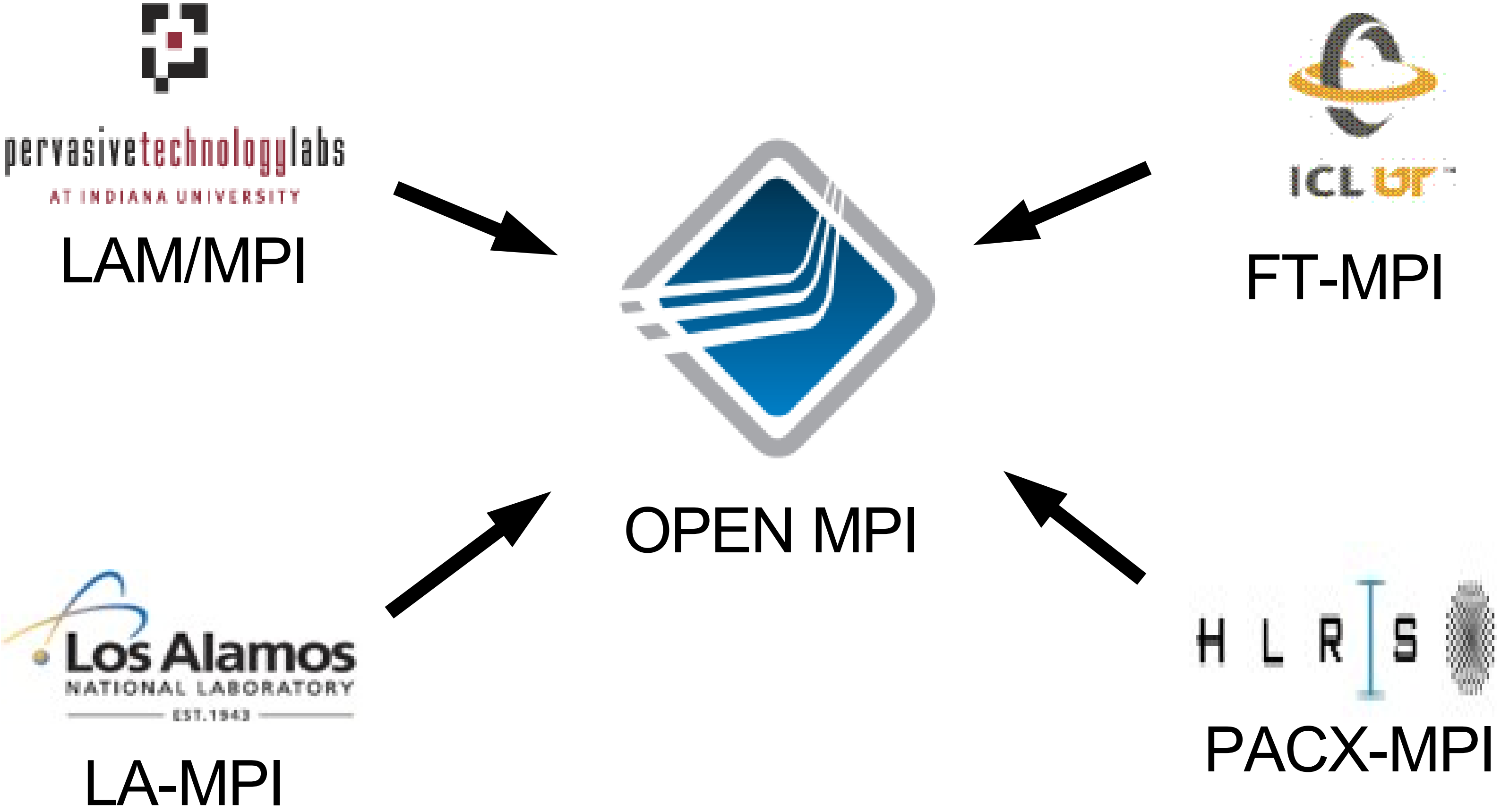
OSGC
May 15, 2008

On Balancing
Open-Source Cluster Development Between Industry and Research

Len Wisniewski
Engineering Manager
Software Developer Tools and Services
Sun Microsystems

In the beginning...



LAM/MPI

FT-MPI

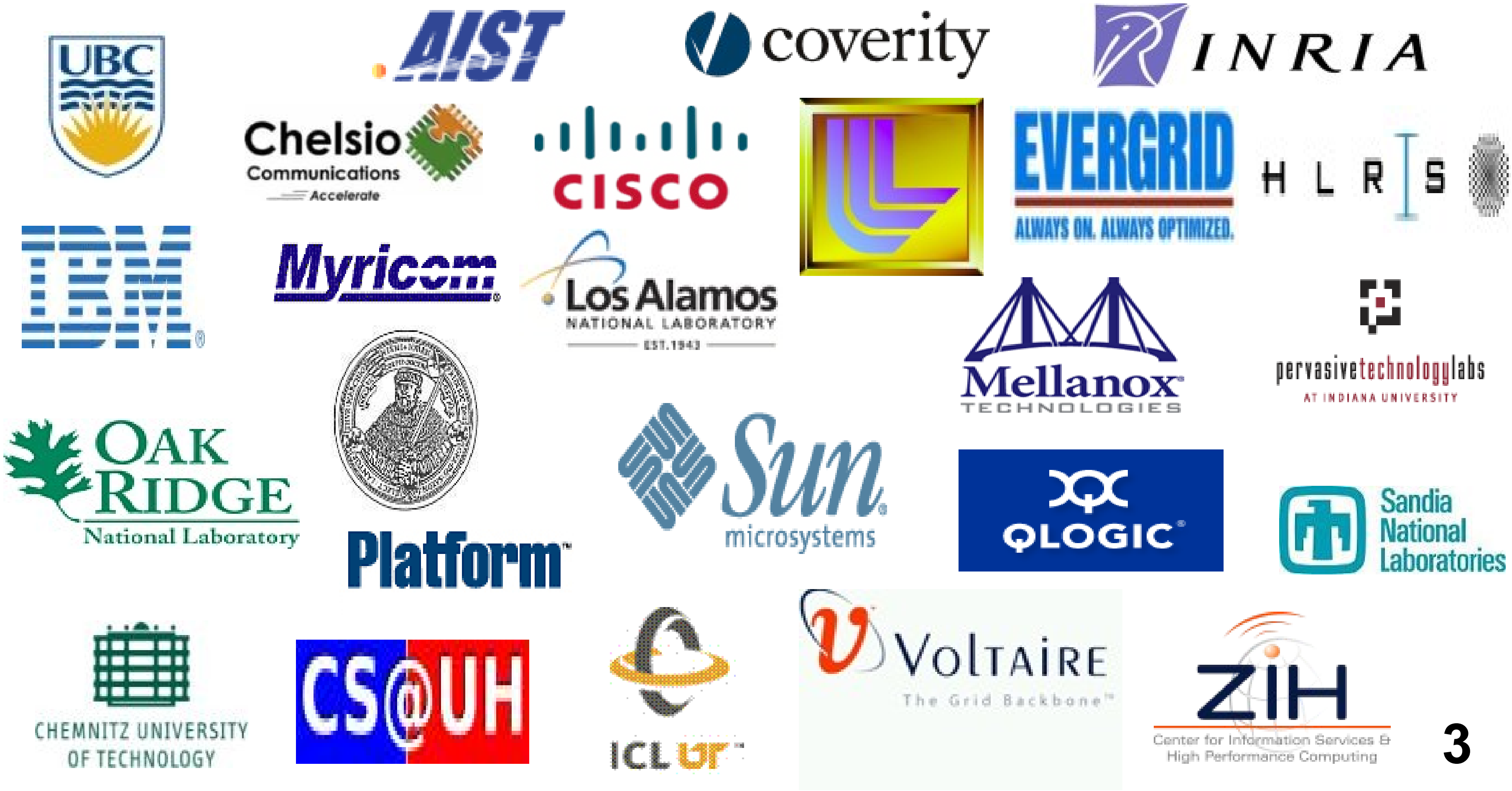
OPEN MPI

LA-MPI

PACX-MPI

Now today...

