Hewlett Packard
Enterprise

DESIGNING THE HPE CRAY MESSAGE PASSING TOOLKIT SOFTWARE STACK FOR HPE CRAY EX SUPERCOMPUTERS

Krishna Kandalla
Kim McMahon
Naveen Ravi
Trey White
Larry Kaplan
Mark Pagel

May 9, 2023

AGENDA

INTRODUCTION & BACKGROUND

HPE CRAY MPI DESIGN AND IMPLEMENTATION DETAILS

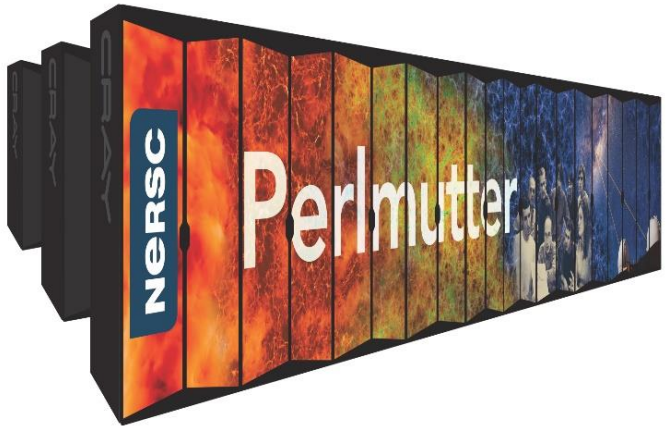
PERFORMANCE EVALUATION

TIPS & TRICKS FOR DEBUGGING AND TUNING HPC APPLICATIONS ON HPE CRAY EX SYSTEMS

Q & A

INTRODUCTION: HPE CRAY EX SUPERCOMPUTERS

- High density compute and high performance networking have ushered in the Exascale computing era
- ORNL's Frontier supercomputer achieved 1.102 Exaflop/s and relies on the HPE Cray EX architecture
- Frontier TDS "Crusher" system, the #1 spot for the GREEN500 list (June 2022), is also an HPE Cray EX
- Other notable large scale HPE EX systems in the Top500 list
  - LUMI (#3), NERSC (#8), CINES (#11), Pawsey (#15), ExxonMobile (#16)



INTRODUCTION: HPE CRAY MPI

- Software stacks on HPC systems must address several critical challenges: *ease-of-programming, portability, communication/synchronization efficiency, scaling*
- Majority of pre-exascale and exascale applications rely on the Message Passing Interface (MPI) programming model
- HPE Cray Message Passing Toolkit offers the primary MPI & SHMEM implementations on HPE Cray EX systems
- HPE Cray MPI was instrumental in achieving the 1.102 Exaflop/s performance barrier on the Frontier supercomputer!
- HPE Cray MPI has also enabled several ECP, DOE, and CAAR applications on Frontier and other systems

INTRODUCTION: HPE CRAY EX NETWORK ARCHITECTURE

- HPE Slingshot Rosetta switch offers industry leading performance and scalability
  - Supports optimized HPC functionality in addition to Ethernet protocols
  - High radix 64-port switch offers 12.8 Tb/s bandwidth, supports 100Gbps and 200 Gbps network adapters
  - Networks based on Rosetta can connect more than 250,000 compute nodes with a maximum of 3 hops
  - Rosetta offers support for congestion management, adaptive routing, quality of service, collective acceleration
- HPE Slingshot-10 consists of NVIDIA ConnectX-5 RoCE (100 Gbps) NICs and HPE Rosetta switches
- HPE Slingshot-11 consists of HPE Cassini (200 Gbps) NICs and HPE Rosetta switches
  - HPE Cassini NICs support reliable, connectionless communication protocols
  - HPE Cassini offers several capabilities in hardware: tag-matching, MPI Rendezvous protocol progression, hardware triggered operations, counting events, atomics, and traffic classification
- Both HPE Slingshot-10 and Slingshot-11 system configurations support multiple NICs per compute node and heterogeneous (CPU/GPU) architectures

AGENDA

INTRODUCTION & BACKGROUND

HPE CRAY MPI DESIGN AND IMPLEMENTATION DETAILS

PERFORMANCE EVALUATION

TIPS & TRICKS FOR DEBUGGING AND TUNING HPC APPLICATIONS ON HPE CRAY EX SYSTEMS

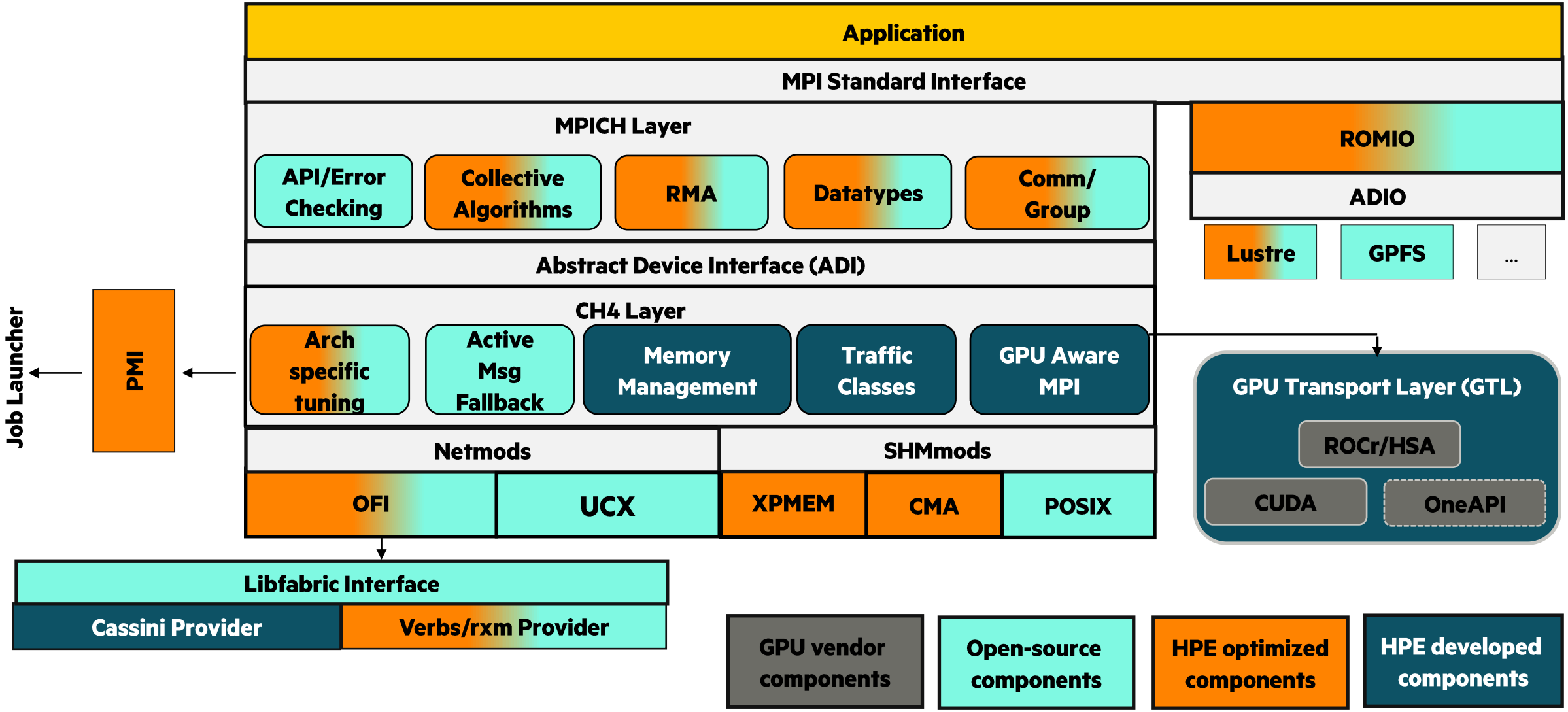
Q & A

HPE CRAY MPI COVERAGE FOR HPE EX AND HPE APOLLO SYSTEMS

- Based on open-source ANL MPICH 3.4a2 (CH4), compliant with the MPI-3.1 Standard
- Current HPE EX/Apollo Release – HPE Cray MPI 8.1.26

FEATURE	COVERAGE
CPU architectures	Intel, AMD, NVIDIA Grace (under development)
Network architectures	Slingshot-10: OFI (verbs; rxm) provider, UCX driver support Slingshot-11: OFI (CXI) provider HPE Apollo InfiniBand clusters: OFI (verbs;rxm), UCX driver support
GPU architectures	AMD, NVIDIA, Intel (under development)
Operating Systems	RHEL/CENTOS, SLES, and COS
PrgEnv and compilers	PrgEnv-cray, PrgEnv-gnu, PrgEnv-nvidia, PrgEnv-amd, PrgEnv-aocc, and PrgEnv-intel
Launcher support	Slurm , PALS, and Flux

HPE CRAY MPI SOFTWARE ARCHITECTURE



HPE CRAY MPI – SOFTWARE FEATURES FOR HPE EX & APOLLO SYSTEMS

- Based on open-source MPICH (CH4 device) from Argonne National Laboratory
- XPMEM (supported only on COS systems) and Cross Memory Attach (CMA) for on-node single-copy transfers
- Tuned and optimized algorithms for select collectives across different payload and communicator sizes
- Robust affinity support for multiple NICs per node (each MPI rank uses one NIC)
- Highly optimized for NVIDIA and AMD GPUs for HPE EX systems (Intel GPU support is in progress)
- Improved MPI_THREAD_MULTIPLE behavior via optimized thread synchronization mechanisms
- Support for different size memory allocations:
  - Standard Linux page sizes including community huge pages – both Transparent Huge Pages and hugetlbfs (preferred)
  - Non-standard huge page sizes (based on hugetlbfs and available only in COS)
- Support for PMI2 and Cray PMI with Slurm and Parallel Application Launch Service (PALS)
- Flexible, intuitive rank reordering feature
- MPI I/O performance enhancements and statistics
- ABI Compatibility:
  - HPE Cray MPI is ABI compatible with open-source MPICH implementation from the Argonne National Laboratory
  - MPIxlate: ABI translator for MPI programs, enables applications compiled using an MPI library that is not binary compatible with HPE Cray MPI, to be run without recompilation on supported HPE Cray systems

KEY FEATURES IN HPE CRAY MPI FOR HPE SLINGSHOT-11 SYSTEMS

- Tuned for HPE Slingshot-11 network architecture: HPE Cassini NICs and HPE Rosetta switches
- Support for Slingshot-11 traffic classes
- Benefits from Slingshot-11 hardware support for congestion management
- Leverages Slingshot-11 hardware offload for MPI tag-matching and rendezvous protocol
- Highly scalable implementation using connectionless protocols with small memory footprint
- Efficient communication protocols for small message data transfers
- HPE Cray MPI will soon leverage Slingshot-11 hardware support for collective operations
- Tight integration between HPE Slingshot-11 network and GPU architectures:
  - Supports RDMA capabilities for inter-node data movement operations involving GPU-attached application buffers
  - GPU-NIC Async prototypes to facilitate development of next generation GPU-enabled applications

SUPPORT FOR TRAFFIC CLASSIFICATION ON SLINGSHOT-11 SYSTEMS

- Traffic Classes control network properties and enable predictable performance and network utilization
- HPE recommends a set of “Best Practice” Traffic Classes for HPC applications (currently max 3 of these available per site):
  - **Low Latency**
    - –High priority, low latency, low jitter class, bandwidth capped
    - –Ideal for barrier synchronization and small-msg collectives ops
  - **Dedicated Access**
    - –Intended for high priority jobs, guaranteed bandwidth allocation
  - **Bulk Data**
    - –Isolates I/O and long transfers from other classifications
  - **Best Effort**
    - –Default shared class, provides ‘fair-share’ within the class
- Users can request a specific traffic class via the MPICH OFI DEFAULT TCLASS environment variable
  - Not every traffic class will be available on every HPE Slingshot-11 system (site dependent)
  - Specific traffic classes also require prior Workload Manager (WLM) authorization

GPU SUPPORT IN HPE CRAY MPI

- Users can request “GPU Aware” MPI support via the MPICH_GPU_SUPPORT_ENABLED runtime variable
  - Most GPU Aware MPI capabilities are available on both HPE Slingshot-10 and HPE Slingshot-11 systems
- Intra-node GPU-GPU Peer-to-Peer IPC
  - Intra-GPU, GPU <-> GPU, and CPU <-> GPU transfers
  - HPE Cray MPI supports AMD and NVIDIA GPUs on Slingshot systems
    - Internal prototypes for Intel GPUs are being evaluated
- Inter-node GPU-NIC RDMA (otherwise known in the industry as “GPU Direct RDMA”)
  - Enables direct transfers between NIC and GPU without requiring copies through CPU-attached staging buffers
  - HPE Cray MPI supports AMD and NVIDIA GPUs on SS-11 systems
    - Support for Intel GPUs is under development
- GPU-NIC Async
  - Decouples CPU / GPU control and data paths
  - Reduces frequency and overheads of CPU / GPU synchronization points
  - Potentially improves utilization of all three critical resources on the compute nodes: CPU, GPU, and NIC
  - New MPI APIs and API extensions are being developed
    - Requires application-level changes
  - Currently in research and early prototype phase only on HPE Slingshot-11 systems

AGENDA

INTRODUCTION & BACKGROUND

HPE CRAY MPI DESIGN AND IMPLEMENTATION DETAILS

PERFORMANCE EVALUATION

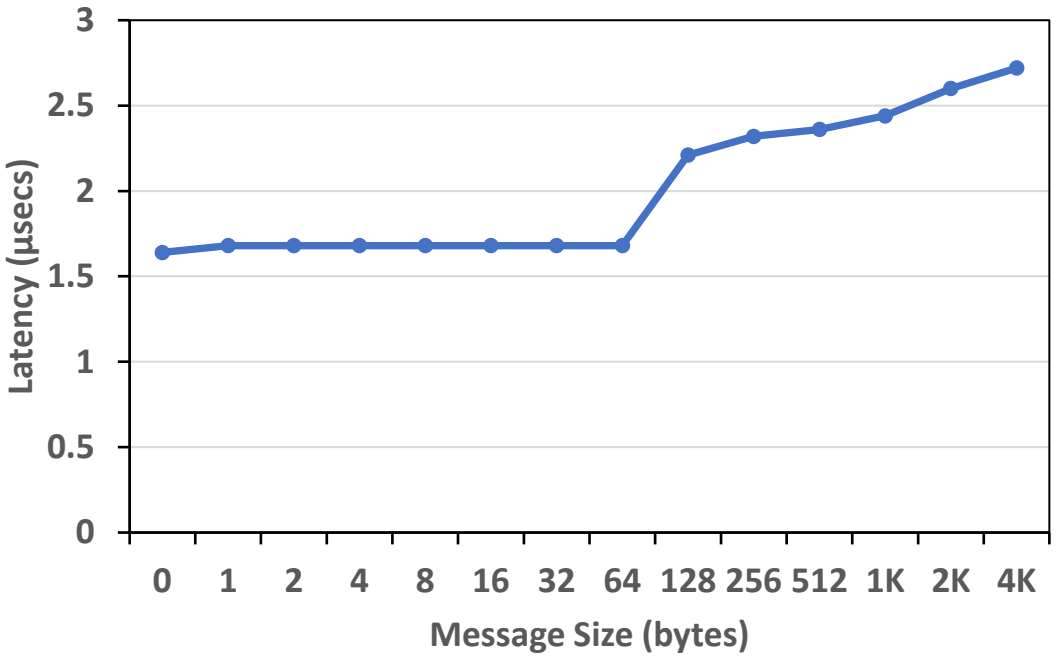
TIPS & TRICKS FOR DEBUGGING AND TUNING HPC APPLICATIONS ON HPE CRAY EX SYSTEMS

Q & A

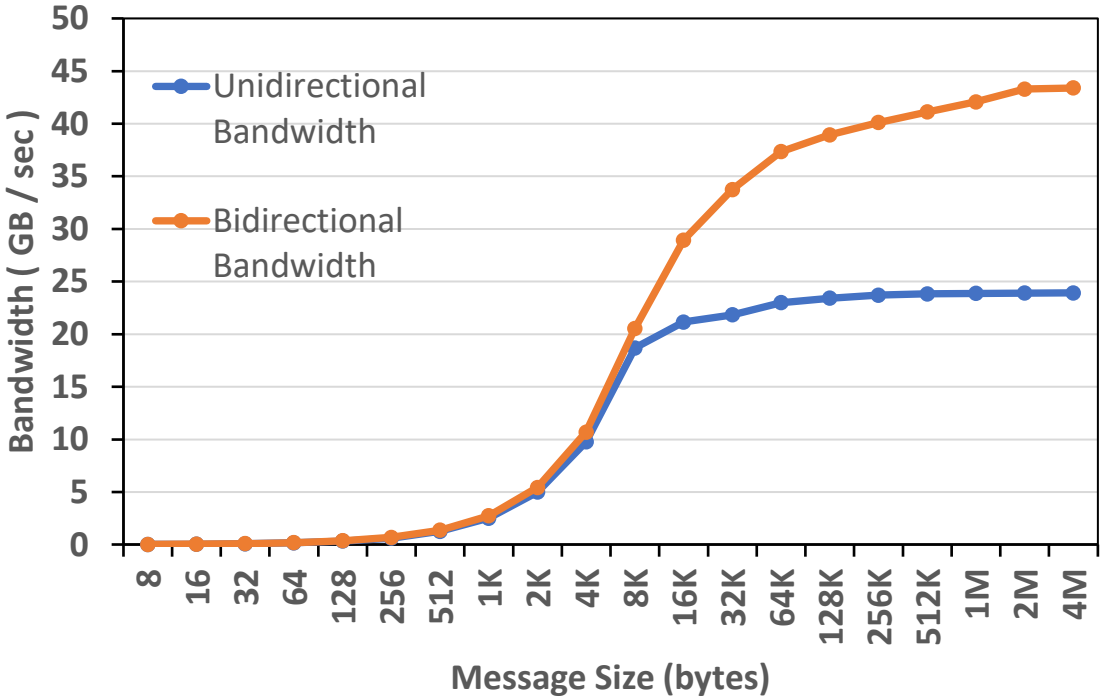
MPI POINT-TO-POINT PERFORMANCE STUDIES ON WINDOM ARCHITECTURE

- HPE Cray MPI behavior on “Windom” systems based on AMD Milan CPUs & HPE Slingshot-11 network
- Experiments use “near” nodes attached to the same switch

