Open MPI for Exascale (OMPI-X)
ECP Project ST-2.3.1.11-OMPIX

Project Team

Institution	PI	Additional Participants
ORNL (Lead)	**David Bernholdt (Lead)**	Manju Gorentla, Terry Jones, Thomas Naughton, Geoffroy Vallee
LANL	Howard Pritchard	Nathan Graham, Nathan Hjelm
LLNL	Ignacio Laguna	Chris Chambreau, Murali Emani, Martin Schulz
SNL	Ron Brightwell	Ryan Grant
UTK	George Bosilca	Aurelian Bouteiller
In collaboration with the Open MPI Community (http://open-mpi.org)		

Project Focus Areas

Focus Area	Topics	Technical Lead
Runtime Interoperability for MPI+X and Beyond	APIs for better sharing of threads between MPI and other thread-based runtimes.	Geoffroy Vallee (ORNL)
Extending the MPI Standard to Better Support Exascale Architectures	Endpoints, Finepoints, Sessions	Ryan Grant (SNL)
Open MPI Scalability and Performance	Memory footprint, collectives, message matching, one-sided, PMIx	Manju Gorentla (ORNL)
Supporting More Dynamic Execution Environments	Intelligent process placement and contention management	Terry Jones (ORNL)
Resilience in MPI and Open MPI	ULFM, ReInit, resilience in PMIx	George Bosilca (UTK)
MPI Tool Interfaces	MPI_T, PMPI replacement	Chris Chambreau (LLNL)
Quality Assurance for Open MPI and New Developments	Test infrastructure deployed to ECP-relevant systems. Regular testing of Open MPI and OMPI-X developments	Howard Pritchard (LANL)

Recent Progress

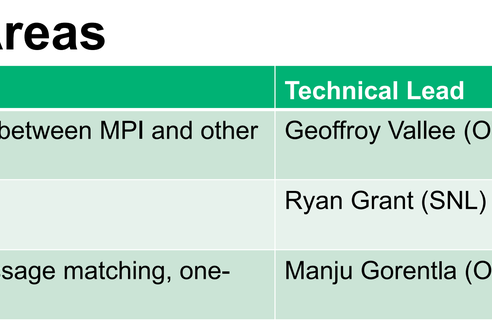
Interoperability for MPI-X and Beyond

- Modify the OpenMP LLVM compiler to interface with PMIx
- Data exchange between the MPI and OpenMP runtimes via PMIx
- Implement a placement policy based on the number of MPI ranks and available cores/HT per node
- Beginning to conduct joint experiments for evaluation and implementation of more advanced policies (collaboration with ECP SOLLVE project)

Connections: SOLLVE, UPC++/GASnet, PMIx

MPI Finepoints - Partitioned Multi-threaded MPI Communication

- Finepoints
  - New MPI multi-threading interface
  - Better efficiency with minimal app changes
- Leverages hardware capabilities
- Allows new type of overlap in communication
- Early prototype demonstrated with ECP mini-app
  - ~5% improvement in runtime
  - ~25% improvement in communication
- Collaborations with ECP Qthreads project and EU Intertwine project



ECP miniapp running on a KNL with MPI procs x threads, with 4 MPI procs to 1 MPI proc. The miniapp is controlled for noise and artificial noise injected to demonstrate good performance in practice (real noise on systems is in the 3% range).

Connections: Qthreads, Intertwine

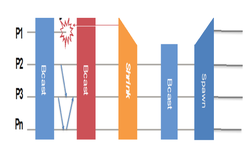
Topology and Congestion Awareness

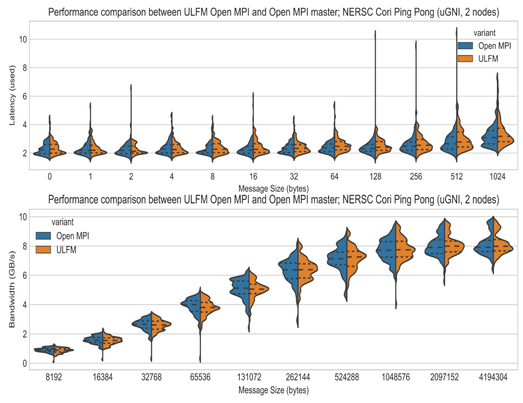
- Developing module to gather the communication weights between processes
  - Capability to distinguish between pt2pt collective file IO or RMA)
- Developing module to reorder processes based on weights
- Initial implementation available in Open MPI GitHub master

User-Level Fault Mitigation (ULFM) in Open MPI

- Resilience in varied application types
  - Malleable applications enjoy a **cheap, tailored recovery procedure**
  - Non-malleable applications can **restore complete MPI** capabilities without redeployment
- Integration of ULFM resilience shows no overhead on raw communication performance on ECP hardware
- Recently delivered
  - **Stable resilience; tested** deployments on ECP hardware with support for job schedulers and accelerated networks
  - Support for **resilience with threads, non-blocking collective operations, RMA operations**
  - State-of-the-art research in **resilient collective algorithms and failure detection implementation**
- Impacts
  - **Large application community** using Open MPI ULFM to explore resilience in HPC
  - **Industry users** (databases, MapReduce) also use Open MPI ULFM to explore non-HPC workloads over MPI
  - **User documentation and education** helps ECP applications move forward on resilience

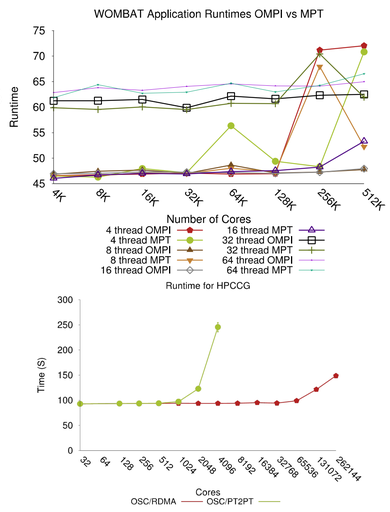
Connections: EXAALT, QMCPACK, PMIx





MPI Performance and Scalability Improvements

- Remote Memory Access
  - New RMA implementation allows scaling
  - Application performance similar to highly tuned vendor impls.
- MPI Message Matching
  - Initial prototype perf. up to 2X
  - Integration plan underway
- Multi-threading
  - Multiple improvements completed
- Non-blocking Collectives



Open MPI+PMIx+SCON for Scalability

- Initial evaluation of scalable startup performance showing 3x improvement in launch time for PMIx Open MPI vs Cray ALPS Open MPI (presented during PMIx BoF at SC'17)
- PMIx event notification API in the PMIx standard document
- Event notification API used in fault-tolerant Open MPI (ULFM)
- Generic failure detector API defined
  - Transfer failure detector capability from Open MPI to PMIx (prototyping started)
- Implement SCON, scalable overlay network library that provides communication capabilities for PMIx (API design in progress)

Connections: PMIx

Continuous Integration/Nightly Testing

- Resolved issues with next generation Nightly tester (MTT) reporting results to community database at AWS
- Added plugin to next generation MTT to test nightly tarball bills from https://open-mpi.org/downloads
- Deployed current MTT on ORNL summit-dev platform
- Participating in the ECP ST facilities WG evaluation of CI RFP responses

Connections: ECP CI Testing, Facilities

Connections: PMIx