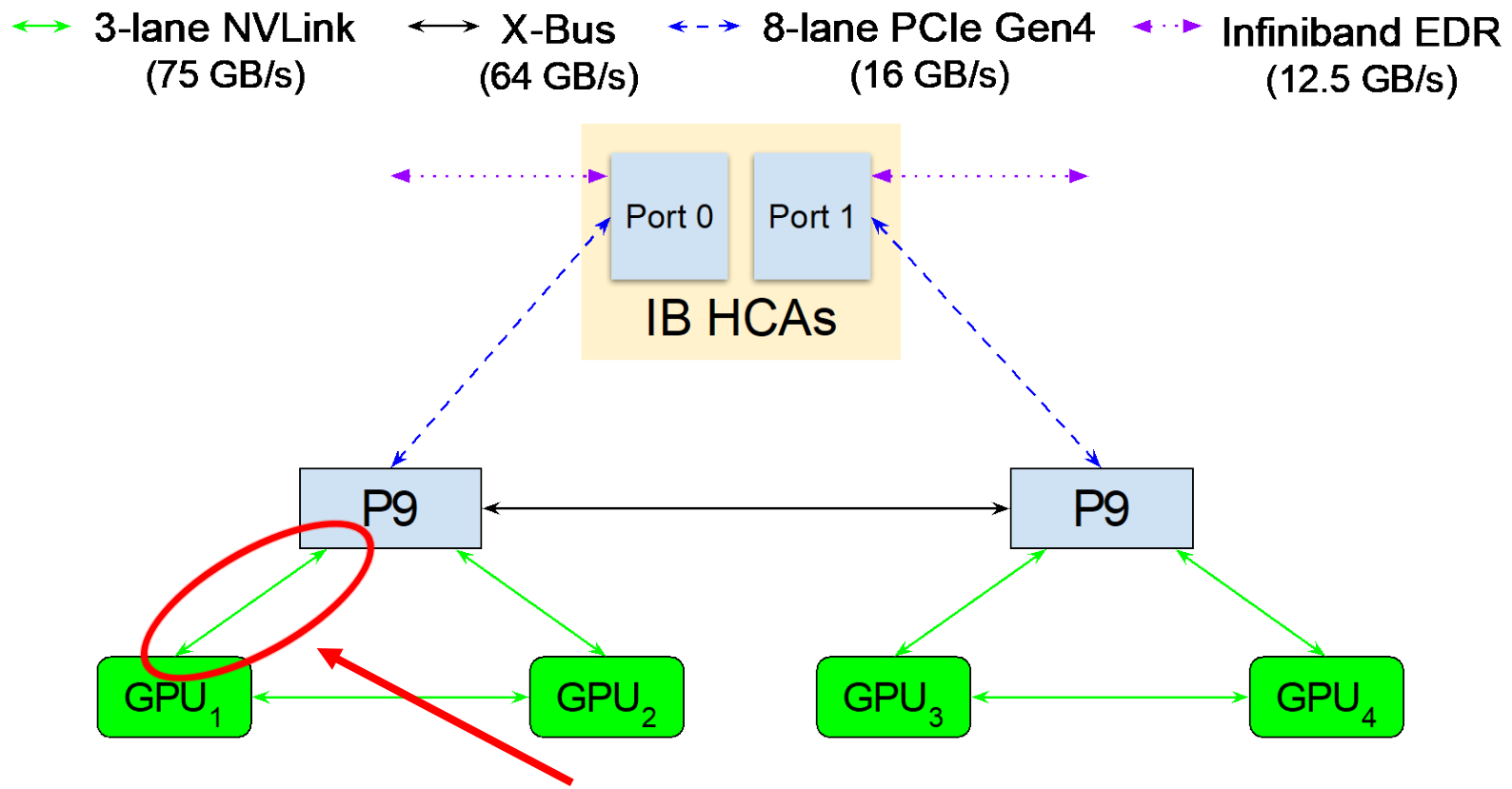
GPU HBM2 (Lassen) - Latency





- Multi-Process Service (MPS) Capabilities in NVIDIA GPUs
- MVAPICH2-GDR **~20x** less than other libraries up to 4KB
- OpenMPI linear trend in latency after 4KB
  - OpenMPI does **not** use IPC for intra-node, intra-GPU communication

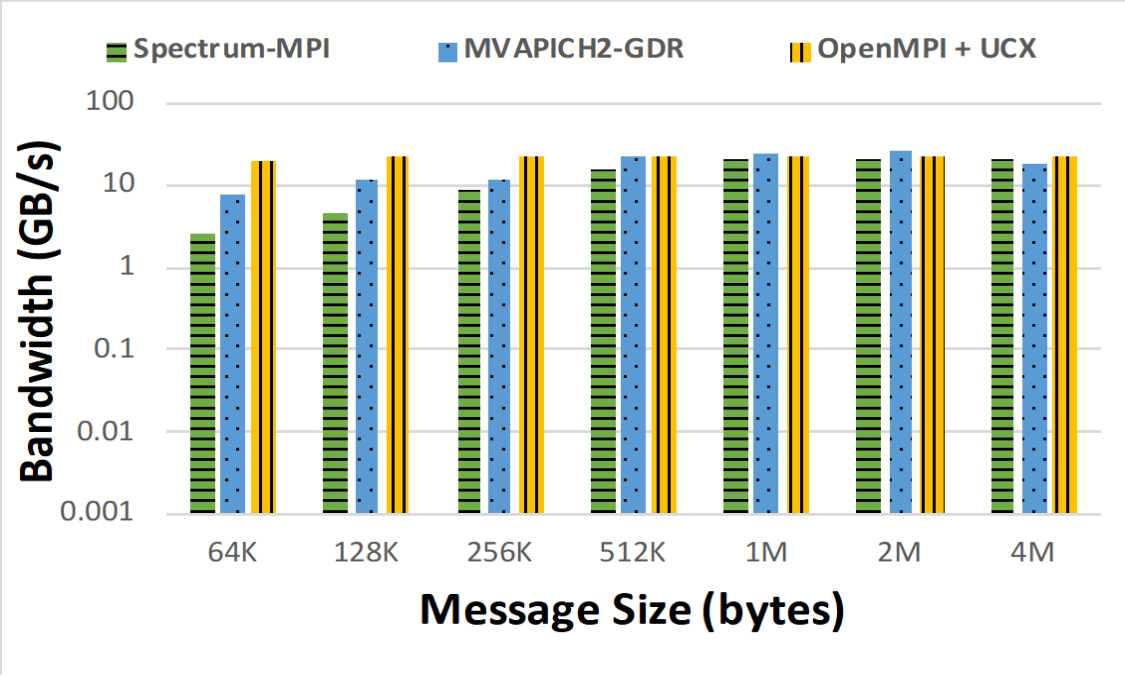
Hardware Configuration – Lassen OpenPOWER System

