GASNet-EX RMA Communication Performance on Recent Supercomputing Systems

Paul H. Hargrove and Dan Bonachea
gasnet.lbl.gov
doi.org/10.25344/S40C7D

Parallel Applications Workshop, Alternatives to MPI+X
Held in conjunction with SC22:
The International Conference for High Performance Computing, Networking, Storage, and Analysis

Outline

- Background
  - **GASNet-EX**
  - Methodology and Systems
- Performance Results
- Closing

SC22
Dallas, TX | hpc accelerates.

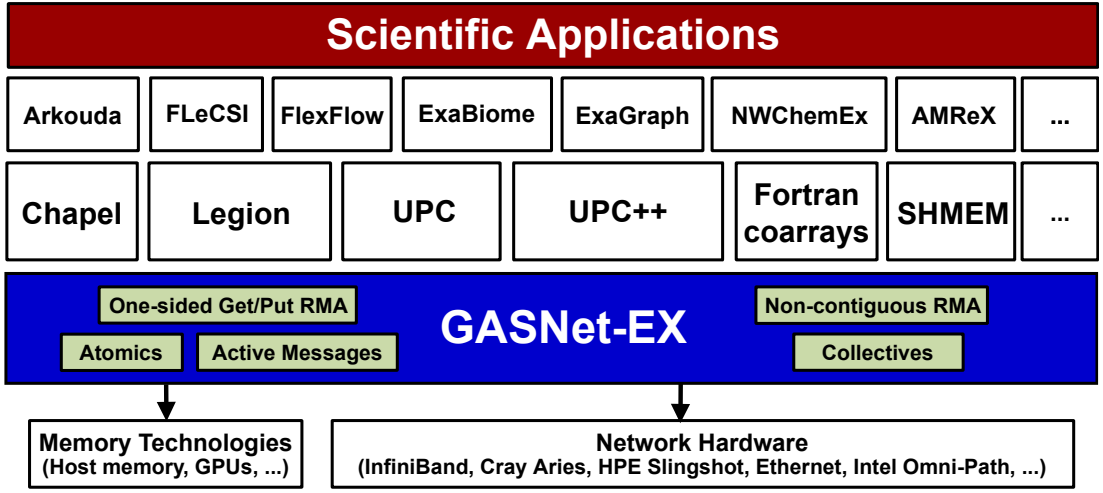
EXASCALE COMPUTING PROJECT
BERKELEY LAB
GASNet-EX

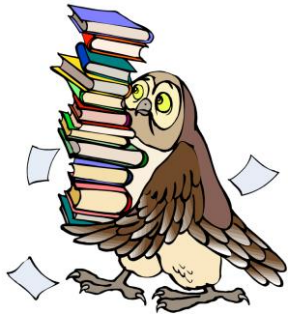
The Pagoda Project

https://go.lbl.gov/pagoda

Support for lightweight communication in exascale applications, frameworks and programming models:

- **GASNet-EX:** low-level communication layer that provides a network-independent interface suitable for Partitioned Global Address Space (PGAS) runtime developers
- **Berkeley UPC:** portable implementation of the UPC language over GASNet-EX
- **UPC++:** C++ PGAS library for application, framework and library developers, a productivity layer over GASNet-EX





GASNet-1: Overview

- Started in 2002 to provide a portable network communication runtime for three PGAS languages:
  - UPC, Fortran Coarrays and Titanium

- Primary features:
  - Non-blocking RMA (one-sided Put and Get)
  - Active Messages (a restricted form of RPC)
- Motivated by semantic issues in (then current) MPI RMA
  - Dan Bonachea, Jason Duell, "Problems with using MPI 1.1 and 2.0 as compilation targets for parallel language implementations", IJHPCN 2004. doi.org/10.25344/S4JP4B

GASNet-EX: Overview

- GASNet-EX is the next generation of GASNet
  - Provides Remote Memory Access (RMA) and Active Message (AM) interfaces for implementing Partitioned Global Address Space (PGAS) programming models
  - Updates GASNet-1 design to address the needs of newer programming models such as UPC++, Legion and Chapel
  - Incorporates 20 years of lessons-learned and focuses on the challenges of emerging exascale systems
  - Provides backward compatibility for GASNet-1 clients
- Motivating goals for GASNet-EX include:
  - Support more client asynchrony
  - Enable more client adaptation
  - Improve memory footprint
  - Improve threading support
  - Increase offload to network hardware
  - Support for device memory

GASNet: Adoption and Portability

Client runtimes

LBNL UPC++	Legion (Stanford/NVIDIA/…)	Caffeine
Berkeley UPC	Titanium	OpenSHMEM reference impl.
GCC/UPC	Rice Co-Array Fortran	Omni XcalableMP
Clang UPC	OpenUH Fortran coarrays	PARADISE++ Devastator
Chapel (Cray/HPE)	OpenCoarrays in GCC Fortran	At least 6 others known to us

