Open MPI: Overview

SC06, November 15, 2006

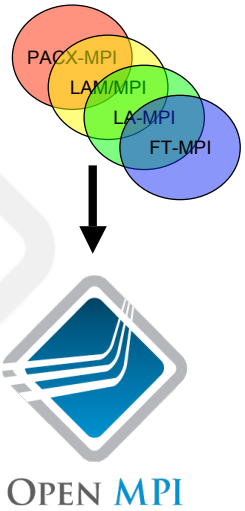
Jeff Squyres, Cisco Systems

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

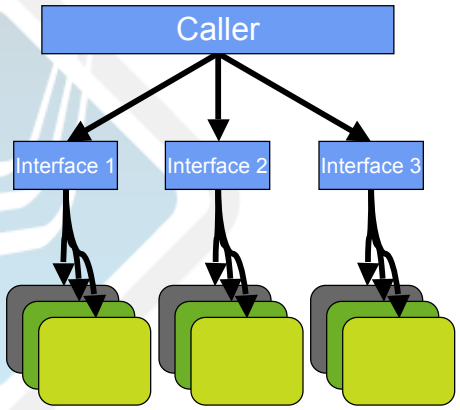
- #6: Sandia Thunderbird cluster
  - Dell PowerEdge 1850
  - InfiniBand
- Linpack result
  - 4347 dual processor nodes
  - 53 teraflops
  - 84.66% network efficiency
- Powered by Open Fabrics / Open MPI

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)

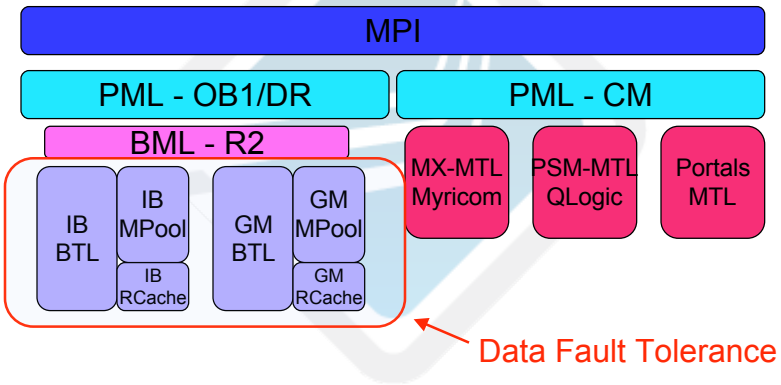
Open MPI: Point-To-Point Communication Fault-Tolerance and Heterogeneity

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

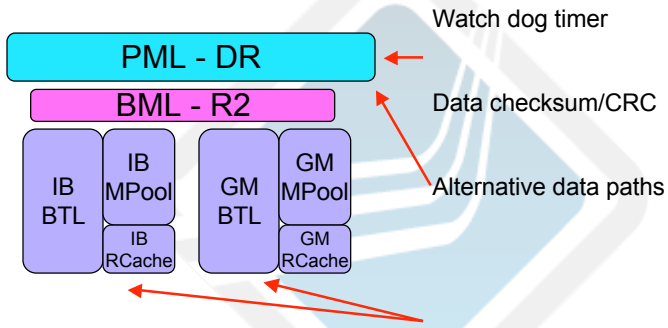
Point - to - Point Architecture

Component Architecture:

- Plug-in's for different Capabilites (networks)
- Run time tunable parameters



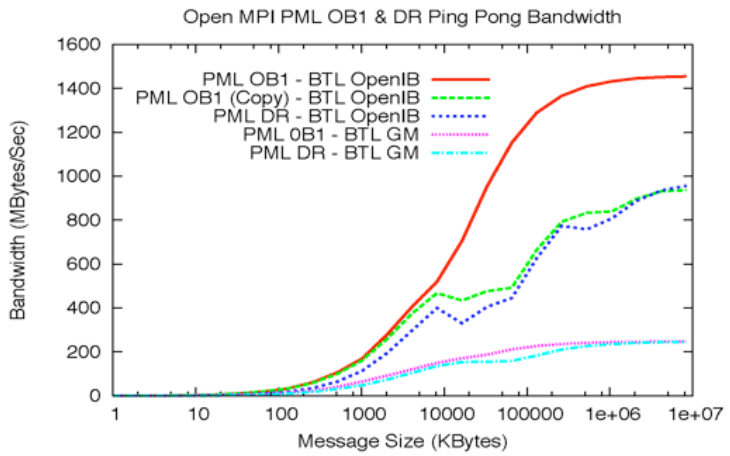
Support for Fault Tolerance



Ping-Pong Latency (usec)

Run Parameters	Latency
Open MPI OB1/OpenIB	2.99
Open MPI DR/GM	6.21
Open MPI OB1/OpenIB	7.59
Open MPI DR/GM	12.10

Ping-Pong Bandwidth (MB/Sec)