- Currently 15 Members, 9 Contributors, 1 Partner
- Plus individual contributors


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

Today's members in categories

Labs

Los Alamos
NATIONAL LABORATORY
EST. 1943

OAK RIDGE
National Laboratory

Sandia
National
Laboratories

INRIA

AIST

Academia

pervasivetechnologylabs
AT INDIANA UNIVERSITY

ICL UT™

HLRS

CS@UH

ZIH
Center for Information Services &
High Performance Computing

CHEMNITZ UNIVERSITY
OF TECHNOLOGY

UBC

Industry

Chelsio
Communications
Accelerate

CISCO

IBM®

Mellanox
TECHNOLOGIES

QLOGIC®

Myricom®

Sun
microsystems

Voltaire
The Grid Backbone™

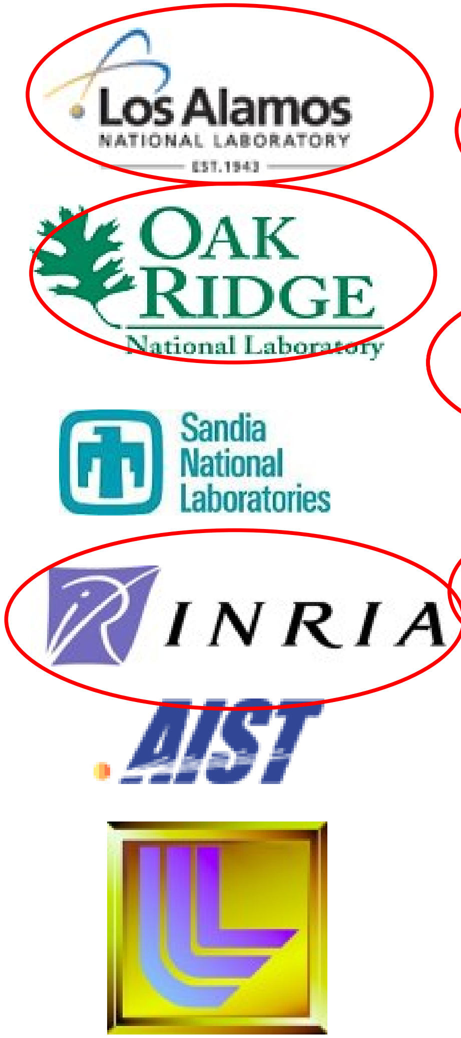
EVERGRID
ALWAYS ON. ALWAYS OPTIMIZED.