Network conduits

OpenFabrics Verbs (**InfiniBand**)	IBM DCMF (BG/P)	Myricom GM (Myrinet)
Mellanox MXM and VAPI (InfiniBand)	IBM LAPI (Colony and Federation)	Dolphin SISCI
Cray uGNI (Gemini and **Aries**)	Cray Portals3 (Seastar)	Sandia Portals4
Intel PSM2 (Omni-Path)	SHMEM (Cray X1 and SGI Altix)	
IBM PAMI (BG/Q and others)	Quadric elan3/4 (QsNet I/II)	
UDP (any TCP/IP network)	**OFI/libfabric (Slingshot, Omni-Path)**	**Shared memory (no network)**
MPI 1.1 or newer	**UCX (multiple)**	

Supported platforms

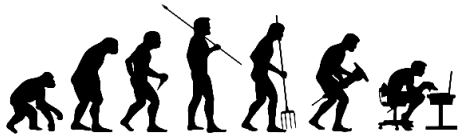
Over 10 compiler families, 15 operating systems and dozens of architectures

* These lists and counts include both current and past support

GASNet-EX: Some New Features

- Subset teams
- Local completion control
  - Explicit control over buffer lifetime to improve overlap
- Immediate-mode injection
  - Avoid stalls in low-resource conditions
- Negotiated-payload Active Messages
  - Construct messages in GASNet's buffers to avoid `memcpy()`
- Remote atomic operations
  - Utilize offload capabilities in modern network interfaces
- Device memory RMA (e.g. GPUs)
  - Offload data movement via PCI peer-to-peer transfer technologies

GASNet-EX: Status



- GASNet-EX is still evolving
  - New features on the previous slide are implemented
  - They bring benefits demonstrated in prior work
  - Additional capabilities to appear in the future
- Several programming models have adopted GASNet-EX
  - UPC++ and Berkeley UPC require GASNet-EX
  - Legion has a new backend to use EX-specific features
  - Chapel embeds GASNet-EX
  - Many others (see `gasnet.lbl.gov`)
- Delivers excellent performance on current systems

Outline

- Background
  - GASNet-EX
  - **Methodology and Systems**
- Performance Results
- Closing

SC22
Dallas, TX | hpc accelerates.

EXASCALE COMPUTING PROJECT
BERKELEY LAB
GASNet-EX

Experimental Methodology

Two RMA performance metrics

- Flood bandwidth
- Latency

GASNet tests contained in the 2022.3.0 release

- **`testlarge`** for flood bandwidth
- **`testsmall`** for latency

MPI tests from the Intel MPI Benchmarks (IMB) v2021.3

- **`IMB-RMA`** subtests **`Unidir_put`** and **`Unidir_get`**
- **`IMB-MPI1`** subtests **`Uniband`** and **`PingPong`**
- Built with each system's default version of the vendor-provided MPI

See also the reproducibility appendix and

Bonachea D, Hargrove P. GASNet-EX: A High-Performance, Portable Communication Library for Exascale, Proceedings of Languages and Compilers for Parallel Computing (LCPC'18). Oct 2018. doi.org/10.25344/S4QP4W

Production HPC Systems Evaluated

OLCF Summit

- IBM Power 9 GPUs — IBM AC922 / Mellanox InfiniBand network
- IBM Spectrum MPI — GASNet-EX ibv-conduit

NERSC Cori Haswell

- Intel Haswell CPUs — Cray XC-40 / Aries network
- Cray MPI — GASNet-EX aries-conduit

NERSC Perlmutter SS-10

- AMD Milan CPUs — HPE Cray EX / Slingshot-10 (Mellanox ConnectX-5 NICs)
- HPE Cray MPI — GASNet-EX ofi-conduit over libfabric verbs provider

NERSC Perlmutter SS-11

- AMD Milan CPUs — HPE Cray EX / Slingshot-11 (HPE Cassini NICs)
- HPE Cray MPI — GASNet-EX ofi-conduit over libfabric cxi provider