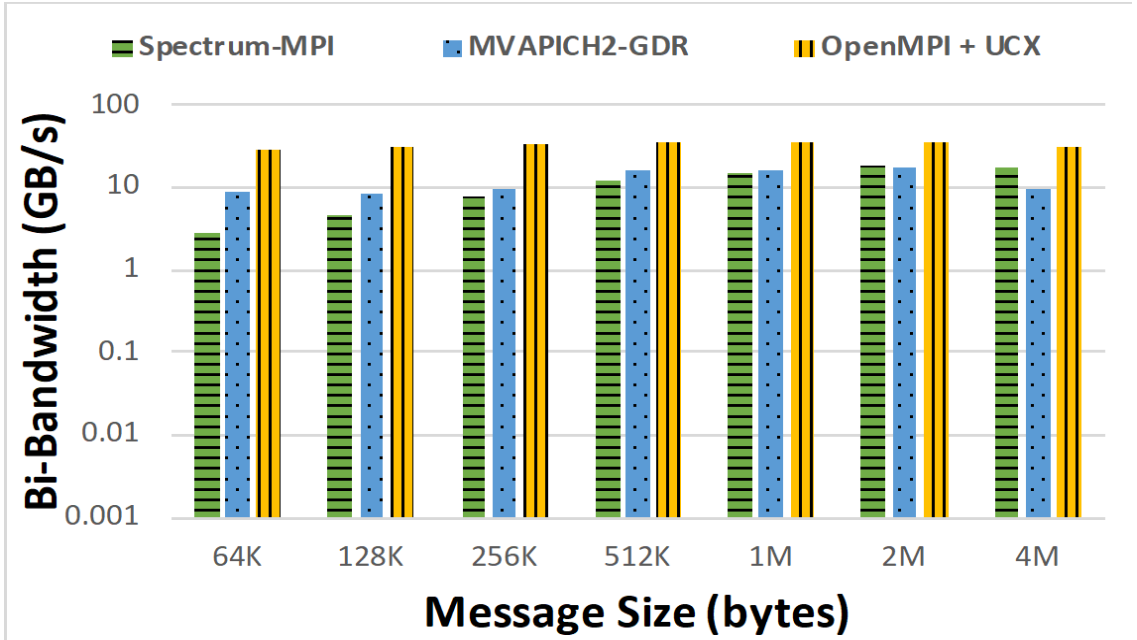


Communication through: **NVLink between CPU and GPU**

- Theoretical Peak Bandwidth – 75 GB/s
- Map a process to a CPU to communicate with a process mapped to a GPU

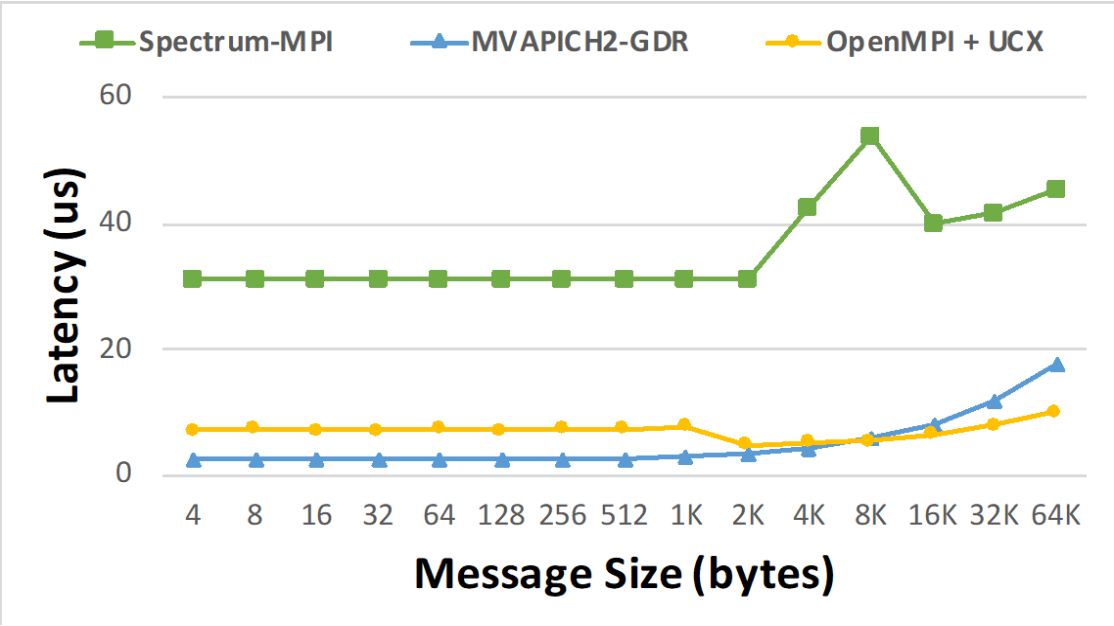
NVLink between CPU and GPU (Lassen) - Bandwidth

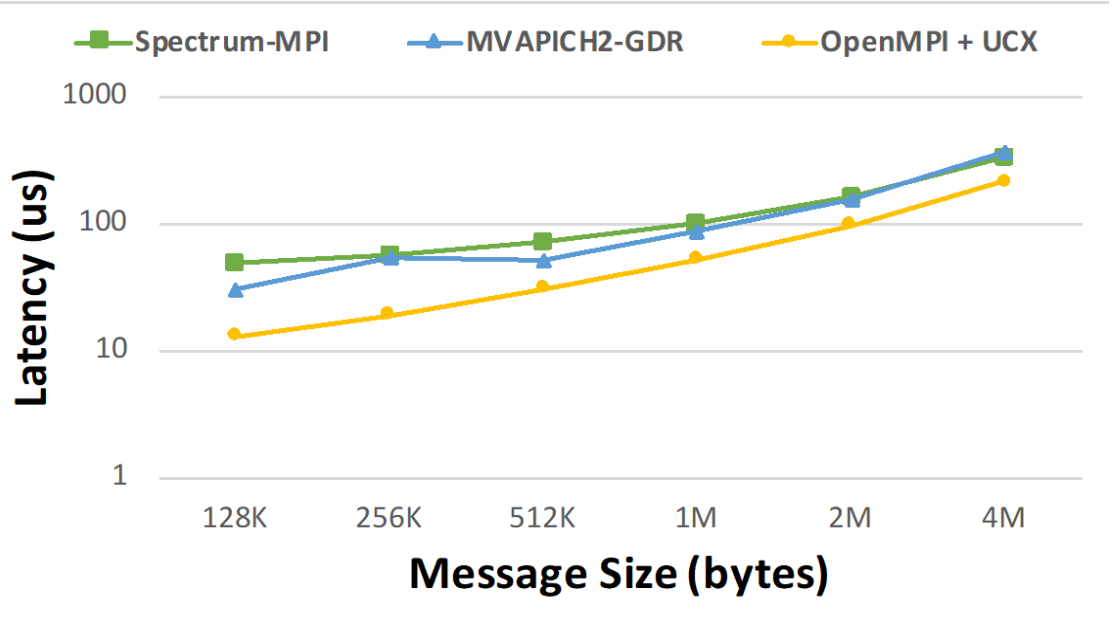




- Achievable Peak Bandwidth – 68.78 GB/s
- Bounded by 8-lane PCIe Gen4 and IB HCA