coverity

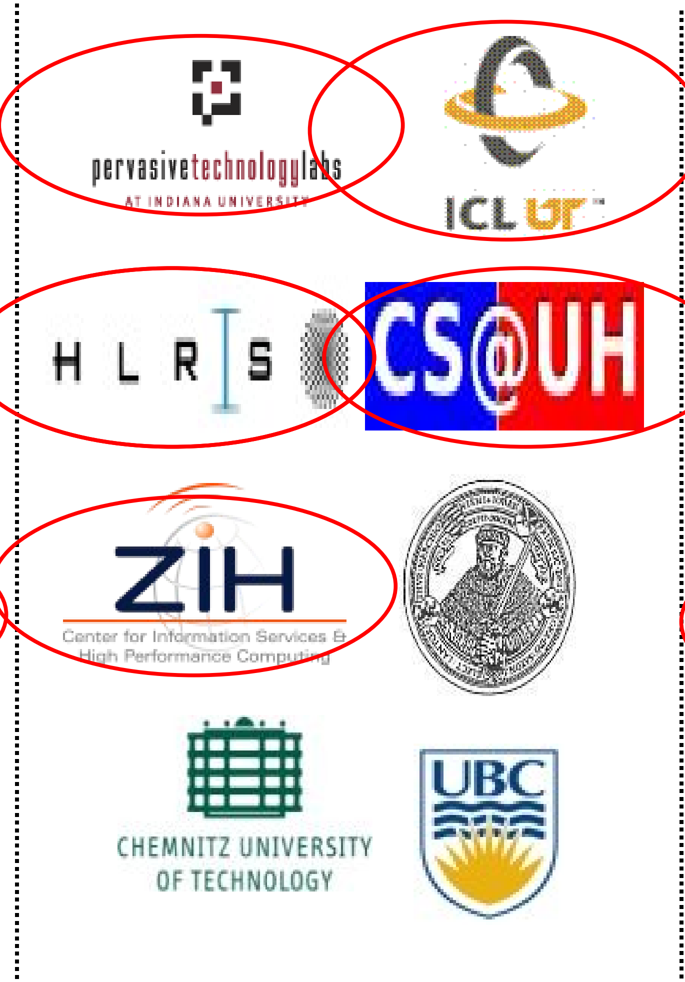
Platform™

Member organizations

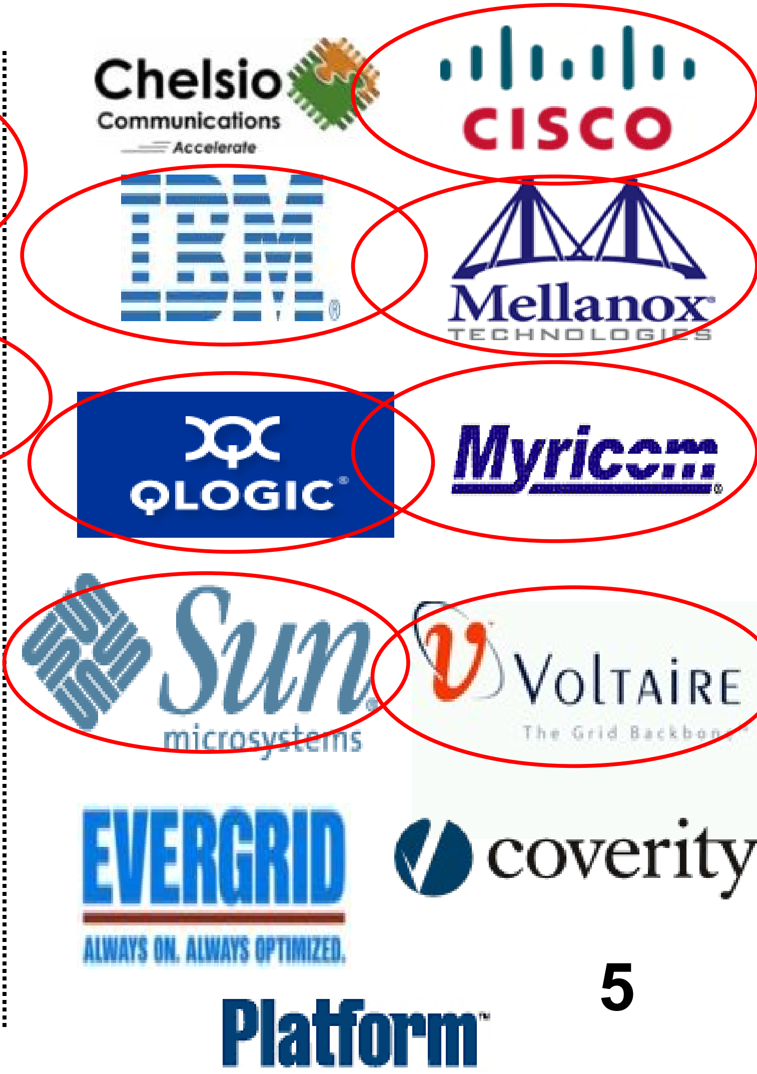
Labs



Academic


pervasivetechnologylabs
AT INDIANA UNIVERSITY
ICL UT
HLRS
CS@UH
ZIH
Center for Information Services & High Performance Computing
CHEMNITZ UNIVERSITY OF TECHNOLOGY
UBC

