Open MPI State of the Union Community Meeting SC '10

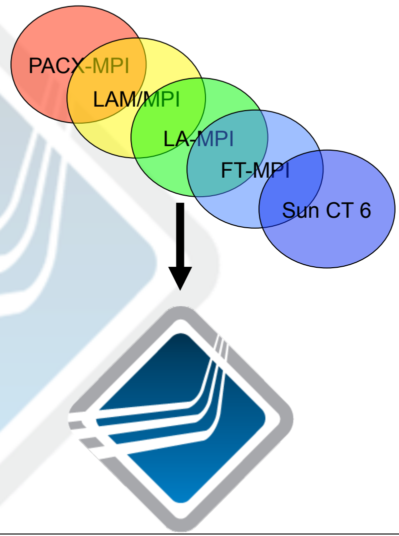
Dr. Jeff Squyres, Dr. George Bosilca,
Dr. Brice Goglin

Agenda

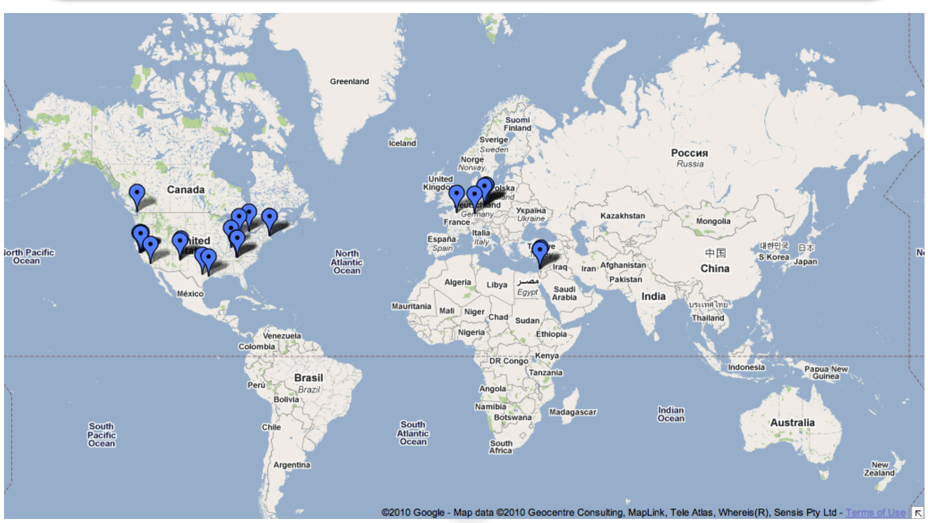
- Open MPI Project / Community
- