Hewlett Packard
Enterprise

HPE CRAY MPI UPDATE

Naveen Ravi
HPE Cray Programming Environments

20 November 2024, SC 24 MPICH BoF

WHAT IS HPE CRAY MPI?

- Enhanced MPI library implementation based on the open-source ANL MPICH implementation
- Proprietary product maintained by HPE Cray MPT team
- Released as part of the HPE Cray Programming Environment software stack
- Optimized and extended for HPE Cray EX and HPE Apollo systems

HPE CRAY MPI ENABLES THE WORLD'S TOP SUPERCOMPUTERS

November 2023 TOP500 List

#1 ORNL Frontier (1.194 EFlops/s)
#2 ANL Aurora
#5 EuroHPC/CSC Lumi
#12 DOE/SC/LBNL NERSC Perlmutter
#17 GENCI-CINES Adastra
#20 KAUST Shaheen III
#24 DOE/NNSA/LANL/SNL Crossroads
#25 Pawsey Supercomputing Centre Setonix

November 2024 TOP500 List

#1 LLNL El Capitan (1.742 EFlops/s)
#2 ORNL Frontier (1.353 EFlops/s)
#3 ANL Aurora
#5 Eni HPC6
#7 CSCS Alps
#8 EuroHPC/CSC Lumi
#10 LLNL Tuolumne
#13 LANL Venado
#19 DOE/SC/LBNL NERSC Perlmutter
#30 GENCI-CINES Adastra
#38 KAUST Shaheen III
#43 DOE/NNSA/LANL/SNL Crossroads
#45 Pawsey Supercomputing Centre Setonix

HPE Cray MPI is used as the primary MPI implementation driving domain experts to scale and tune scientific applications on these systems

HPE CRAY MPI SUPPORT MATRIX

CPU Architectures	Intel CPUs (Intel SPR), AMD CPUs (AMD Milan) NEW Nvidia Grace CPUs (Under Development)
GPU Architectures	AMD GPUs (MI250X), Nvidia GPUs (A100) NEW AMD MI300A, Nvidia H100, Intel GPUs (Under Development)
Network Architectures	HPE Slingshot SS10 HPE Slingshot SS11 (200 Gpbs) HPE Apollo InfiniBand clusters NEW HPE Slingshot (400 Gbps) (Under Development)
Operating Systems	RHEL / CENTOS, SLES, and COS NEW TOSS
Supported Job Launchers	Slurm, PALS NEW Flux
Supported Programming Envs and Compilers	PrgEnv-cray, PrgEnv-gnu, PrgEnv-nvidia, PrgEnv-amd, PrgEnv-aocc, and PrgEnv-Intel

KEY FEATURES IN HPE CRAY MPI – 1

- Highly optimized for low latency and high bandwidth point-to-point and collective communications
- Scalable initialization and launch cost using optimized Cray PMI interface with Slurm and PALS

- GPU support
  - GTL – GPU Transport Layer – HPE developed library for handling GPU-attached communication buffers
  - Provides basic GPU-aware P2P, RMA, and collective communication operations

- Advanced GPU support
  - GPU Direct Async communication schemes
  - NEW GPU stream triggered communication
  - GPU kernel triggered communication

- Collective communication performance
  - Optimized small payload on-node communication
  - Tuned and optimized algorithms for select operations
  - GPU kernel-based reductions

- MPI I/O performance enhancements and stats

KEY FEATURES IN HPE CRAY MPI – 2

- HPE Slingshot-11 features
  - Library tuned specifically for HPE Slingshot-11 networks – HPE Cassini NICs and HPE Rosetta Switches
  - Support for traffic classes
  - NEW NIC offloaded collectives for small payloads
  - NEW NIC accelerated non-blocking collectives using triggered operations
  - Hardware offloaded MPI tag-matching and rendezvous protocols
  - Hardware assisted congestion management
  - Small memory footprint for network resource management using connectionless protocols
  - Tight integration for enabling GPU-aware communication

- Usability
  - HPE Slingshot counter statistics
  - Support for options enabling efficient GPU-NIC affinity on multi-NIC systems
  - ABI compatibility with different MPI implementations
  - MPIxlate : HPE developed ABI translator for MPI programs
  - Flexible, intuitive rank reordering features

HPE CRAY MPI PLANS FOR 2025

- Rebase with ANL MPICH 4.0 to support the MPI 4.0 standard
- Additional optimizations leveraging HPE Slingshot-11 NIC hardware capabilities
- Support next-generation Slingshot hardware (HPE Slingshot 400 Gpbs)
- Optimize for Nvidia Grace and Hopper architecture
- Tune for AMD MI300A architecture
- Collaborate with MPI Forum participants for introducing GPU-NIC Async features

THANK YOU

nravi@hpe.com