Industry



Blurred lines of self-interest

- Labs
  - A. Support running MPI jobs on production clusters
  - B. Support MPI research concepts
  - C. Support vendor platforms and tools
- Academia
  - B. Support MPI research concepts
  - A. Support running MPI jobs on production clusters
  - C. Support vendor platforms and tools
- Industry
  - A. Support running MPI jobs on production clusters
  - C. Supporting vendor platform
  - B. Support MPI research concepts

Open MPI goals

- Create a free, open source, peer-reviewed, production-quality complete MPI-2 implementation.
- Provide extremely high, competitive performance (latency, bandwidth, ...pick your favorite metric).
- Directly involve the HPC community with external development and feedback (vendors, 3rd party researchers, users, etc.).
- Provide a stable platform for 3rd party research and commercial development.
- Help prevent the "forking problem" common to other MPI projects.
- Support a wide variety of HPC platforms and environments.

www.open-mpi.org

Self-interest (specifically)

- Labs
  - Supporting large clusters (like LANL Road Runner, Sandia Thunderbird, and ORNL Cray machines)
- Academia
  - Research projects
    - Fault tolerance (Univ of Tennessee)
    - Checkpoint / restart (Indiana Univ)
    - Hierarchical collectives (Univ of Houston)
  - Supporting large clusters (like Red Storm)
  - Supporting additional OS types like Mac OS and Windows
- Industry
  - Network vendors are most interested in ensuring network stack works properly
  - Sun interested in ensuring all components of Sun HPC stacks work properly
  - Systems vendors interested in supporting large customer configs (like TACC Ranger and Road Runner)

Self-interest translated into community roles

- Labs
  - Drive super-scale issues
    - e.g., Scalable job startup and collectives
- Academia
  - Drive research projects
- Industry
  - Drive platform support
    - e.g., Sun drives support of Sun Grid Engine, Solaris, and Sun Studio, and supported third party tools
    - http://www.sun.com/clustertools
    - e.g., network vendors drive development of OFED Verbs Byte Transfer Layer (BTL)

Back to the goals...

- What do all Open MPI members need?
  - Production quality code
  - Stability
    - Super-scale cannot impact lower scale
    - Research cannot regress existing functionality
    - Platform support cannot regress other platforms or research or super-scale
- Important ways to achieve these
    - Sound software engineering process / practice
    - MPI Testing Tool !

A Day in the Life of the Open MPI Community

- Direct collaboration
- Weekly Open MPI concalls
- Shared Bug Database / Wiki
- Subversion (web-based) source control
  - Shared source code, tests, documents
- Community Releases
  - Release Managers
  - Gatekeepers
- In-person meetings
  - Quarterly meetings
  - Euro PVM / MPI conference
  - MPI Forum
- Community mailing lists
- Coordinated Supercomputing presence

www.open-mpi.org

Nano Template Copyright 2005 DecalGirl.com. Allow 1/8" from cut lines for bleed region.

