



Experience with MVAPICH2 on Mellanox Hardware

Gilad Shainer

MUG 2013



Leading Supplier of End-to-End Interconnect Solutions



Comprehensive End-to-End Software Accelerators and Managment

MPI

SHMEM

PGAS

MXM
Mellanox Messaging Acceleration

FCA
Fabric Collectives Acceleration

Management

UFM
Unified Fabric Management

Storage and Data

VSA
Storage Accelerator (iSCSI)

UDA
Unstructured Data Accelerator

Comprehensive End-to-End InfiniBand and Ethernet Solutions Portfolio

ICs

Adapter Cards

Switches/Gateways

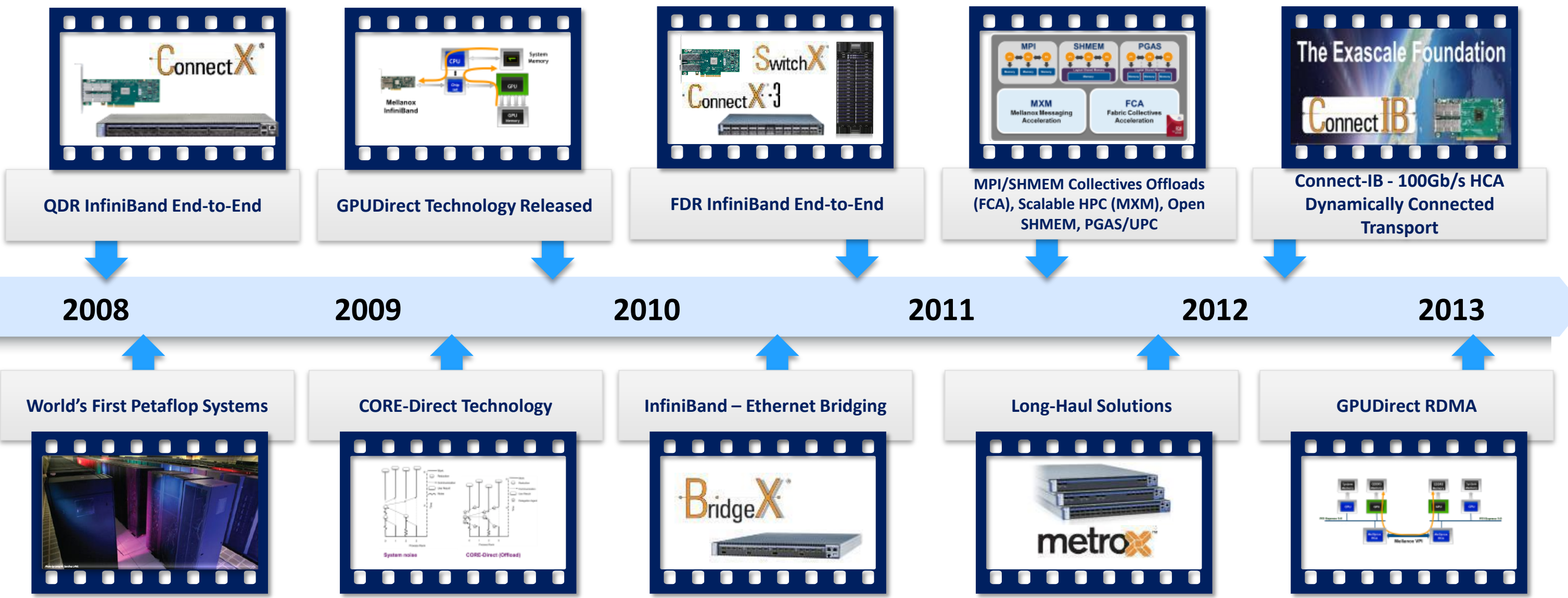
Long-Haul Systems

metroX™

Cables/Module



Mellanox Interconnect Development Timeline



Technology and Solutions Leadership, EDR InfiniBand Expected in 2014-2015

Mellanox InfiniBand Paves the Road to Exascale



Connect-IB

Architectural Foundation for Exascale Computing

- World's first 100Gb/s interconnect adapter
 - PCIe 3.0 x16, dual FDR 56Gb/s InfiniBand ports to provide >100Gb/s
- Highest InfiniBand message rate: 137 million messages per second
 - 4X higher than other InfiniBand solutions
- <0.7 micro-second application latency
- Supports GPUDirect RDMA for direct GPU-to-GPU communication
- New Innovative Transport – Dynamically Connected Transport Service
- Supports Scalable HPC with MPI, SHMEM and PGAS/UPC offloads