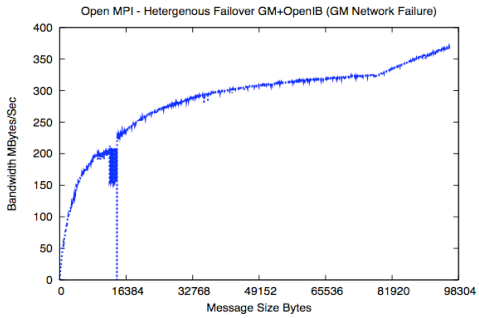


Device failover

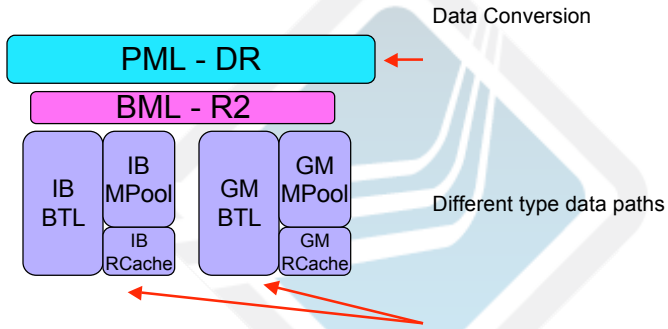
gshipman@boxtop1:~/ompi-test/simple/ping

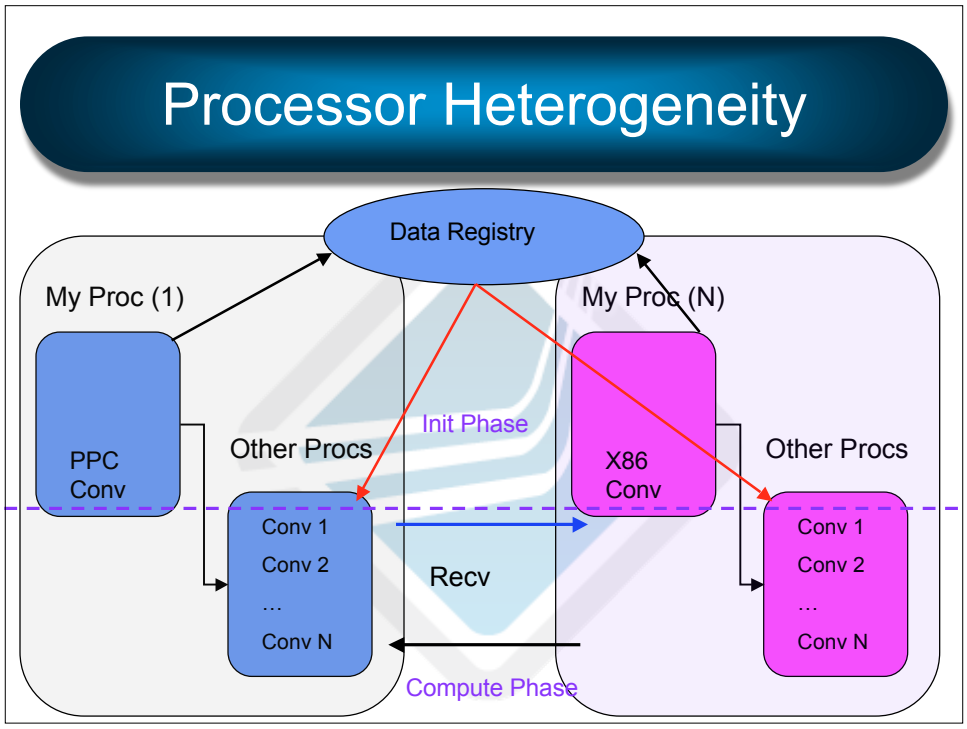
```
  0 pinged   1:    15264 bytes     96.08 uSec   158.87 MB/s
  0 pinged   1:    15296 bytes     96.22 uSec   158.97 MB/s
  0 pinged   1:    15328 bytes     72.16 uSec   212.42 MB/s
  0 pinged   1:    15360 bytes     96.77 uSec   158.72 MB/s
  0 pinged   1:    15392 bytes     96.67 uSec   159.23 MB/s
  0 pinged   1:    15424 bytes     72.76 uSec   211.98 MB/s
[boxtop2.lanl.gov:03305] ../../../../../../ompi_europvm/ompi/mca/pml/dr/pml_dr_v
frag.c:83:mca_pml_dr_vfrag_wdog_timeout: failing BTL: gm
[boxtop2.lanl.gov:03305] ../../../../../../ompi_europvm/ompi/mca/pml/dr/pml_dr_v
frag.c:167:mca_pml_dr_vfrag_reset: selected new BTL: openib
[boxtop1.lanl.gov:03148] ../../../../../../ompi_europvm/ompi/mca/pml/dr/pml_dr_v
frag.c:83:mca_pml_dr_vfrag_wdog_timeout: failing BTL: gm
[boxtop1.lanl.gov:03148] ../../../../../../ompi_europvm/ompi/mca/pml/dr/pml_dr_v
frag.c:167:mca_pml_dr_vfrag_reset: selected new BTL: openib
  0 pinged   1:    15456 bytes  52295.24 uSec     0.30 MB/s
  0 pinged   1:    15488 bytes     64.81 uSec   238.97 MB/s
  0 pinged   1:    15520 bytes     64.50 uSec   240.62 MB/s
  0 pinged   1:    15552 bytes     64.31 uSec   241.83 MB/s
  0 pinged   1:    15584 bytes     64.69 uSec   240.90 MB/s
  0 pinged   1:    15616 bytes     64.54 uSec   241.98 MB/s
  0 pinged   1:    15648 bytes     64.72 uSec   241.78 MB/s
  0 pinged   1:    15680 bytes     64.62 uSec   242.63 MB/s
  0 pinged   1:    15712 bytes     64.78 uSec   242.53 MB/s
  0 pinged   1:    15744 bytes     64.74 uSec   243.17 MB/s
```