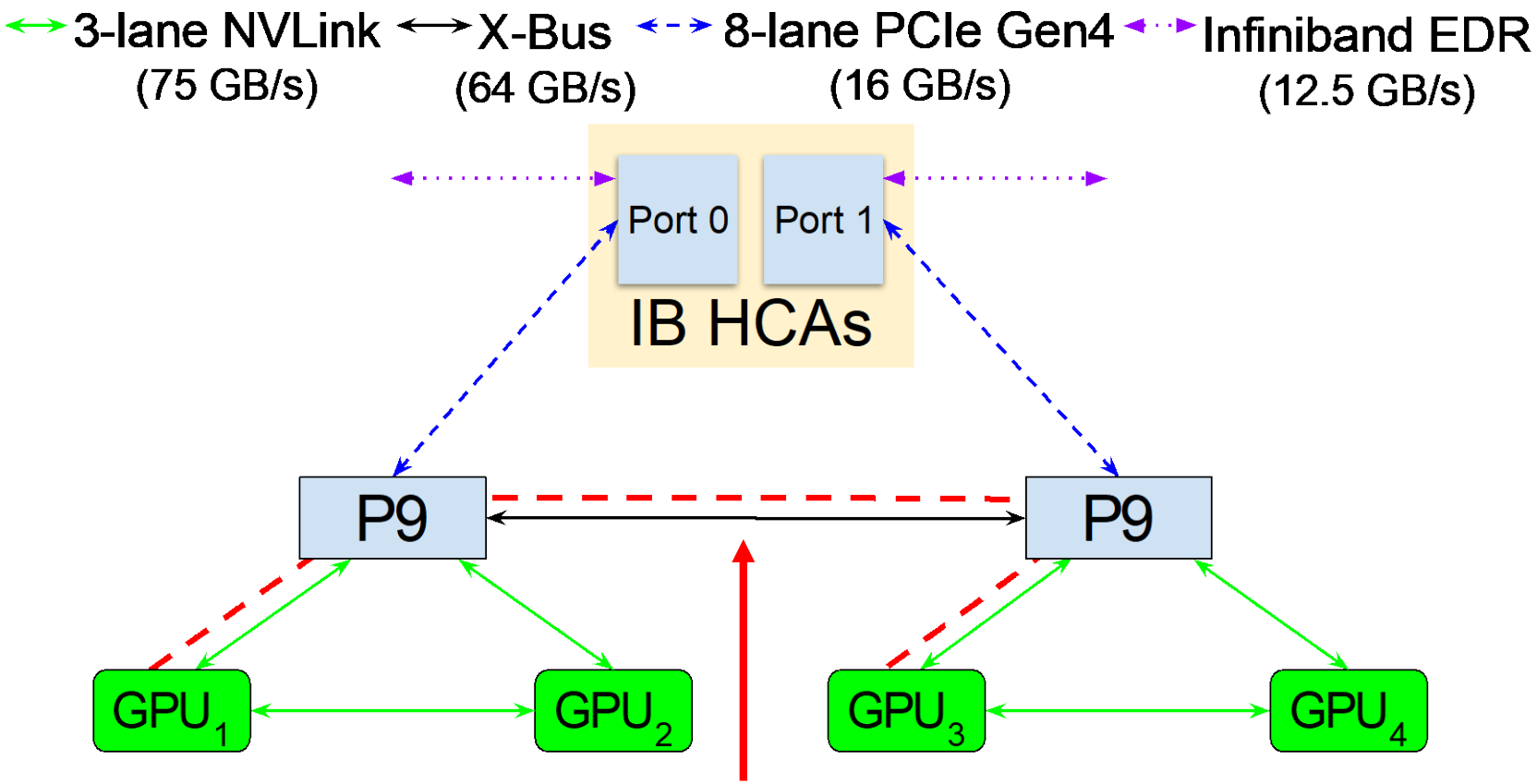
NVLink between CPU and GPU (Lassen) - Latency





- Exploit both CPU and GPU to maximize parallelism
- Place MPI processes on the same NUMA node
- MVAPICH2-GDR & OpenMPI similar small message latency up to 2KB
  - Spectrum-MPI **~10x** higher

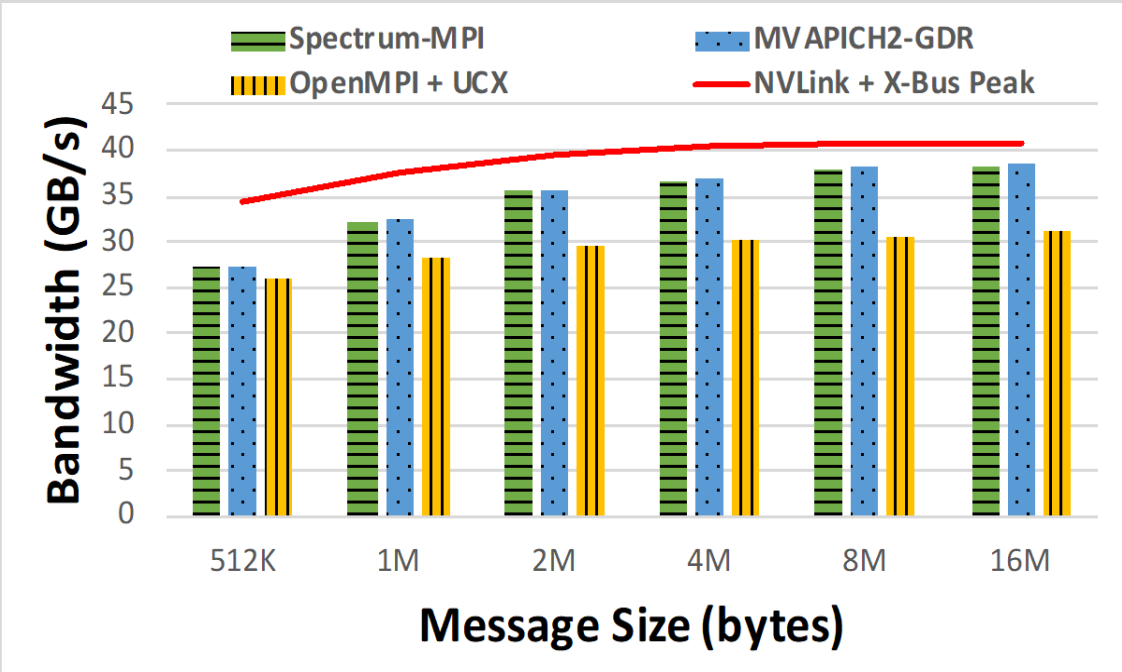
Hardware Configuration – Lassen OpenPOWER System

