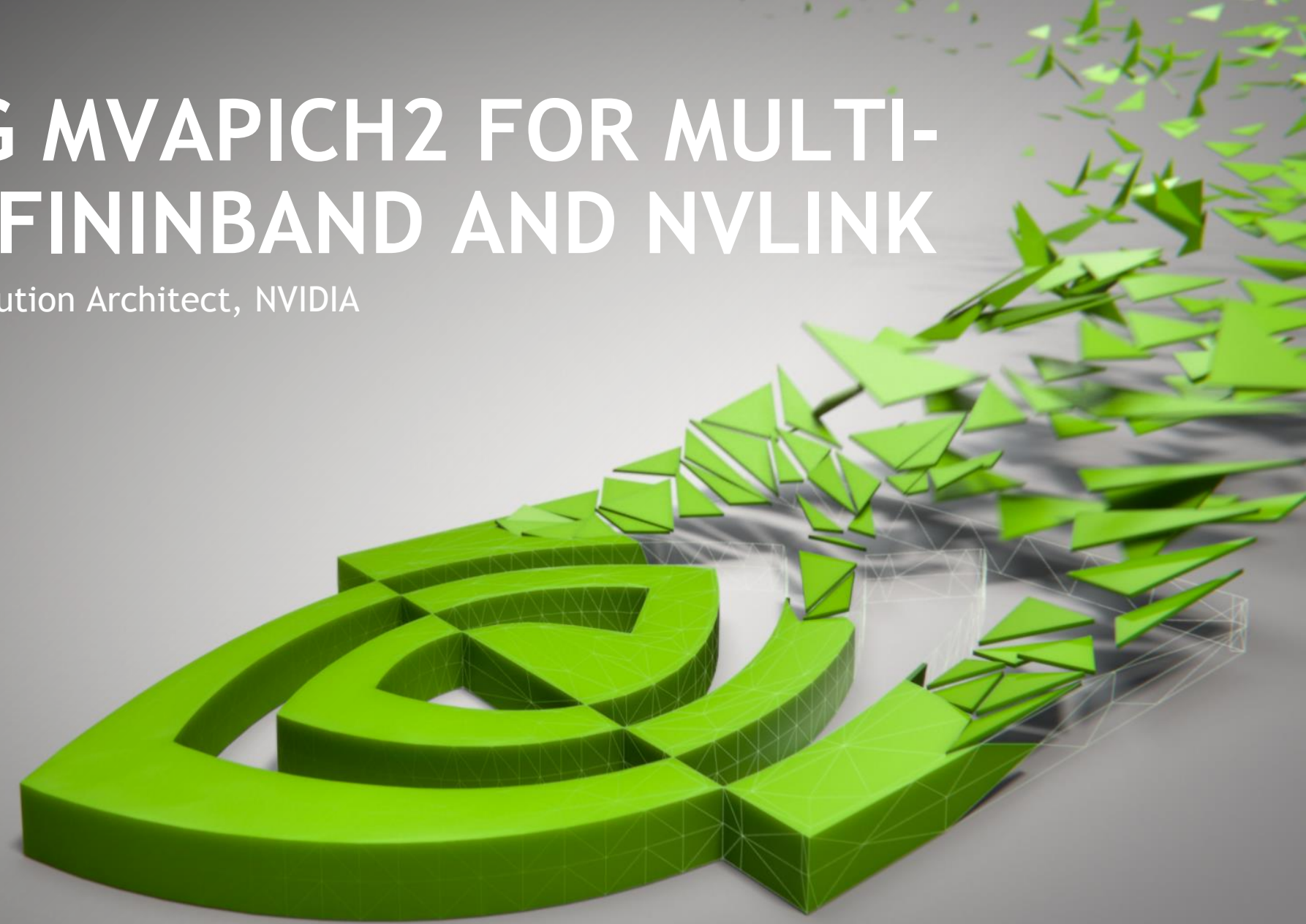
TUNING MVAPICH2 FOR MULTI-RAIL INFININBAND AND NVLINK

Craig Tierney, Solution Architect, NVIDIA

August 15, 2017

NVIDIA



Agenda

- **Overview**
- Multi-rail
- Intranode performance with NVLink

GPUDIRECT FAMILY[1]

Technologies, enabling products !!!

GPUDIRECT SHARED GPU-SYSMEM

GPU pinned memory shared with other RDMA capable devices
Avoids intermediate copies

GPUDIRECT P2P

Accelerated GPU-GPU memory copies
Inter-GPU direct load/store access

GPUDIRECT RDMA[2]

Direct GPU to 3rd party device transfers
E.g. direct I/O, optimized inter-node communication

GPUDIRECT ASYNC

Direct GPU to 3rd party device synchronizations
E.g. optimized inter-node communication

https://developer.nvidia.com/gpudirect

http://docs.nvidia.com/cuda/gpudirect-rdma

GPUDIRECT FAMILY[1]

Technologies, enabling products !!!

GPUDIRECT SHARED GPU-SYSMEM

GPU pinned memory shared with other RDMA capable devices
Avoids intermediate copies

GPUDIRECT P2P

Accelerated GPU-GPU memory copies
Inter-GPU direct load/store access

Each platform has different performance characteristics, unique tuning required

GPUDIRECT RDMA[2]

Direct GPU to 3rd party device transfers
E.g. direct I/O, optimized inter-node communication

GPUDIRECT ASYNC

Direct GPU to 3rd party device synchronizations
E.g. optimized inter-node communication

https://developer.nvidia.com/gpudirect

http://docs.nvidia.com/cuda/gpudirect-rdma

OVERVIEW

- GPUs provide significantly more performance than CPUs for many HPC workloads

Application	Domain	CPU Node Equivalence
HOOMD-Blue	Particle Dynamics	31
SPECFEM 3D	Seismic Wave Propagation	78
VASP	Quantum Chemistry	24
LSMS	Materials Code	36
QUDA	Lattice Quantum Chromo Dynamics	71

Number of nodes necessary to achieve same performance of 8-way P100 server.
http://www.nvidia.com/object/application-performance-guide.html

OVERVIEW

- GPUs provide significantly more performance than CPUs for many HPC workloads

Application	Domain	CPU Node Equivalence
HOOMD-Blue	Particle Dynamics	32
SPECFEM 3D	Seismic Wave Propagation	78
VASP	Quantum Chemistry	24
LSMS	Materials Code	36
QUDA	Lattice Quantum Chromo Dynamics	71

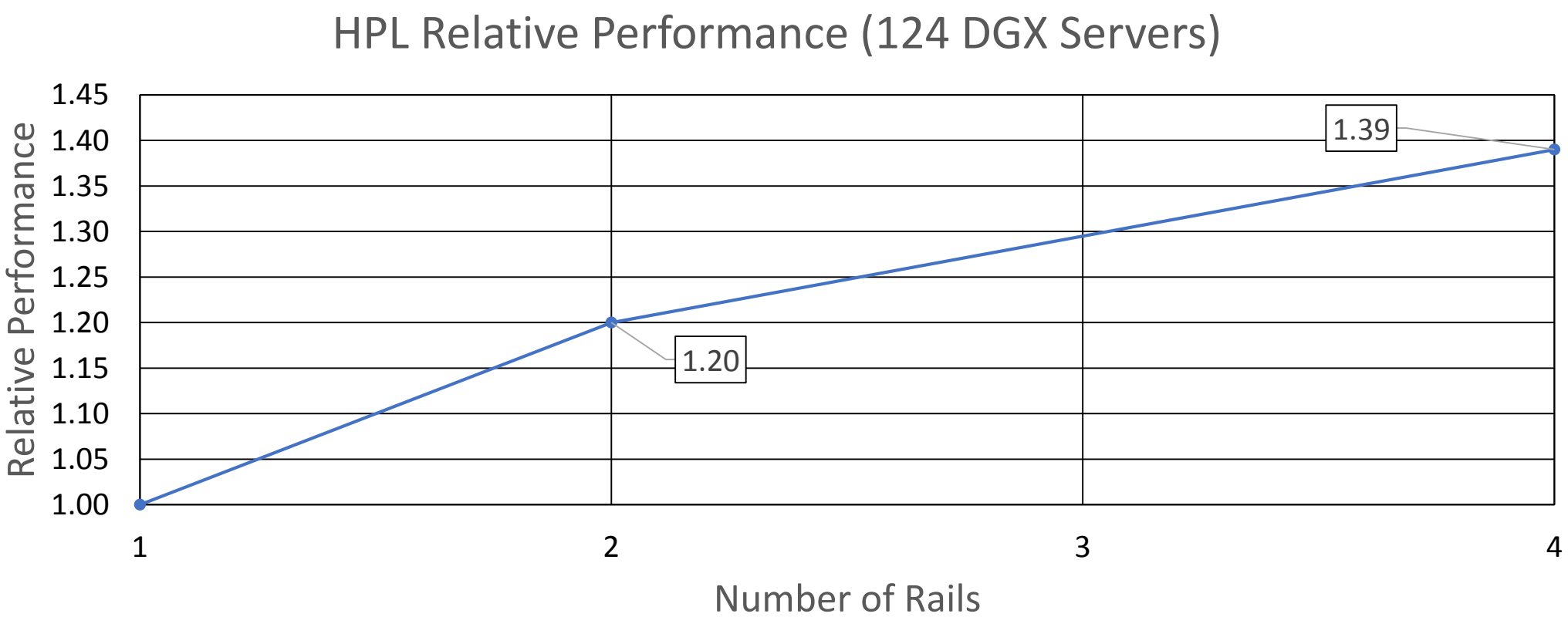
Takeaway:

Network communications have to be highly efficient to applications to scale

Single-rail networks may not be enough

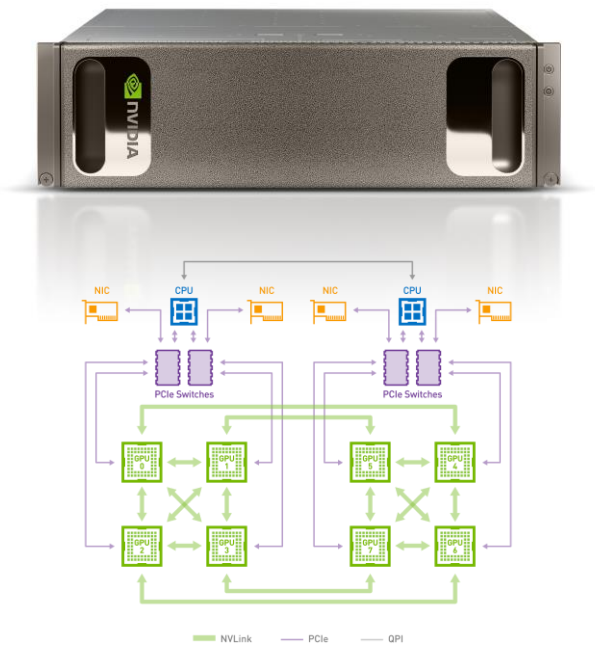
Number of nodes necessary to achieve same performance of 8-way P100 server.
http://www.nvidia.com/object/application-performance-guide.html

MULTI-RAIL APPLICATION PERFORMANCE



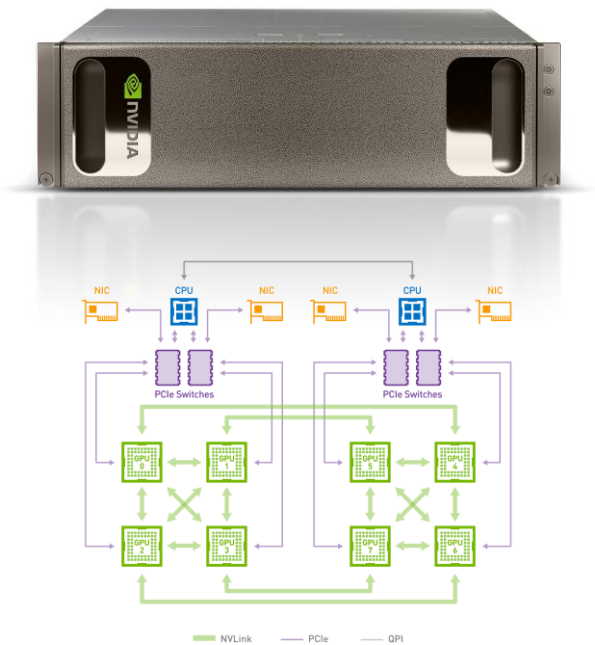
TEST CONFIGURATION

- Servers
  - 8 NVIDIA P100 GPUs /w NVLink
  - Intel Broadwell E5 v2697
  - Quad-rail Mellanox EDR Infiniband
- Switch
  - Mellanox SB7790, 36-port EDR switch
- MPI – Mvpaich2-GDR 2.2 w/ GCC