Connect[™]IB

Enter the World of Boundless Performance

GPUDirect

Receive

Transmit

Non GPUDirect

InfiniBand

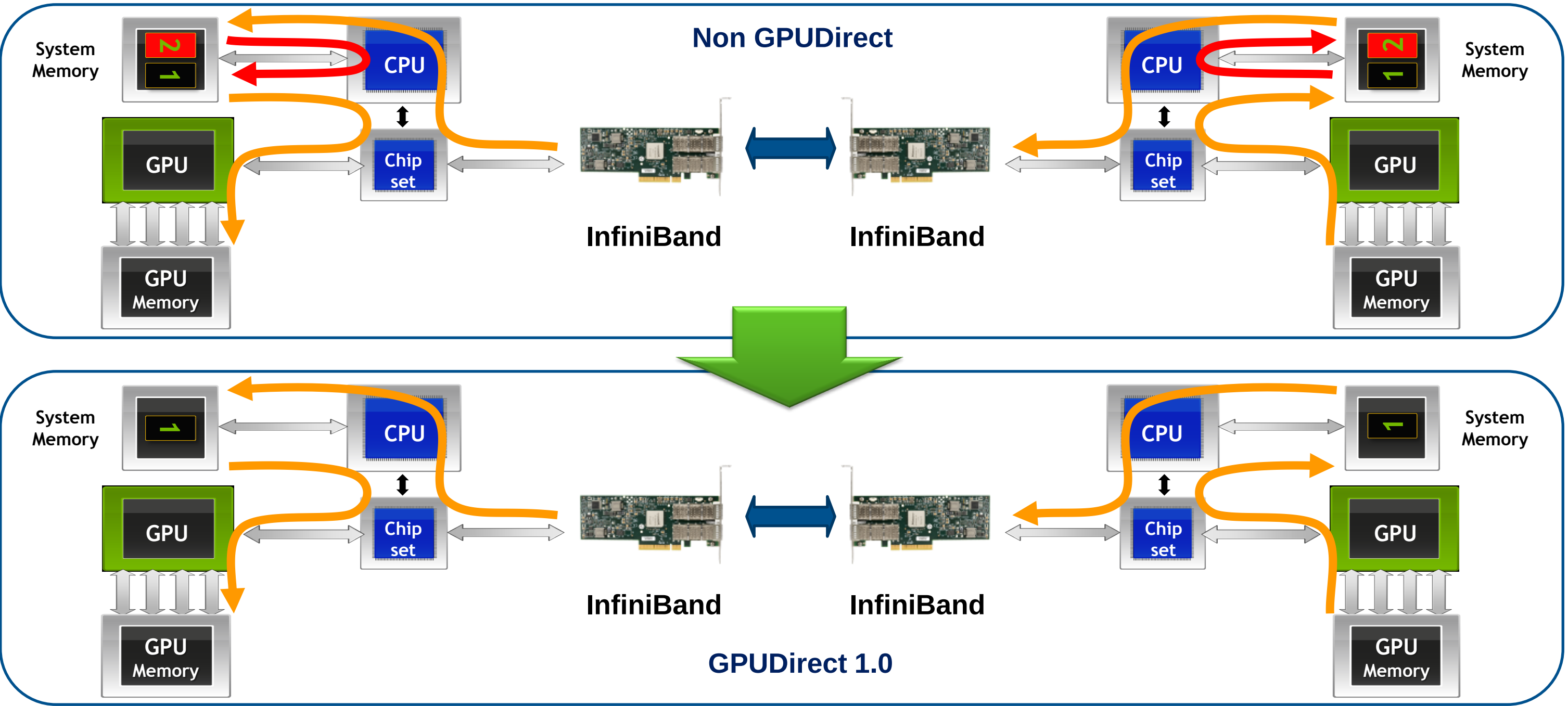
InfiniBand



GPUDirect 1.0

InfiniBand

InfiniBand