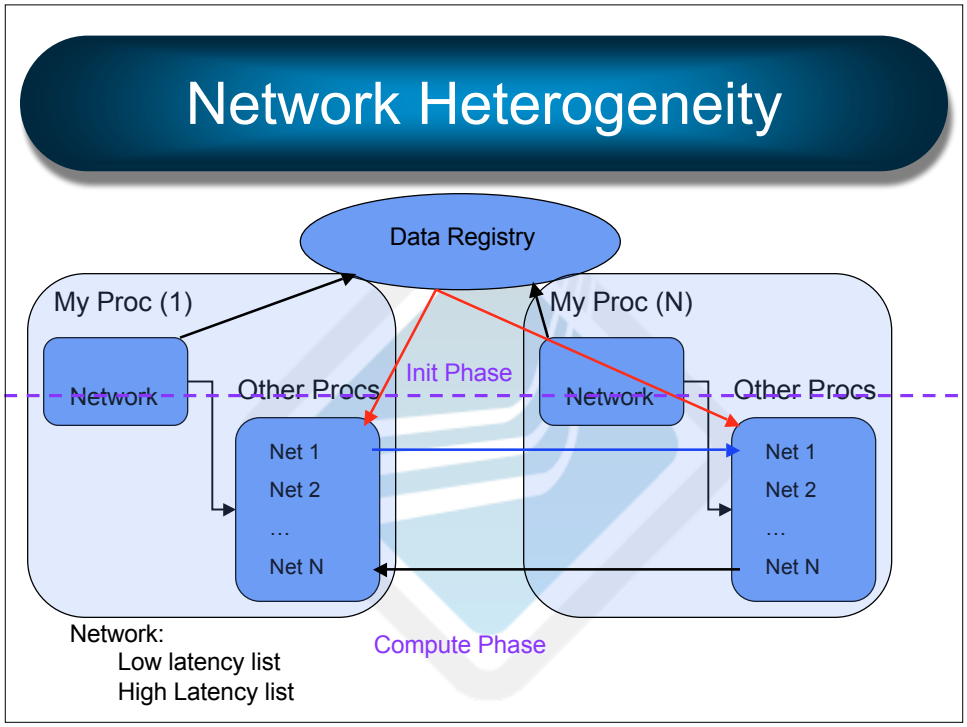
NIC Failover: Ping-Pong (MB/sec)



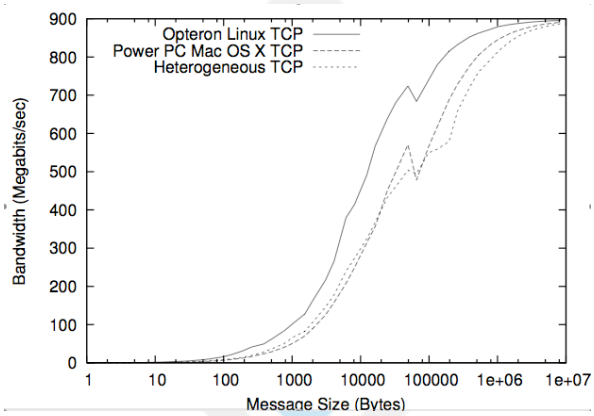
Support for Data and Network Heterogeneity




Processor Heterogeneity
Data Registry
My Proc (1)
My Proc (N)
PPC
Conv
Other Procs
Init Phase
X86
Conv
Other Procs
Conv 1
Conv 2
…
Conv N
Recv
Conv 1
Conv 2
…
Conv N
Compute Phase


Network Heterogeneity
Data Registry
My Proc (1)
My Proc (N)
Network
Other Procs
Init Phase
Network
Other Procs
Net 1
Net 2
…
Net N
Net 1
Net 2
…
Net N
Network:
Low latency list
High Latency list
Compute 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/