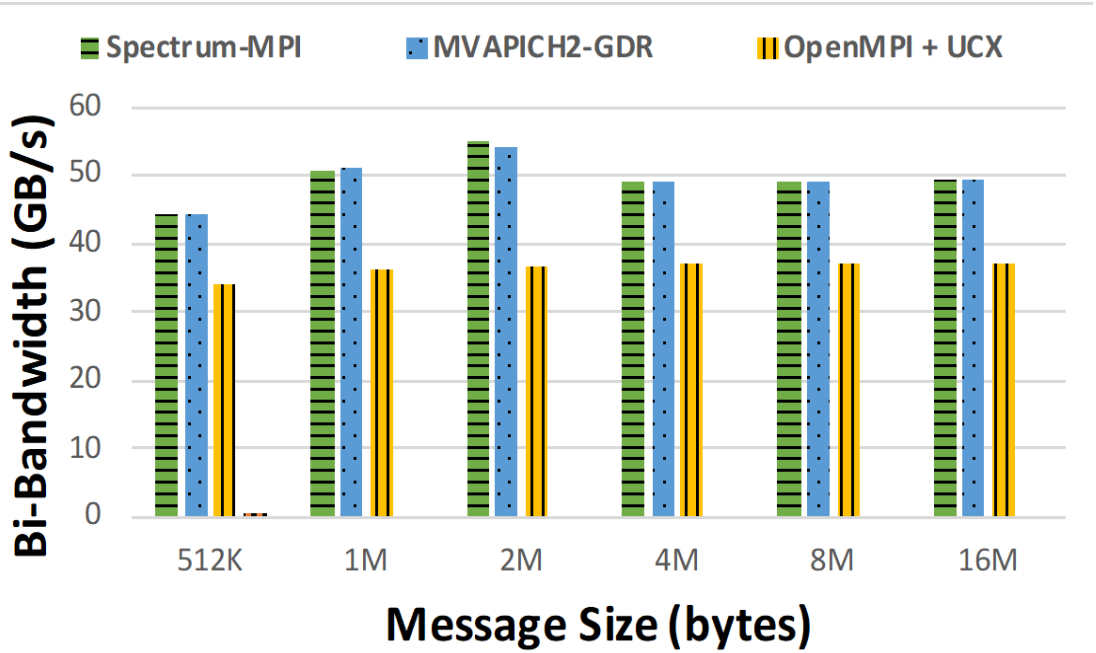


Communication through: **NVLink and X-Bus**

- Theoretical Peak Bandwidth – 64 GB/s
- Map two processes to different NUMA nodes (CUDA_VISIBLE_DEVICES=0,2)

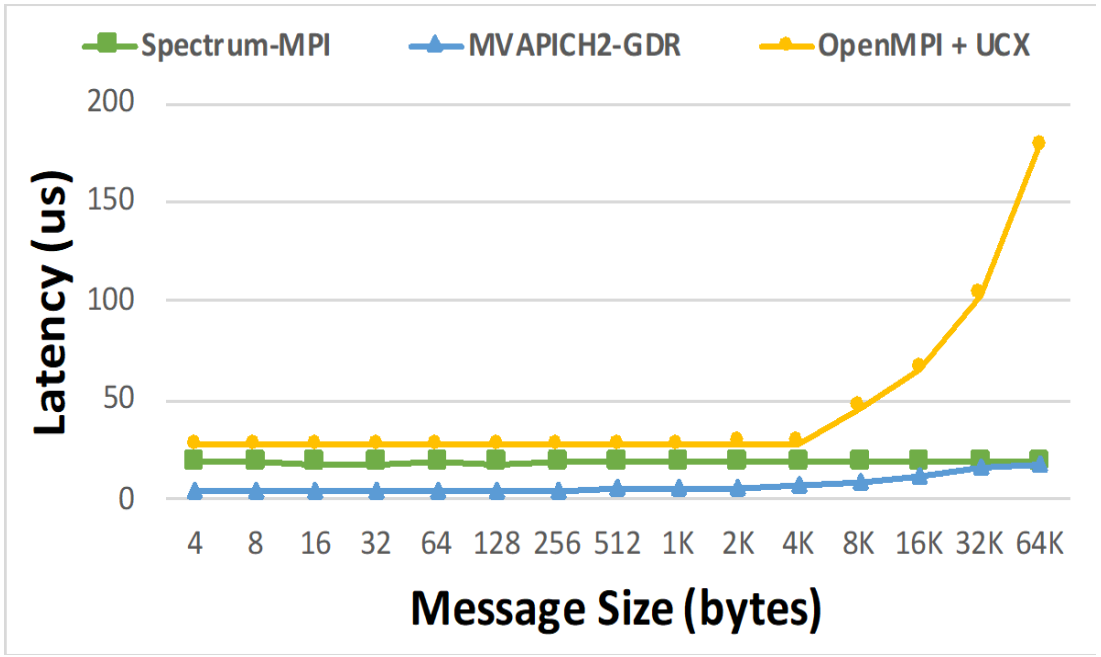
NVLink and X-Bus (Lassen) - Bandwidth

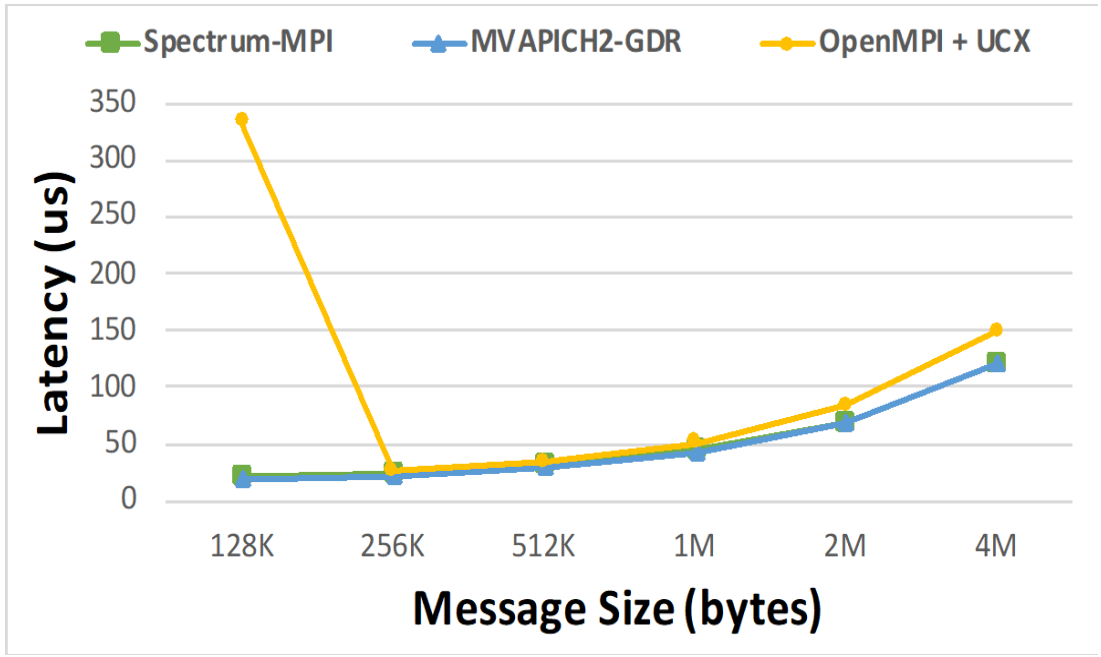




- Achievable Peak Bandwidth – 58.01 GB/s
  - Peak bandwidth MPI libraries can achieve is only around 80% of X-Bus bandwidth
- OpenMPI yields **~76%** achievable bandwidth
- Spectrum-MPI and MVAPICH2-GDR yields **~95%** achievable bandwidth

