Open MPI on the XT3

Brian Barrett, Ron Brightwell,
Jeff Squyres, and
Andrew Lumsdaine
May 11, 2006

Overview

- What is Open MPI?
- Why is it running on the XT3?
- How well does it run?
- Lessons learned / Porting Issues
- Future work

What is Open MPI

- Complete MPI-2 implementation
- Designed for large-scale clusters
- Highly optimized datatype engine
- Optimized collective routines
- Run-time loadable component architecture
    - Well defined abstraction points
    - Simplifies customizing to a platform

Cluster Features

- Multiple NIC support with message striping
- Message error detection and recovery
- Process fault tolerance
- MPI_THREAD_MULTIPLE support
- Thread-based asynchronous progress
- Rich run-time support for clusters

Open MPI Collaborators

Los Alamos
NATIONAL LABORATORY
EST.1943
HLRS
Mellanox
TECHNOLOGIES
CISCO SYSTEMS
ICL UT
Sun
microsystems
CS@UH
pervasivetechnologylabs
AT INDIANA UNIVERSITY
Voltaire
The Grid Backbone™

Why port to the XT3?

- Application developers only wanted to complain about one MPI implementation
- Interesting “big iron” machine
  - Sane network
  - Developer experience on similar architecture
- Test framework abstractions

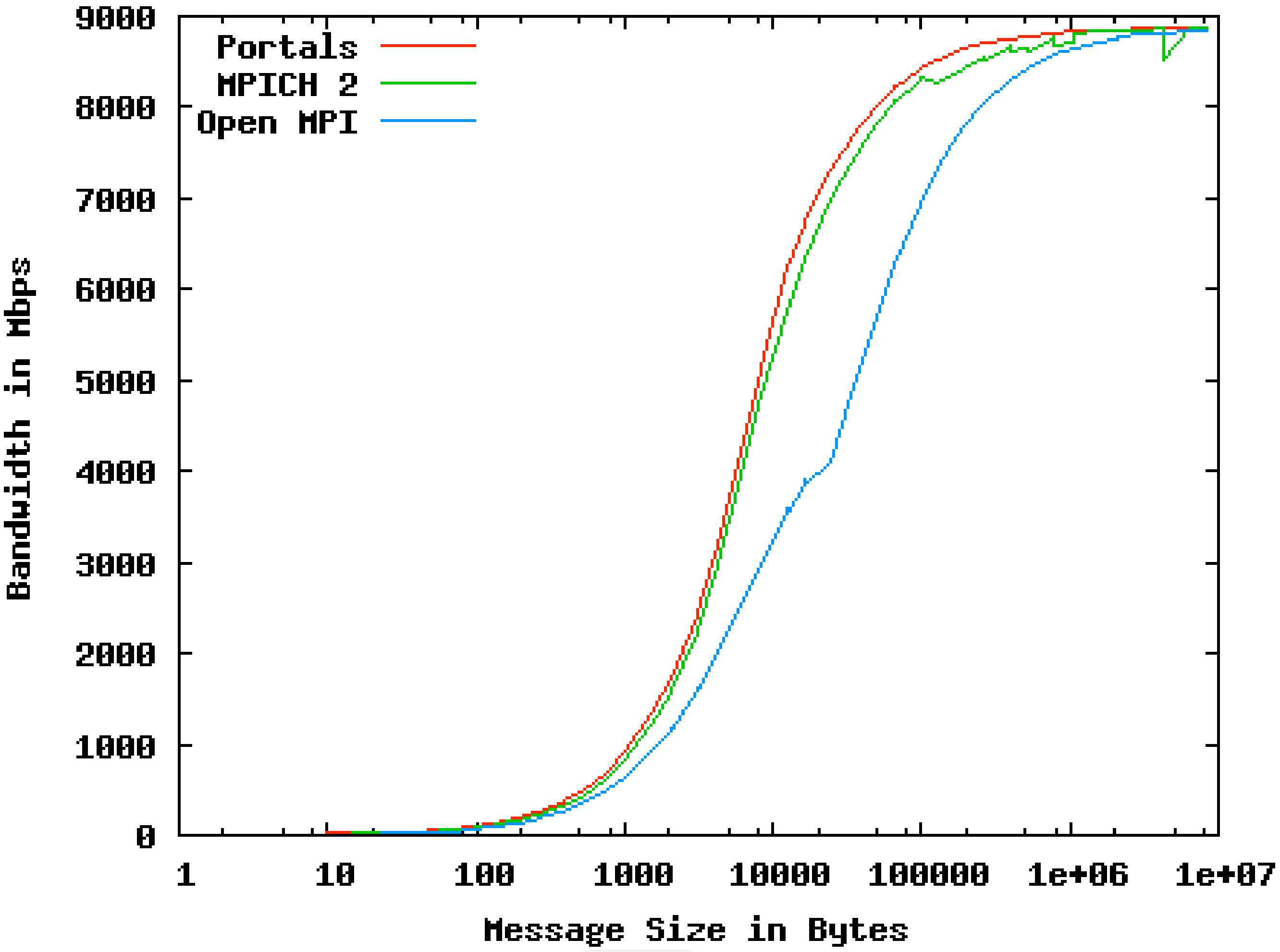
How well does it run?