TEST CONFIGURATION

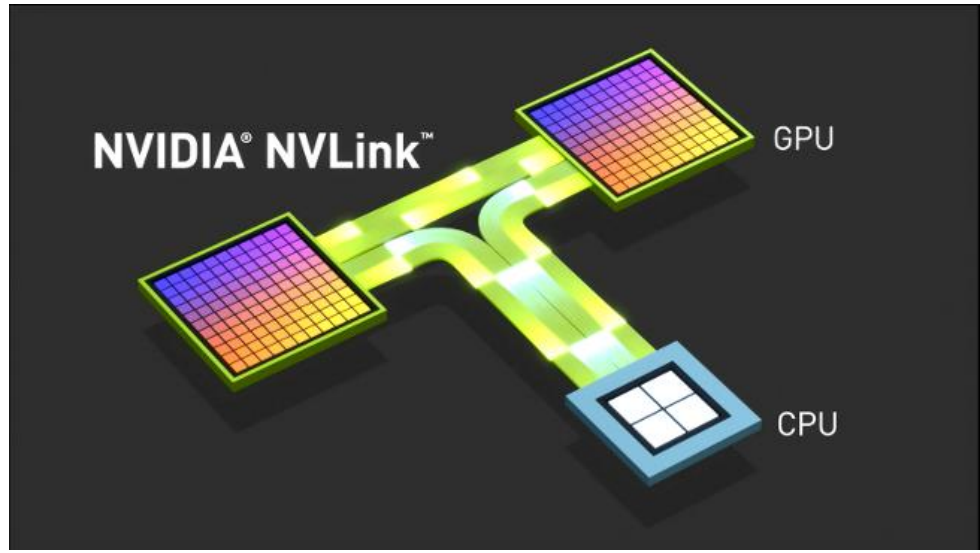
- Servers
  - 8 NVIDIA P100 GPUs /w NVLink
  - Intel Broadwell E5 v2697
  - Quad-rail Mellanox EDR Infiniband
- Switch
  - Mellanox SB7790, 36-port EDR switch
- MPI – Mvpaich2-GDR 2.2 w/ GCC



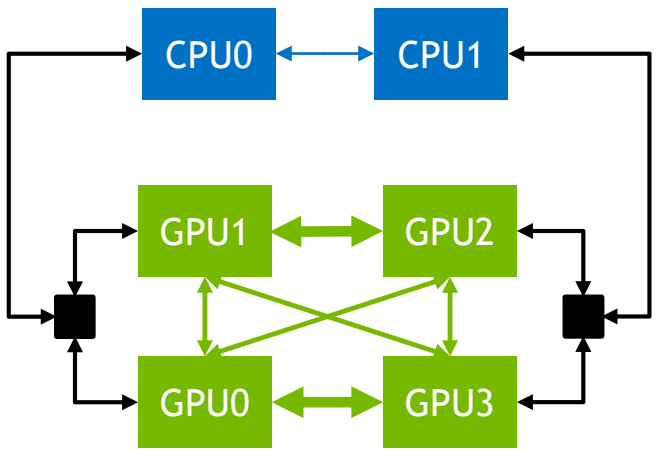
Results presented here are intended to show general trends and not guarantee performance for any particular system or configuration.

NVLINK

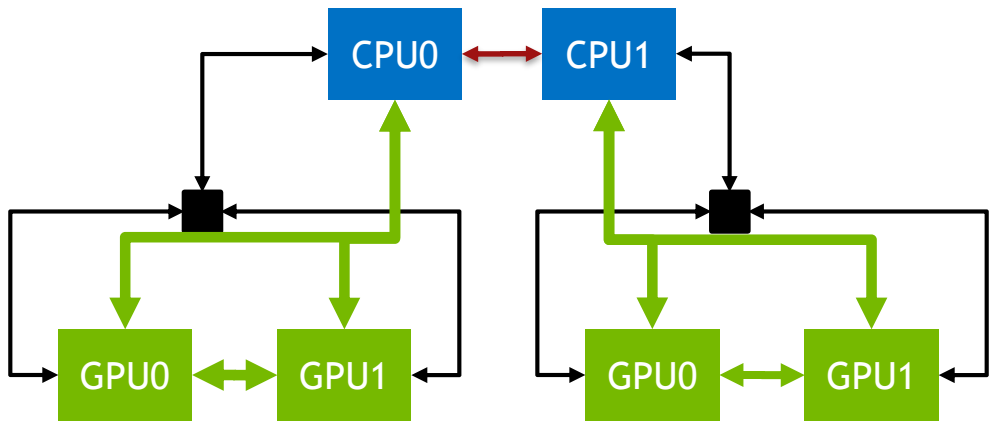
- NVLink is an energy-efficient, high-bandwidth path between the GPUs and the CPU
- NVLink Gen 1, Provides up to 160 GB/s
- NVLink Gen 2, Provides up to 300 GB/s
- IBM Power 8 (and future) systems connect GPUs directly to the CPU, removing the need for PCIe to communicate to the GPU



