MPICH: A High-Performance Open-Source MPI Implementation

SC11 Birds of a Feather Session
Schedule

- MPICH2 status and plans
- Presentations from developers, partners and collaborators
  - Cray (Larry Kaplan)
  - Microsoft (Fab Tillier)
  - Intel (Bill Magro)
  - The Ohio State University (DK Panda)
  - IBM (Robert Wisniewski)
- Wrap up
Goals of the MPICH2 project

- Be the MPI implementation of choice for the highest-end parallel machines
  - 4 of 5 of the fastest machines in the June 2011 Top500 list use MPICH2-based implementations
- Carry out the research and development needed to scale MPI to exascale
  - Optimizations to reduce memory consumption
  - Fault tolerance
  - Efficient multithreaded support for hybrid programming
  - Performance scalability
- Work with the MPI Forum on standardization and early prototyping of new features
MPICH2-Based Implementations on the Highest End Machines

Several vendors and other groups use MPICH2 as the basis for their own MPI implementation

- Cray XT and XE series
- IBM for Blue Gene
- Intel MPI
- Microsoft MPI
- Myricom MPI
- OSU MVAPICH2 for InfiniBand
- University of British Columbia
Tracking MPI Forum Work

- MPICH has always been the first to provide fully compliant implementations
  - MPI-1, MPI-2.0, MPI-2.1, MPI-2.2
  - MPICH has been MPI-2.2 compliant for more than two years
  - Intend to continue this trend
- Already working on implementing the MPI-3 draft standard
  - Nonblocking collectives ✔ done
  - Tools ✔ done
  - RMA
  - Hybrid programming MPI+Threads
  - Fault tolerance
MPICH2 Release Model

- Pre-releases: alphas and RCs
  - Get feedback from
    - ISVs
    - Derivative implementations
    - Package maintainers
- Major releases
  - One per year
  - ABI won’t change between major releases
  - New major features
- Minor releases
  - Bug fixes
  - Performance improvements
  - Minor features
- Patch releases
  - Critical bug fixes
Previous MPICH2 Releases

1.0.8

1.1.x series
MPI 2.1
Blue Gene/P
Nemesis
Hierarchical collectives

1.1.1

1.1.x series
MPI 2.1
Blue Gene/P
Nemesis
Hierarchical collectives

1.2.x series
MPI 2.2
Hydra introduced

1.2

1.2.x series
MPI 2.2
Hydra introduced

1.2.1

1.3.x series
Hydra default
Async progress
Checkpointing
FTB

1.3

1.3.x series
Hydra default
Async progress
Checkpointing
FTB

1.3.1

1.3.x series
Hydra default
Async progress
Checkpointing
FTB

1.3.2

1.4.x series
Improved process FT
ARMCI API

1.4

1.4.x series
Improved process FT
ARMCI API

1.4.1

1.4.x series
Improved process FT
ARMCI API
Next MPICH2 Release

- Current Stable Release: mpich2-1.4.1p1
- Current Active Development Effort: mpich2-1.5.x series
  - Major Features:
    - MPI 3 Nonblocking Collectives
    - MPI 3 Noncollective groups
      - Create communicator noncollectively
    - MPI 3 RMA
    - MPI 3 Fault tolerance
    - MPI 3 Tools interface
    - Complete revamp of the build system
    - Many more features
  - Preview releases at SC11 (GA in Q2/2012)
- Following release: mpich2-1.6.x series
  - Full MPI 3 compliance
  - Revamp CH3 device
New Features

- **MPI 3 RMA**
  - Unified model
    - Relaxes consistency management requirements
  - Dynamic windows
    - Allows users to dynamically attach and detach memory
  - Atomic operations
    - Allow lock-free data structures
  - Request-based one-sided operations
    - Enable fine-grained communication-computation overlap

- **MPI 3 Fault Tolerance**
  - Failed processes don’t affect p-to-p communication with other processes
  - Communicators can restore collective communication
  - Query for failed processes
  - Callbacks for failed process notification
Collaborators/Partners

- Core MPICH developers
  - IBM
  - INRIA
  - Microsoft
  - Intel
  - University of Illinois
  - University of British Columbia

- Derivative implementations
  - Cray
  - Myricom
  - Ohio State University

- Other Collaborators
  - Absoft
  - Pacific Northwest National Laboratory
  - QLogic
  - Queen’s University, Canada
  - Totalview Technologies
  - University of Utah
Developer, Partner and Collaborator Presentations
Acknowledgements

- Thanks to the following for equipment loans
  - Intel
  - Microsoft
  - Myricom
  - QLogic

- And to the following for financial support
  - Department of Energy
  - Microsoft
Questions or Comments?
MPICH2 Quiz

- How many source code management systems has the MPICH project used so far?
- When was MPICH formed?
- What does ROMIO stand for?
- Where did the name Nemesis come from?
MPICH2 Quiz

- What does the \textit{CH} in MPICH stand for?
- When was MPICH formed?
- What does ROMIO stand for?
- Where did the name Nemesis come from?
Backup Slides
Support for High-speed Networks
- 10-Gigabit Ethernet iWARP, Qlogic PSM, InfiniBand, Myrinet (MX)
- Extended to support proprietary platforms
  - BlueGene/L, BlueGene/P, Cray
- Distribution with ROMIO MPI/IO library
- Profiling and visualization tools (MPE, Jumpshot)
Nemesis Performance

- 207 nanoseconds 0-byte MPI latency
- 2.6GHz Intel Clovertown, dual-core, dual processor machine
Tools Within MPICH2

- MPE library for tracing MPI and other calls
- Scalable log file format (slog2)
- Jumpshot tool for visualizing log files
  - Supports threads
- Collchk library for checking that the application calls collective operations correctly

Performance Tuning Tool - Jumpshot

Each line represents 1000’s of messages

Detailed view shows opportunities for optimization
MPICH as a Research Vehicle for Exascale MPI

- Scalability
  - Scalable Memory Use in MPI: A Case Study with MPICH2. *EuroMPI 2011*
  - **MPI on Millions of Cores.** *Parallel Processing Letters, 2011*
- Hybrid MPI+Threads
  - **Fine-Grained Multithreading Support for Hybrid Threaded MPI Programming.** *International Journal of High Performance Computing Applications, 2010*
  - **Enabling Concurrent Multithreaded MPI Communication on Multicore Petascale Systems.** *EuroMPI 2010*
- Fault tolerance
  - **Scalable Distributed Consensus to Support MPI Fault Tolerance.** *Submitted to IPDPS 2012*
  - **Fault Tolerance in MPI Programs.** *International Journal of High Performance Computer Applications 2004*