Open MPI

Roy Kim

September 17, 2007

CISCO

Overview

- The Open MPI Project
- Cisco’s role in Open MPI

The Open MPI Project

Low Latency and MPI

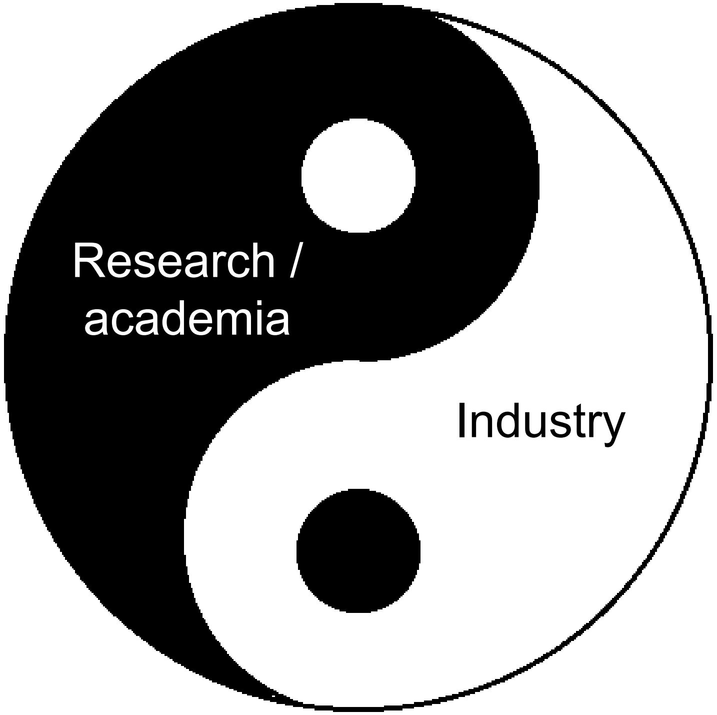
	Sockets API					MPI	
	TCP			SDP		Open MPI	
	IP	IPoIB					
	Gigabit Ethernet	SDR IB	DDR IB	SDR IB	DDR IB	SDR IB	DDR IB
Latency (us)	45.7	20.3	14.8	10	8.8	3.6	3.2
Bandwidth MB/s	118	560	584	896	1033	960	1350
CPU	9%	23%	26%	27%	28%	25%	25%

Open MPI Is…

- Open source project
- Consolidation and evolution of several prior MPI implementations
- All of MPI-1 and MPI-2
- Production quality
- Vendor-friendly
- Research- and academic-friendly

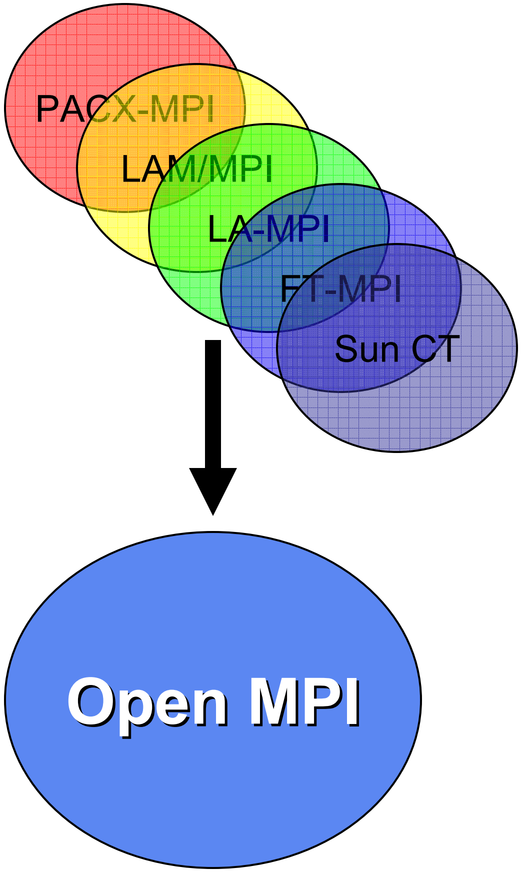
Why Does Open MPI Exist?

- Maximize all MPI expertise
  - Research / academia
  - Industry
  - …elsewhere
- Capitalize on [literally] years of MPI research and implementation experience
- The sum is greater than the parts



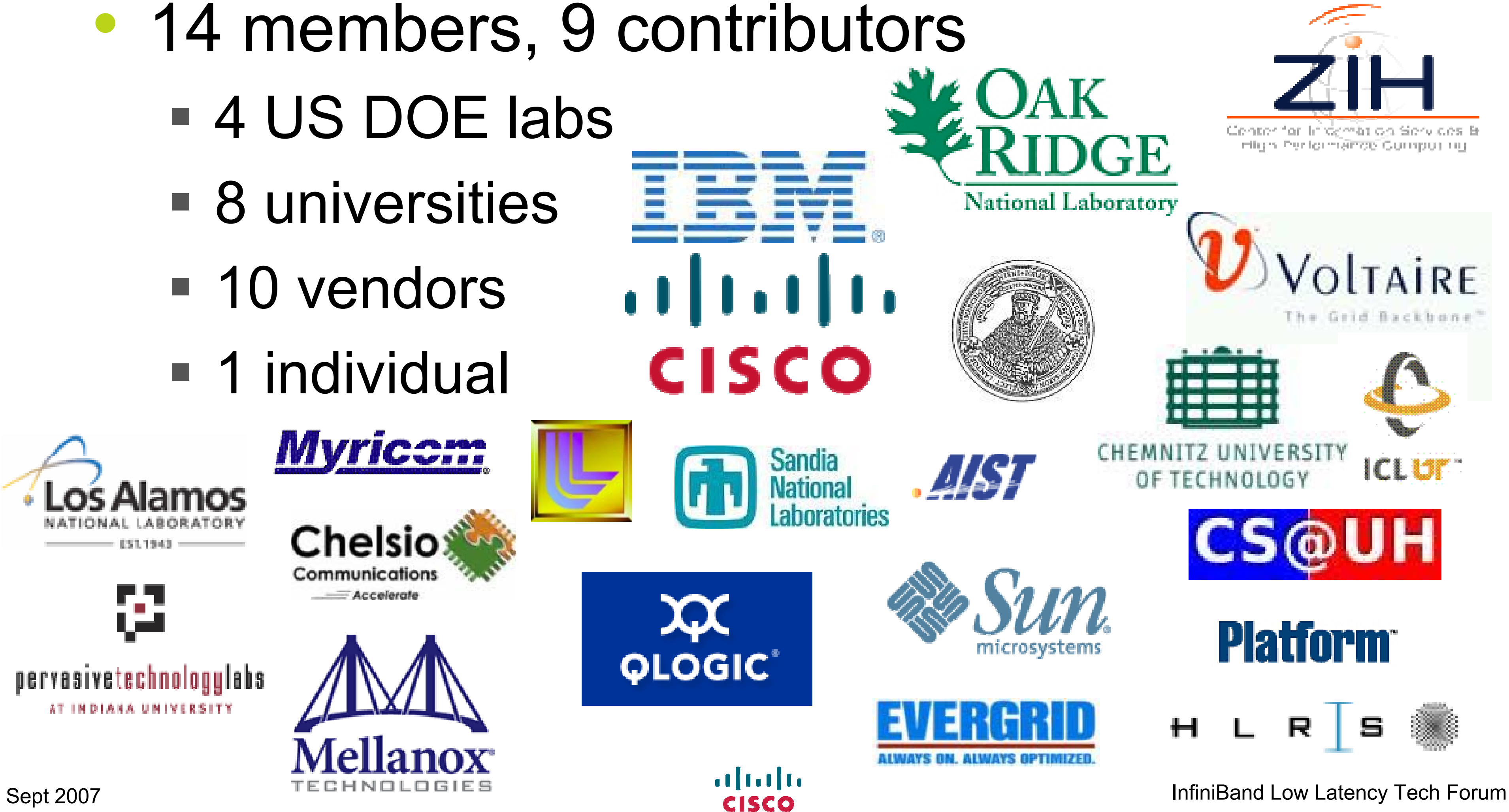
Prior MPI Projects

- Merger of experience from
  - FT-MPI (U. of Tennessee)
  - LA-MPI (Los Alamos)
  - LAM/MPI (Indiana U.)
  - PACX-MPI (HLRS, Stuttgart)
  - Sun ClusterTools
  - …