Receive

Transmit

GPUDirect 1.0

InfiniBand

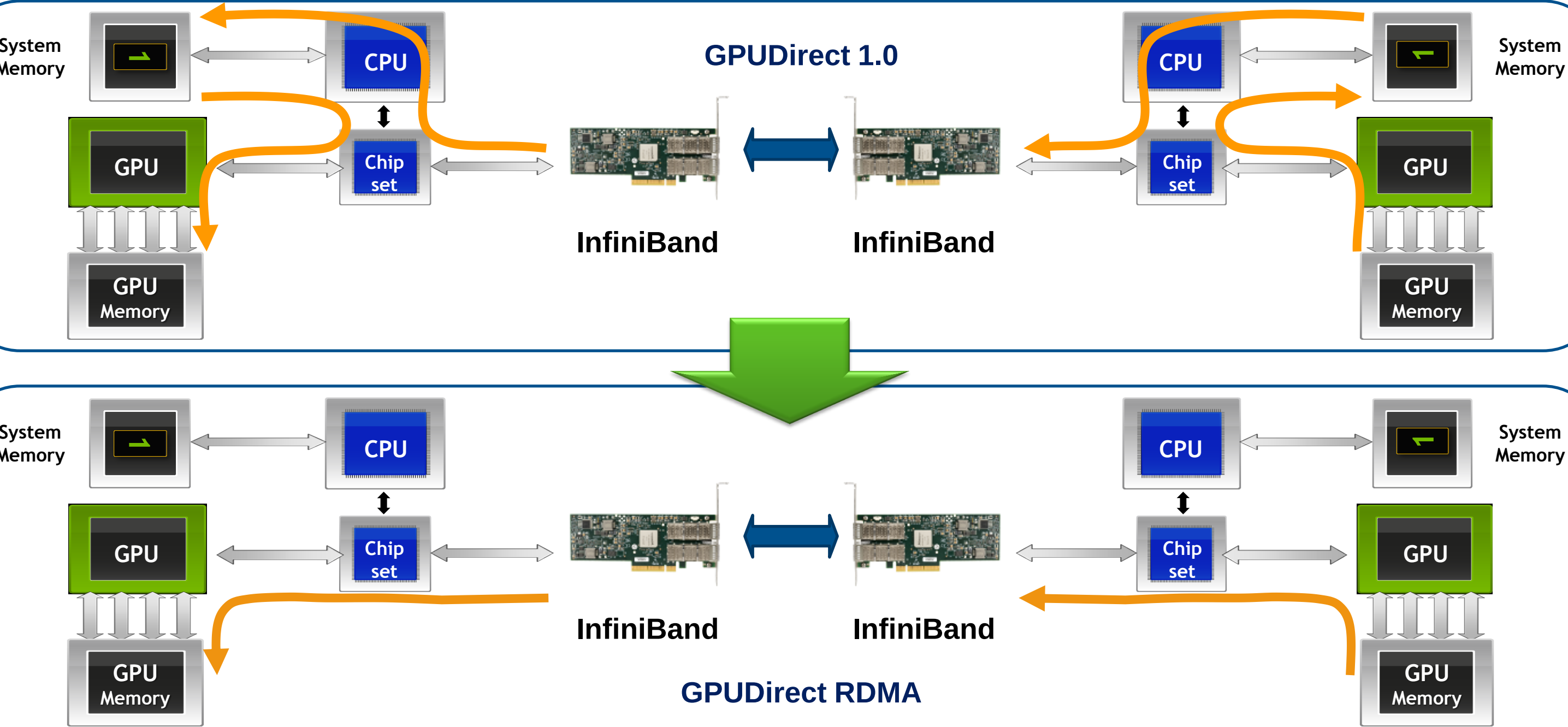
InfiniBand



GPUDirect RDMA

InfiniBand

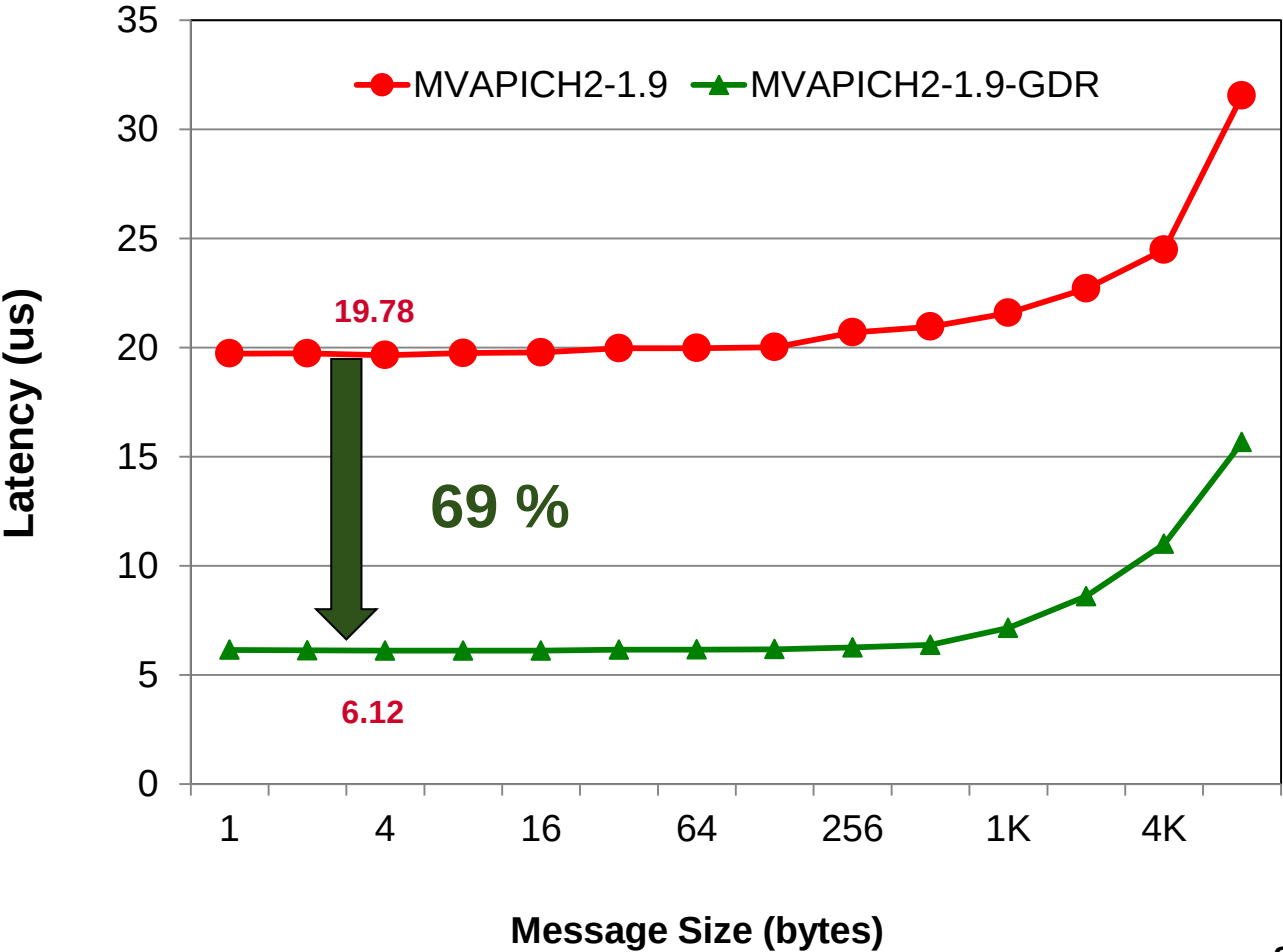
InfiniBand