Outline

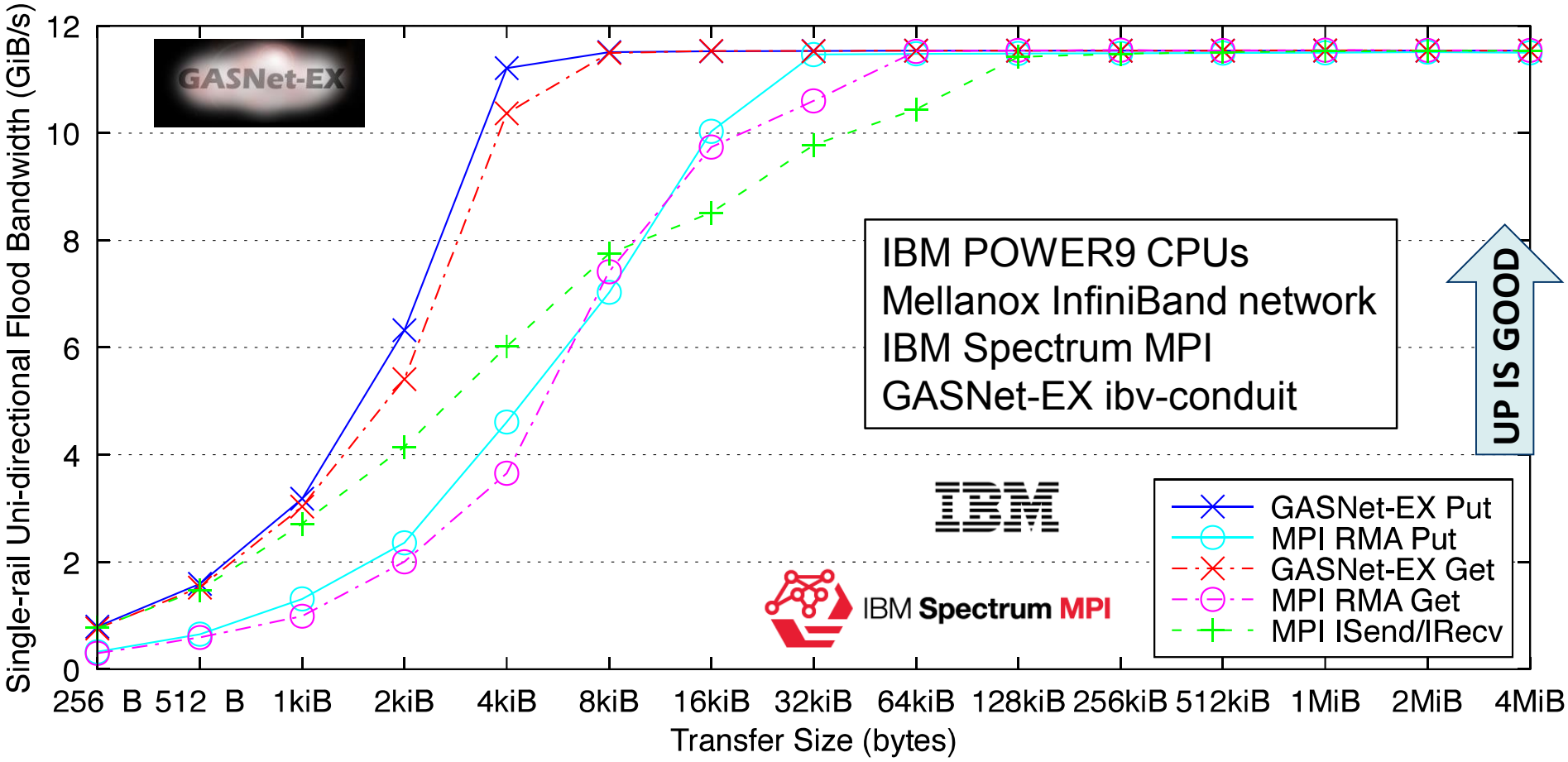
- Background
  - GASNet-EX
  - Methodology and Systems
- **Performance Results**
- Closing

SC22
Dallas, TX | hpc accelerates.

