

CUDA DEVELOPMENTS IN UCX

November 13, 2018



FEATURES

Inter-node and intra-node GPU->GPU support for:

- Pinned Device Memory

- Managed Memory

Mechanisms:

- GPUDirectRDMA, cudaMemcpy/GDRcopy for inter-node transfers

- CUDA-IPC for intra-node transfers

Optimizations:

- Cache to avoid cost of `culpcOpenMemHandle/cuPointerGetAttribute`

INTER-NODE RESULTS

Experiment setup:

Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz

Tesla V100-SXM2 (CUDA 9.2.88, driver version 410.48)

ConnectX-4 Mellanox HCAs (OFED-internal-4.0-1.0.1)

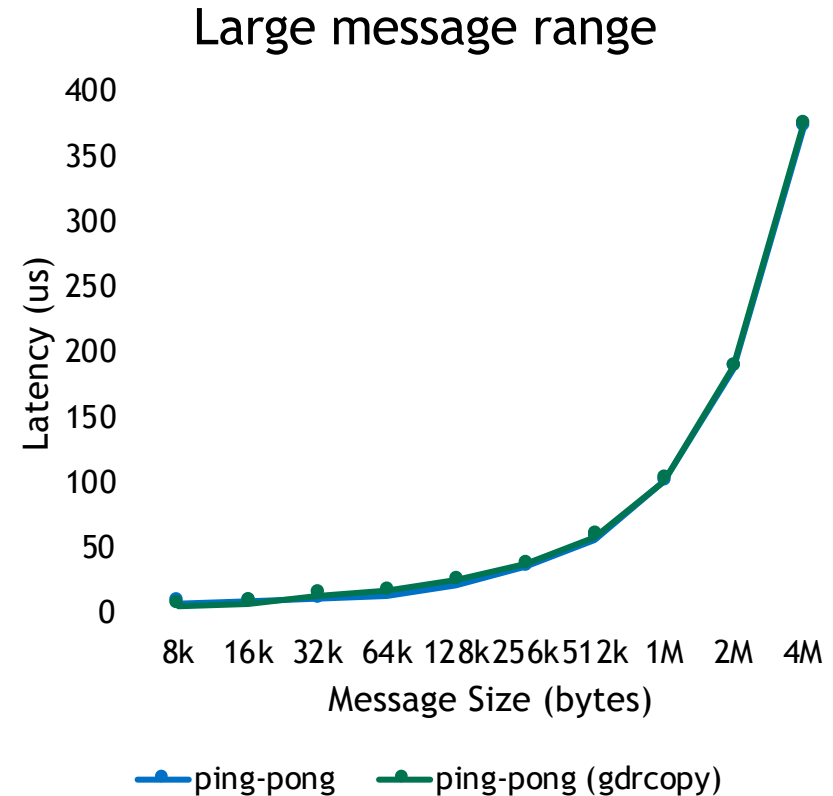
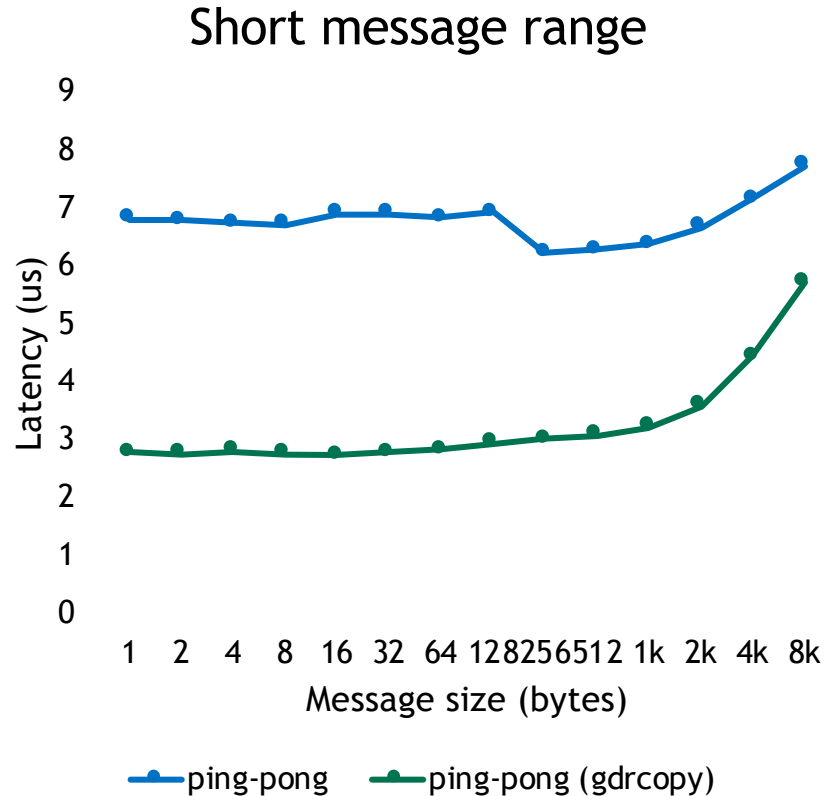
Network Software:

OpenMPI 4.1.0a1

UCX 1.4

DEVICE BUFFER LATENCY

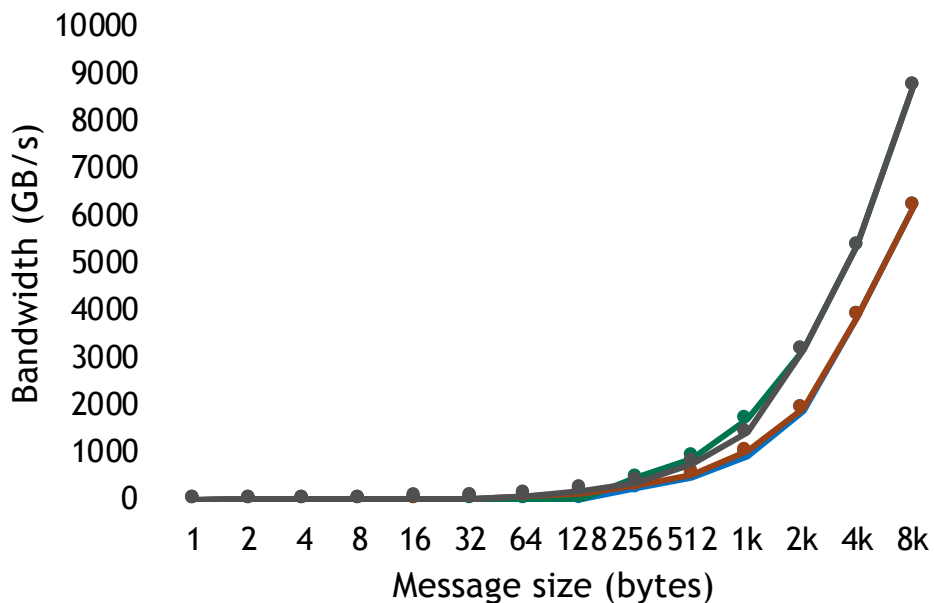
`osu_latency D D; UCX_RNDV_SCHEME=put_zcopy`