MPI Point to Point Latency
AMD EPYC 7763 64-Core Processor (Milan)
Single Cassini NIC

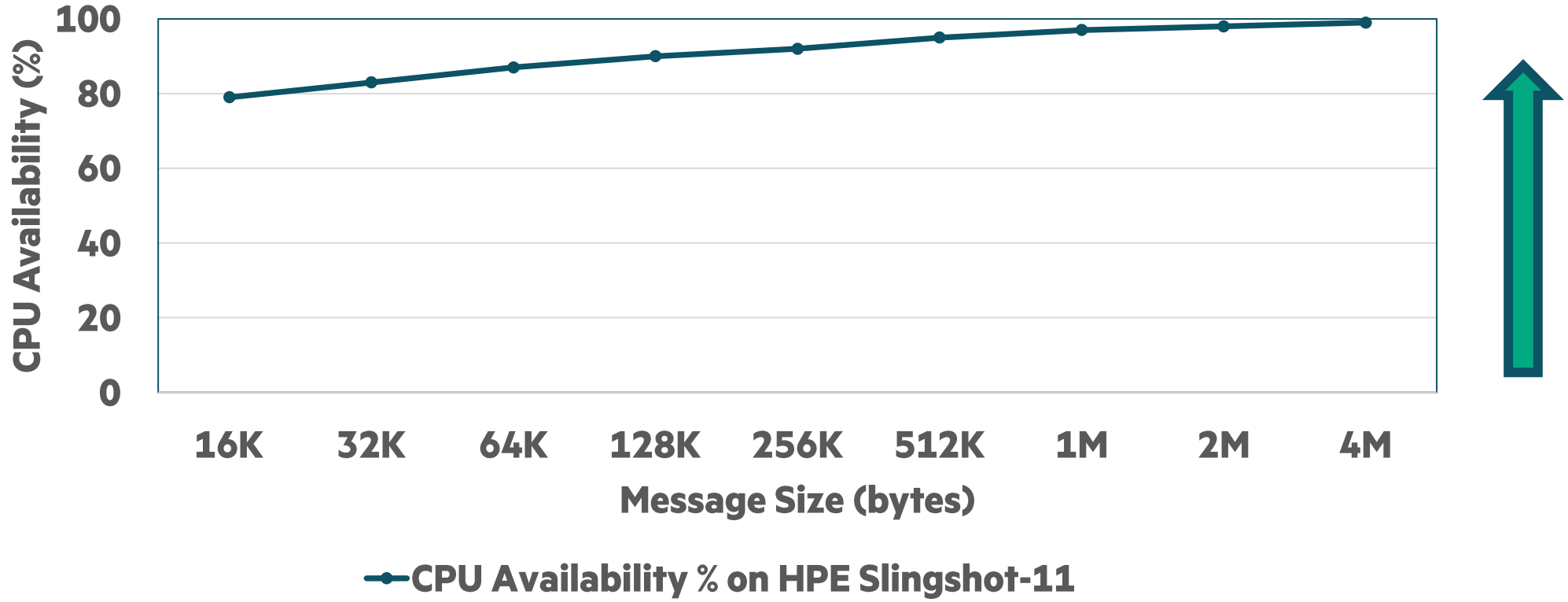


MPI Point to Point Bandwidth
AMD EPYC 7763 64-Core Processor (Milan)
Single Cassini NIC



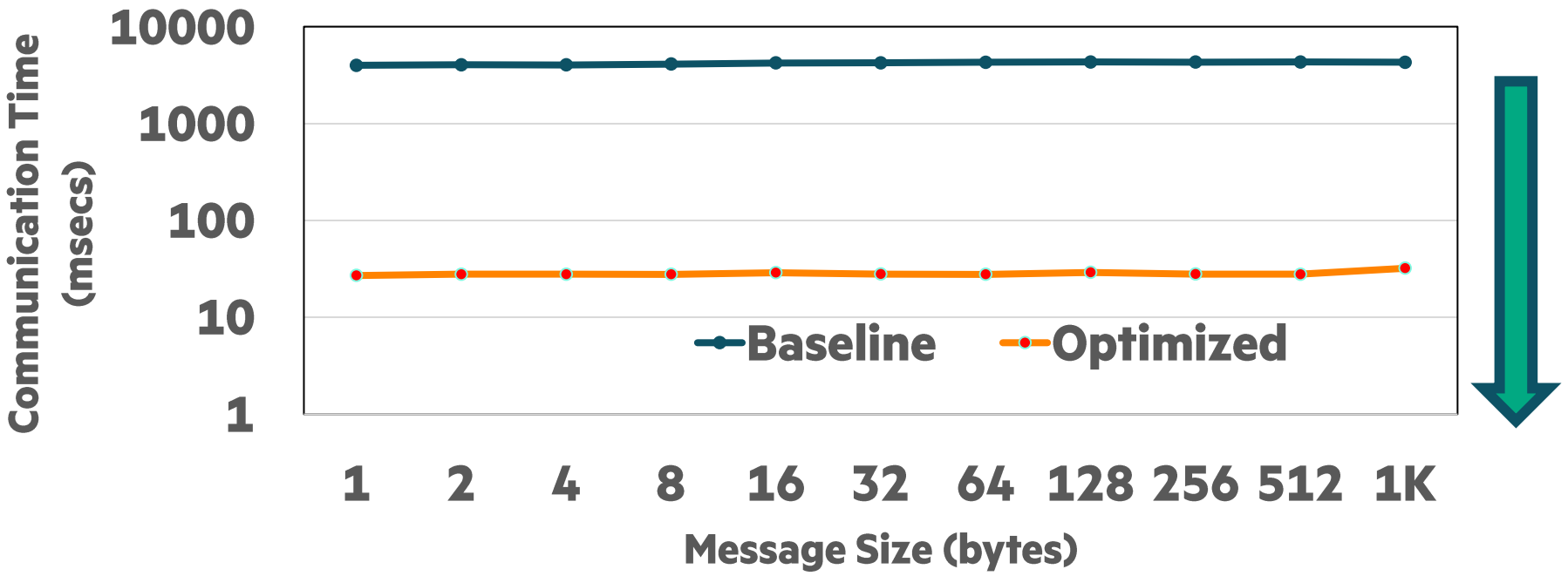
COMPUTATION/COMMUNICATION OVERLAP (HPE SLINGSHOT-11)