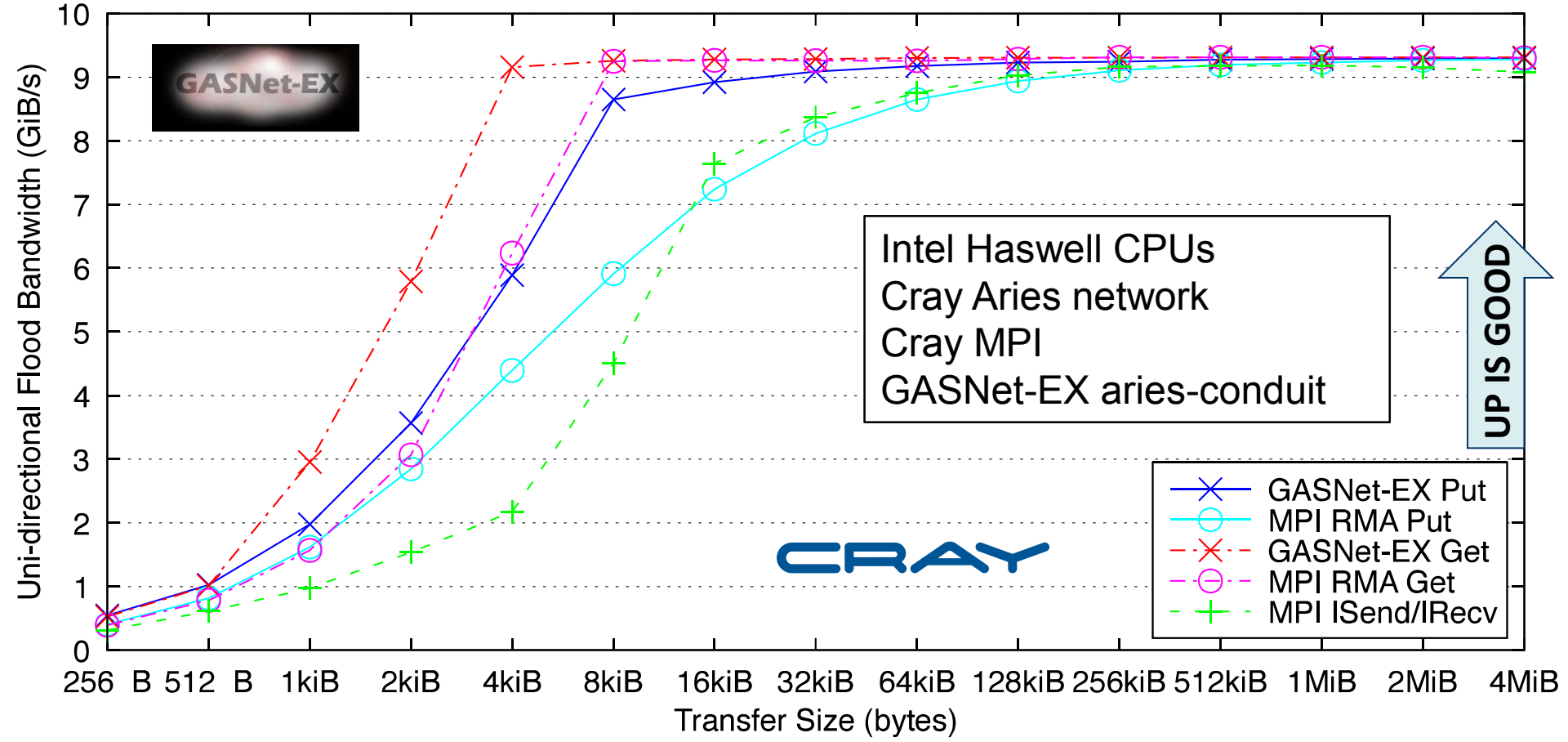
EXASCALE COMPUTING PROJECT
BERKELEY LAB
GASNet-EX

Bandwidth on OLCF Summit

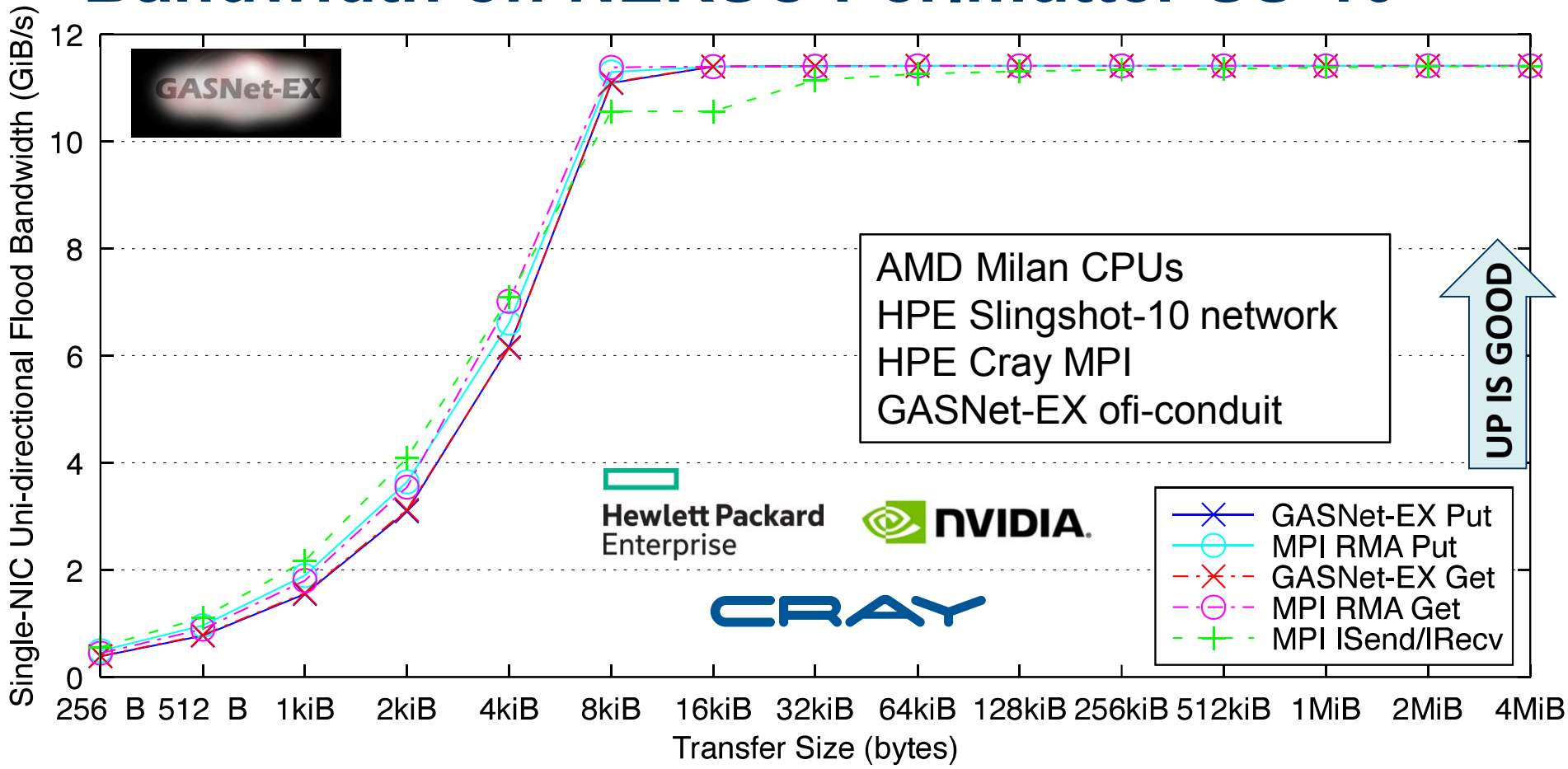
OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY



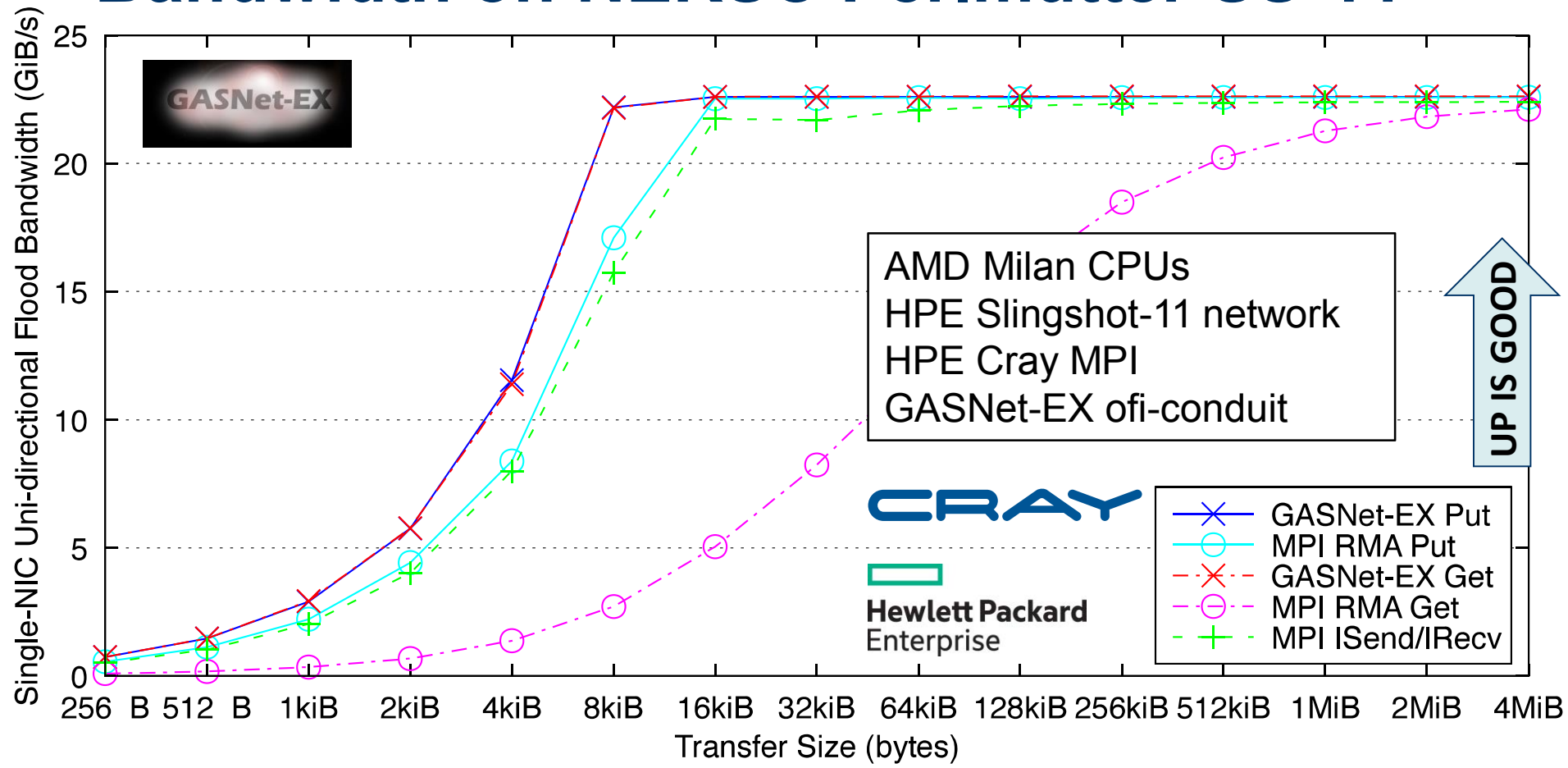
Bandwidth on NERSC Cori Haswell



Bandwidth on NERSC Perlmutter SS-10



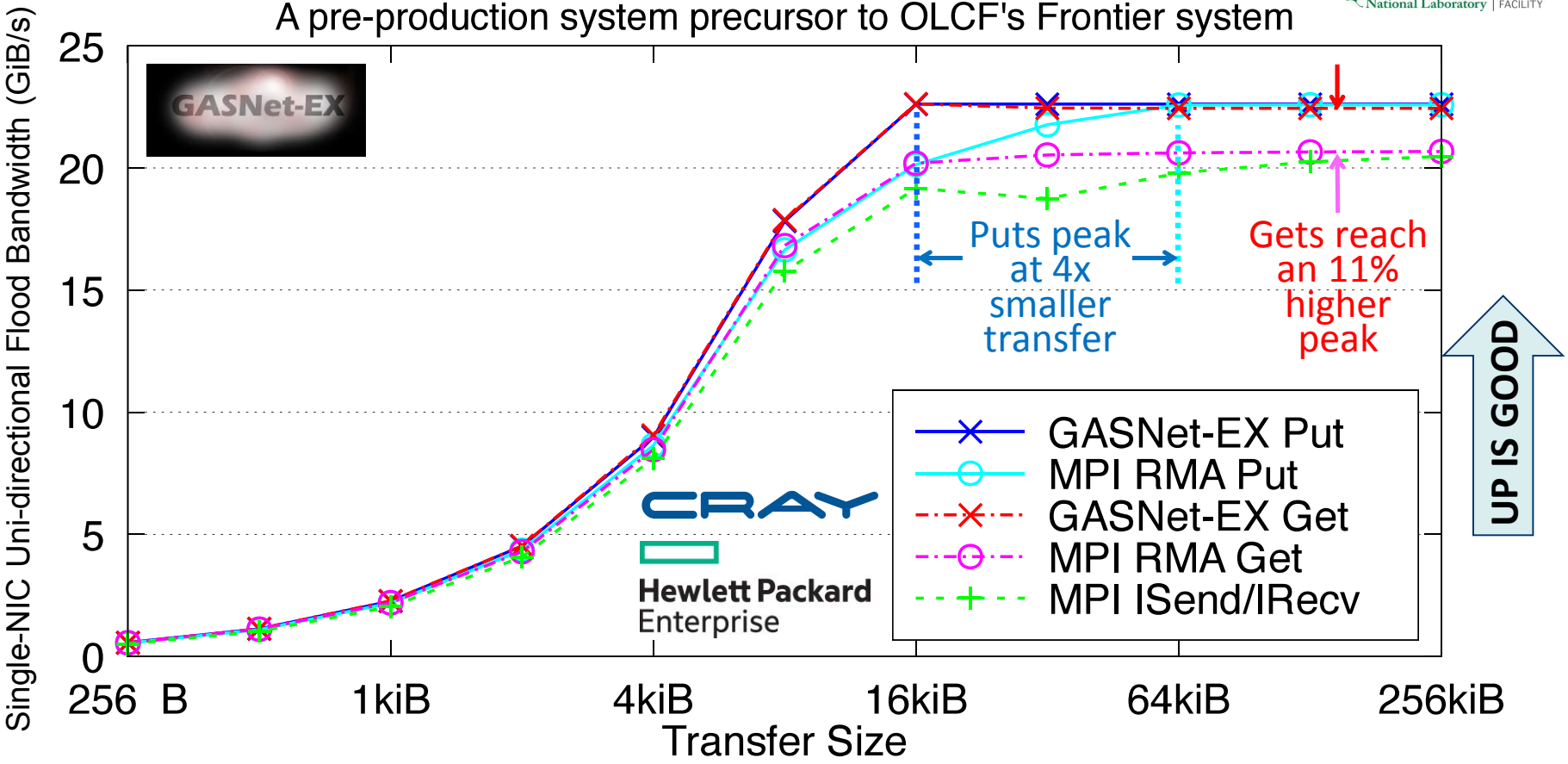
Bandwidth on NERSC Perlmutter SS-11



Crusher: HPE Cray EX / Slingshot-11, HPE Cray MPI

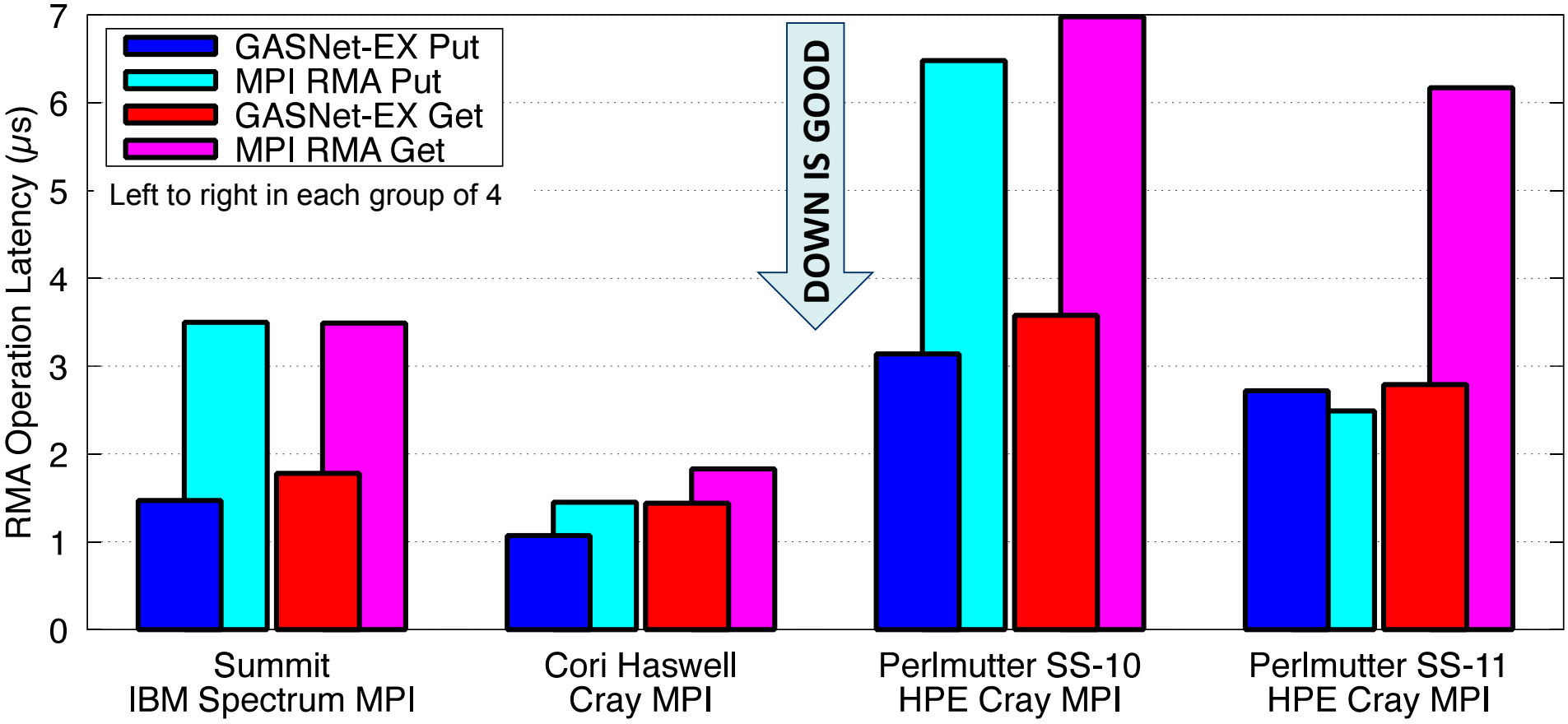
A pre-production system precursor to OLCF's Frontier system

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY



A comparison of uni-directional point-to-point host-memory flood bandwidth benchmarks, run March 2022 on OLCF's Crusher system.
Shows the performance of RMA (Put and Get) operations using GASNet-EX and both RMA and message-passing (Isend/Irecv) using HPE Cray MPI.
Results were obtained using then current GASNet tests and Intel MPI Benchmarks, respectively.

Eight-byte RMA Latency



Outline

- Background
  - GASNet-EX
  - Methodology and Systems
- Performance Results
- **Closing**

SC22
Dallas, TX | hpc accelerates.

EXASCALE COMPUTING PROJECT
BERKELEY LAB
GASNet-EX

In Conclusion…

- GASNet-EX is a widely adopted communication library in PGAS programming model implementations
- RMA microbenchmarks comparing GASNet-EX to vendor's MPI on four representative systems show†
  - GASNet-EX bandwidth outperforms the equivalent MPI RMA operations
    - Up to 2.7x faster for Puts and up to 3.1x faster for Gets
    - GASNet-EX reaches peak bandwidth at up to 8x smaller transfer sizes
  - For small-transfer latency
    - GASNet-EX RMA outperforms MPI RMA by up to 2.38x

† The anomalously poor outlier behavior of MPI RMA Get on Perlmutter SS-11 has been excluded from this summary.