SYSTEM ARCHITECTURE DIAGRAMS W/ NVLINK

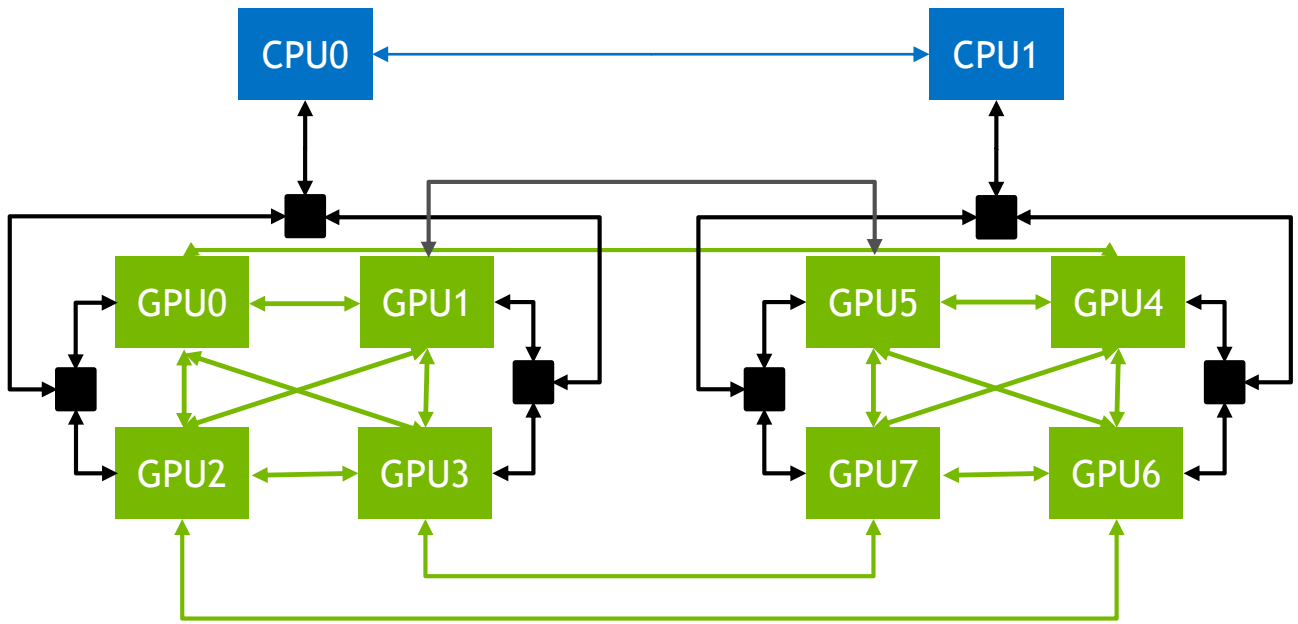


Ex: Supermicro 1028GQ-TRX



IBM Power8 Minsky

SYSTEM ARCHITECTURE DIAGRAMS W/ NVLINK

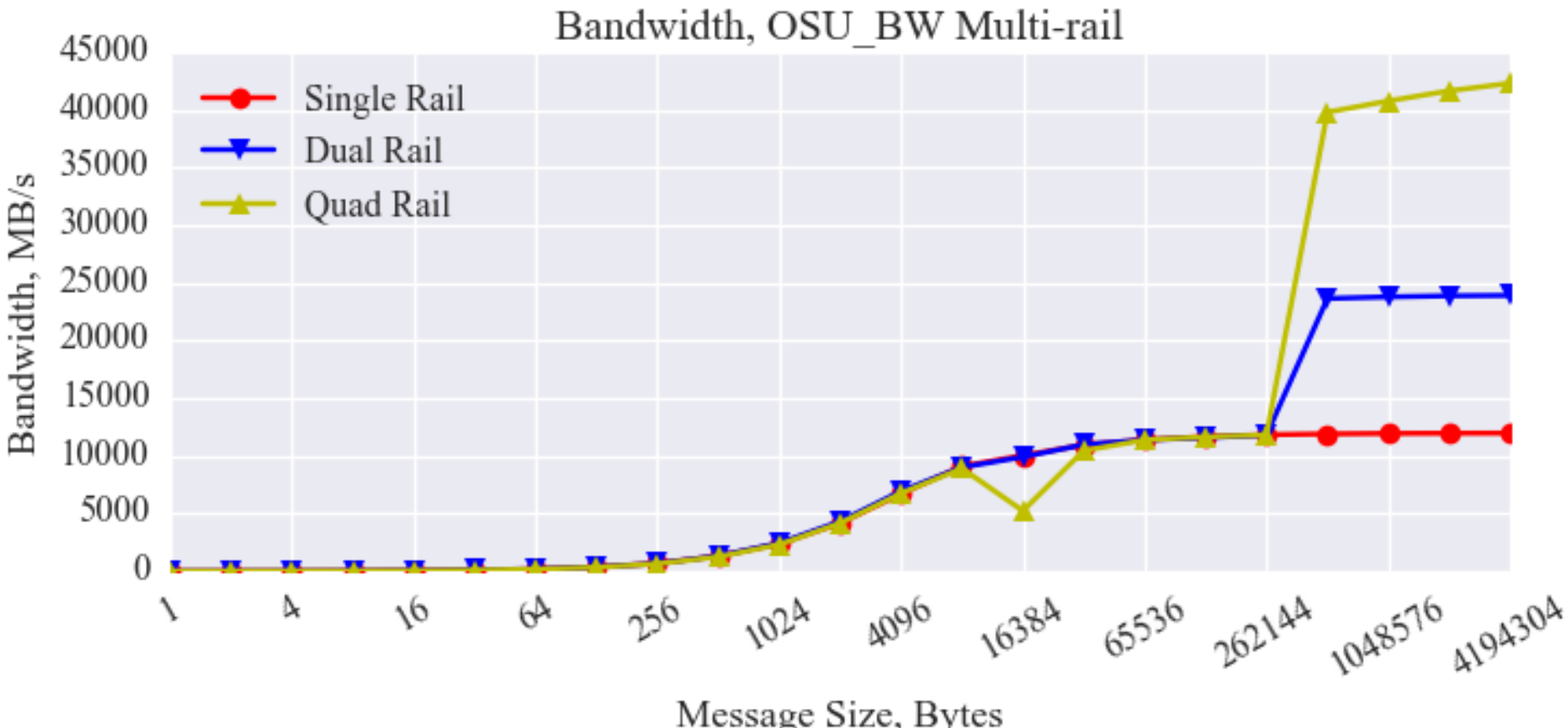


EX: Nvidia DGX Server

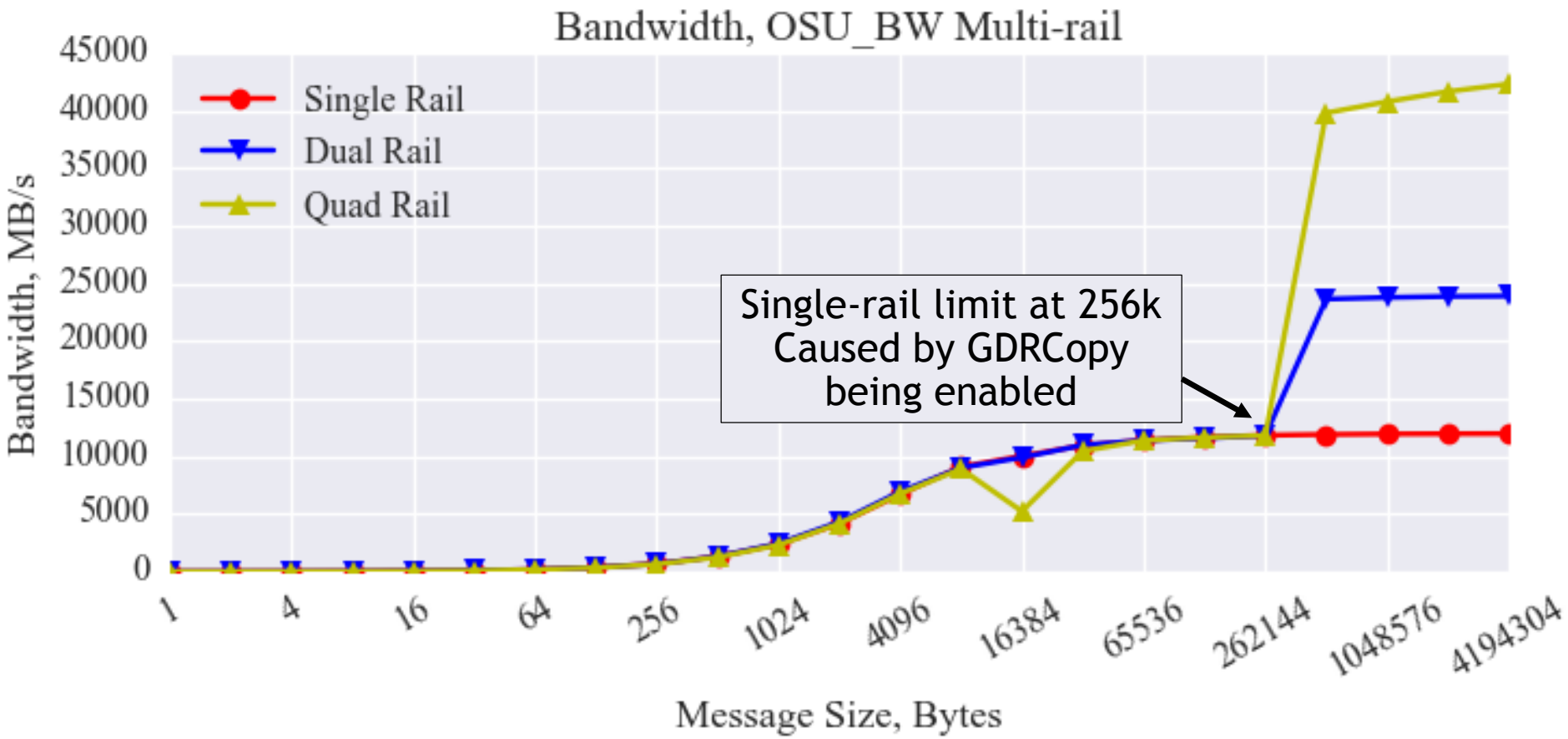
Agenda

- Overview
- **Multi-rail**
- Intranode performance with NVLink

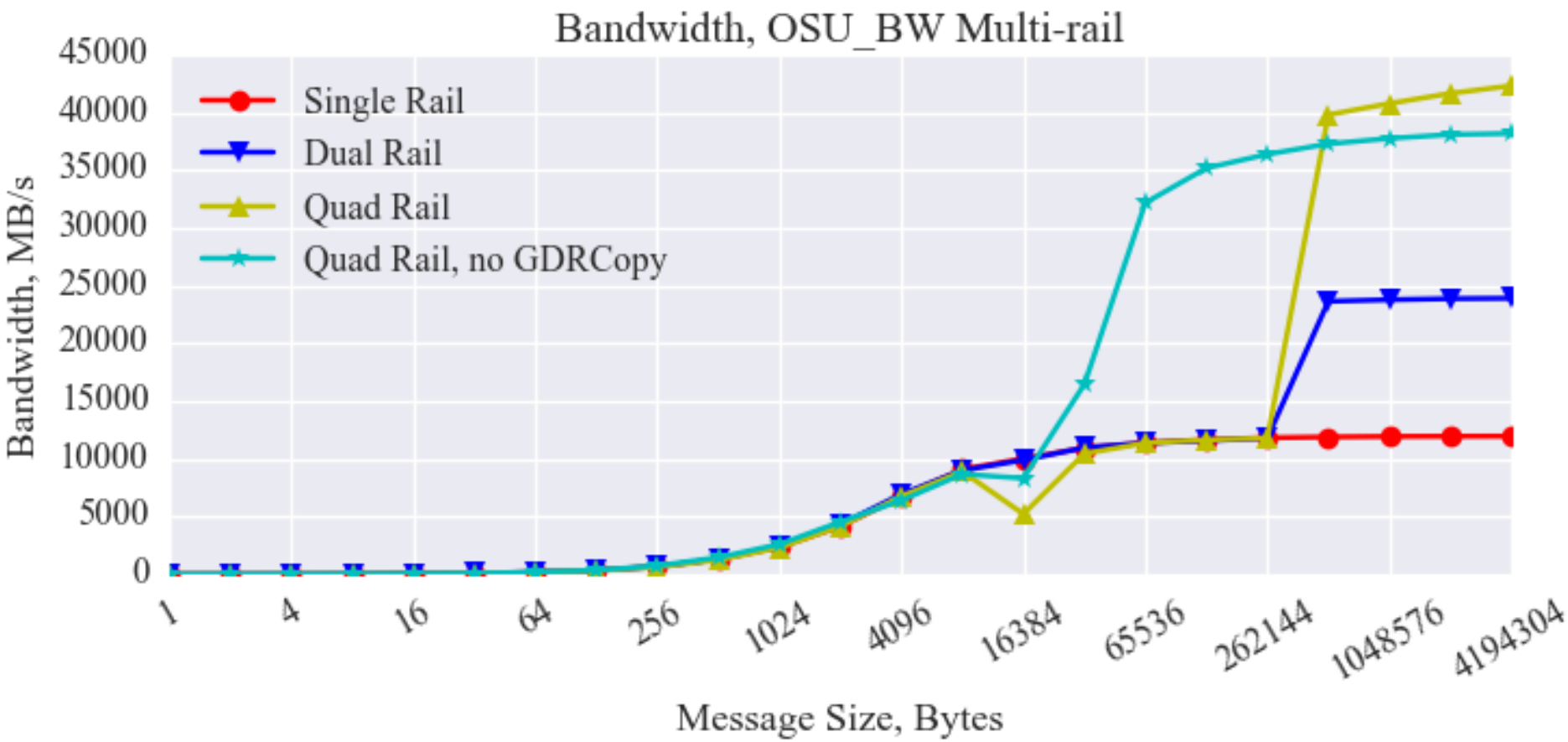
IB PERFORMANCE, OSU_BW



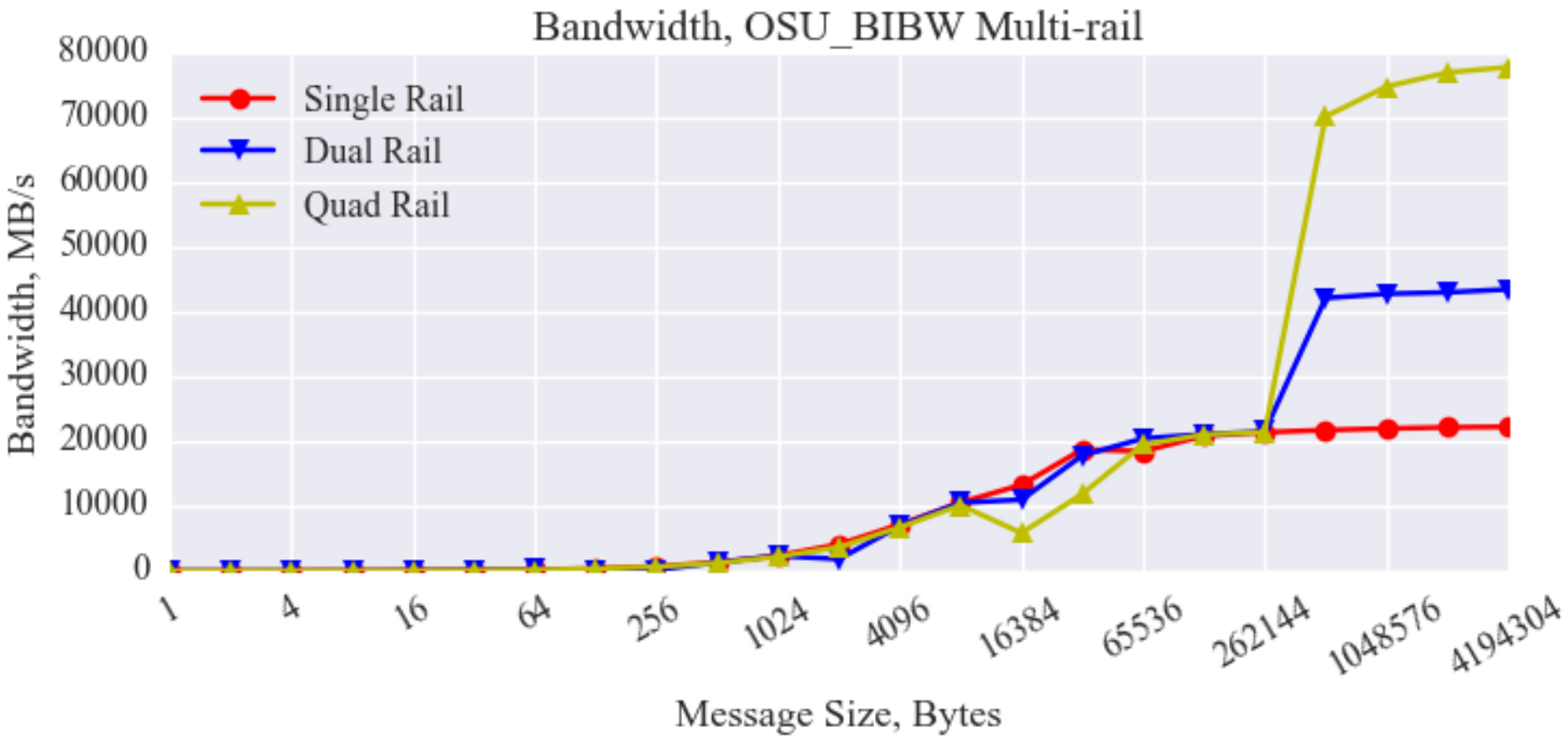
IB PERFORMANCE, OSU_BW



IB PERFORMANCE, OSU_BW



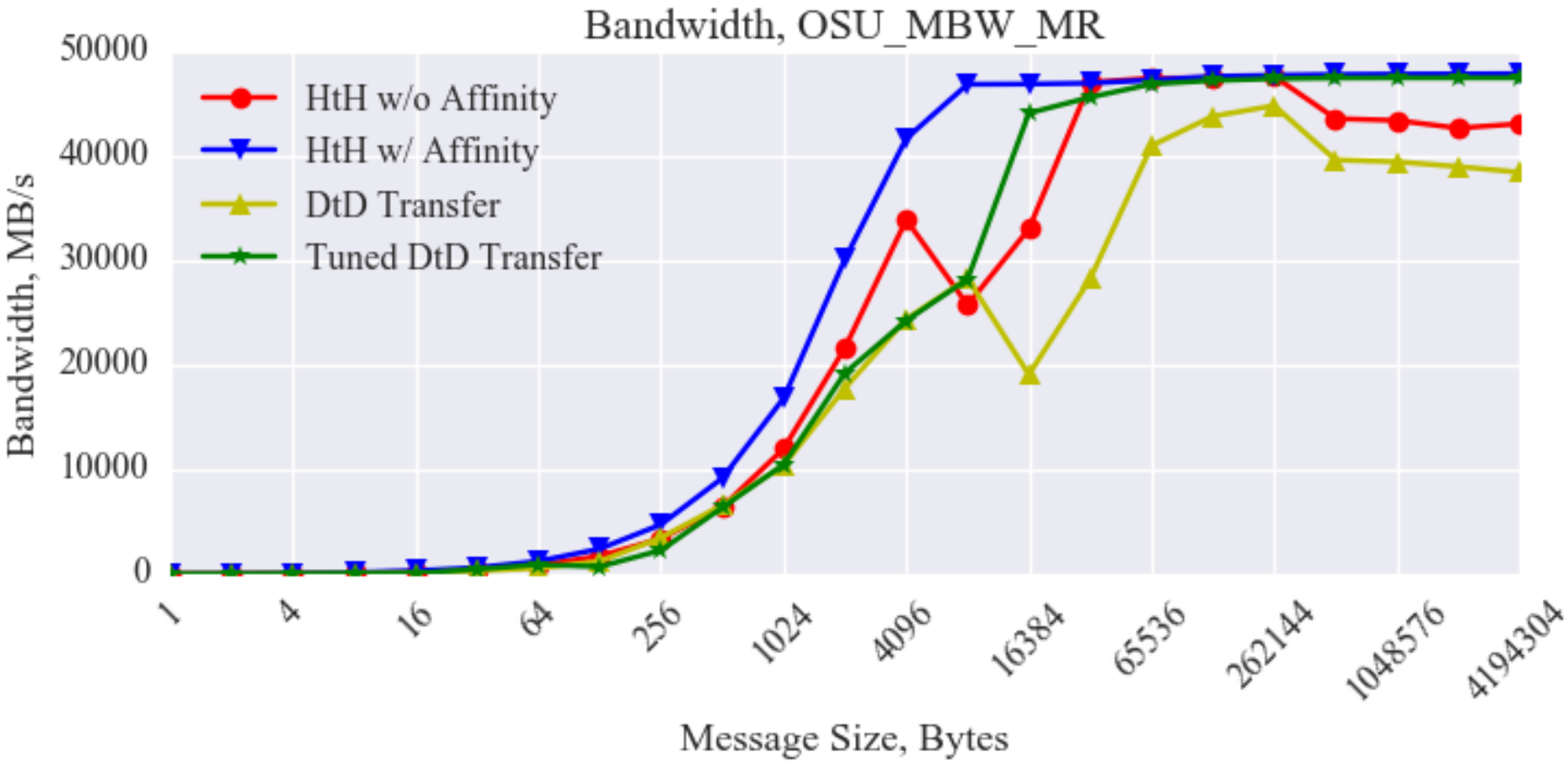
IB PERFORMANCE, OSU_BIBW



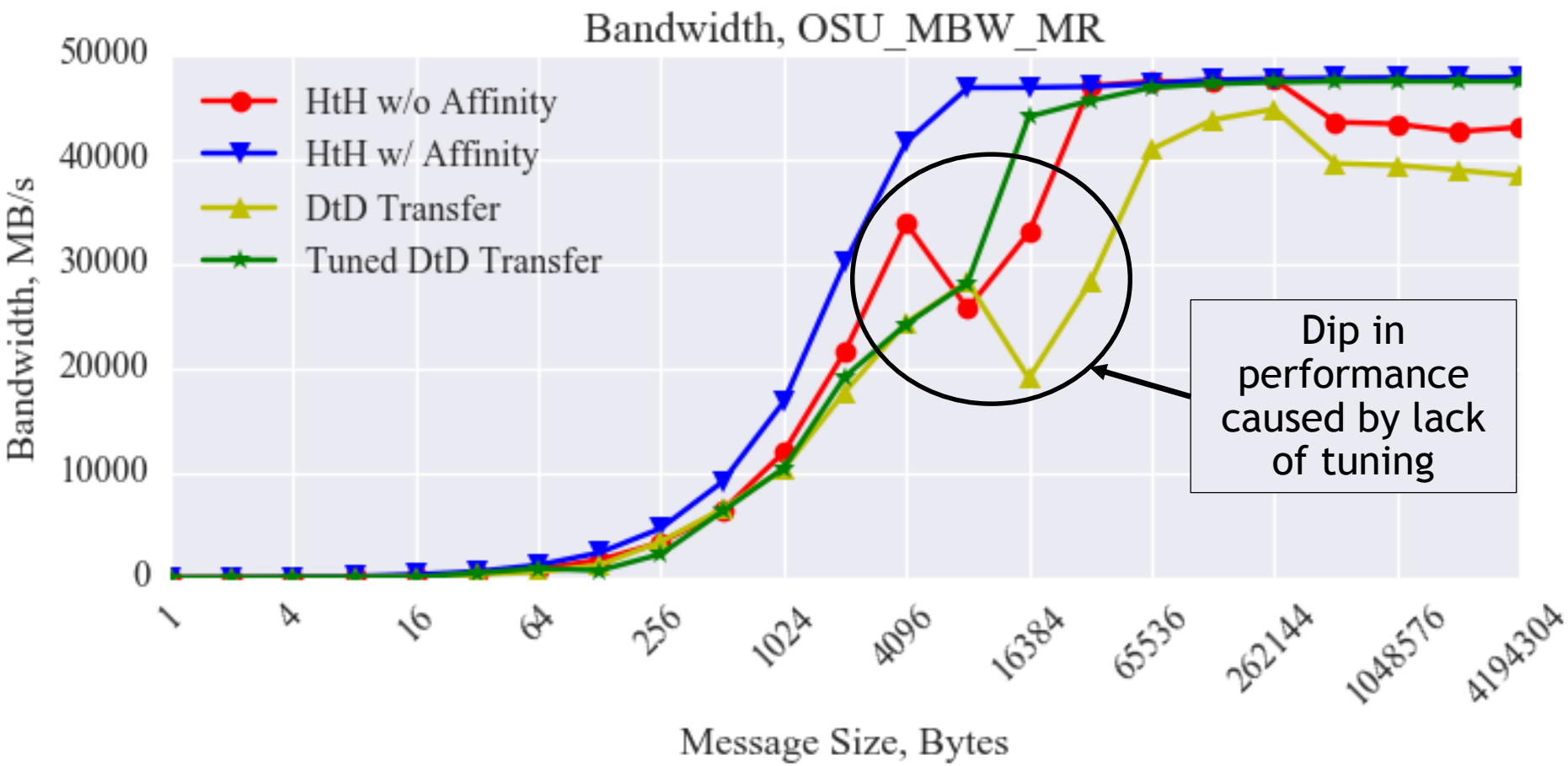
RANK LAYOUT ON MULTI-RAIL GPU SYSTEMS

- While multi-rail performance is impressive, in practice there will be multiple ranks per node, most likely 1 rank per GPU
- A better test is to measure bandwidth is to have one MPI rank per GPU