- Computation/communication overlap with HPE Cray MPI on systems using HPE Slingshot-11 network
  - Results are based on Sandia Overlap Benchmark
  - Hardware offload capabilities in HPE Cassini NICs results in 100% CPU availability for processes performing large payload MPI_Irecv operations



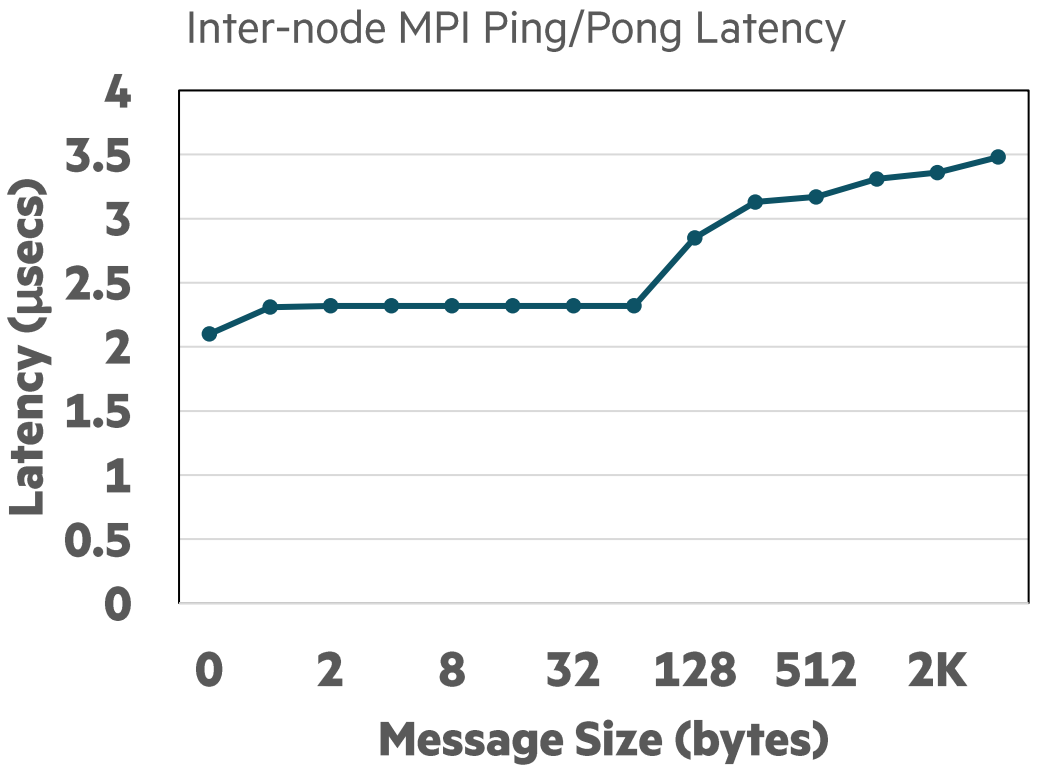
MPI_IGATHERV OPTIMIZATIONS

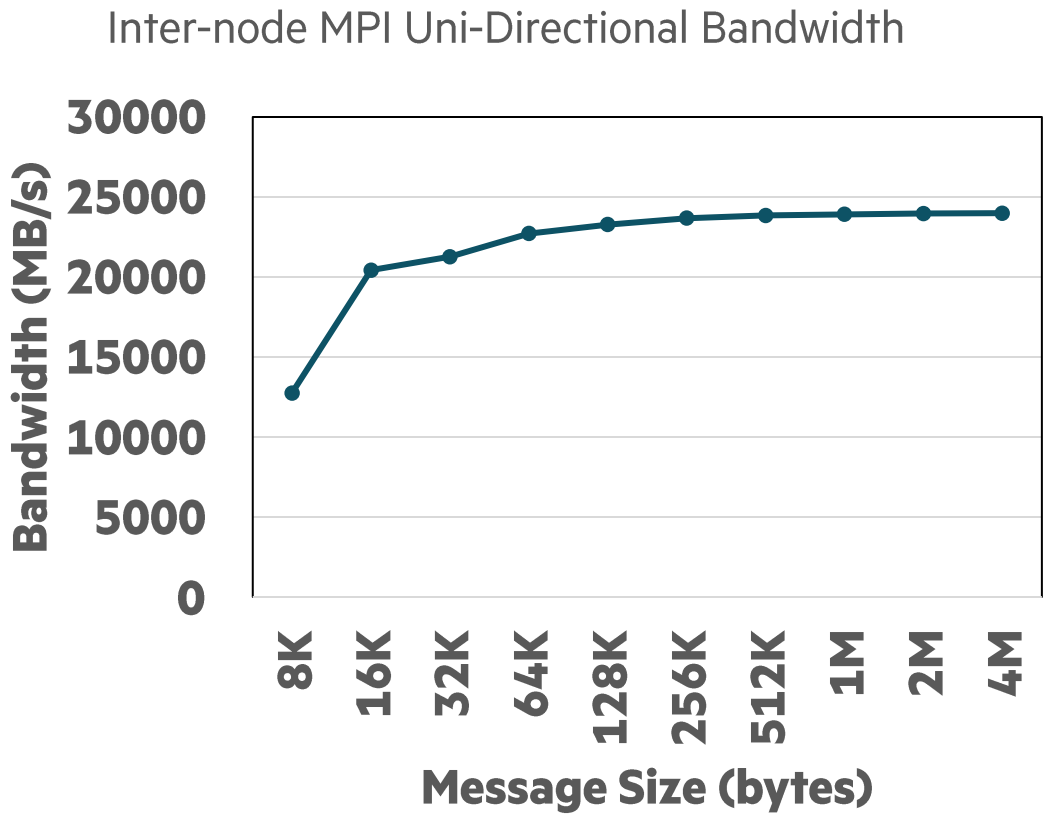
- HPE Cray MPI offers a highly optimized software implementation for MPI_Igatherv
  - Available on both HPE Slingshot-10 and HPE Slingshot-11 systems
- Chart compares MPI_Igatherv latency on a 512-node system with 128 Processes Per Node
  - Each compute node has 2 AMD Milan CPU sockets and 1 HPE Slingshot-11 Cassini NIC
- Optimized implementation improves performance by about **130X** (enabled by default)