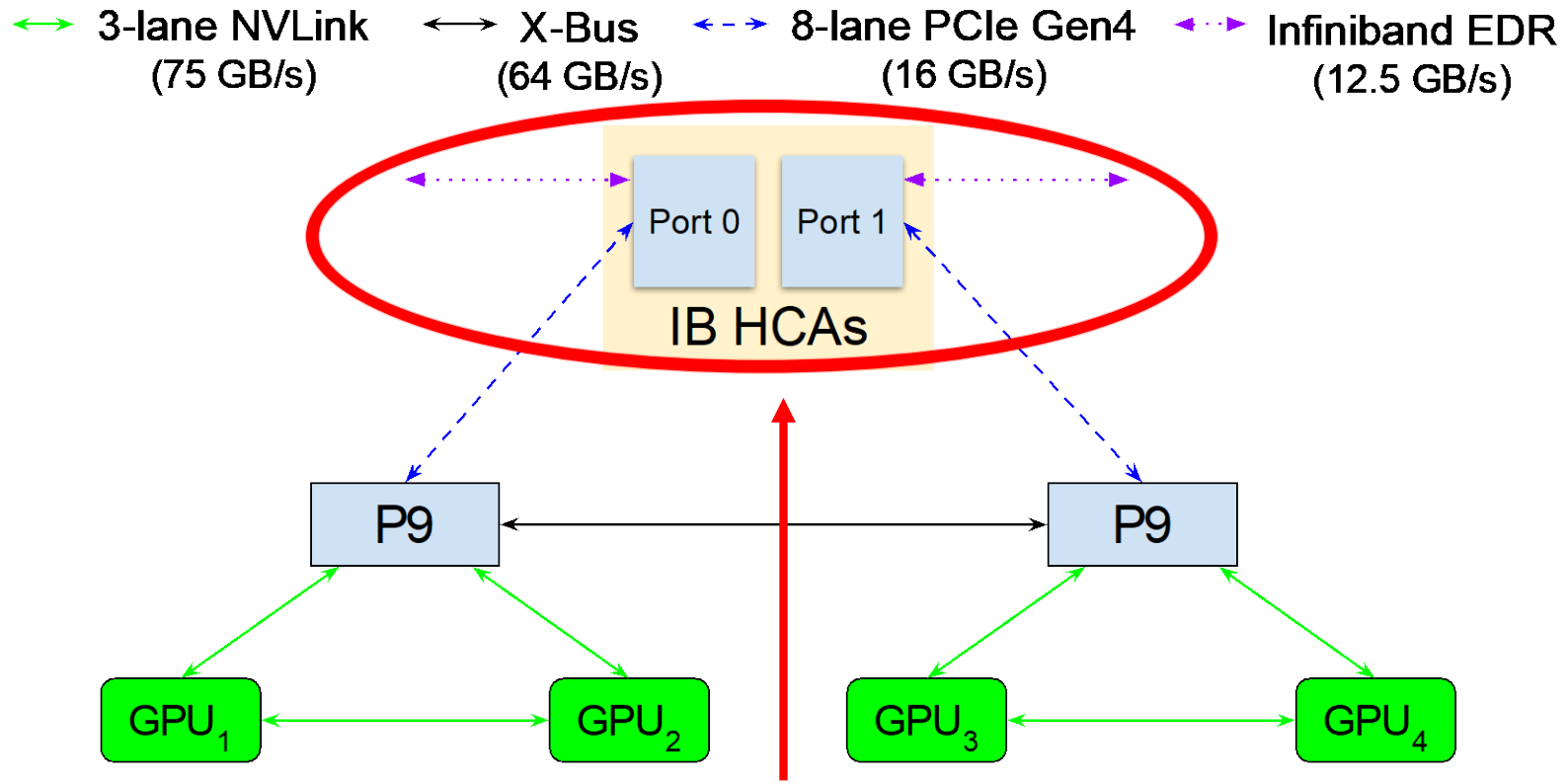
NVLink and X-Bus (Lassen) - Latency





- X-Bus dominates the performance
- OpenMPI degradation between 4KB and 256KB
  - setting the thresholds for selecting communication protocols
- MVAPICH2-GDR **~4x** lower than Spectrum-MPI & **~5x** lower than OpenMPI up to 4K

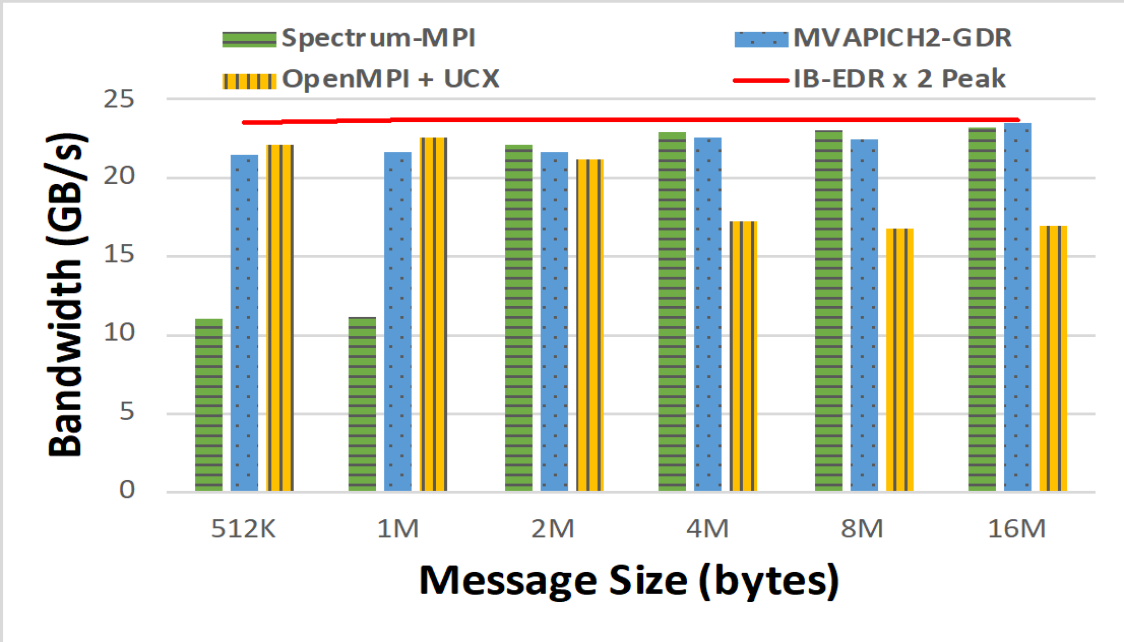
Hardware Configuration – Lassen OpenPOWER System

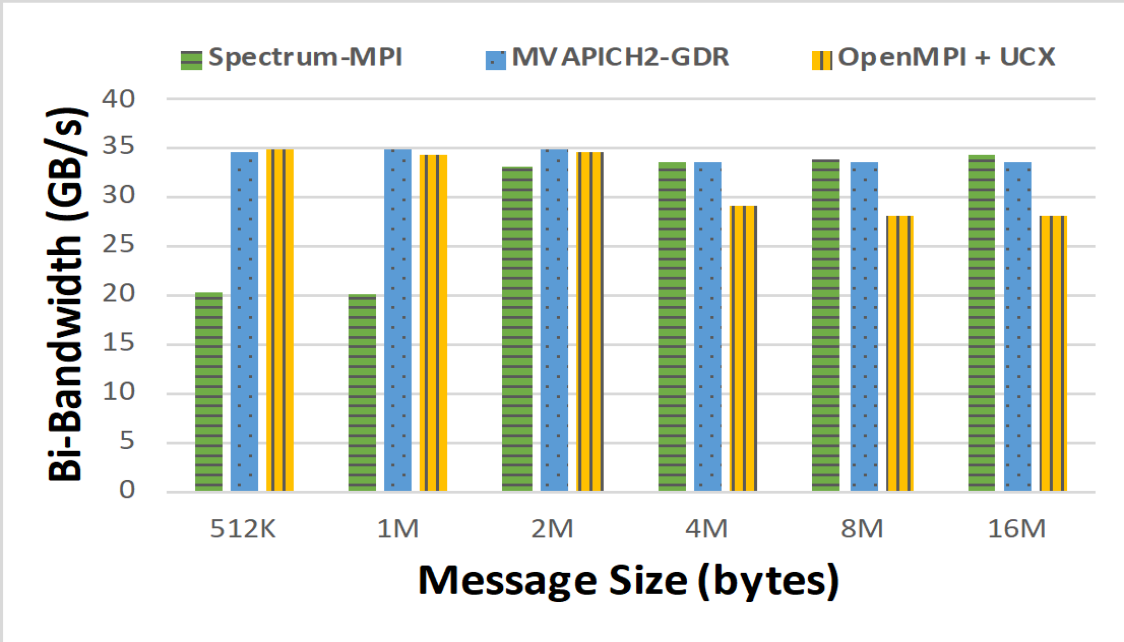


Communication through: **InfiniBand Socket-Direct Dual Port EDR Network**

- Theoretical Peak Bandwidth – 25 GB/s
- MPI processes launched on different nodes