GPU-GPU Internode MPI Latency

Small Message Latency



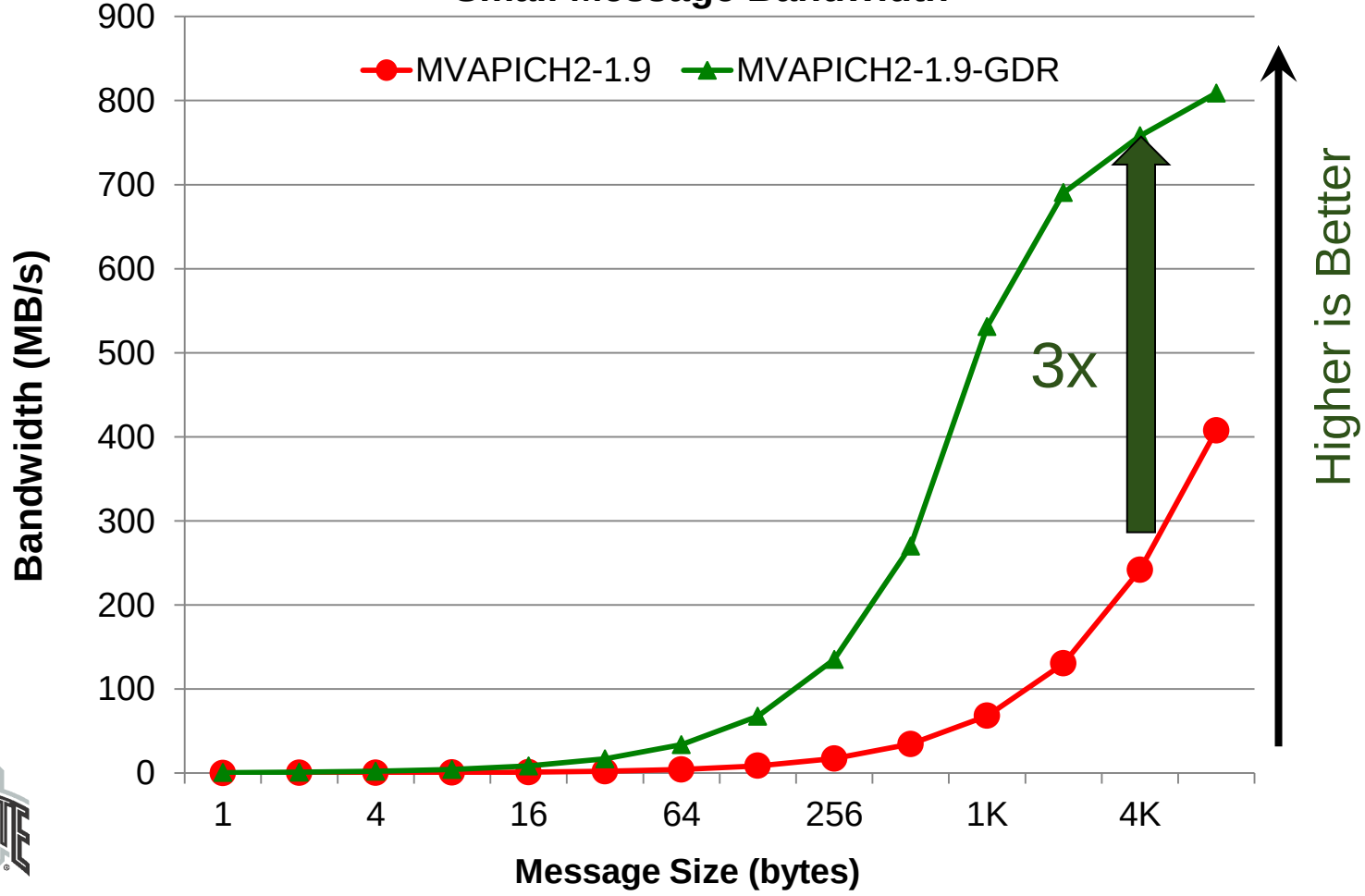
Lower is Better



Source: Prof. DK Panda

GPU-GPU Internode MPI Bandwidth

Small Message Bandwidth



Higher is Better