MPI POINT-TO-POINT PERFORMANCE STUDIES ON BARD PEAK

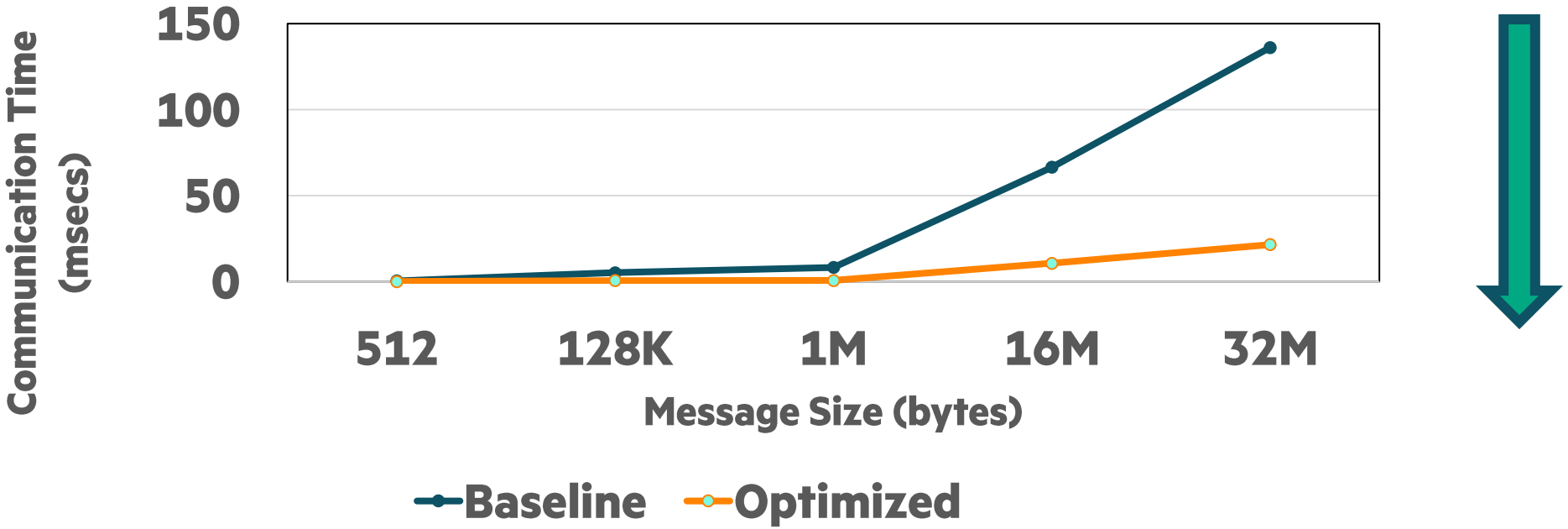
- HPE Cray MPI behavior on “Bard Peak” node architecture and HPE Slingshot-11 network
  - AMD Trento CPU + 4xMI250x GPUs
  - Communication buffers are on GPU-attached High Bandwidth Memory
  - Experiments use “near” nodes attached to the same switch





MPI_REDUCE_SCATTER_BLOCK OPTIMIZATIONS

- HPE Cray MPI offloads compute intensive phases of Reductions to GPU devices
  - Available on both HPE Slingshot-10 and HPE Slingshot-11 systems that also offer GPUs on compute nodes
- MPI_Reduce_scatter_block latency on an 8-node “Bard Peak” system with 8 Processes Per Node
  - 8 nodes are attached to the same HPE Slingshot switch
- Optimized implementation improves performance by about **7X** (enabled by default)
  - Similar optimization also available on systems with NVIDIA GPUs



AGENDA

INTRODUCTION & BACKGROUND

HPE CRAY MPI DESIGN AND IMPLEMENTATION DETAILS

PERFORMANCE EVALUATION

TIPS & TRICKS FOR DEBUGGING AND TUNING HPC APPLICATIONS ON HPE CRAY EX SYSTEMS

Q & A

TIPS & TRICKS FOR TUNING AND DEBUGGING HPE CRAY MPI

Problem	HPE Cray MPI Recommendations
Slingshot-11 Network Timeouts Events • Link flaps may cause network timeouts, packets are automatically reissued • Some applications may experience lower than expected MPI performance • MPI tracks network timeout events for each job • If timeouts are observed, HPE Cray MPI displays following message during MPI_Finalize: **"[MPICH Slingshot Network Summary: *N* network timeouts]"**	• HPE Cray MPI enables collection of additional Cassini hardware counter info (Refer to MPICH_OFI_CXI_COUNTER_REPORT in HPE Cray MPI man pages) • Analyzing these statistics can explain performance issues due to link flaps
Cassini Hardware Resource Exhaustion (Message Matching) • For specific communication patterns, a few MPI processes may run out of Cassini hardware resources used for message matching • Such applications fail with the following error message: **"PtlTE NN LE resources not recovered during flow control. FI_CXI_RX_MATCH_MODE = [hybrid | software] is required"**	• These applications may benefit from the hybrid match mode • If applicable, reorganizing communication pattern to reduce use of many-to-one patterns may help lower unexpected messages for specific MPI processes

AGENDA

INTRODUCTION & BACKGROUND

HPE CRAY MPI DESIGN AND IMPLEMENTATION DETAILS

PERFORMANCE EVALUATION

TIPS & TRICKS FOR DEBUGGING AND TUNING HPC APPLICATIONS ON HPE CRAY EX SYSTEMS

Q & A

THANK YOU

Larry Kaplan (lkaplan@hpe.com)

BACKUP SLIDES

MULTI-NIC SUPPORT IN HPE CRAY MPI

- All NICs are configured on the same network
  - Each MPI rank will be assigned to a single NIC
  - MPI does not a support a single rank striping data across multiple NICs
  - Multiple ranks and/or applications can share a single NIC

- HPE Cray MPI supports several user methods for mapping MPI rank to NIC
  - Ensuring each process uses the “right” NIC is critical towards achieving best performance
  - Users can customize NIC selection logic via MPI environment variables