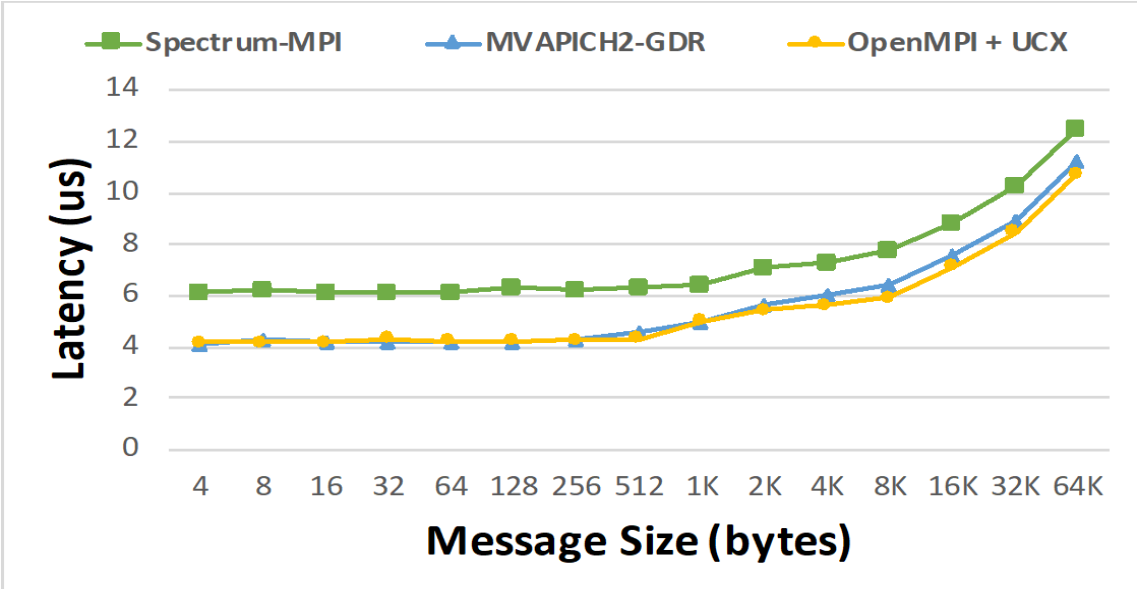
InfiniBand Network (Lassen) - Bandwidth

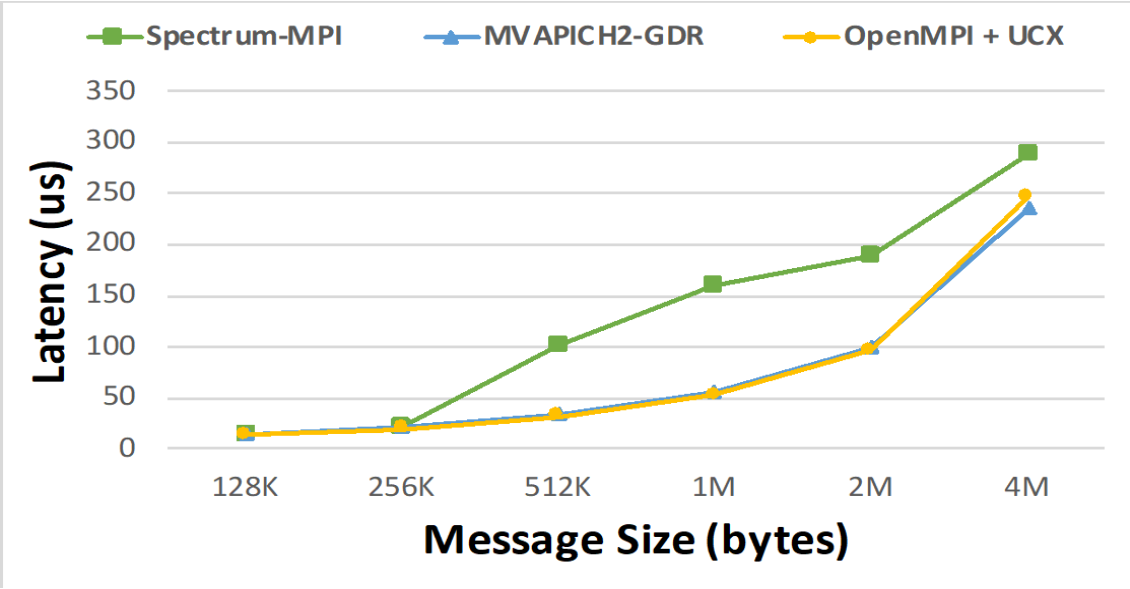




- Achievable Peak Bandwidth – 23.64 GB/s
- Multi-rail support pertinent for peak performance