DEVICE BUFFER BANDWIDTH

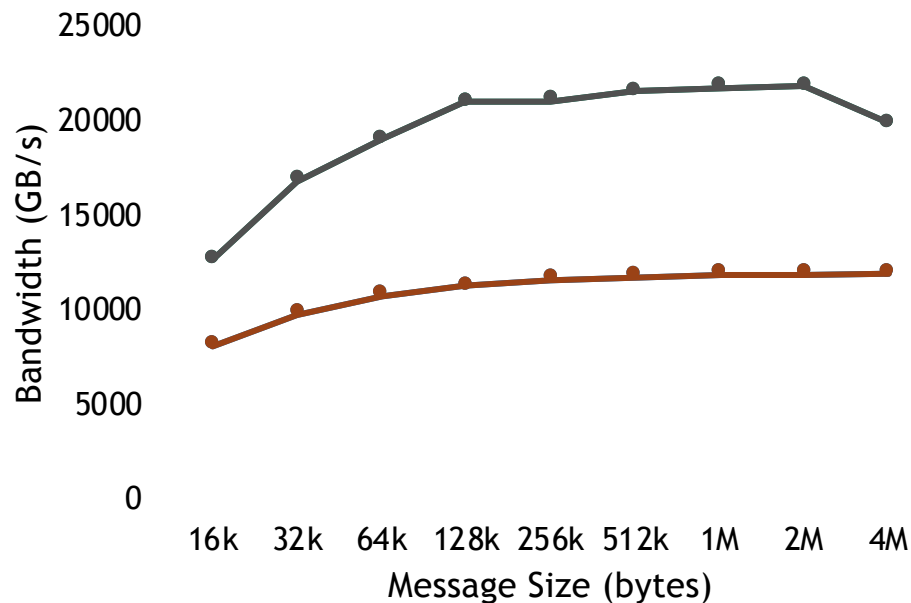
osu_bw, osu_bibw; UCX_RNDV_SCHEME=put_zcopy

Short message range



— unidirectional — bidirectional