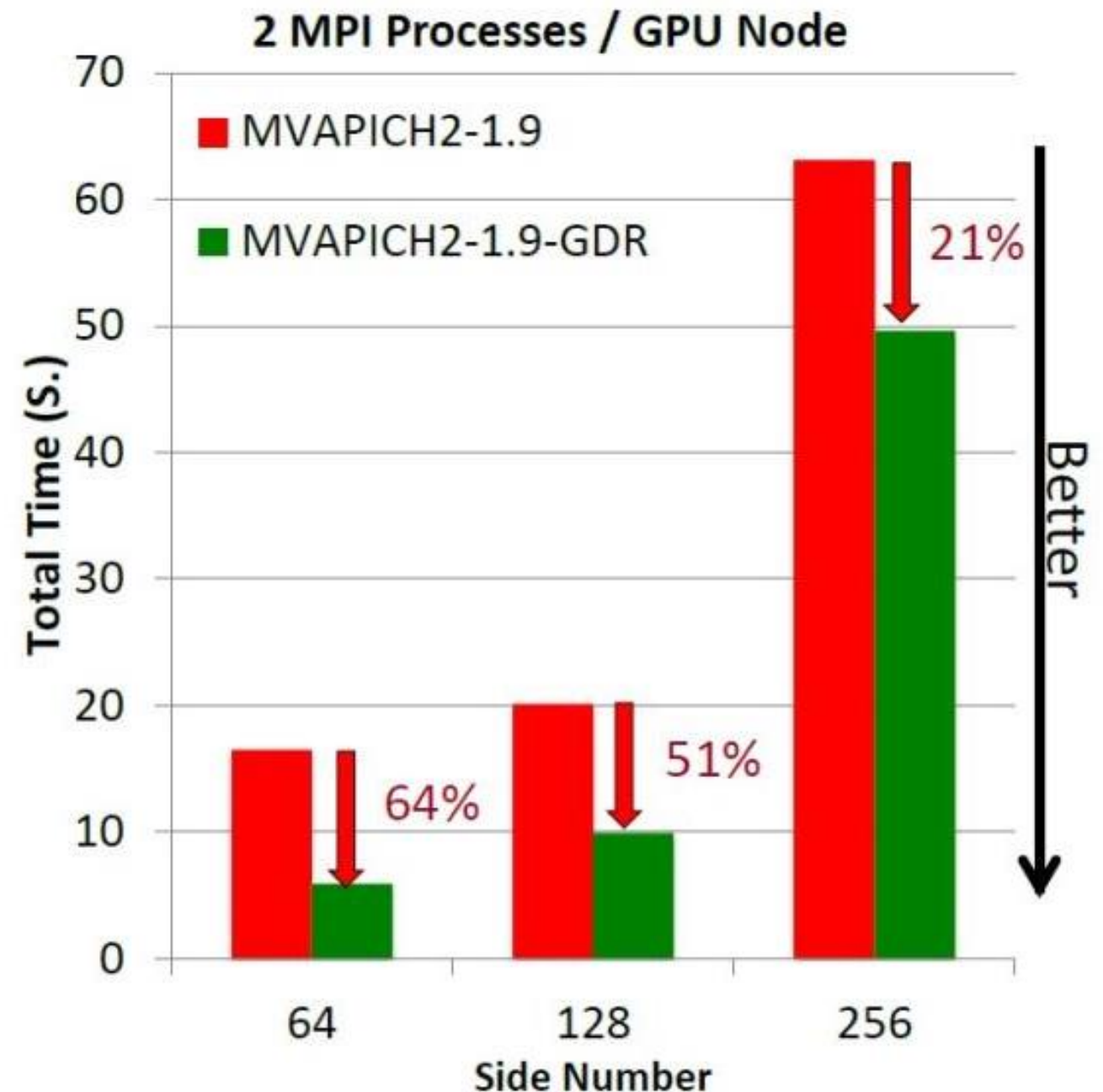
69% Lower Latency

3X Increase in Throughput

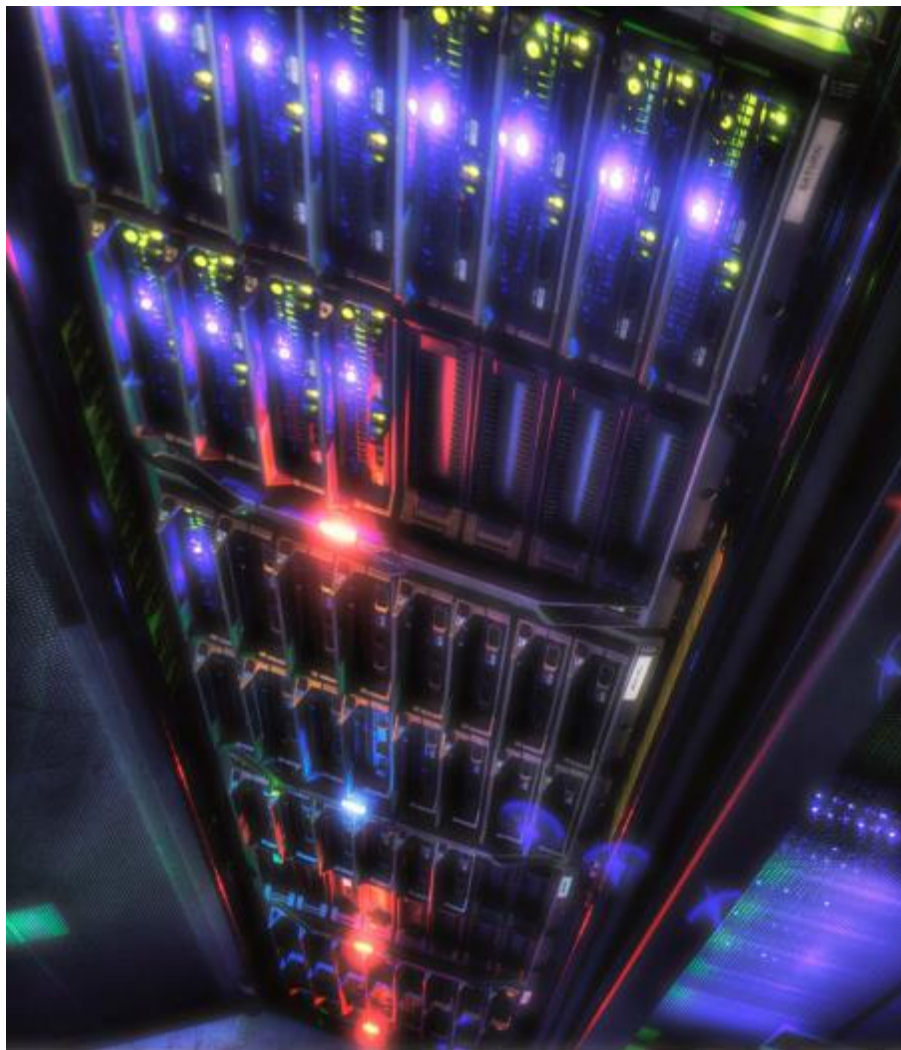
Execution Time of HSG (Heisenberg Spin Glass) Application with 2 GPU Nodes