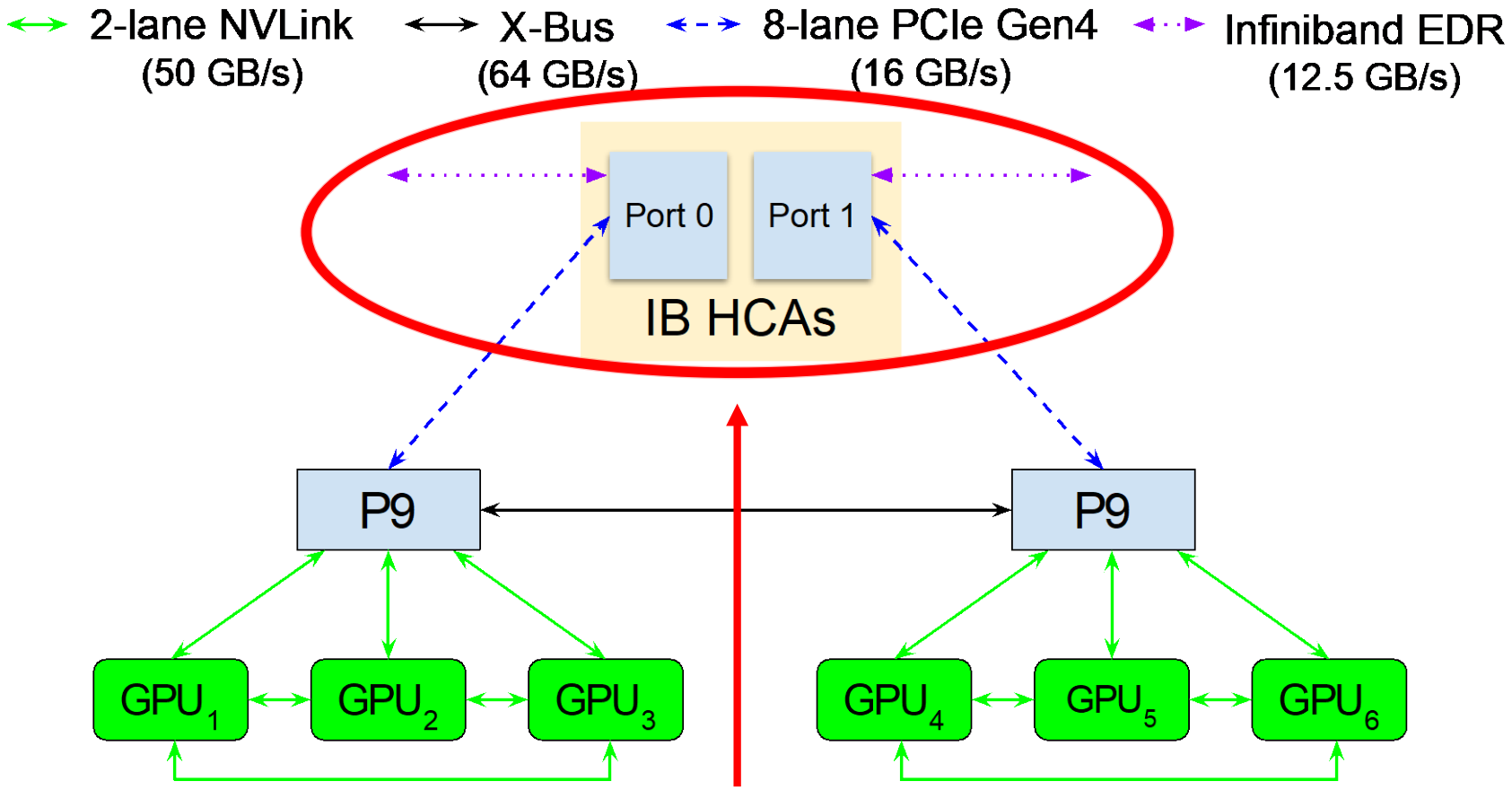
InfiniBand Network (Lassen) - Latency





- MVAPICH2-GDR & OpenMPI in similar range
- SpectrumMPI **~6us** small message latency compared to **~4us** for OpenMPI & MVAPICH2-GDR
- Possible change in communication protocol for SpectrumMPI degradation after 256KB

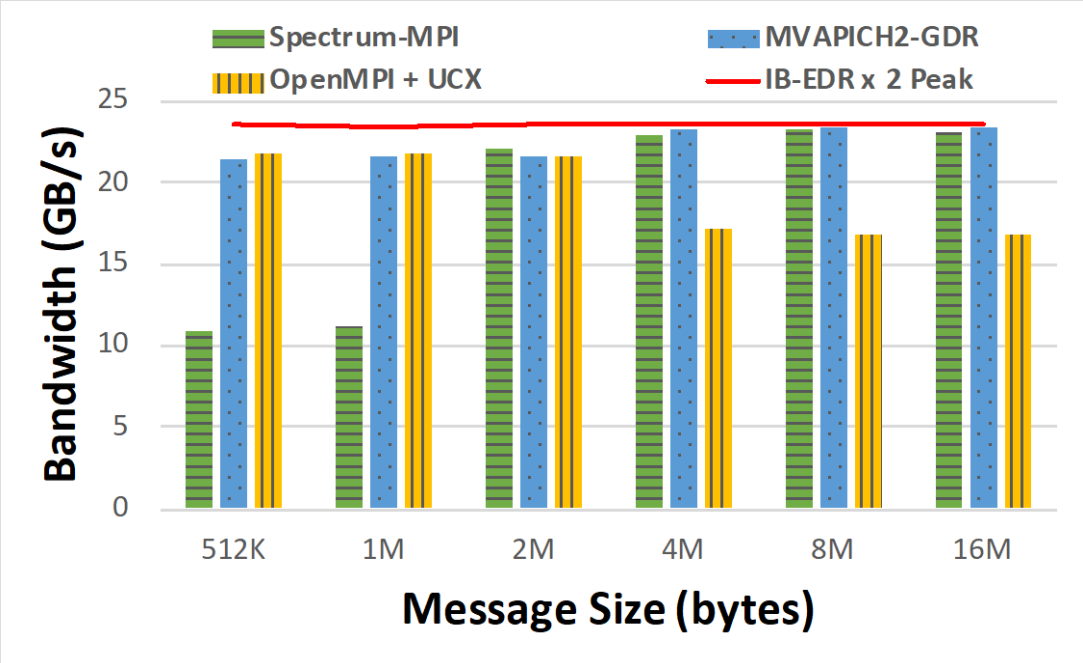
Hardware Configuration – Summit OpenPOWER System

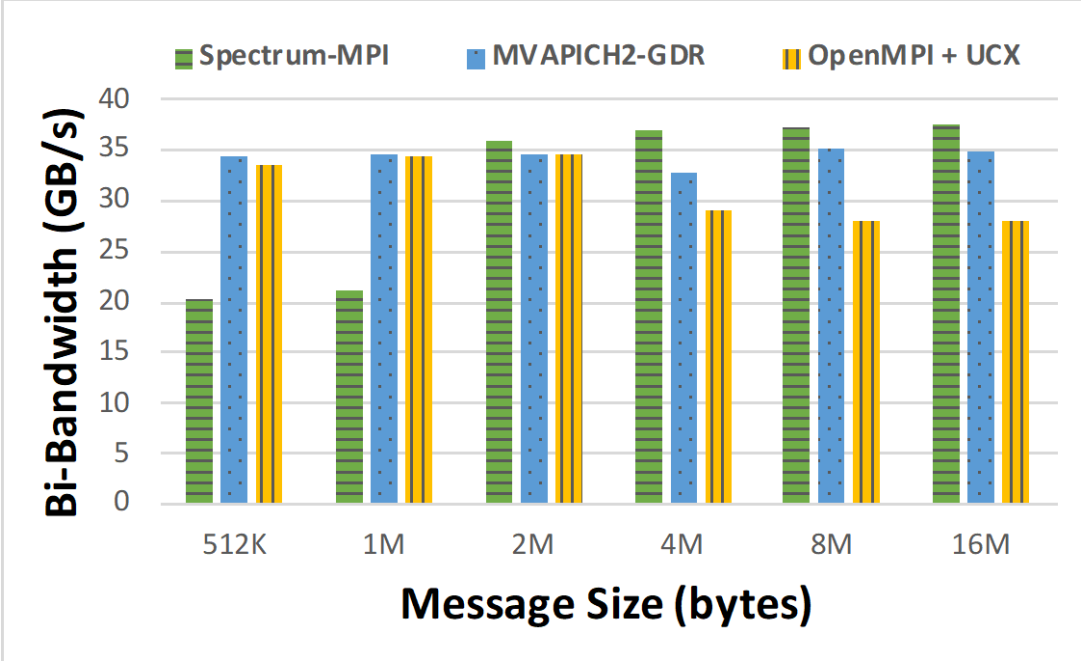


Communication through: **InfiniBand Socket-Direct Dual Port EDR Network**

- Theoretical Peak Bandwidth – 25 GB/s
- MPI processes launched on different nodes