  - This is the OSU_MBW_MR test
- The test has been modified to support device-to-device (DtD) as well as host-to-host (HtH) transfers

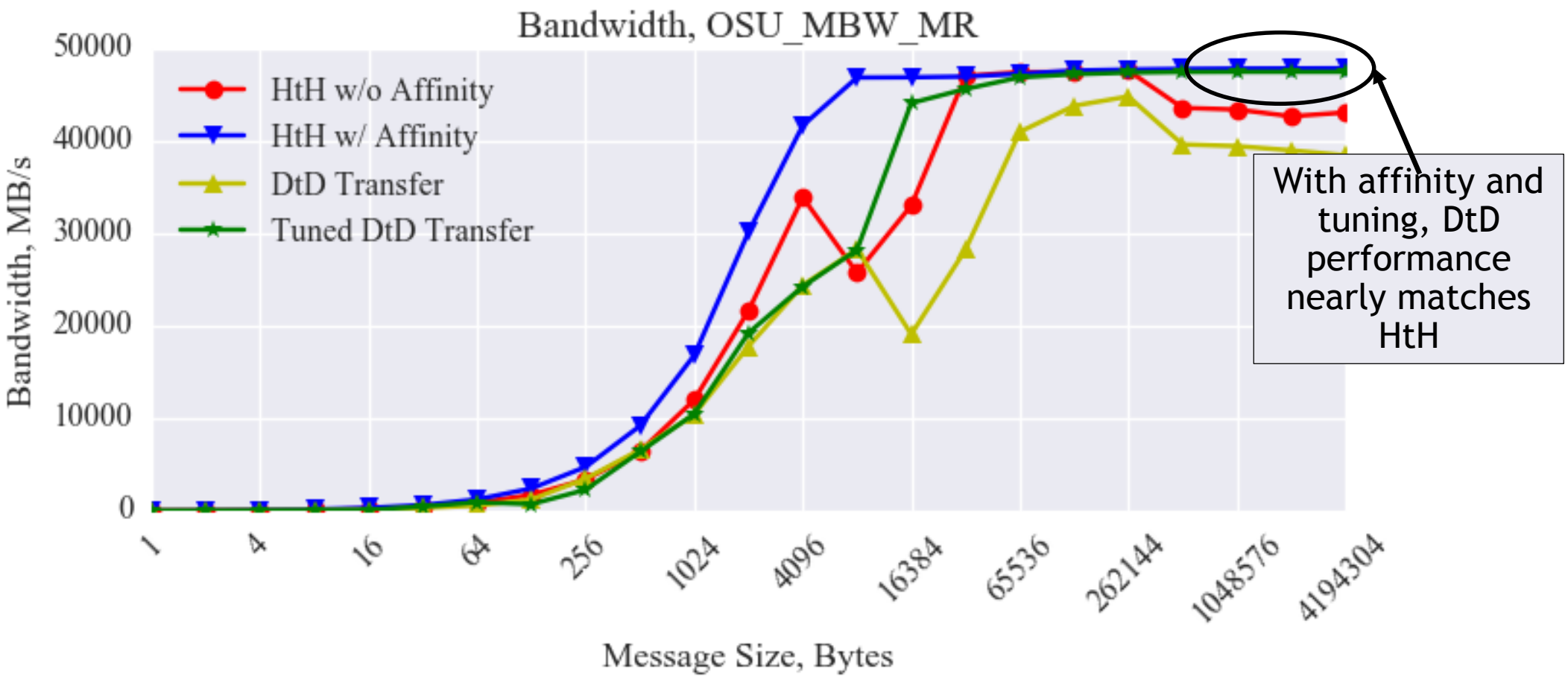
IB PERFORMANCE, OSU_MBW_MR



IB PERFORMANCE, OSU_MBW_MR



IB PERFORMANCE, OSU_MBW_MR



TUNABLE PARAMETERS

MVAPICH2 Tunable	Default Value	Proposed Value	Reason
MV2_CUDA_IPC_THESHOLD	32768	262144	Improve DtD Intranode transfers at 32K and above.
MV2_USE_GPUDIRECT_GDRCOPY_LIMIT	8192	32768	Improve HtD Intranode transfers between 16K and 32K
MV2_GPUDIRECT_LIMIT	8192	4194304*	Improve Internode transfers
MV2_USE_SMP_GDR	1	0	Offset performance degradation to Intranode transfers when setting MV2_GPDIRECT_LIMIT
MV2_GPUDIRECT_RECEIVE_LIMIT	131072	4194304*	Improve all Intranode transfers at 1M and above

Note, the search of this space was not exhaustive. More optimization can be done.