- (Almost) Complete MPI-level support:
  - MPI-2 Dynamics don’t work…
- Performance:
  - Very good for first attempt
  - Still needs to go faster
    - We’ve identified the issues
    - Deciding how best to fix them
- Some performance numbers…

Performance (Latency)

Implementation	1 Byte Latency
Native Portals	5.30us
MPICH-2	7.14us
Open MPI	8.50us

Performance

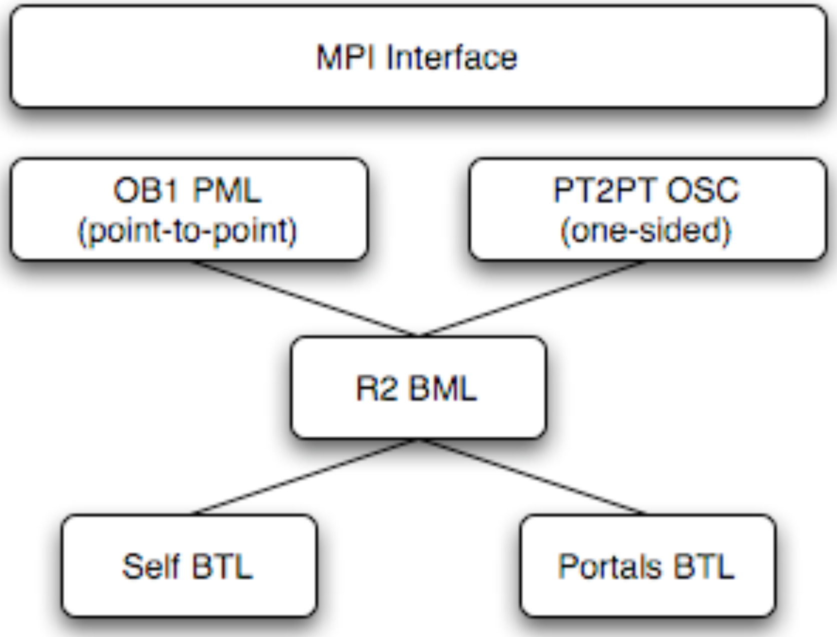


Lessons Learned

- Cross-compilation challenging
- Cluster run-time overkill
  - Component framework allows run-time to “get out of the way”
  - Surprising amount of work
- Performance expectations
  - Designed around InfiniBand and Myrinet/GM
  - Hardware matching growing pains

Portals Point-to-Point

- Choice of abstraction layer for implementation
- BTL design chosen:
  - Performance impact thought to be low
  - Quicker time to completion
  - One-sided support uses BTL layer



Future Work

- Point-to-point performance
  - Early PML work provides performance comperable to MPICH-2
  - Myrinet/MX presents identical challenge
  - Hope to find middle ground
- MPI topology functions
- Collectives performance
  - Need tuning parameters for platform
  - Topology awareness

More Information

- BTL implementation detailed in paper
- Publicly available in Open MPI SVN

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