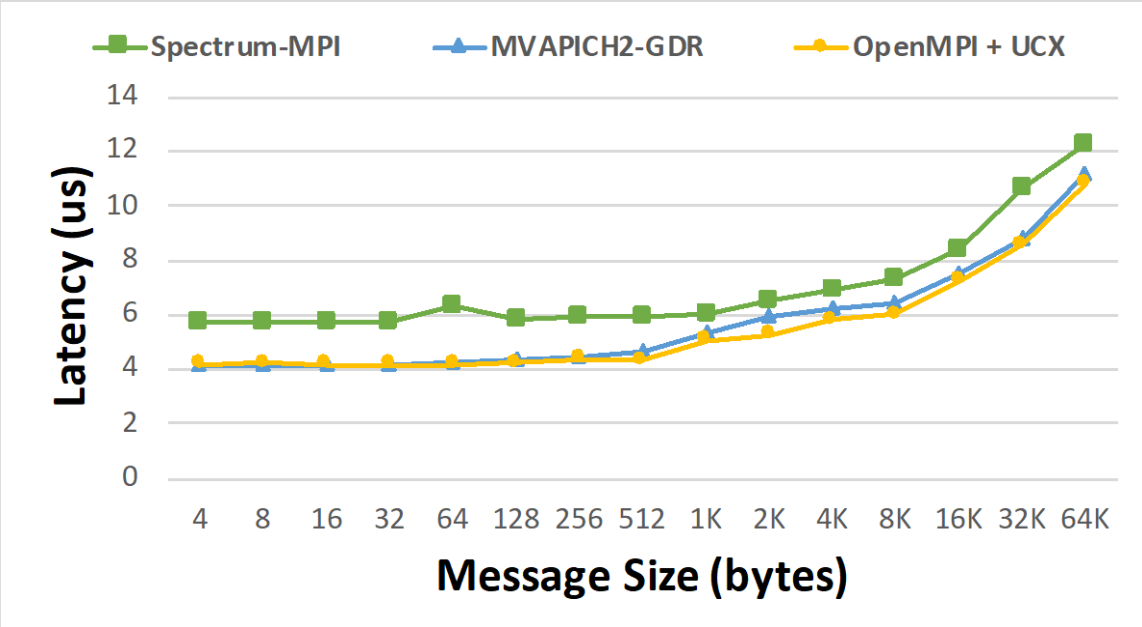
InfiniBand Network (Summit) - Bandwidth

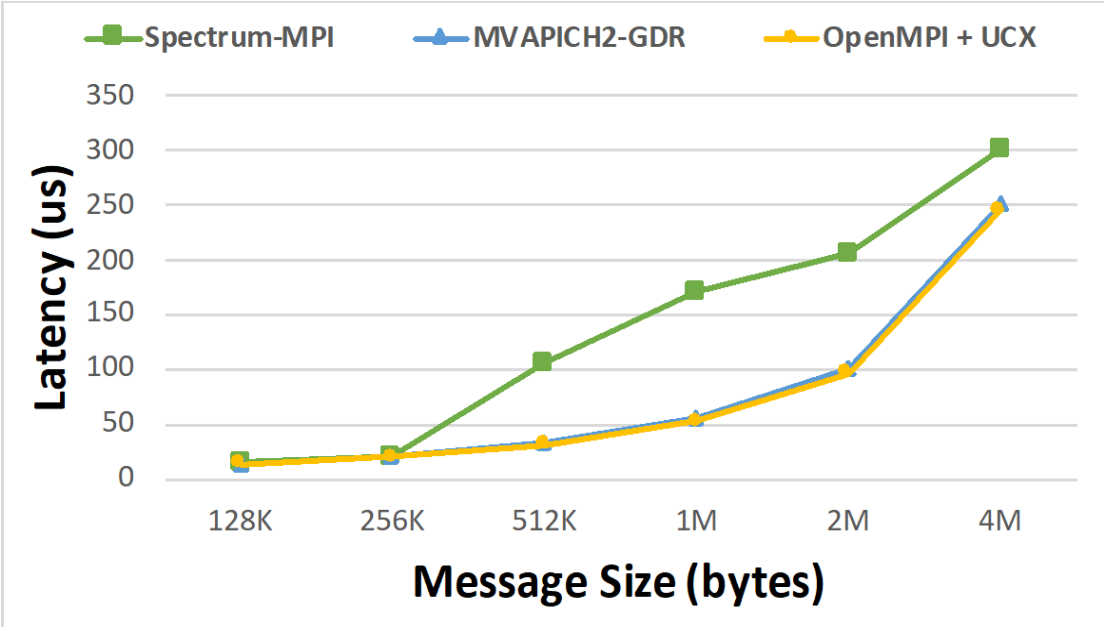




- Achievable Peak Bandwidth – 23.64 GB/s

InfiniBand Network (Summit) - Latency





- Similar Performance on Summit system to Lassen System
- SpectrumMPI **~6us** small message latency compared to **~4us** for OpenMPI & MVAPICH2-GDR

Summary - Performance of Interconnects

Achievable peak bandwidth of MPI libraries and fraction of peak over Interconnects on the Lassen GPU-enabled OpenPOWER System

- MPI Libraries perform similarly for **NVLink GPU-GPU** communications
- MVAPICH2-GDR achieves highest performance through **HBM2**
- MVAPICH2-GDR & SpectrumMPI achieve ~99% peak bandwidth for **IB EDR**

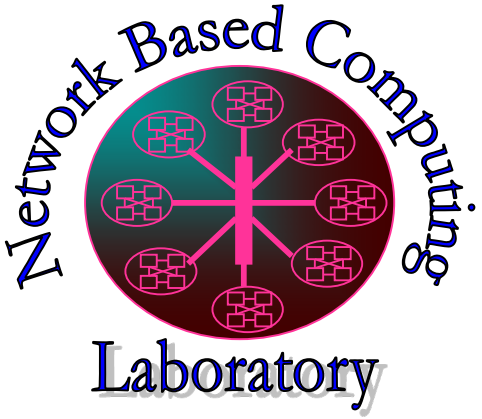
	GPU HBM2	3-lane NVLink2 CPU-GPU	3-lane NVLink2 GPU-GPU	X-Bus	InfiniBand EDR x 2
SpectrumMPI	329 GB/s	21.74 GB/s	67.14 GB/s	39.16 GB/s	23.45 GB/s
	36.55%	31.61%	95.20%	94.60%	99.20%
OpenMPI	0.457 GB/s	23.63 GB/s	67.22 GB/s	31.77 GB/s	22.55 GB/s
	0.05%	34.35%	95.40%	76.73%	95.40%
MVAPICH2-GDR	390.88 GB/s	26.84 GB/s	67.15 GB/s	39.28 GB/s	23.56 GB/s
	43.43%	39.02%	95.30%	94.97%	**99.70%**

Conclusion

Comprehensive performance evaluation of CUDA-aware MPI libraries in terms of latency, bandwidth, and bi-bandwidth on GPU-enabled OpenPOWER systems

- Communication through NVLink between two GPUs on same socket
  - MVAPICH2-GDR, Spectrum-MPI, and OpenMPI + UCX deliver ~**95%** achievable bandwidth
- Communication through IB Network achievable bandwidth:
  - MVAPICH2-GDR ~**99%,** Spectrum-MPI ~**99%**, and Open-MPI + UCX ~**95%**

Thank You!



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

Kawthar Shafie Khorassani
shafiekhorassani.1@osu.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/