* May need to be larger depending on your largest transfer.

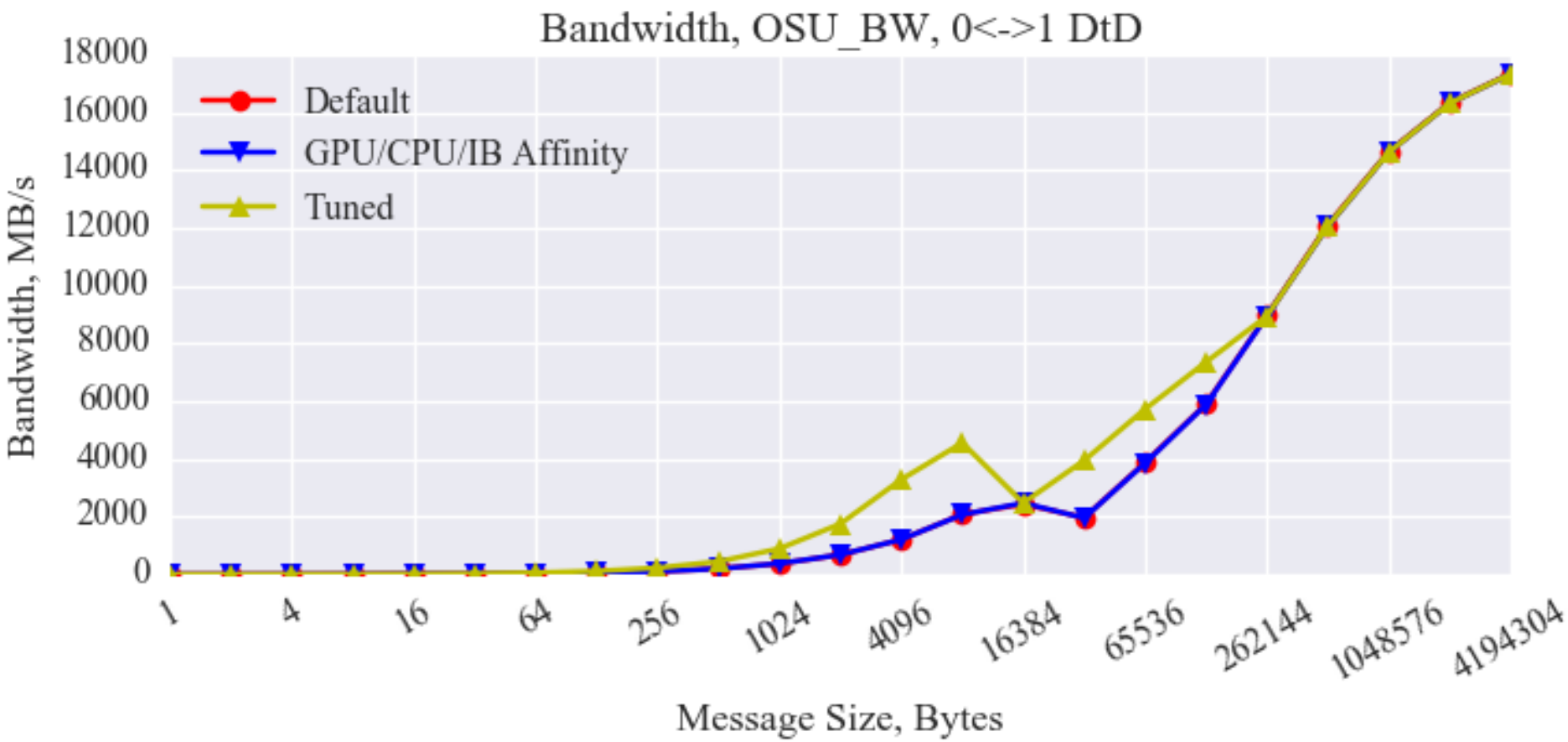
Agenda

- Overview
- Multi-rail
- **Intranode performance with NVLink**

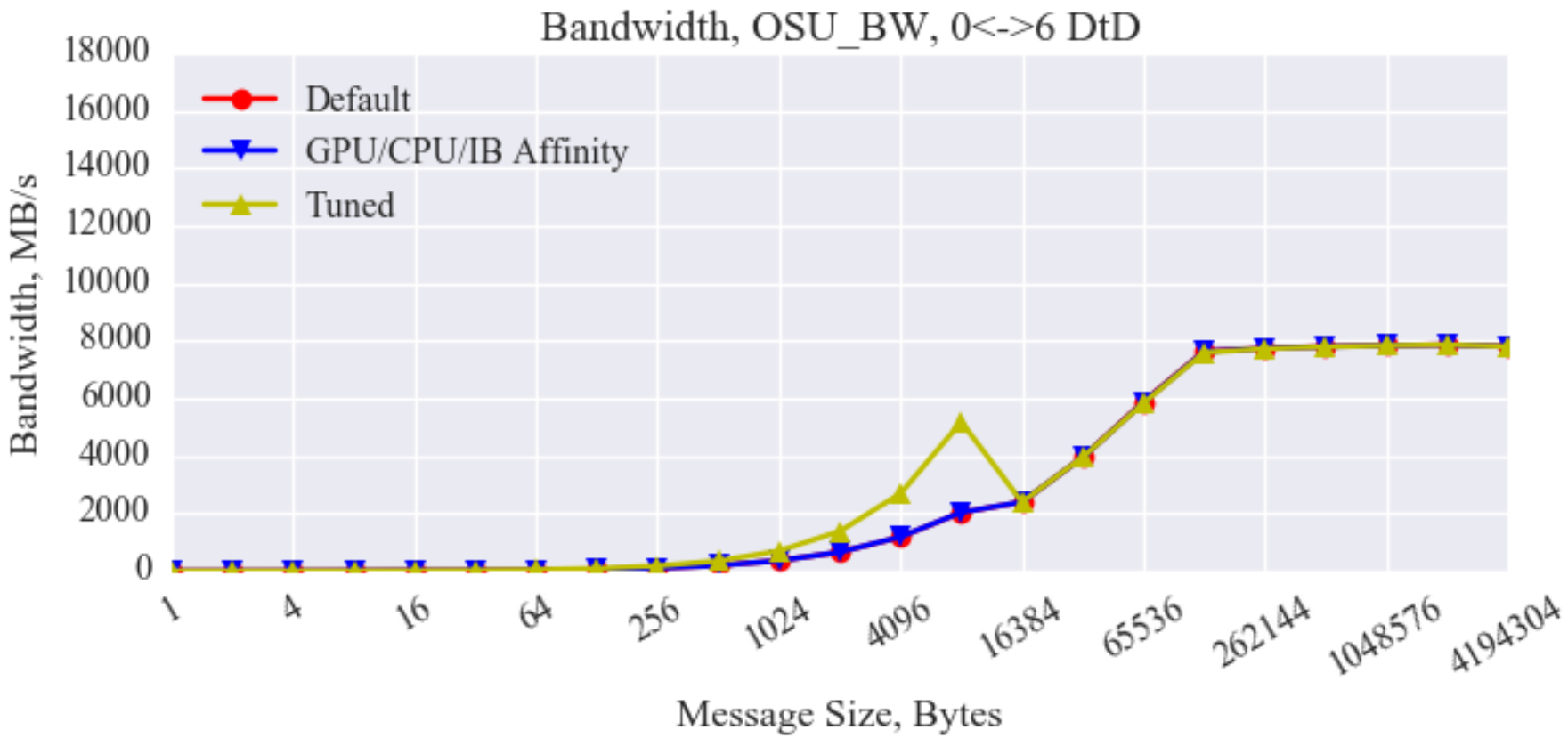
INTRANODE MPI PERFORMANCE

- With the tunables above, what is the performance of intranode transfers?
- Testing pairs:
  - GPUs 0 and 1 – NVLink connected, Same PCIe switch
  - GPUs 0 and 4 – NVLink connected, Different CPU socket
  - GPUs 0 and 6 – Not NVLink connected, Different CPU socket