Open MPI Bug data

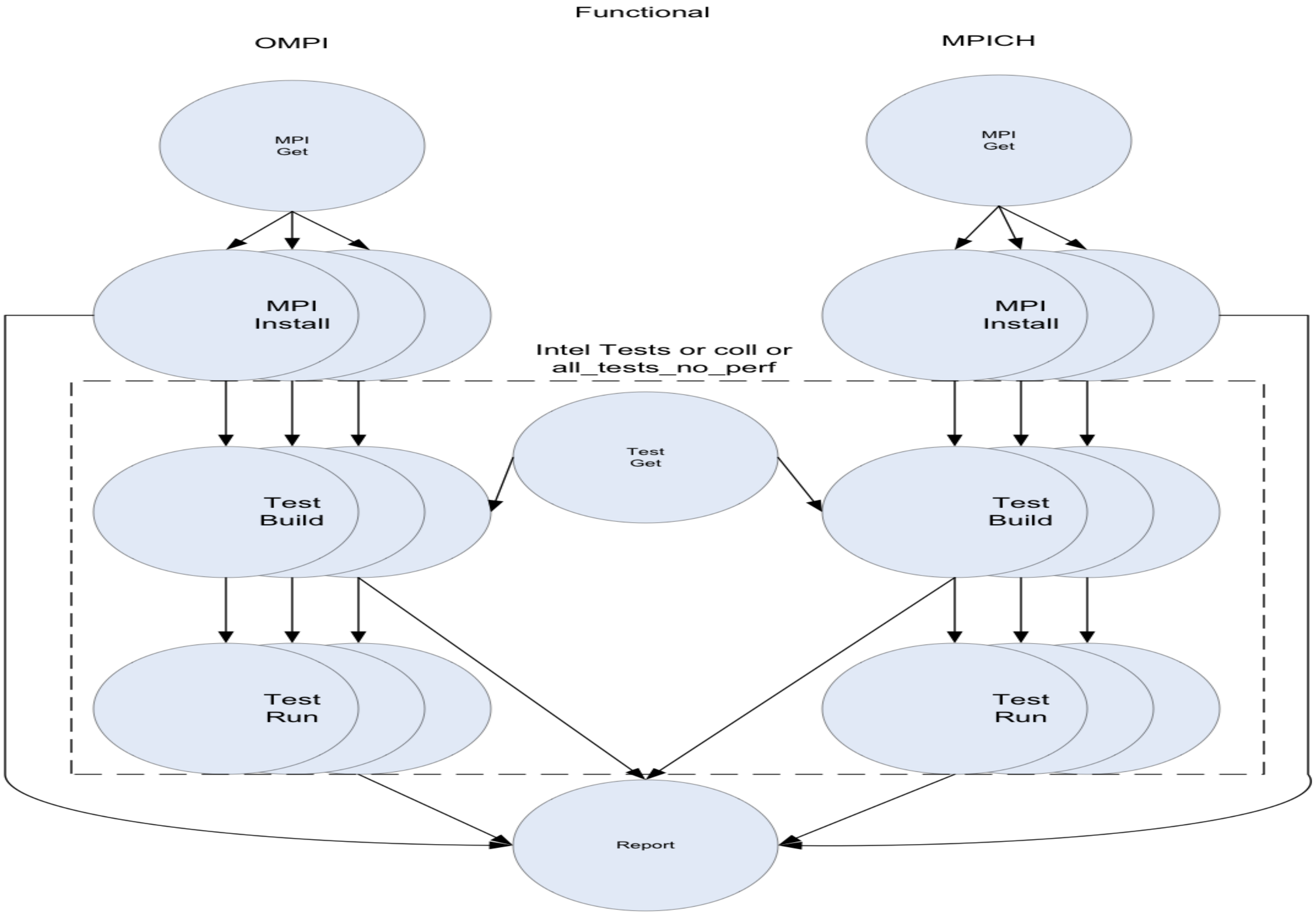
- 1.2 series
  - Submitted
    - Labs 23%, Academia 14%, Industry 63%
  - Fixed
    - Labs 48%, Academia 20%, Industry 31%

- Upcoming 1.3 series
  - Submitted
    - Labs 12%, Academia 21%, Industry 67%
  - Fixed
    - Labs 12%, Academia 31%, Industry 57%

