Open MPI: Overview

SC06, November 15, 2006
Jeff Squyres, Cisco Systems

Open MPI Sponsors

- DoE
  - ASC
  - LANL CCS-1
  - NNSA
- HLRS
- Lilly Endowment
- Microsoft
- NSF
Open Source High Performance Computing

- Open source implementation of MPI-2
- Combined expertise from 4+ previous MPIs
- High performance & robust
- Works with most interconnects
- Modular Component Architecture
  - Combinatorial capabilities
  - Function pointers faster than shared library calls

History of Collaboration

- 9/2003 Euro PVM/MPI
  - Principals meet
- 10/2003 LACSI Symposium
  - Principals agree to collaborate
- 11/2003 SC’03
  - Collaborators agree to start with a “Blank Sheet of Paper”
- 1/2004
  - Design and implementation begin
- 10/2004
  - Linpack
  - ASC/NNSA/DOE, DOE Office of Science, and Eli Lily foundation fund project startup
Current Members

- Academia / Research
- HLRS
- Indiana U.
- Sandia National Lab
- Los Alamos National Lab
- U. of Dresden
- U. of Houston
- U. of Tennessee

- Industry
- Cisco
- IBM
- Mellanox
- Myricom
- QLogic
- Sun
- Voltaire

Current Status

- Stable release: v1.1.2
  - v1.1.3 expected “soon”
- Upcoming release: v1.2
  - Stability and scalability improvements
  - Sun / Solaris / N1GE / uDAPL support
  - Better MX support
  - InfiniPath support
  - TotalView message queue support
  - …and more
Top 500

- #6: Sandia Thunderbird cluster
  - Dell PowerEdge 1850
  - InfiniBand
- Linpack result
  - 4347 dual processor nodes
  - 53 teraflops
  - 84.66% network efficiency
- Powered by Open Fabrics / Open MPI

LINPACK on SNL’s Thunderbird

- Collaboration between
  - Sandia National Laboratory
  - Los Alamos National Laboratory
  - Cisco Systems
  - University of Tennessee
  - (…and others via source code contributions)
- Problem: Stabilize system for full-system runs
Components - Diversity of Implementations Choices

- Formalized interfaces
  - Specifies “black box” implementation
  - Different implementations available at run-time
  - Can compose different systems on the fly
  - Multiple options in a single build

Current Support

- Operating Systems
  - AIX
  - Catamount
  - Linux
  - OS X (BSD)
  - Solaris
  - MS Windows

- Schedulers
  - BJS (LANL Bproc Clustermatic)
  - BProc / XCPU
  - N1GE
  - PBS / Torque
  - Rsh/ssh
  - SLURM
  - Xgrid
  - YOD (Red Storm)

- Networks
  - TCP
  - Shared memory
  - Myrinet
    - GM, MX
  - Infiniband
    - mVAPI, OpenIB
  - InfiniPath
  - Portals (flow control)
Open MPI: Point-To-Point Communication
Fault-Tolerance and Heterogeneity

Richard L. Graham
Advanced Computing Laboratory
Los Alamos National Laboratory
LA-UR-06-xxxx

Point-to-Point Architecture

Component Architecture:
- Plug-in’s for different Capabilities (networks)
- Run time tunable parameters

MPI

PML - OB1/DR

PML - CM

BML - R2

IB BTL

IB MPool

IB RCache

IB BTL

GM MPool

GM RCache

MX-MTL Myricom

PSM-MTL QLogic

Portals MTL

Data Fault Tolerance
Support for Fault Tolerance

- Watch dog timer
- Data checksum/CRC
- Alternative data paths

Ping-Pong Latency (usec)

<table>
<thead>
<tr>
<th>Run Parameters</th>
<th>Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open MPI OB1/OpenIB</td>
<td>2.99</td>
</tr>
<tr>
<td>Open MPI DR/GM</td>
<td>6.21</td>
</tr>
<tr>
<td>Open MPI OB1/OpenIB</td>
<td>7.59</td>
</tr>
<tr>
<td>Open MPI DR/GM</td>
<td>12.10</td>
</tr>
</tbody>
</table>
Ping-Pong Bandwidth (MB/Sec)

Device failover

<table>
<thead>
<tr>
<th>Message Size (KBytes)</th>
<th>Bandwidth (MB/Sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>15256 bytes</td>
</tr>
<tr>
<td>1</td>
<td>15256 bytes</td>
</tr>
<tr>
<td>2</td>
<td>15256 bytes</td>
</tr>
<tr>
<td>3</td>
<td>15256 bytes</td>
</tr>
<tr>
<td>4</td>
<td>15256 bytes</td>
</tr>
</tbody>
</table>

ping: 0 pinged 1 15256 bytes 159.37 MB/s
ping: 0 pinged 1 15256 bytes 150.02 MB/s
ping: 0 pinged 1 15256 bytes 122.42 MB/s
ping: 0 pinged 1 15256 bytes 70.76 MB/s
ping: 0 pinged 1 15256 bytes 211.36 MB/s
NIC Failover: Ping-Pong (MB/sec)

Support for Data and Network Heterogeneity
Processor Heterogeneity

Data Registry

My Proc (1)
PPC Conv

Other Procs
 Conv 1
 Conv 2
 ... 
 Conv N

Recv

My Proc (N)
X86 Conv

Other Procs
 Conv 1
 Conv 2
 ... 
 Conv N

Compute Phase

Network Heterogeneity

Data Registry

My Proc (1)
Network

Other Procs
Net 1
Net 2
...
Net N

Network:
Low latency list
High Latency list

My Proc (N)
Network

Other Procs
Net 1
Net 2
...
Net N

Compute Phase

Init Phase
Netpipe B/W Measurement (TCP/IP)

How Do I Get Involved?

• Source code access:
  ▪ www.open-mpi.org
  ▪ Anonymous read-only repository
  ▪ Tar ball distributions
  ▪ Mailing lists
  ▪ Papers
• Want to become part of the team?
  ▪ www.open-mpi.org/community/contribute
• A lot more work to meet user requirements
Come Join Us!

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