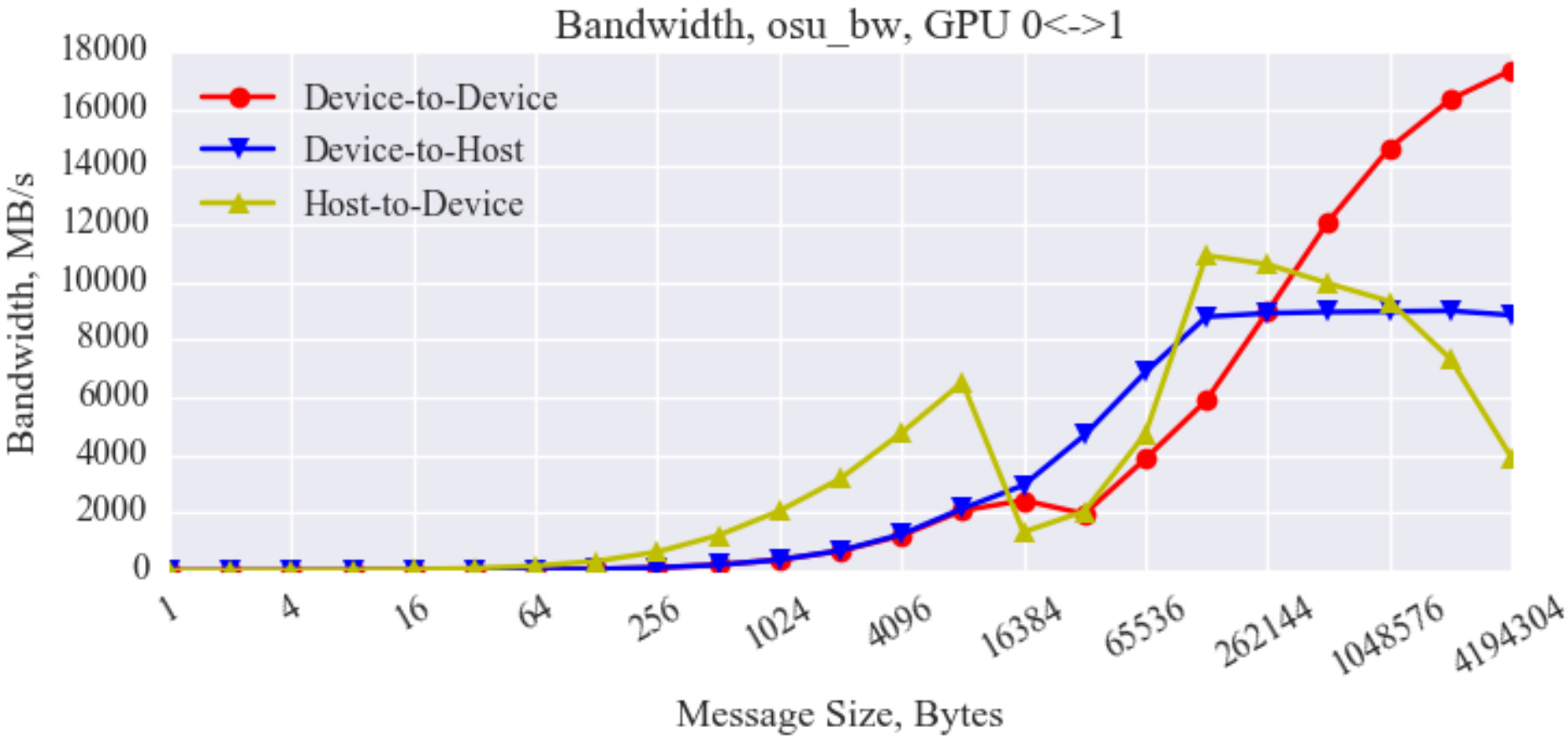
INTRANODE TRANSFERS



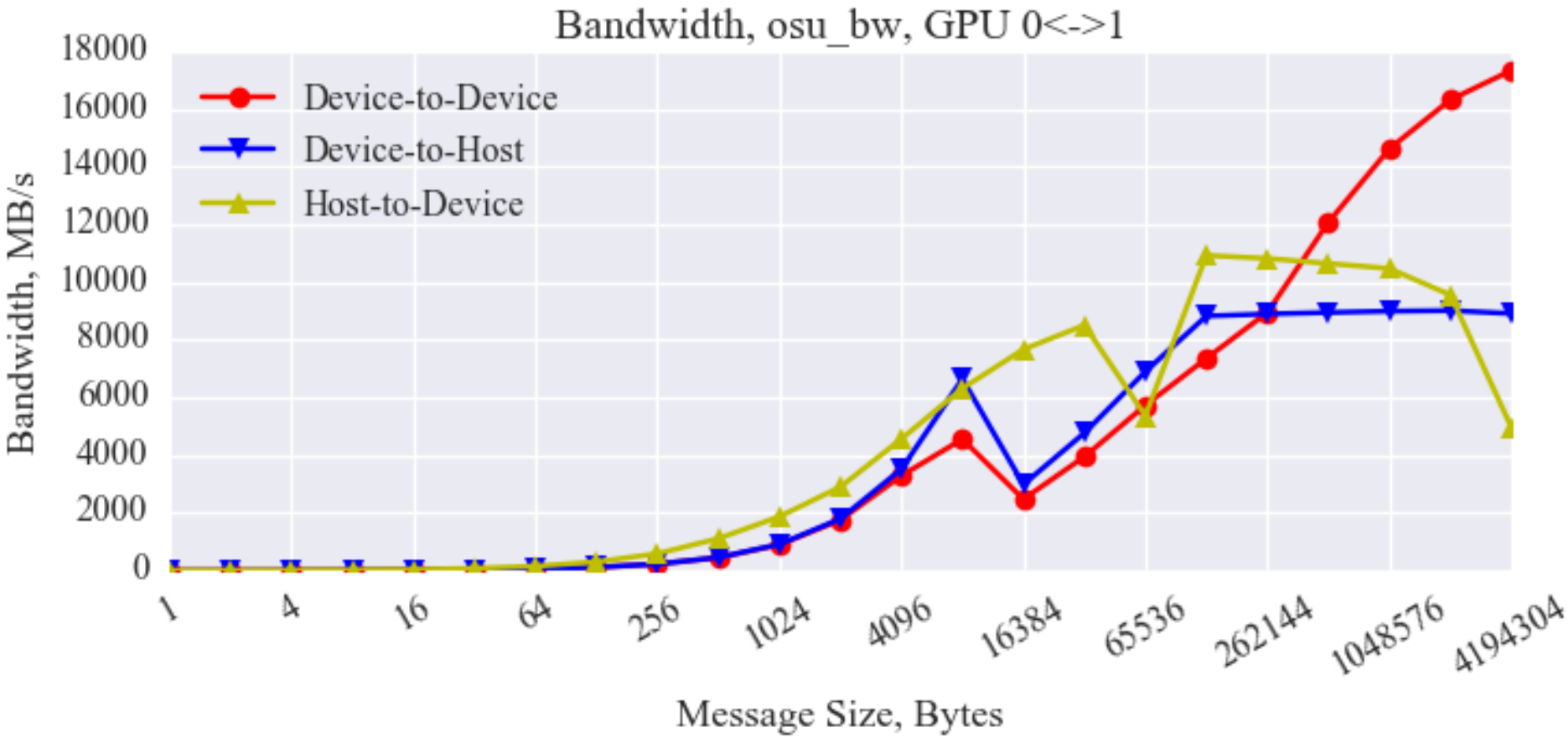
INTRANODE TRANSFERS



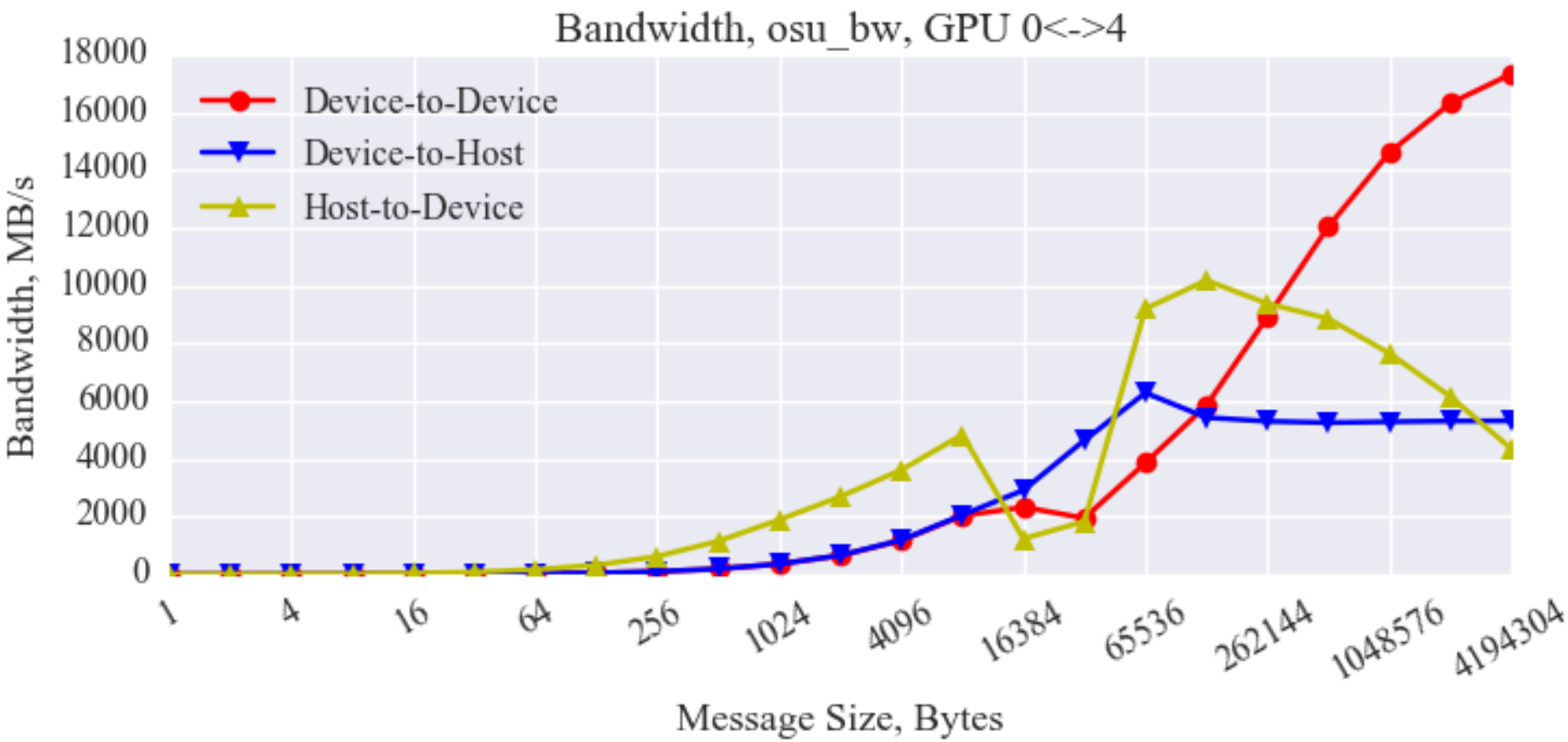
INTRANODE TRANSFERS - DEFAULTS



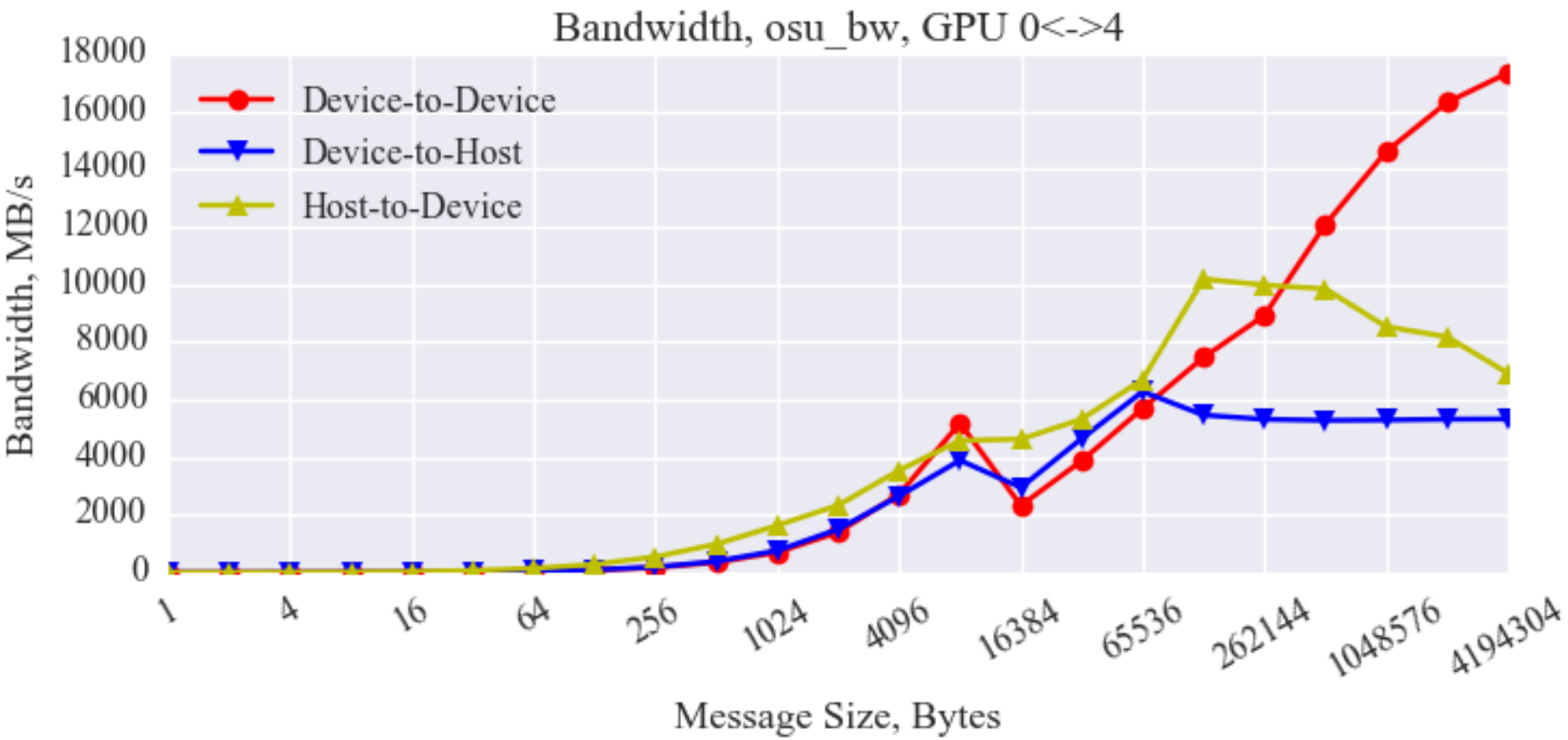
INTRANODE TRANSFERS - TUNED



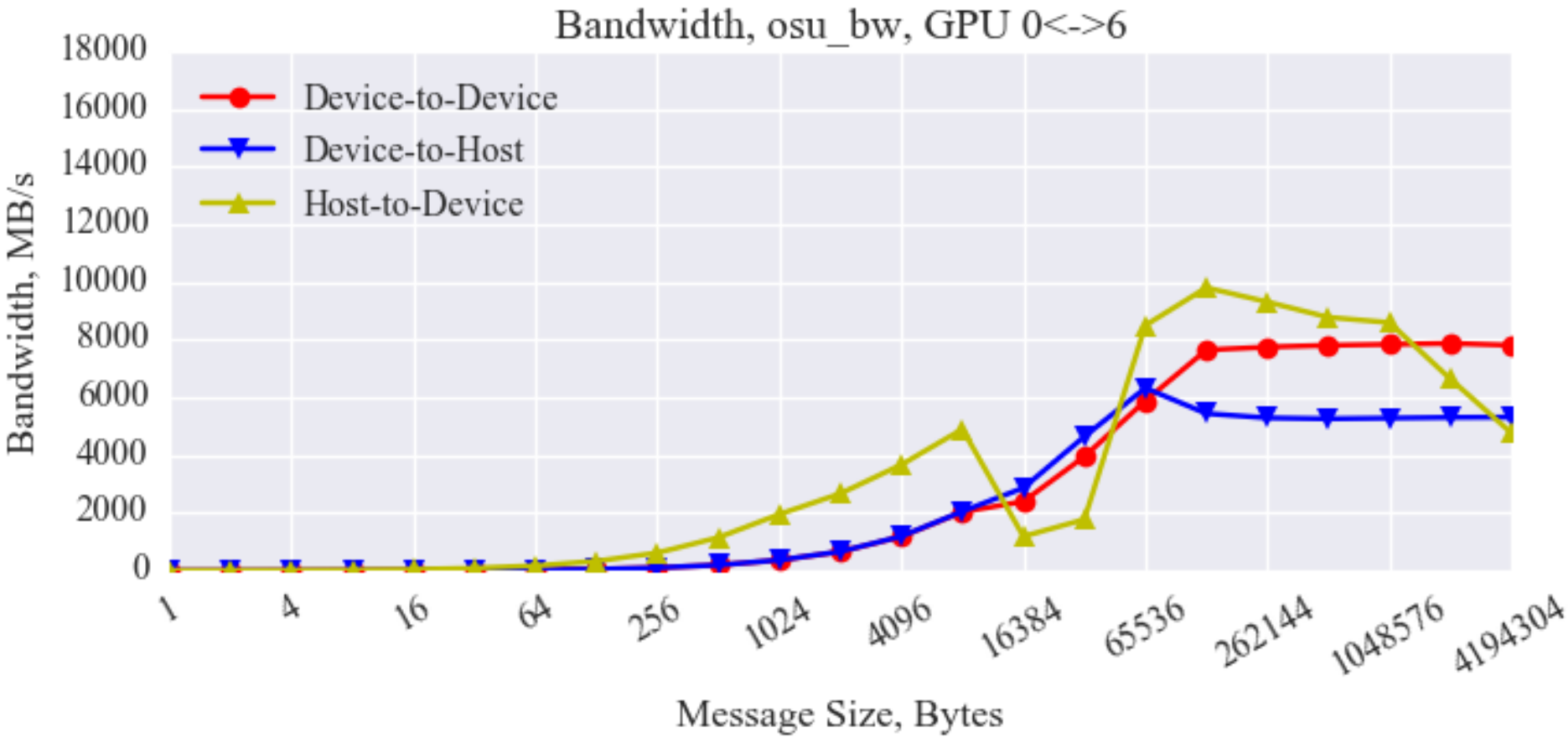
INTRANODE TRANSFERS - DEFAULTS



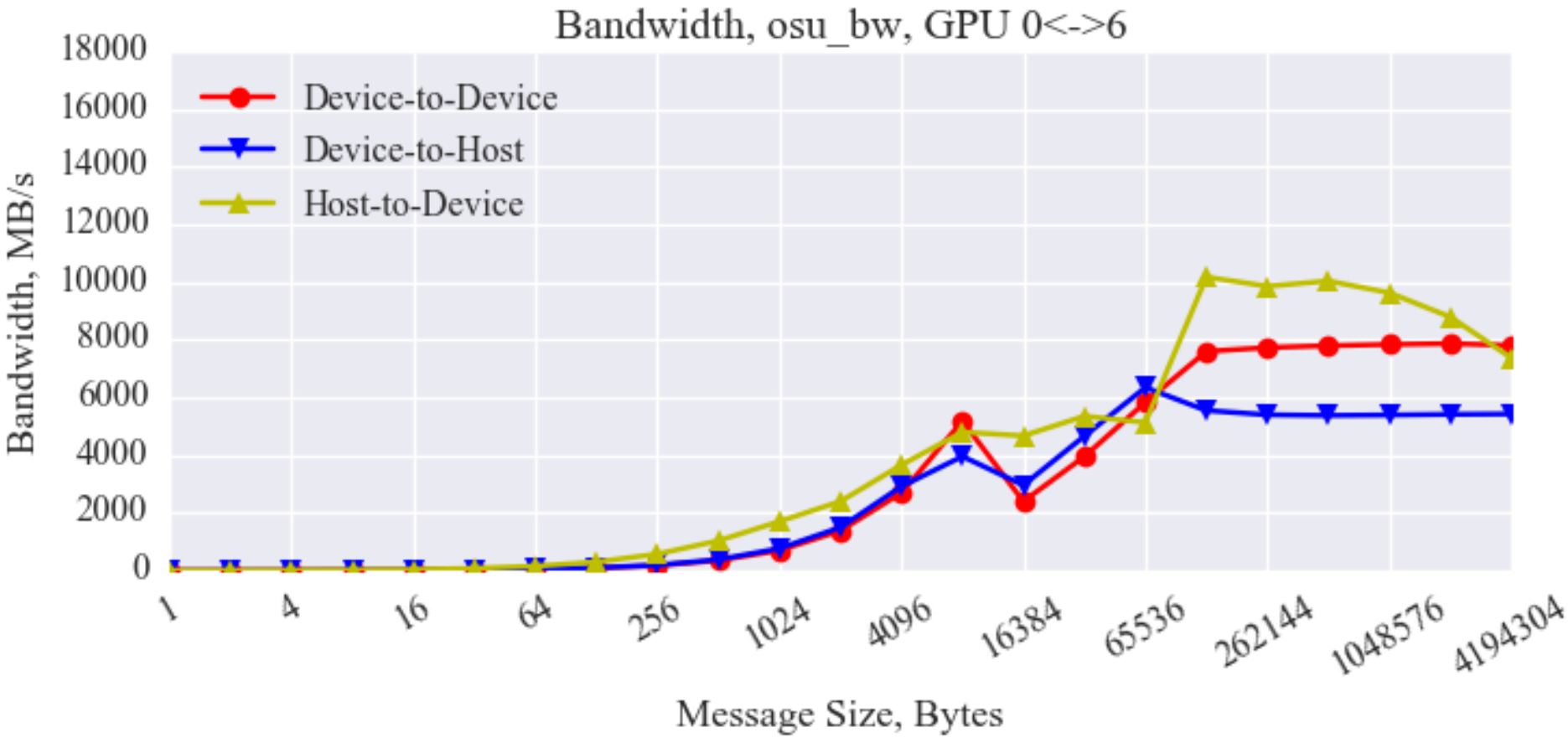
INTRANODE TRANSFERS - TUNED



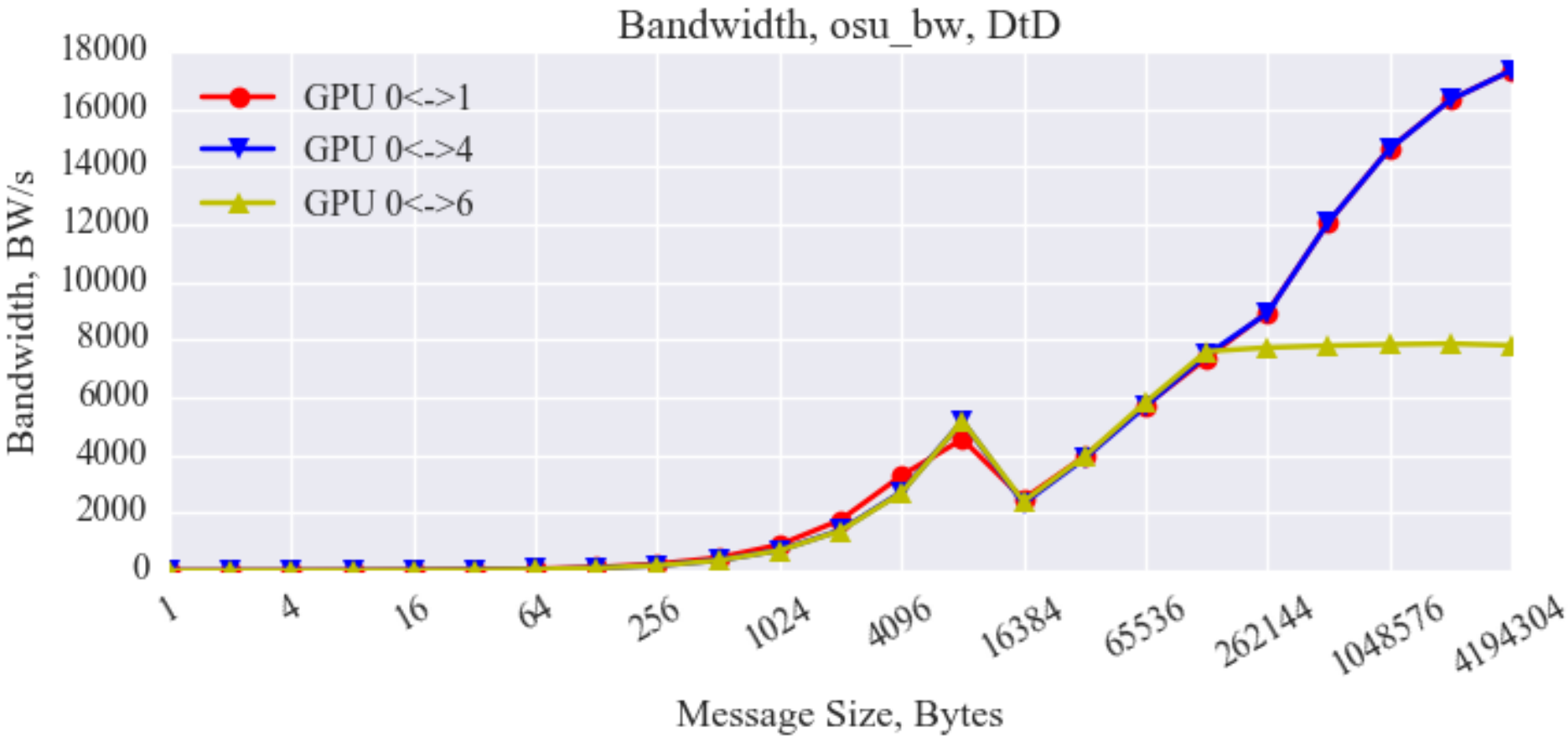
INTRANODE TRANSFERS - DEFAULTS



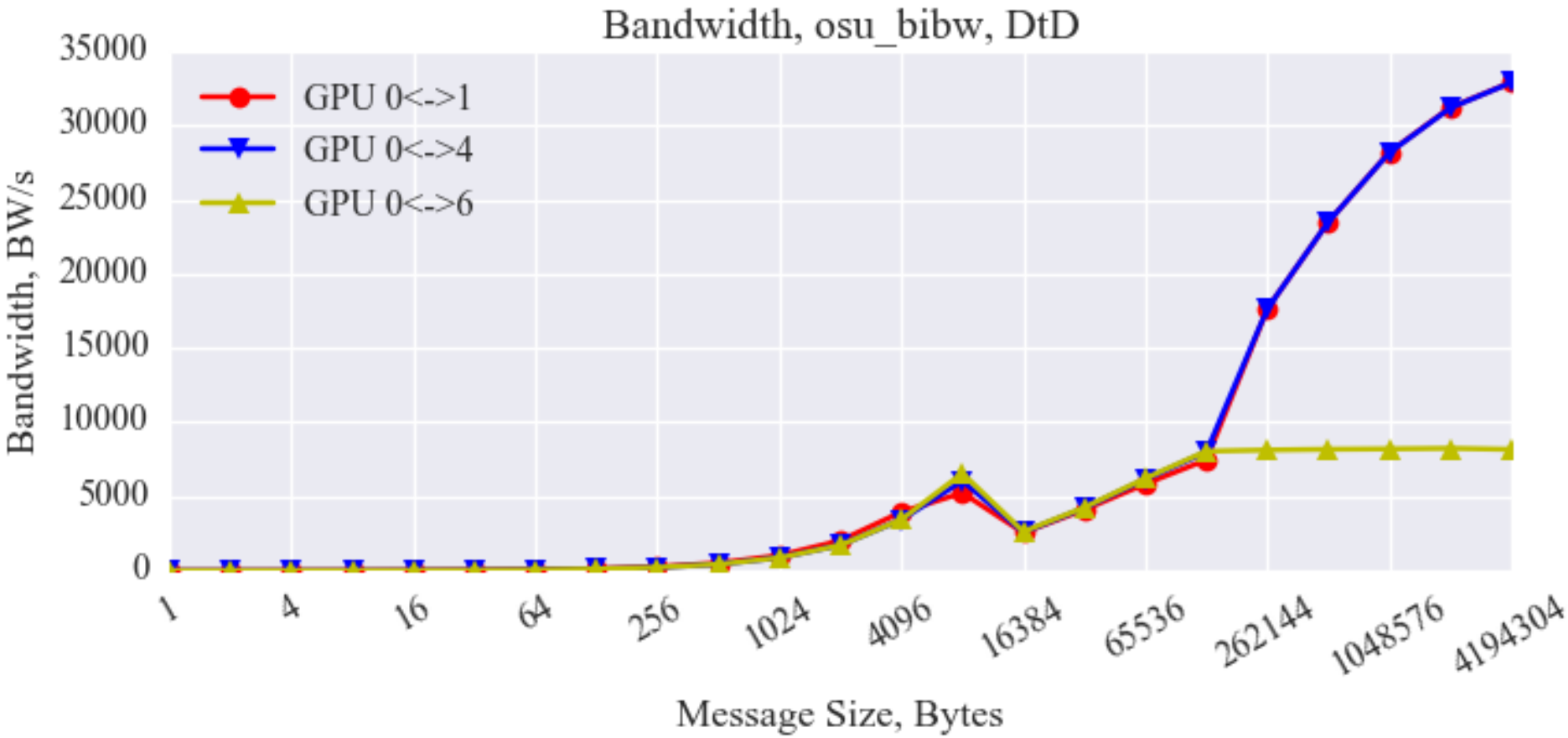
INTRANODE TRANSFERS - TUNED



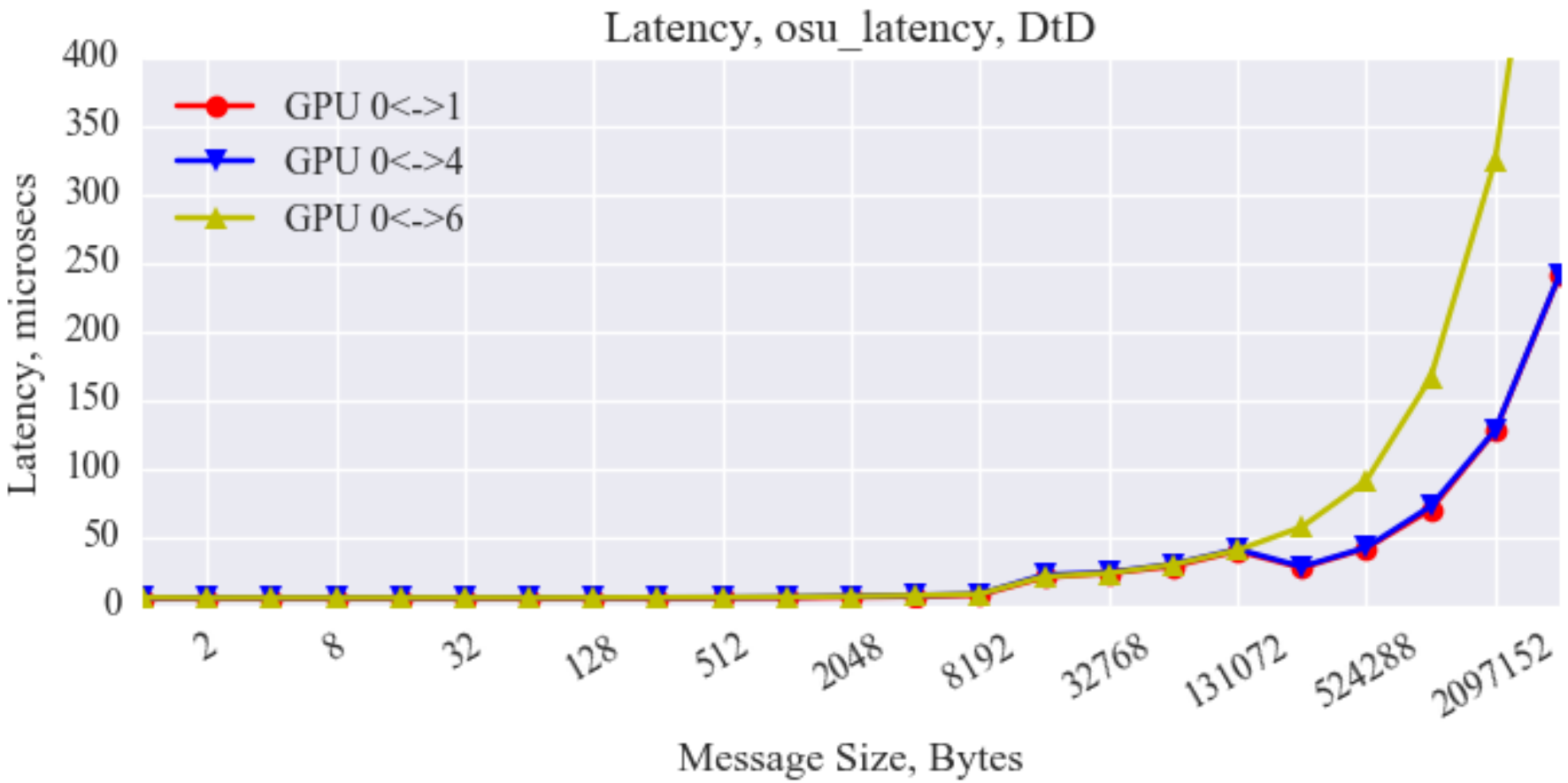
INTRANODE TRANSFERS – BY GPU PAIR



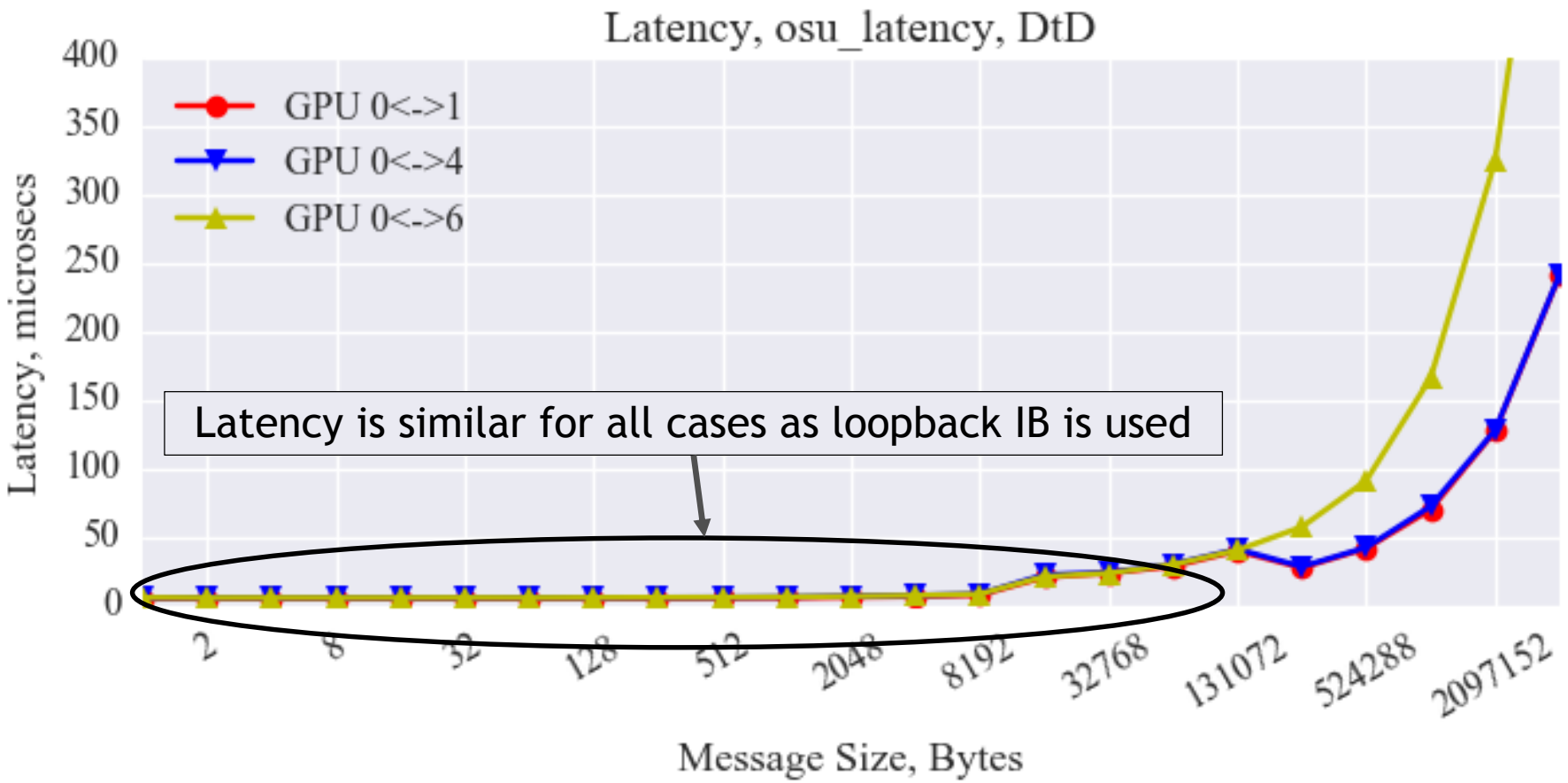
INTRANODE TRANSFERS – BY GPU PAIR