Current Membership

- 14 members, 9 contributors
  - 4 US DOE labs
  - 8 universities
  - 10 vendors
  - 1 individual


ZIH
Center for Information Services & High Performance Computing
OAK RIDGE National Laboratory
IBM
VOLTAIRE
The Grid Backbone
CISCO
CHEMNITZ UNIVERSITY OF TECHNOLOGY
ICL UT
Los Alamos NATIONAL LABORATORY EST.1943
Myricom
Sandia National Laboratories
AIST
CS@UH
Chelsio Communications Accelerate
QLOGIC
Sun microsystems
Platform
pervasivetechnologylabs AT INDIANA UNIVERSITY
Mellanox TECHNOLOGIES
EVERGRID ALWAYS ON. ALWAYS OPTIMIZED.
H L R I S
CISCO

Cisco’s Role in Open MPI

Open Source

- Open source community can be useful
  - Lots of “free” work for Cisco
  - …even for features that we care about!
- But not a one-way street
  - [Cisco] Must give in order to receive
  - Cannot expect free work for nothing

Cisco Open MPI Participation

- Cisco is an active leader in Open MPI
  - [Pro]Active community involvement
- Resources
  - 40 node quad core MPI development cluster
  - 128 + 64 node dual core regression testing clusters
  - ~100,000 MPI regression tests run every night
  - Extensive QA testing during IB driver releases

Cisco Open MPI Employees

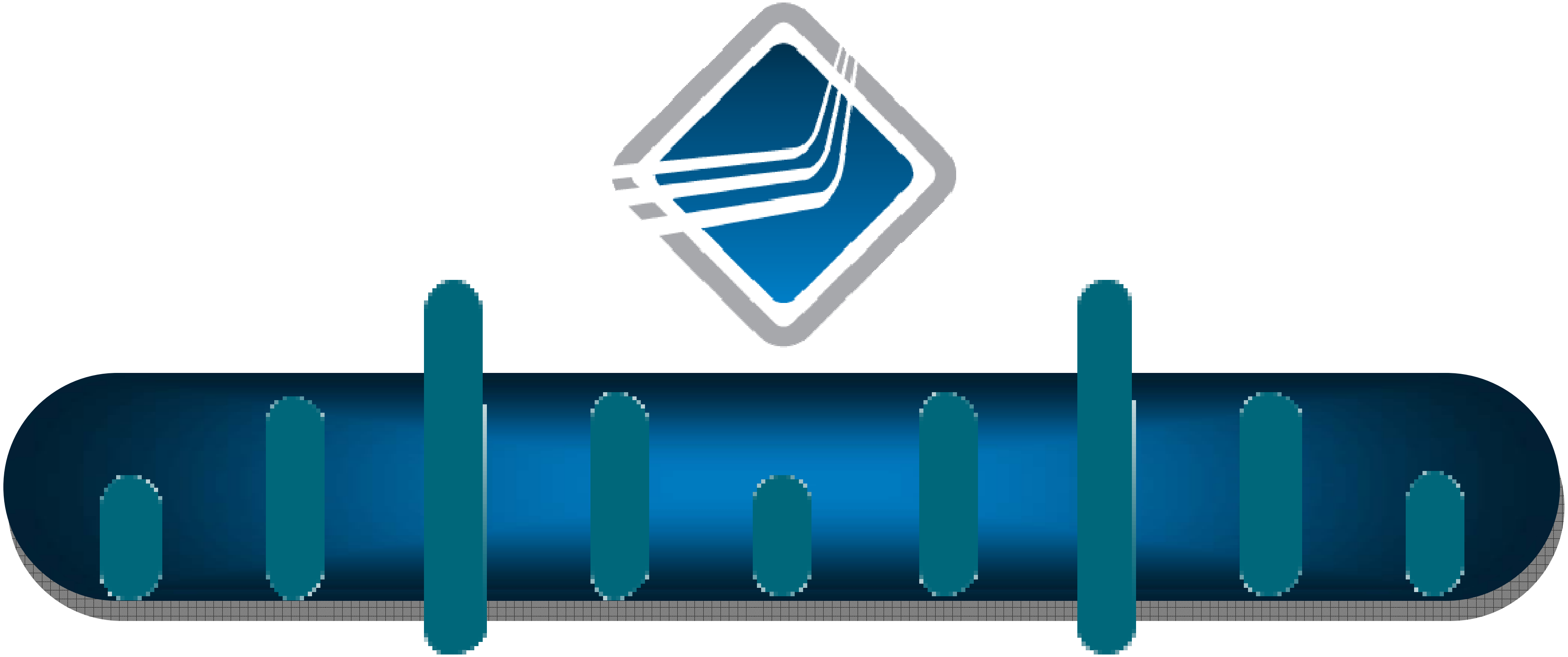
- Full-time employee
  - Jeff Squyres, Ph.D.
  - Co-founder of Open MPI
  - Chapter author in MPI-2 standard
- Summery 2007 intern
- ***Cisco is hiring to expand its MPI team!***
  - Go to www.cisco.com / Career Opportunities
  - Search for keyword “MPI”