- Environment variables for multi-NIC support in HPE Cray MPI
  - MPICH_OFI_NIC_VERBOSE
  - MPICH_OFI_NIC_POLICY
  - MPICH_OFI_NIC_MAPPING
  - MPICH_OFI_NUM_NICS
  - MPICH_OFI_SKIP_NIC_SYMMETRY_TEST

- Multi-NIC support is available for both HPE Slingshot-10 and HPE Slingshot-11 systems

HPE CRAY MPI FEATURES FOR HPE CRAY EX SYSTEMS: COLLECTIVES

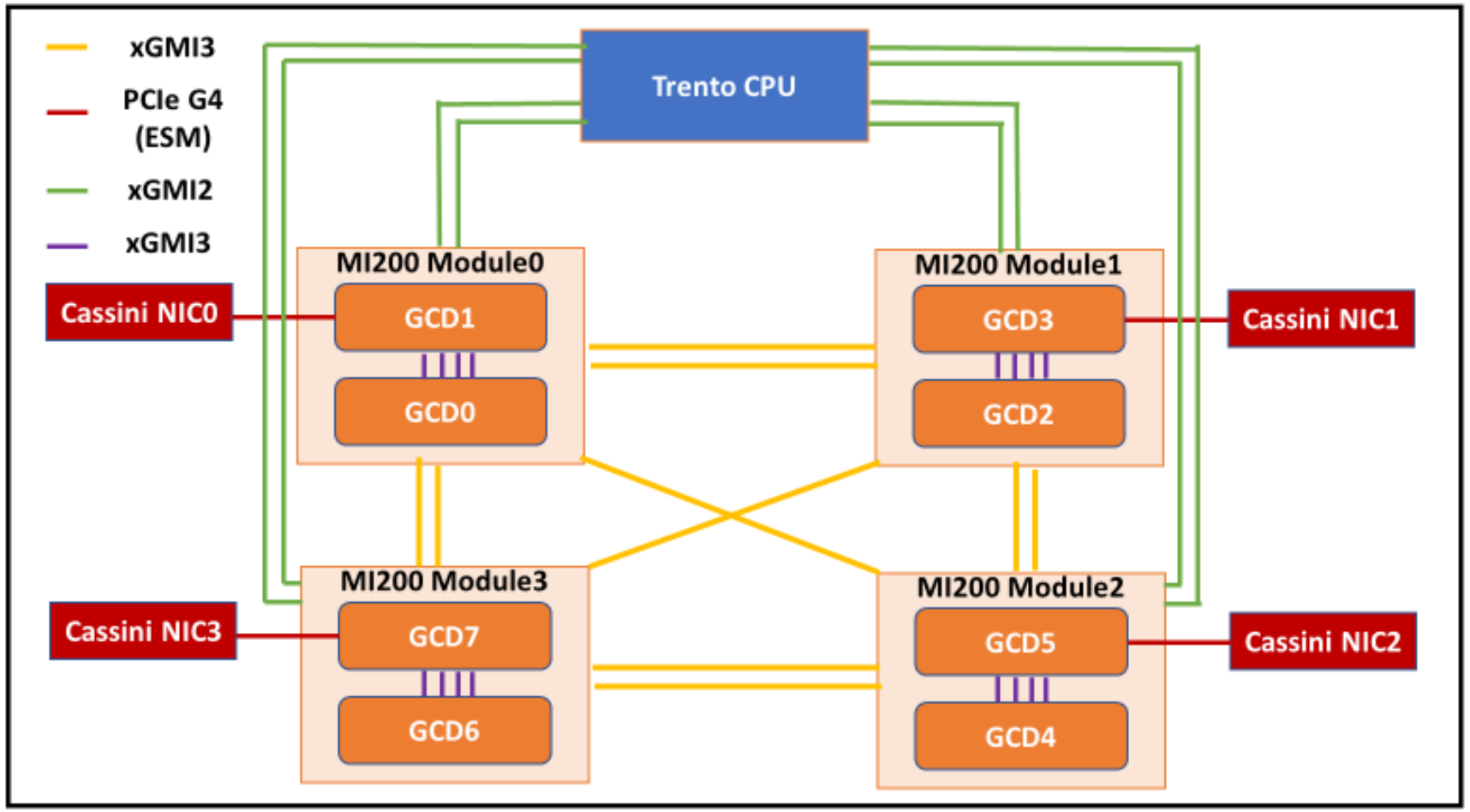
- HPE Cray MPI offers software optimizations for various collectives:
  - MPI_Allreduce, MPI_Bcast, MPI_Alltoall(v), MPI_Allgather(v), MPI_Barrier, MPI_Gatherv, MPI_Igatherv, MPI_Reduce_scatter_block, and MPI_Scatterv
- HPE Cray MPI offers non-blocking collectives for communication/computation overlap
- GPU-specific optimizations for various MPI collectives (e.g., large payload global reductions)
- Optimizations are enabled by default
  - Env. variables to tune performance are documented in HPE Cray MPI man pages ($ man mpi)
- For other collectives, HPE Cray MPI will utilize implementations of collective operations inherited from open-source MPICH from Argonne National Lab (ANL)

HPE CRAY MPI FEATURES FOR HPE CRAY EX SYSTEMS: RMA

- MPI-3 RMA operations are fully supported on both CPU-only and CPU/GPU Slingshot-10 and Slingshot-11 systems
- Intra-node RMA operations rely on XPMEM (supported only on COS systems), GPU Peer2Peer IPC (on CPU/GPU systems), CMA, or POSIX shared memory
- Inter-node RMA operations rely on the OFI layer and the Cassini provider for SS-11 systems
  - RMA operations can leverage rich set of multi-NIC and Cassini hardware capabilities
- HPE Cray MPI handles different Libfabric (OFI) completion semantics (FI_DELIVERY_COMPLETE and FI_TRANSMIT_COMPLETE) for RMA operations
  - Semantics are implemented as a combination of software and hardware optimizations
- HPE Cray MPI is investigating the use of hardware atomics to optimize select MPI RMA operations on HPE Slingshot-11 systems

