Open MPI:
A Research Platform

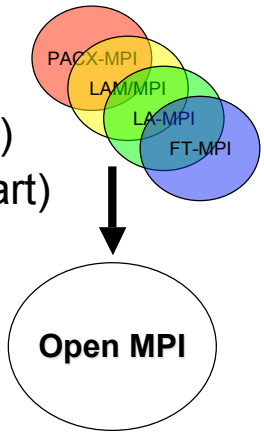
Tim Mattox, Ph.D.
Open Systems Lab
Pervasive Technology Labs
Indiana University

What is Open MPI?

- Open source implementation of MPI-2
- High performance & robust
- Works with most interconnects
- Combined expertise from 4+ previous MPIs

Parentage of Open MPI

- LAM/MPI (Indiana U.)
- FT-MPI (U. of Tennessee)
- LA-MPI (Los Alamos, Sandia)
- PACX-MPI (HLRS, U. Stuttgart)



Motivation for a new MPI

- Each prior project had different strong points
  - Could not easily combine into one code base
- New concepts could not easily be accommodated in old code bases
- Start over with a blank sheet of paper
  - Harnessing our many years of combined implementation experience

Modular Component Architecture

- Logical progression of prior MPI component architecture research (LAM/MPI)
  - More component types
  - More services provided to components
  - Decentralized management
- Combinatorial capabilities
- Function pointer faster than shared library call
- End result is a “highly pluggable” MPI

Open MPI was born in 2003

- Developers of FT-MPI, LA-MPI, LAM/MPI
  - Kept meeting at conferences in 2003
  - Culminated at SC 2003: Let’s start over
  - Open MPI was born
- Started serious design and coding work January 2004
  - All of MPI except one-sided operations
  - First release 1Q 2005

Open MPI Today

- Version 1.0 released in November 2005
- Version 1.1 released in June 2006
  - Added support for one-sided communications
  - Version 1.1.2 released in October 2006
  - Version 1.1.3 soon
- Version 1.2b1 (beta) released today
  - Library level matching: Myrinet/MX & Portals
  - First set of Tuned Collectives
  - Sun N1 Grid Engine
  - Data Reliability

HPC “Users” of Open MPI

- Sysadmins
- Vendors
  - Software
  - Network
  - Cluster/Machine
- Researchers
  - Scientists
  - Developers

HPC “Users” of Open MPI

- Sysadmins
- Vendors
  - Software
  - Network
  - Cluster/Machine
- Researchers
  - Scientists
  - Developers

Why do Research with Open MPI?

- Open source
- Modular design
  - Treat “uninteresting parts” as black boxes
  - Mix & match components
- The leverage effect
  - Stand on the shoulders of giants
  - Can experiment inside a production quality MPI

Fault Tolerance Research

- Data Reliability
- Checkpoint/Restart
- Process Migration
- Batch & Gang Scheduling

See Josh's talk tomorrow at 11am

Multicore Optimization Research

- Processor Affinity
- Memory Affinity
- Process Mapping
- Shared Memory Collectives

Collectives Research

- Non-blocking Collectives
- Topology Aware Collectives
  - Hierarchical Networks
  - Flat Neighborhood Networks (FNNs)
- Hardware Collectives
  - BlueGene
  - Aggregate Function Networks (AFNs)

Conclusions

- MCA lets you play inside Open MPI
- Clear path from research to production
- Vibrant research community today

http://www.open-mpi.org/

Open MPI Events @ IU Booth

- Right Now… 4:30 - 4:45pm

  Open MPI: Collective Communication research at UT
  By George Bosilca

- Thursday 11:00 - 11:50am

  Dealing with disaster: Fault Tolerance in Open MPI

Questions?