— unidirectional (gdrcopy) — bidirectional (gdrcopy)

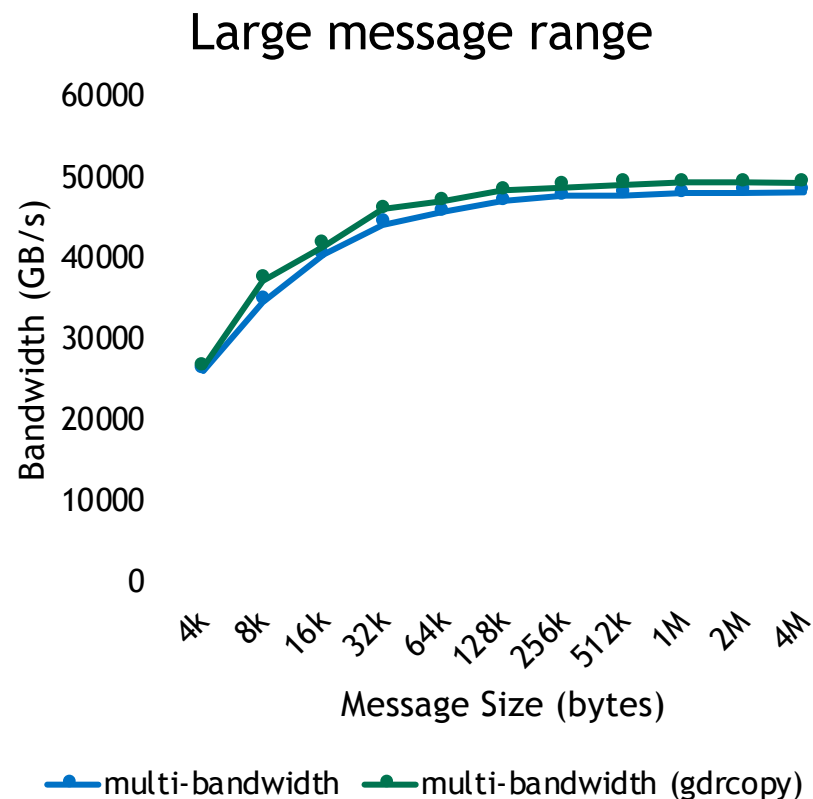
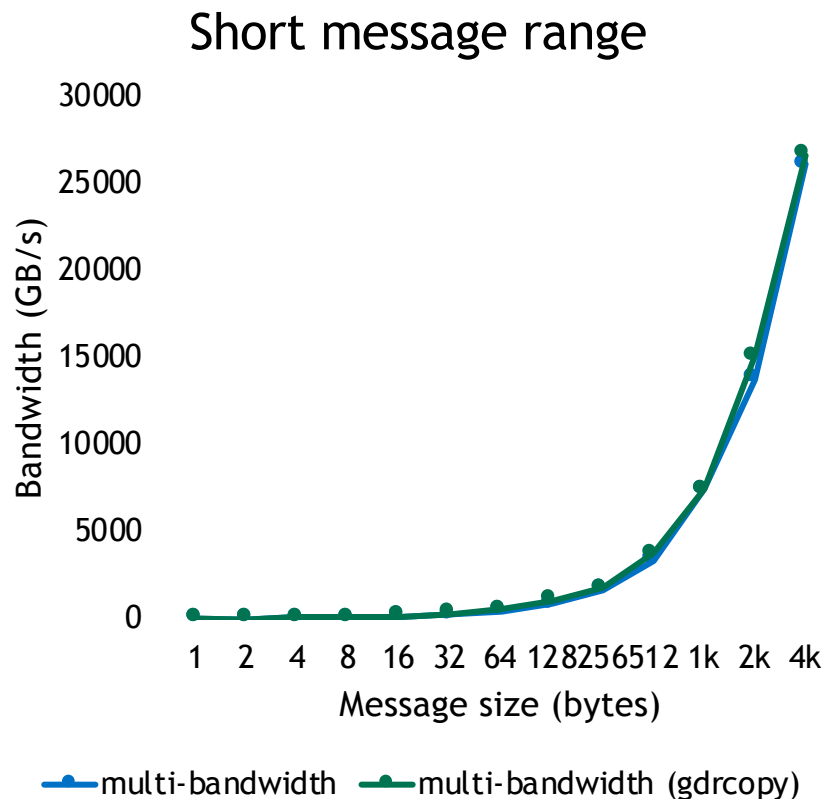
Large message range