HPE CRAY MPI ON BARD PEAK NODE ARCHITECTURE

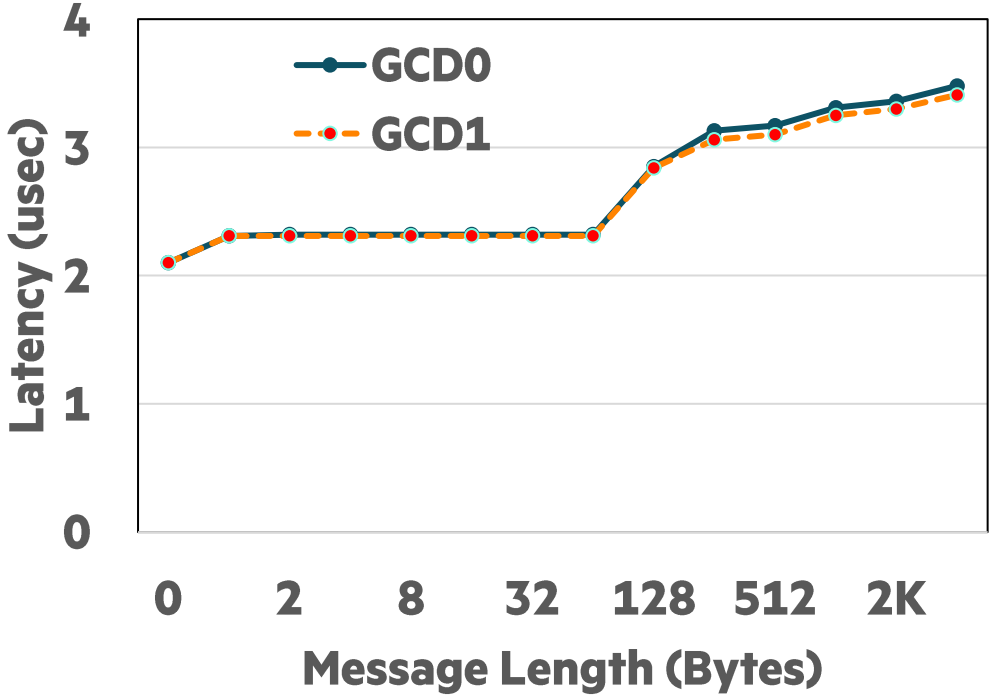
- HPE Cray MPI is highly optimized for the Bard Peak node architecture



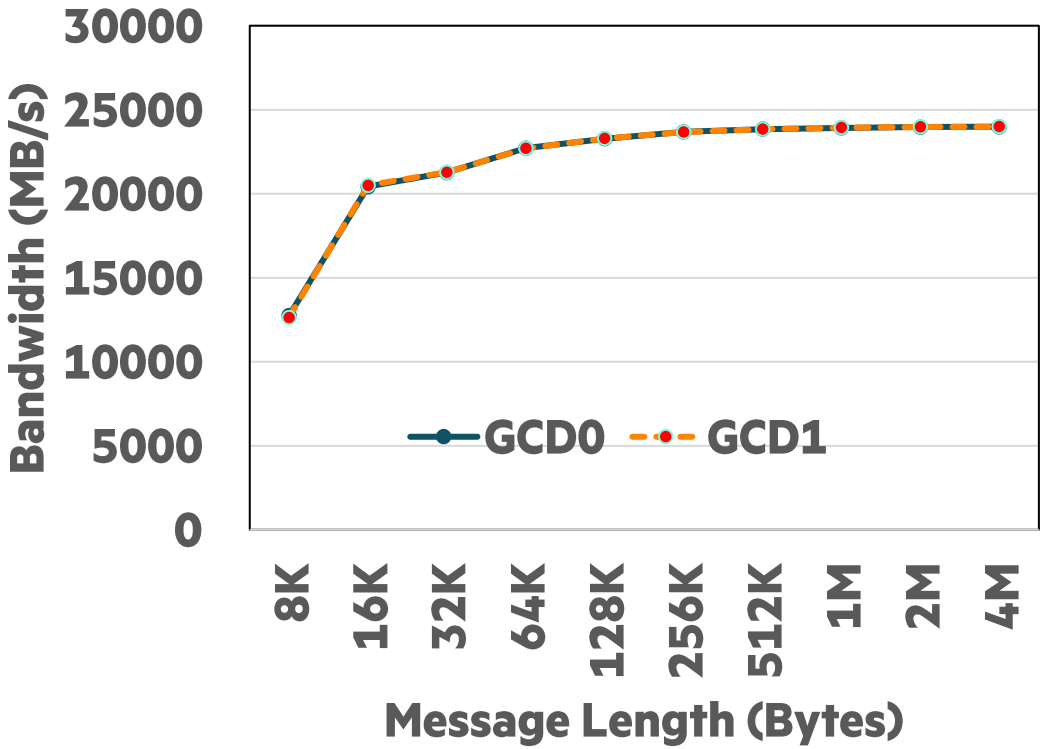
MPI POINT-TO-POINT PERFORMANCE STUDIES ON BARD PEAK

- HPE Cray MPI behavior on “Bard Peak” node architecture and HPE Slingshot-11 network
  - Communication buffers are on GPU-attached High Bandwidth Memory
  - Experiments use “near” nodes attached to the same switch

Inter-node MPI Ping/Pong Latency



Inter-node MPI Uni-Directional Bandwidth



TIPS & TRICKS FOR TUNING AND DEBUGGING HPE CRAY MPI (CONTINUED)

Problem	HPE Cray MPI Recommendations
Fork() issues: • After fork(), a child processes attempts to access memory owned by the parent process	• Users are advised to set following runtime variables: CXI_FORK_SAFE=1, CXI_FORK_SAFE_HP=1, FI_CXI_DISABLE_CQ_HUGETLB=1 • With SLES15 SP4 and newer Linux kernels, some of the fork() issues have been addressed in the Linux kernel • COW semantics are changing from *no_copy* to *eager_copy*
GPU NIC RDMA Corner cases: • Some GPU-enabled applications may hang and the following messages are seen in dmesg logs: *"cxi_core: cass_vma_write_flag:22 VMA does not have write permissions"*	This Is typically due to following user errors: • MPICH_GPU_SUPPORT_ENABLED=1 is not set • On NVIDIA GPU systems, "module load cudatoolkit" is missing in job submission scripts • Managed memory support is disabled. HPE Cray MPI supports Managed memory by default and users may override this via runtime variables documented in HPE Cray MPI man pages