Open MPI: A High-Performance, Heterogeneous MPI

Richard L. Graham, Galen M. Shipman, Brian W. Barrett, Ralph H. Castain, George Bosilca

LA-UR-06-3453

Los Alamos NATIONAL LABORATORY EST.1943

NNSA

Open MPI Collaboration

- The University of Tennessee
- Indiana University
- HLRS
- The University of Huston
- Sandia National Laboratory
- LANL
- Cisco
- Mellanox
- Voltaire
- Sun Microsystems
- Myricom
- IBM
- QLogic

Los Alamos NATIONAL LABORATORY EST.1943

NNSA

Contributors

- LANL
  - Ralph Castain
  - David Daniel
  - Tim Woodall
- U. of Tennessee
  - George Bosilca
  - Graham Fagg
- Indiana University
  - Brian Barrett
- Cisco Systems
  - Jeff Squyres

Los Alamos NATIONAL LABORATORY EST.1943

NNSA

Outline

- Introduction
- Design for automation
  - Run time layer (Open RTE)
  - High performance communications layer (Open MPI)
- Future directions

Los Alamos NATIONAL LABORATORY EST.1943

NNSA

Goal of Heterogeneous Support

- Focus on library functionality
  - Job startup
  - Communications
- Reliable run-time
- High performance where required
  - Job initialization/termination
  - Communications

Los Alamos
NATIONAL LABORATORY
EST.1943

NNSA

Aspects of Heterogeneity

- Processor
- Network
- Run-time environment
- Application

Los Alamos
NATIONAL LABORATORY
EST.1943

NNSA

Aspects of Heterogeneity in Open MPI

- Run-time library (ORTE)
- High performance communications Library ==> Open MPI

Los Alamos NATIONAL LABORATORY EST.1943

NNSA

Design

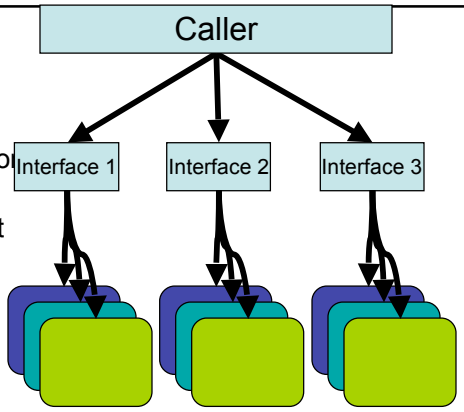
Los Alamos NATIONAL LABORATORY EST.1943

NNSA

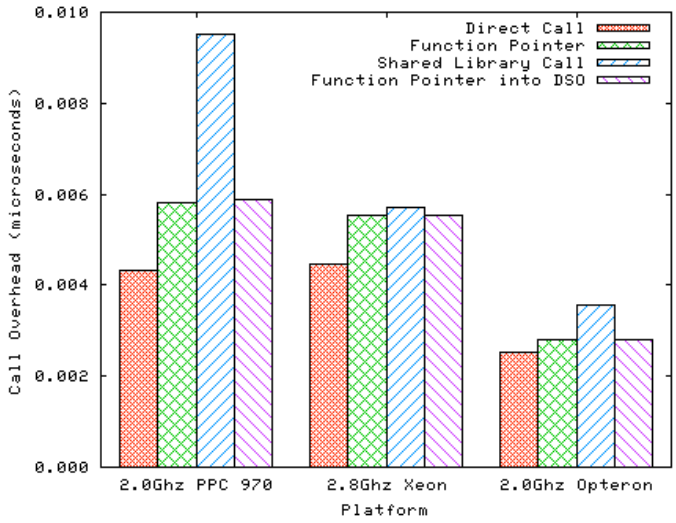
Components

- Formalized interfaces
  - Specifies “black box” implementation
  - Different implementations available at run-time
  - Can compose different systems on the fly

Heterogeneity Is Optional



Performance Impact

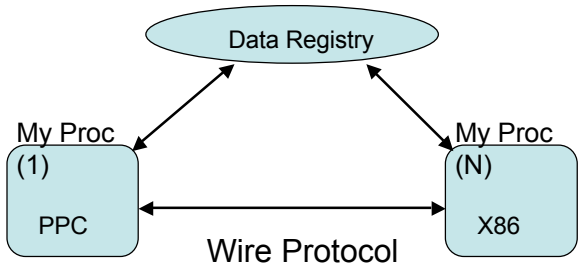


Run-Time

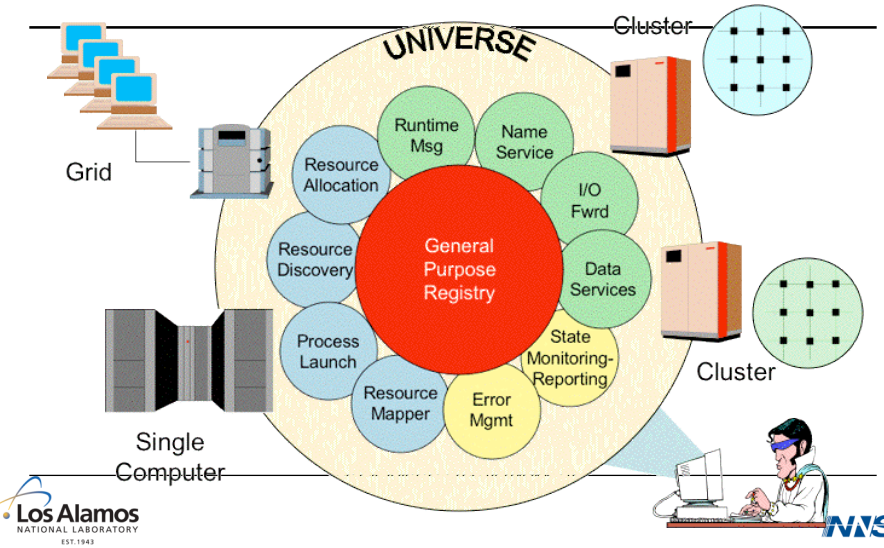
- Processor
- Multi-cell
- Application

A Key Idea

- Wire protocol (network byte order) used to bootstrap the run-time system



OpenRTE Architecture



General Purpose Registry

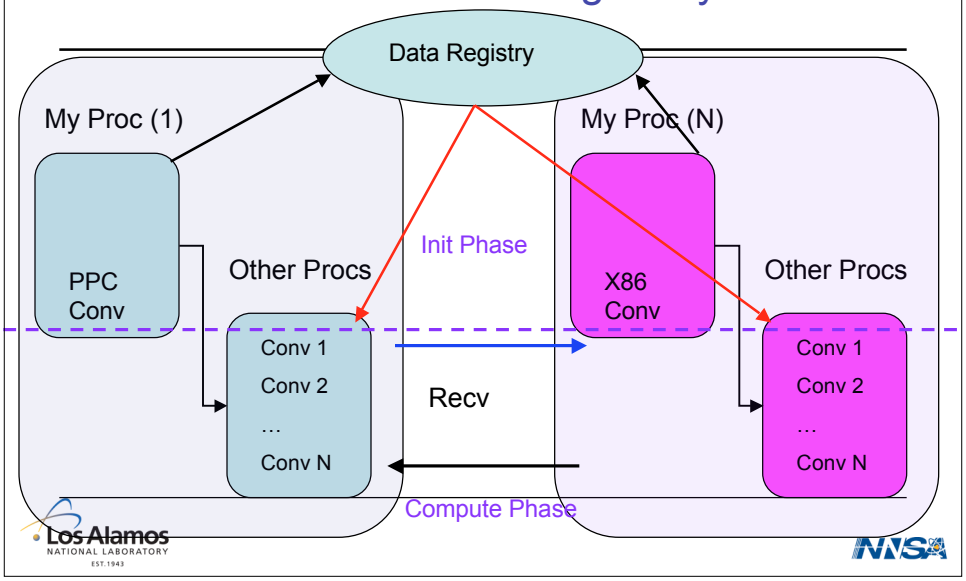
- Distributed data storage/retrieval system
  - All common data types plus user-defined
  - Heterogeneity between storing process and recipient automatically resolved
- Publish/subscribe
  - Support event-driven coordination and notification
  - Subscribe to individual data elements, groups of elements, wildcard collections
  - Specify actions that trigger notifications, information to be returned

Accessible to application programmers

MPI

- Processor
- Network
- Application

Processor Heterogeneity



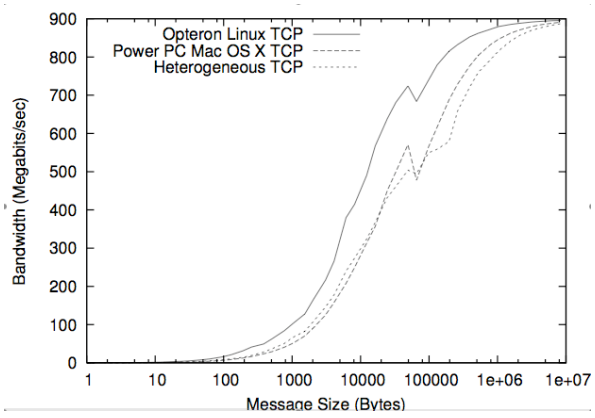
Machine Description

Byte	Bits	Description
1	1 - 2	Always 00, allowing recognition of endian encoding
	3 - 4	endian: 00 = little, 01 = big
	5 - 6	reserved: Always 00
	7 - 8	reserved: Always 00
2	1 - 2	length of long: 00 = 32, 01 = 64
	3 - 4	reserved for length of long long: Always 00
	5 - 6	length of C/C++ bool: 00 = 8, 01 = 16, 10 = 32
	7 - 8	length of Fortran LOGICAL: 00 = 8, 01 = 16, 10 = 32
3	1 - 2	length of long double: 00 = 64, 01 = 96, 10 = 128
	3 - 4	number of bits in the exponent of a long double: 00 = 01, 01 = 14
	5 - 7	number of bits of mantissa in a long double: 000 = 53, 001 = 64, 010 = 105, 011 = 106, 100 = 107,101 = 113
	8	Intel or SPARC representation of mantissa: 0 = SPARC, 1 = Intel
4	1 - 2	Always 11, allowing recognition of endian encoding
	3 - 4	reserved: Always 11
	5 - 6	reserved: Always 11
	7 - 8	reserved: Always 11

Data Conversions

- Endianess
- Size of data type (In progress)
- Data Representation (planned)

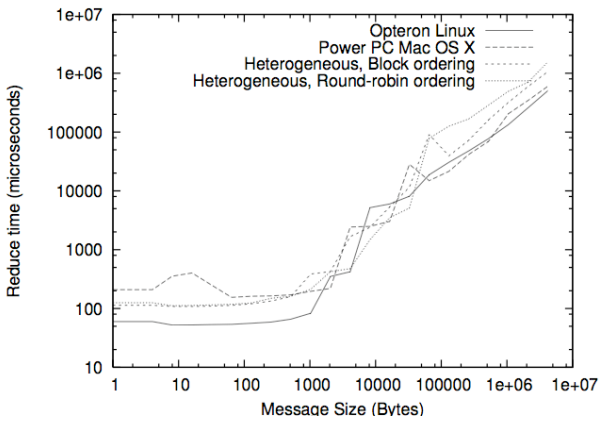
Netpipe B/W Measurement (TCP/IP)



Los Alamos NATIONAL LABORATORY EST.1943

NNSA

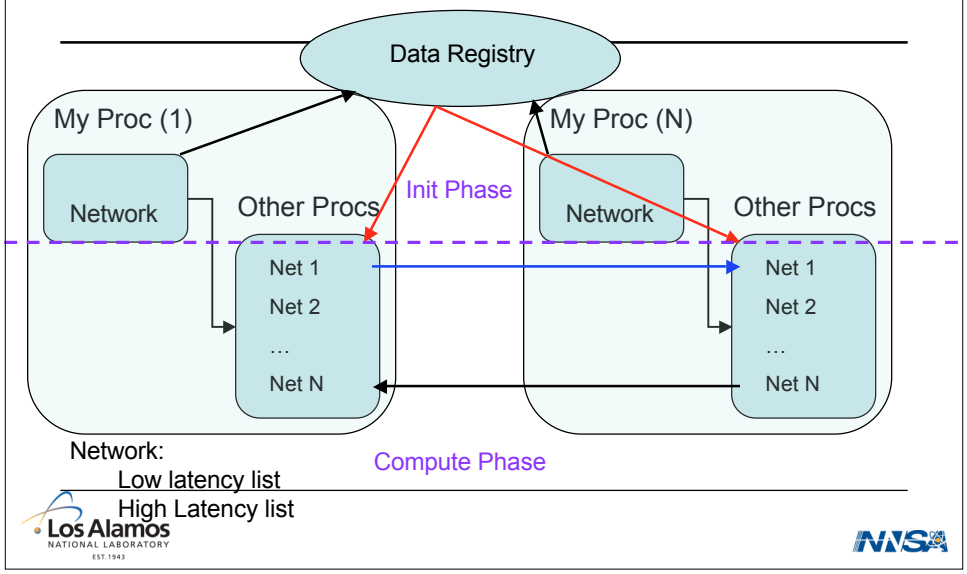
MPI_Reduce Data



Los Alamos NATIONAL LABORATORY EST.1943

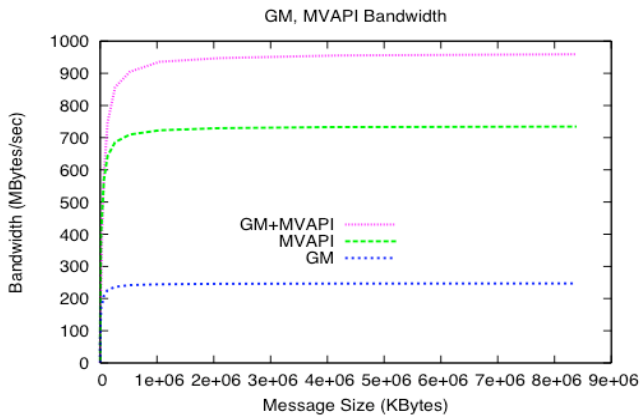
NNSA

Network Heterogeneity



Network:
- Low latency list
- High Latency list

Multi-NIC Ping-Pong Bandwidth



Visualization Display Benchmark (Paraview simmulation)

Network	Total time
GM only	24.92 sec
MVAPI only	8.53 sec
GM+MVAPI	6.55 sec

Application Heterogeneity

- Low level communication library does not assume any “symmetry” in the application
- Applications need to use library in a consistent manner

Future Work

- Continue to define/refine the multi-cell run-time environment
- Performance enhancements to the high performance communications library
- Scalability of the data registry
- Alternative implementations of the registry (DB's being investigated)

Los Alamos
NATIONAL LABORATORY
EST.1943

NNSA