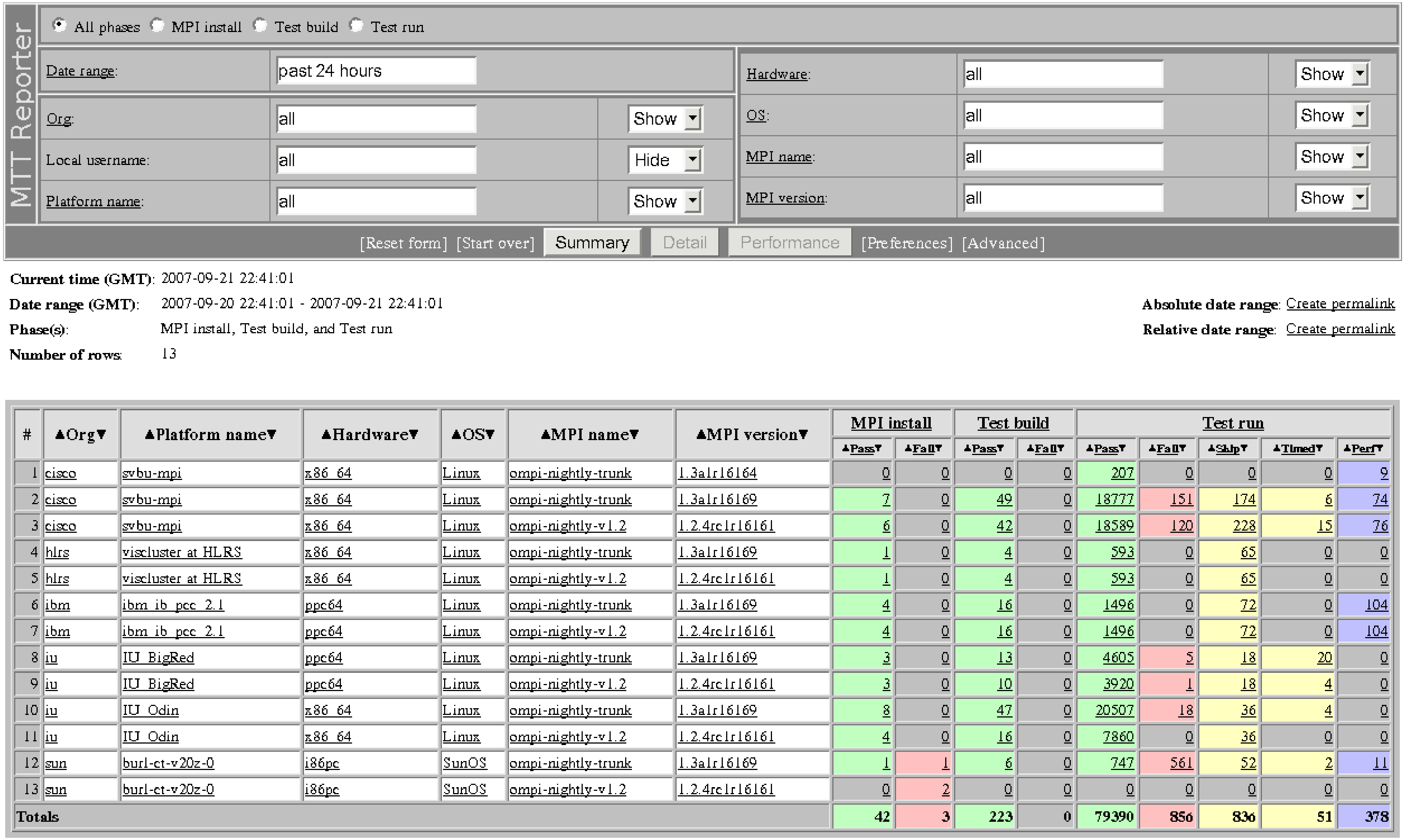
- Sample sizes small, can be skewed by individual's current affiliation, or by differing engineering habits
- Most industry members new to codebase for 1.2
- Industry taking on greater share of fixing in 1.3

MPI Testing Tool

- An Extensible Framework for Distributed Testing of MPI Implementations. Joshua Hursey (Indiana U), Ethan Mallove (Sun), Jeffrey M. Squyres (Cisco) and Andrew Lumsdaine (Indiana U) 14th PVM / MPI Conference, Paris, France, October 2007.