— unidirectional — bidirectional
— unidirectional (gdrcopy) — bidirectional (gdrcopy)

DEVICE BUFFER MULTI-BANDWIDTH

osu_mbw_mr; EXPERIMENTAL UCX GPU-HCA affinity feature



INTRA-NODE RESULTS

Network Hardware:

1 NVLINK2 between GPU 0 -> GPU 1 (CVD=0,1)

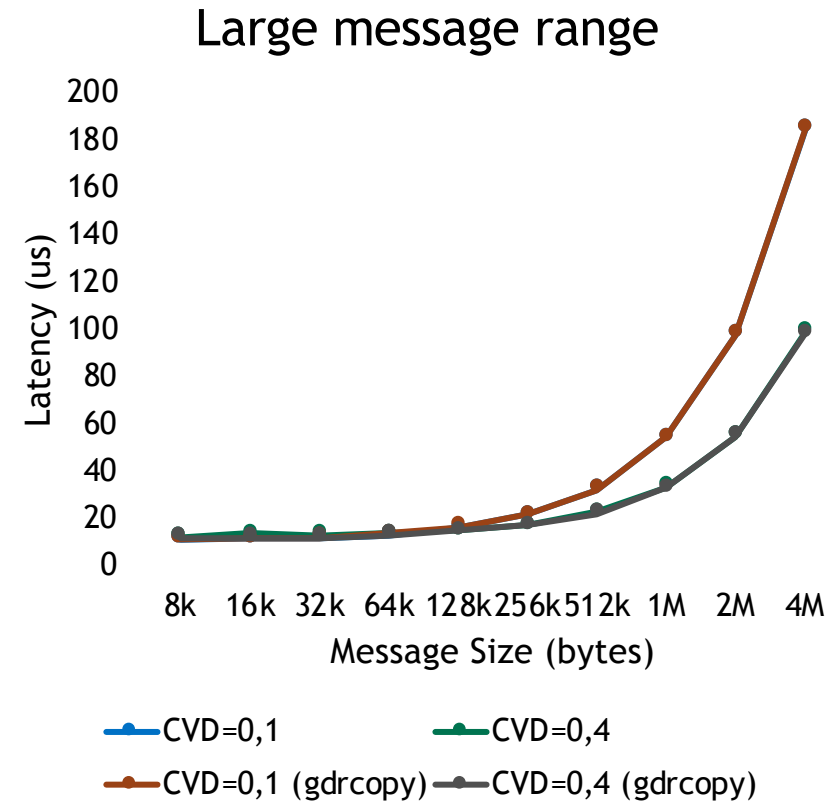
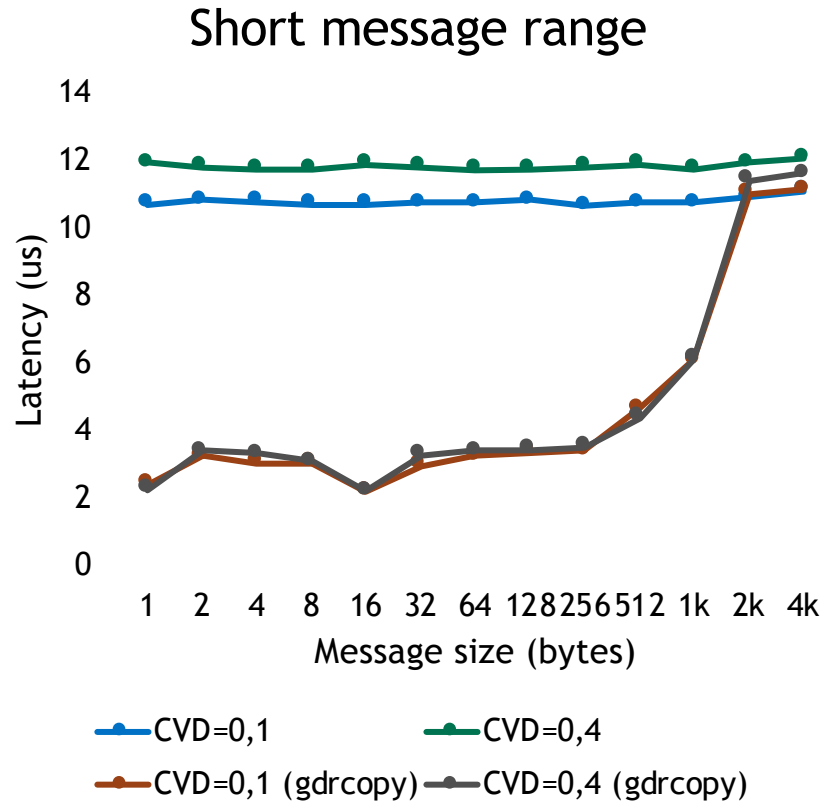
2 NVLINK2 between GPU 0 -> GPU 4 (CVD=0,4)

Network Software:

UCX 1.4 (CUDA-IPC, GDRCopy and SM UCTs used)