Other Cisco Open MPI Efforts

- Cisco Research Center (CRC)
  - Provides funding for academic research
  - Up to $100K grants (unrestricted) each
- Cisco Open MPI development current funding
  - Indiana University
  - University of Houston

Conclusion

- Cisco dedicated tackling the problem of latency
- Cisco believes in the future direction of Open MPI



CISCO

Backup (Software Architecture)

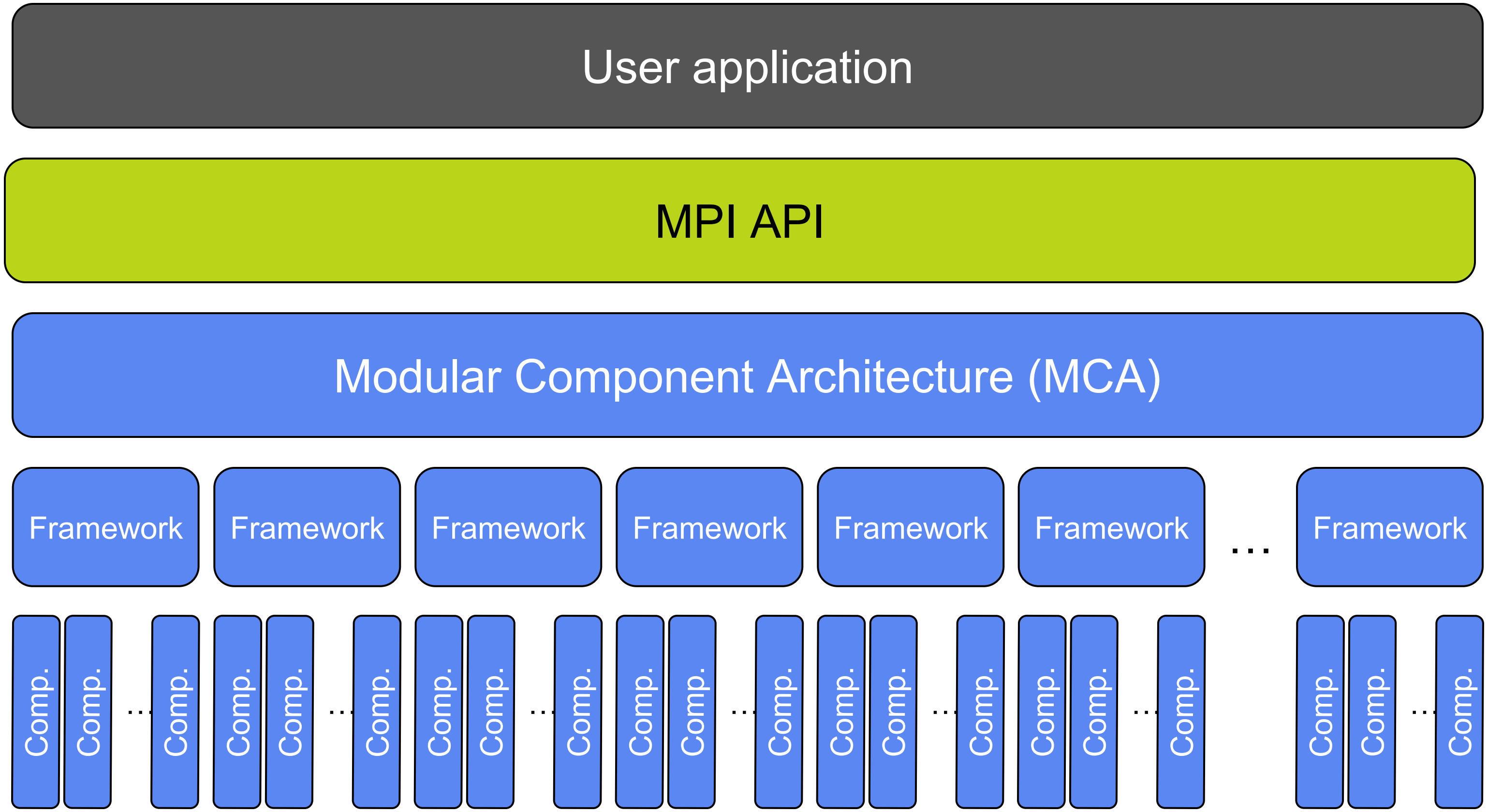
Technology Overview

- A few core libraries
  - Core functionality written in C
  - Some assembly for speed
  - External bindings provided for C++, F77, F90
- Most functionality is in plugins
  - Plugins can be compiled / distributed separately from main code base
  - Plugins can be distributed as open or closed source

Plugins

- Lots and lots of plugin types
  - Back-end network
  - Resource manager support
  - Operating system support
    - Processor Affinity, Memory Affinity, Assembly Language
- All can be loaded (or not) at run-time
  - Choice of network is a run-time decision
  - *User applications no longer linked against network libraries (e.g., libibverbs)*
  - Companion concept: run-time parameters

Plugin High-Level View



Network Agnostic

- Networks currently supported (plugins):
  - OpenFabrics, TCP, shared memory, mVAPI, …others
- True multi-device support
  - Multi-network, multi-device
  - Any mix of any network in a single job – each is just a run-time loaded plugin
- Open MPI will try to use any network that you have (unless told not to)

Resource Manager Agnostic

- Resource managers currently supported (plugins):
  - PBS/Torque, SLURM, BProc, N1GE / SGE, rsh/ssh, Xgrid, Yod, LoadLeveler
  - Work starting on LSF
- Currently assumes running in a scheduled job (no need to provide a hostfile)
  - Future directions include submitting a job

Operating System Agnostic

- 100% user-space code
  - Extremely comprehensive “configure” script
- Linux
  - [Just about] Any distribution, any version
- Others
  - OS X, Solaris, Windows (!)
  - AIX no longer supported
  - No demand for others [yet]

Compiler Agnostic

- Mostly
  - Portability problems usually have to do with assembly code and are easily solved
- Regularly test with (Linux / OS X):
  - GNU 3.x, 4.x
  - Intel 8.x, 9.x
  - PGI 6.x, 7.x (5.x *probably* works)
  - Pathscale 3.0