INTRANODE TRANSFERS – BY GPU PAIR



INTRANODE TRANSFERS - BY GPU PAIR



SUMMARY

- Mvapich2 is already very efficient with dual-rail and quad-rail configurations
- Setting affinity properly for HCAs and GPUs is key to maximize total node BW
- NVLink bandwidth provides great potential for improving scalability over PCIe
- MPI point-to-point message passing can be greatly improved with the tunables provided in the Mvapich2 stack
  - More tuning will improve results
  - More tuning may discover parameters need to be scoped for intra- and inter- node transfers differently
  - Results should be rerun with mvpaich2-2.3a to see what changes have been made

Thank you!

```
#!/bin/bash

if [ $MV2_COMM_WORLD_LOCAL_RANK -eq 0
]; then
    export LOCAL_RANK=$1
elif [ $MV2_COMM_WORLD_LOCAL_RANK -eq
1 ]; then
    export LOCAL_RANK=$2
else
    echo "THIS TOOL ONLY SUPPORTS 2
RANKS, EXITING"
    exit
fi

shift
shift

case ${LOCAL_RANK} in
[0])
        export MV2_IBA_HCA=mlx5_0
        export CORELIST=0-4
        ;;
[1])
        export MV2_IBA_HCA=mlx5_0
        export CORELIST=5-9
        ;;
[2])
        export MV2_IBA_HCA=mlx5_1
        export CORELIST=10-14
        ;;
…
[6])
        export MV2_IBA_HCA=mlx5_3
        export CORELIST=30-34
        ;;

[7])
        export MV2_IBA_HCA=mlx5_3
        export CORELIST=35-39
        ;;
esac

export MV2_ENABLE_AFFINITY=0

numactl --physcpubind=$CORELIST $*
```