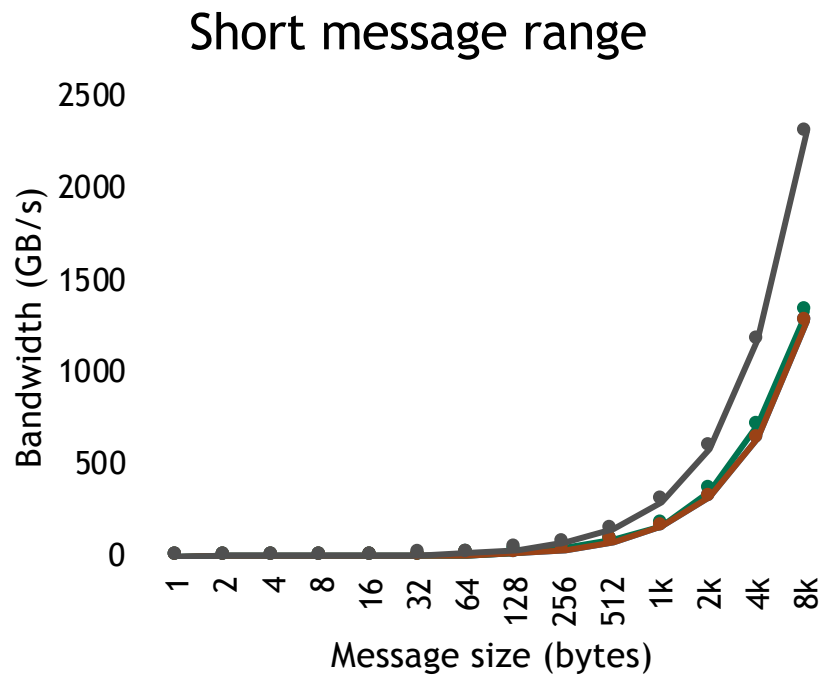
DEVICE BUFFER LATENCY

osu_latency; UCX_RNDV_SCHEME=put_zcopy

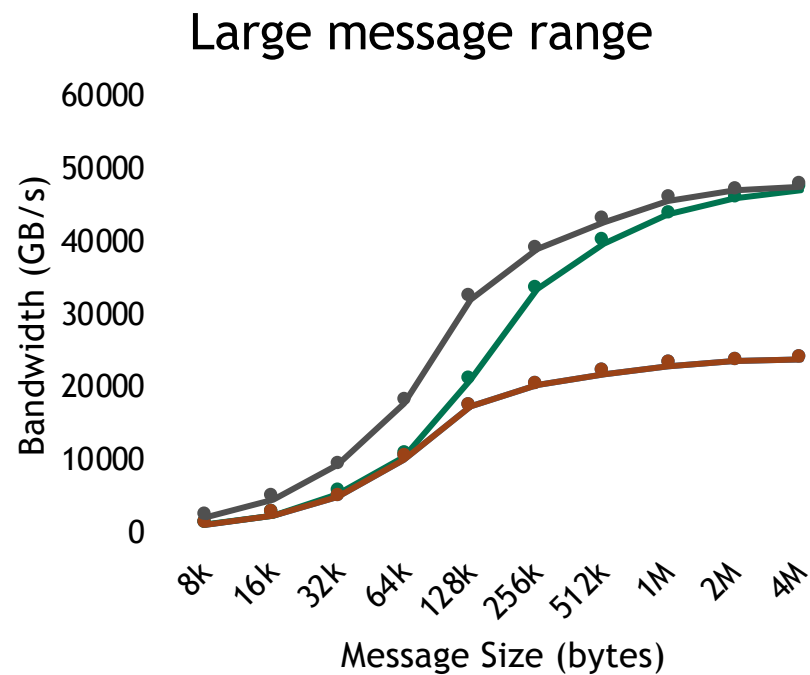


DEVICE BUFFER UNIDIRECTIONAL BANDWIDTH

osu_bw; UCX_RNDV_SCHEME=put_zcopy



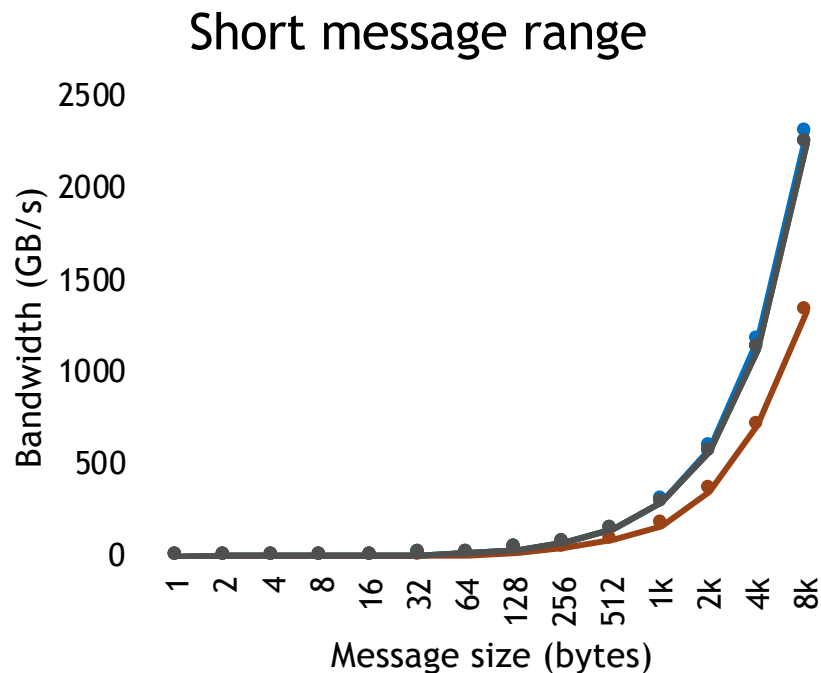
— CVD=0,1 — CVD=0,4
— CVD=0,1 (gdrcopy) — CVD=0,4 (gdrcopy)



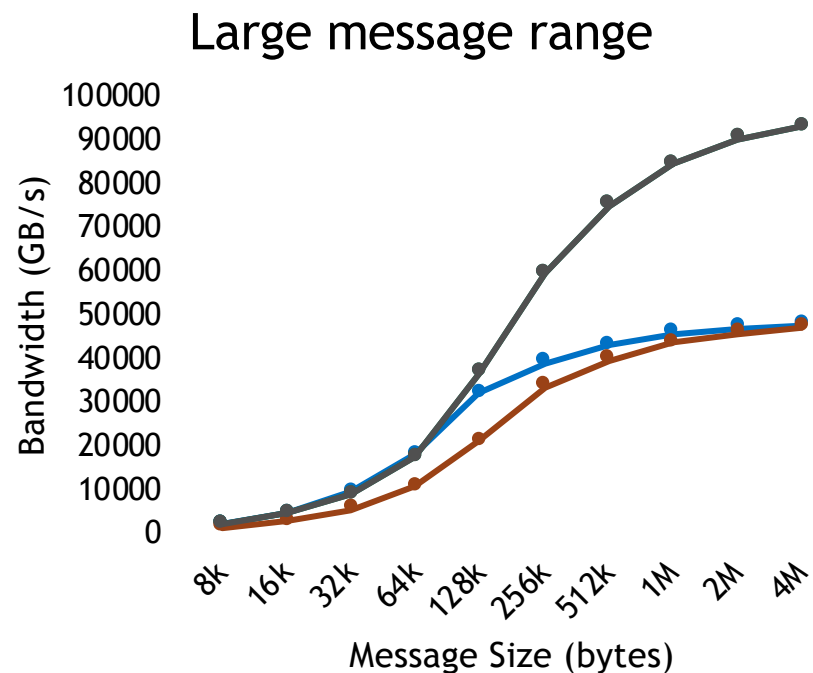
— CVD=0,1 — CVD=0,4
— CVD=0,1 (gdrcopy) — CVD=0,4 (gdrcopy)

DEVICE BUFFER BIDIRECTIONAL BANDWIDTH

`osu_bw; UCX_RNDV_SCHEME=put_zcopy`



— CVD=0,1 — CVD=0,4
— CVD=0,1 (gdrcopy) — CVD=0,4 (gdrcopy)



— CVD=0,1 — CVD=0,4
— CVD=0,1 (gdrcopy) — CVD=0,4 (gdrcopy)

FUTURE FEATURES

MPI Support:

- MPI-3 RMA

- Support MPI-3 non-blocking collectives like `lallreduce`/`lreduce`

Optimizations:

- GPU-HCA affinity for better multi-rail performance

- Managed Memory performance improvements