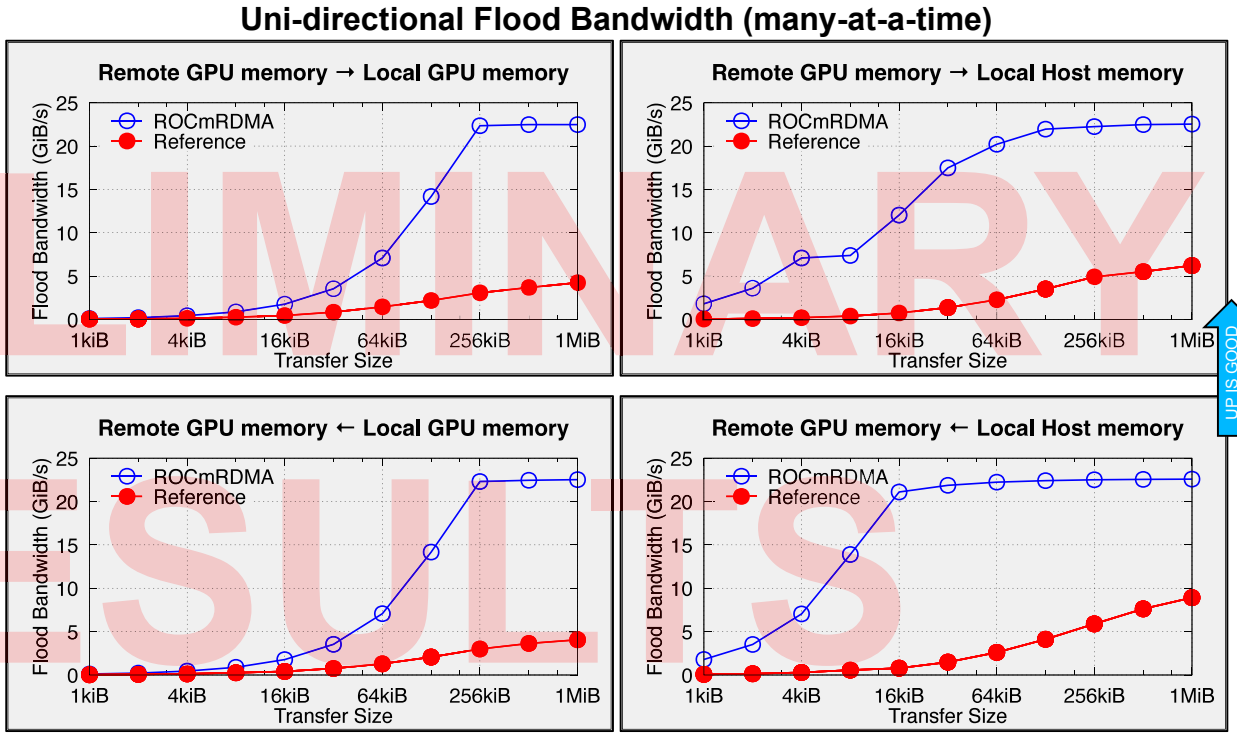
Future Work

- Continued development on the Slingshot-11 network
  - The least mature networking stack
  - Anomalous performance of MPI Get on Perlmutter SS-11
    - Similar on Crusher (though less severe)
  - MPI Puts slightly faster than GASNet-EX
    - Investigate this difference

- Benchmarking of device memory RMA
  - Preliminary results on Crusher on next slide

PREVIEW: GPU Memory RMA on OLCF Crusher†

- GASNet-EX 2022.9.0 added support for zero-copy communication involving GPU memory over HPE Slingshot networks
  - Direct data movement between the NIC+GPU (PCI peer-to-peer transfer)
  - Prior "reference" implementation staged transfers through host memory
- A UPC++-level benchmark shows the benefits of this optimization relative to staging through host memory:
  - Left-hand plots: bandwidth to/from local GPU memory 2.9x to 7.4x better
  - Right-hand plots: bandwidth to/from local host memory 2.4x to 34x better



Results were collected using the `gpu_microbenchmark` test from the 2022.9.0 release of UPC++, run between two nodes of OLCF Crusher, over its Slingshot-11 network using one process per node and one NIC per process.

† Crusher is a preproduction system precursor to OLCF's Frontier system.

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

CRAY

Hewlett Packard Enterprise

ECP EXASCALE COMPUTING PROJECT

BERKELEY LAB

GASNet-EX

Acknowledgements

- This research was funded in part by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration.
- This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.
- This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

THANK YOU

gasnet.lbl.gov

doi.org/10.25344/S40C7D

Upcoming roundtable discussion:
"Pagoda: UPC++ and GASNet-EX for Lightweight Communication and Global Address Space Support"
Tue Nov 15 2:00p CST in DOE Booth (#1600)