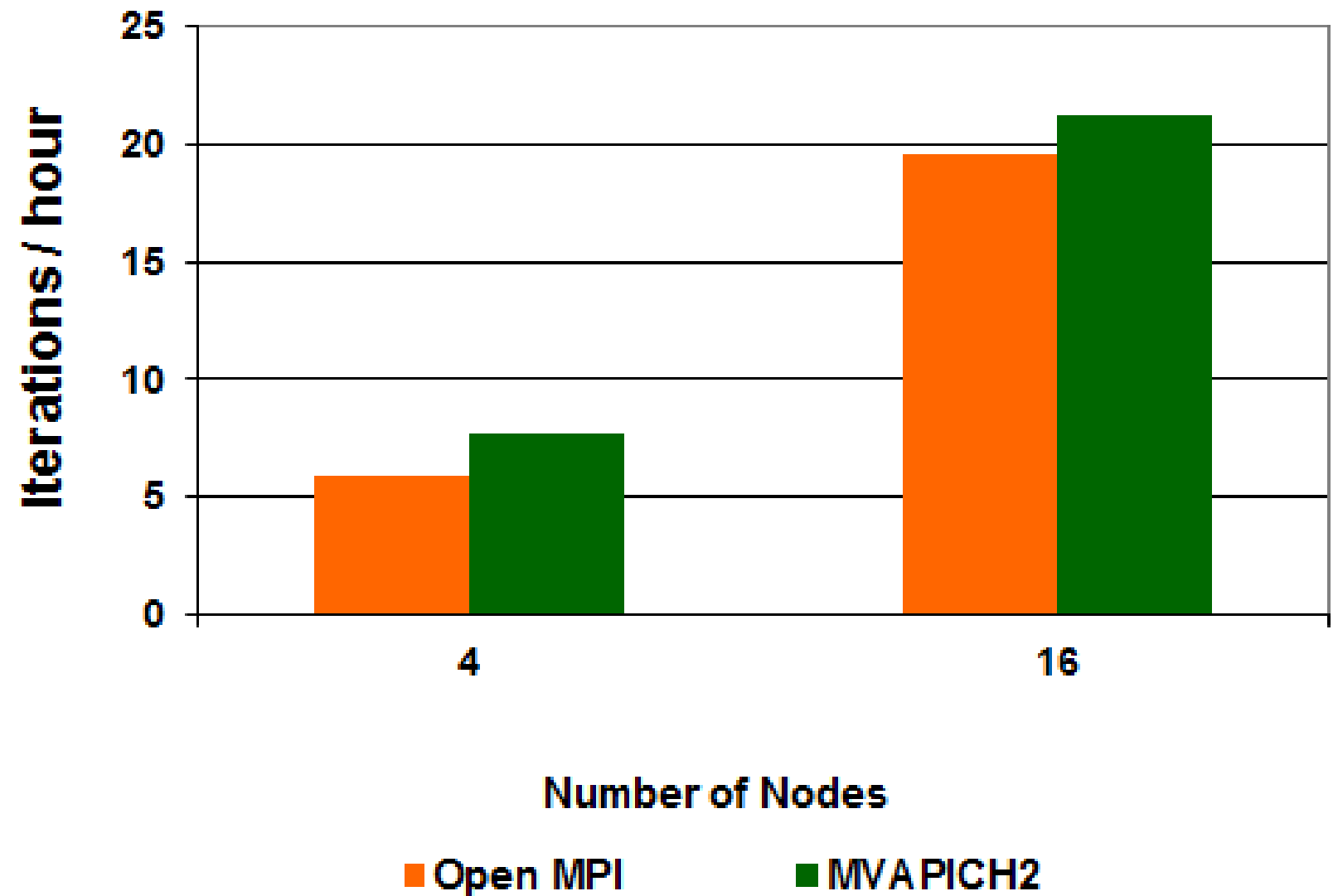
Source: Prof. DK Panda



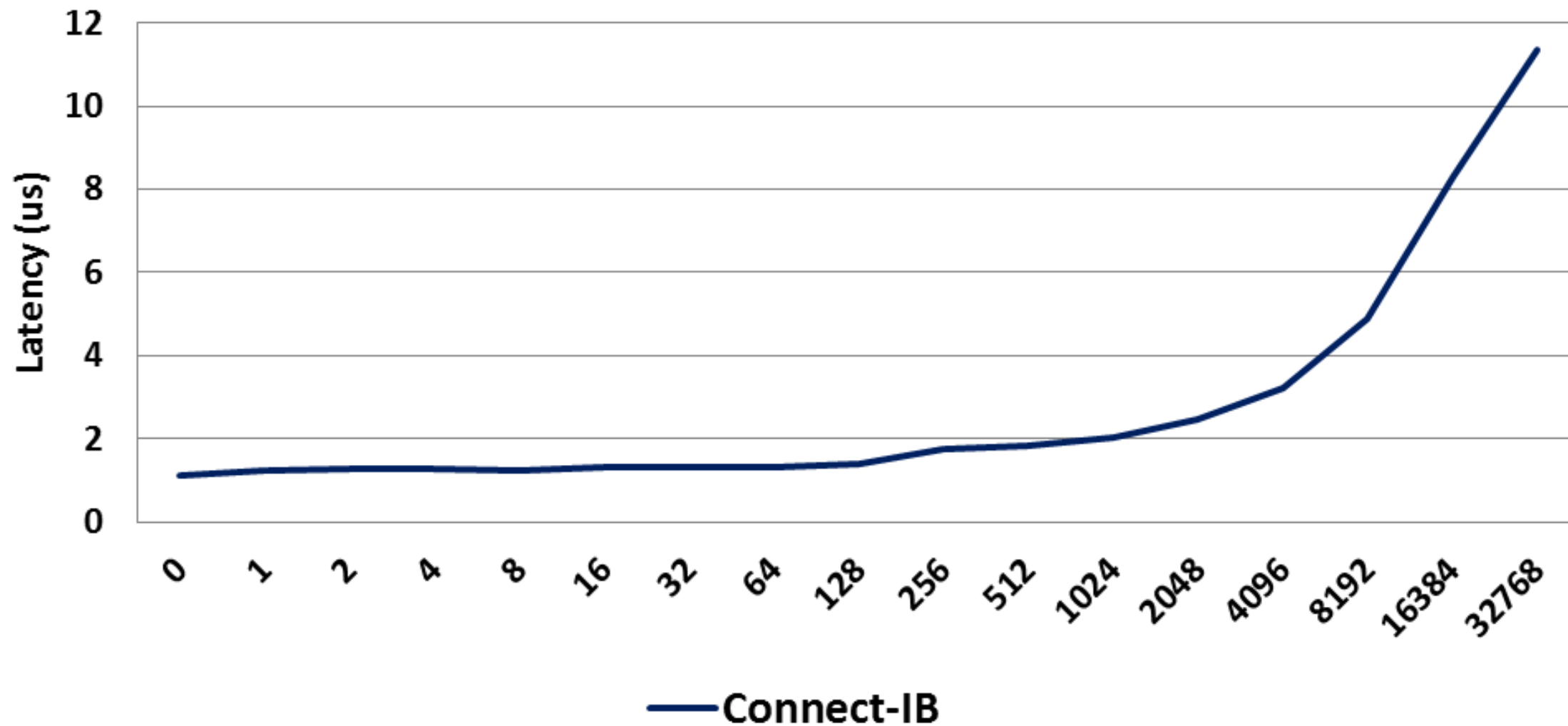
Some Performance Data