MTT Results Summary example



MTT Reporter

● All phases ○ MPI install ○ Test build ○ Test run

Date range:	past 24 hours	
Org:	all	Show
Local username:	all	Hide
Platform name:	all	Show
Hardware:	all	Show
OS:	all	Show
MPI name:	all	Show
MPI version:	all	Show

[Reset form] [Start over] Summary Detail Performance [Preferences] [Advanced]

Current time (GMT): 2007-09-21 22:41:01

Date range (GMT): 2007-09-20 22:41:01 - 2007-09-21 22:41:01

Phase(s): MPI install, Test build, and Test run

Number of rows: 13

Absolute date range: Create permalink

Relative date range: Create permalink

#	▲Org▼	▲Platform name▼	▲Hardware▼	▲OS▼	▲MPI name▼	▲MPI version▼	MPI install		Test build		Test run				
							▲Pass▼	▲Fail▼	▲Pass▼	▲Fail▼	▲Pass▼	▲Fail▼	▲Skip▼	▲Timed▼	▲Perf▼
1	cisco	svbu-mpi	x86_64	Linux	ompi-nightly-trunk	1.3a1r16164	0	0	0	0	207	0	0	0	9
2	cisco	svbu-mpi	x86_64	Linux	ompi-nightly-trunk	1.3a1r16169	7	0	49	0	18777	151	174	6	74
3	cisco	svbu-mpi	x86_64	Linux	ompi-nightly-v1.2	1.2.4rc1r16161	6	0	42	0	18589	120	228	15	76
4	hlrs	viscluster at HLRS	x86_64	Linux	ompi-nightly-trunk	1.3a1r16169	1	0	4	0	593	0	65	0	0
5	hlrs	viscluster at HLRS	x86_64	Linux	ompi-nightly-v1.2	1.2.4rc1r16161	1	0	4	0	593	0	65	0	0
6	ibm	ibm_ib_ppc_2.1	ppc64	Linux	ompi-nightly-trunk	1.3a1r16169	4	0	16	0	1496	0	72	0	104
7	ibm	ibm_ib_ppc_2.1	ppc64	Linux	ompi-nightly-v1.2	1.2.4rc1r16161	4	0	16	0	1496	0	72	0	104
8	iu	IU_BigRed	ppc64	Linux	ompi-nightly-trunk	1.3a1r16169	3	0	13	0	4605	5	18	20	0
9	iu	IU_BigRed	ppc64	Linux	ompi-nightly-v1.2	1.2.4rc1r16161	3	0	10	0	3920	1	18	4	0
10	iu	IU_Odin	x86_64	Linux	ompi-nightly-trunk	1.3a1r16169	8	0	47	0	20507	18	36	4	0
11	iu	IU_Odin	x86_64	Linux	ompi-nightly-v1.2	1.2.4rc1r16161	4	0	16	0	7860	0	36	0	0
12	sun	burl-ct-v20z-0	i86pc	SunOS	ompi-nightly-trunk	1.3a1r16169	1	1	6	0	747	561	52	2	11
13	sun	burl-ct-v20z-0	i86pc	SunOS	ompi-nightly-v1.2	1.2.4rc1r16161	0	2	0	0	0	0	0	0	0
Totals							**42**	**3**	**223**	**0**	**79390**	**856**	**836**	**51**	**378**

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

Applying MTT to cluster development

- MTT could be used for tests other than just MPI tests
- Central MTT repository with open results
    - Hosted by Indiana University
- Nightly runs on broad range of configs
- Future possibilities
    - Automatic regression searchs
        - Find which revision errors started occurring
        - Analyze least common denominator of failures
        - Determine at what scale (nodes, processes)
        - Extend nightly runs to center in on failure cases
    - Extend to grid testing?
        - Heterogeneous testing
            - Open MPI supports heterogeneous clusters
        - One test launch utilizes full range of grid resources to track down issues

Conclusions

- Competing self-interest can allow a code base to become more robust while innovation proceeds when handled openly with careful monitoring
- Open-source development relies on the willingness of its contributors to reach consensus and explore alternative solutions when conflicting self-interest arises
- Consensus on minimum test qualifications and the ability to share and search test data is important for achieving stability in a timely manner
- As clusters (and grids) become more prominent, automated tools for quickly identifying coding errors in a scalable and/or broadly diverse support base become as important as tools for identifying faulty hardware
- Creating an open environment enables new active participation from interested parties with new ideas and tools to contribute
- Being part of a vibrant community is fun and cool!

Len Wisniewski

leonard.wisniewski@sun.com