Quantum ESPRESSO Benchmark (AUSURF112)

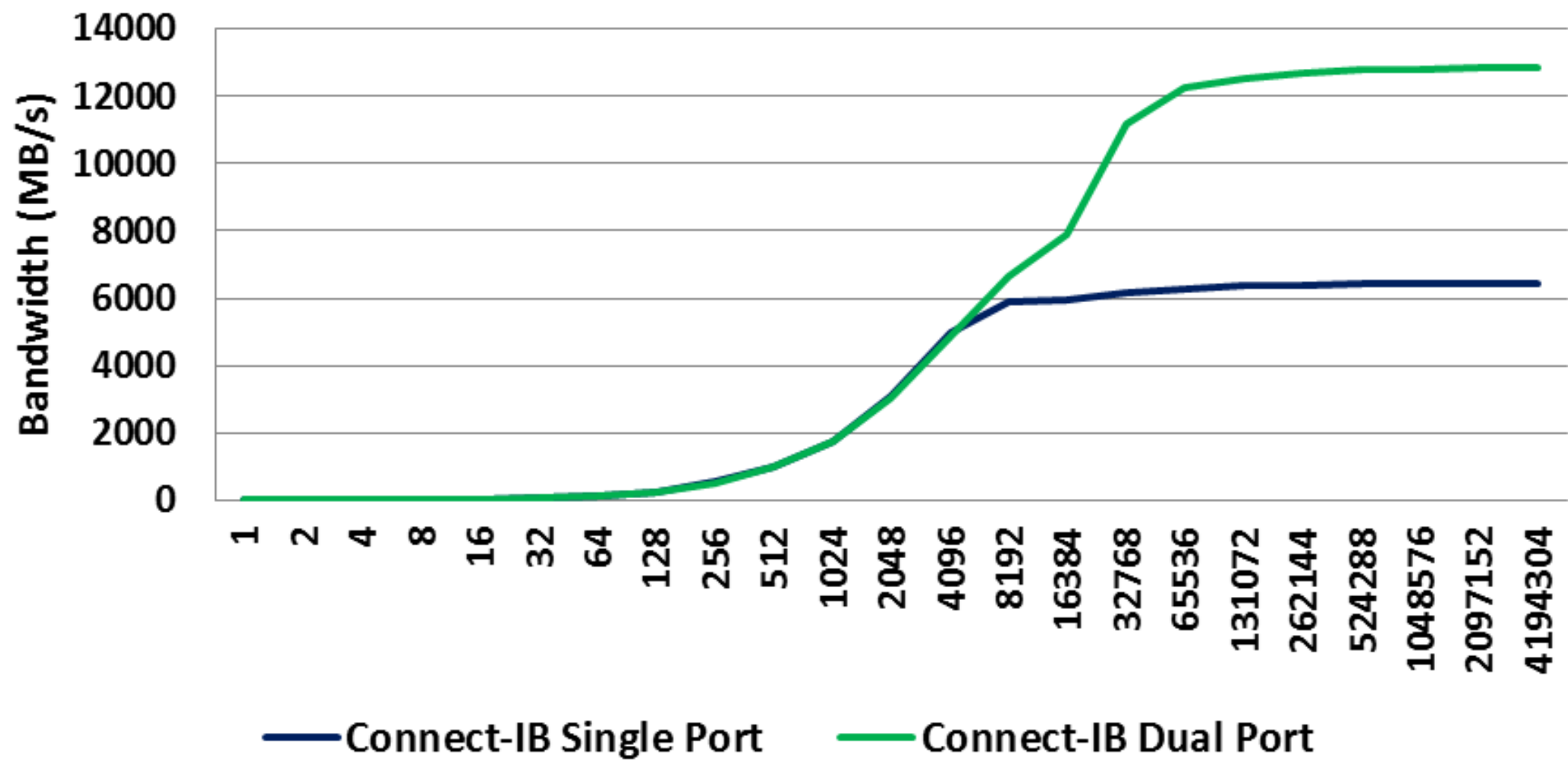


MVAPICH2 1.9 Performance (osu_latency)

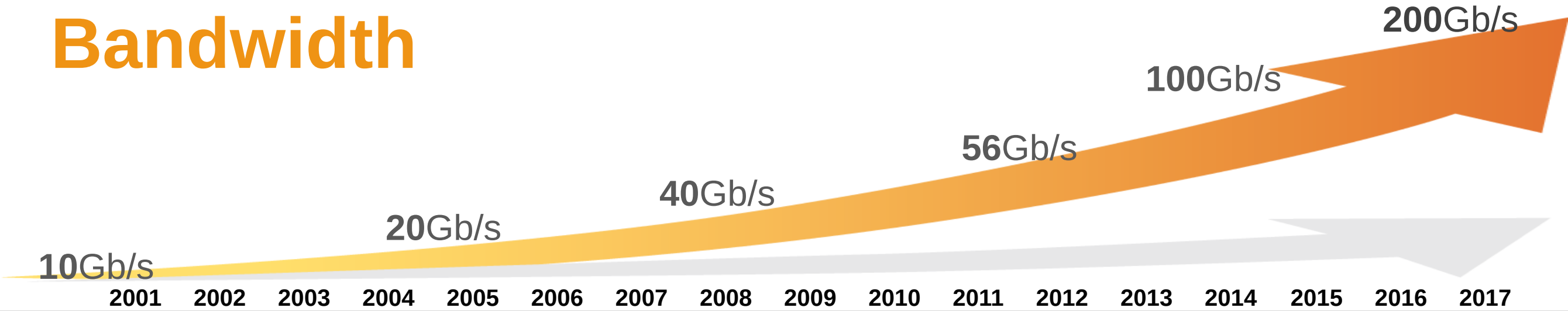




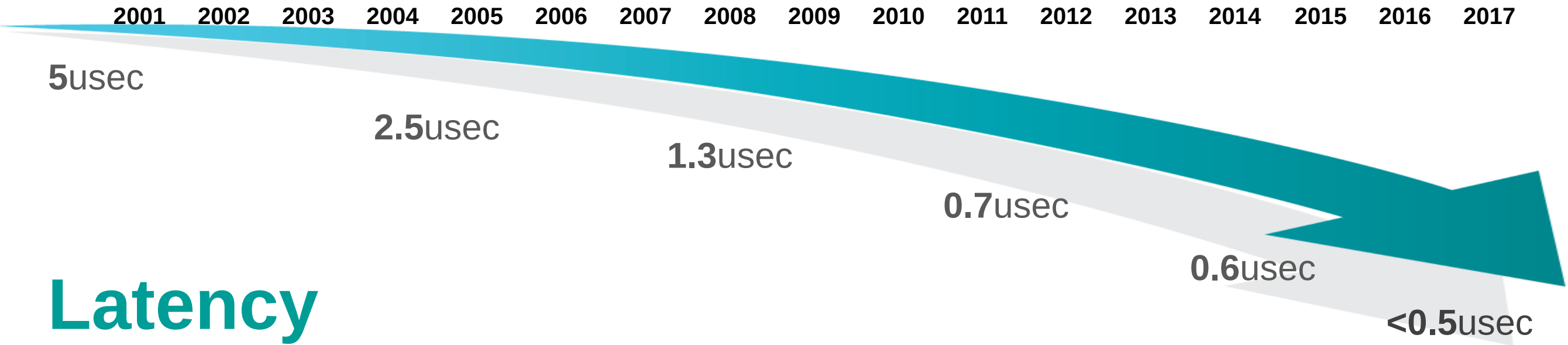
MVAPICH2 1.9 Performance
(osu_bw)



Bandwidth

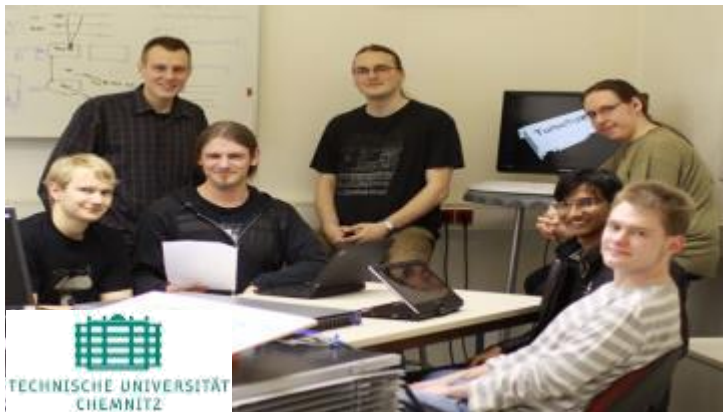


Same Software Interface



Latency

ISC'13 – Student Cluster Challenge Teams



ISC'13 Student Cluster Challenge





Submission for Participation is Open

Thank You

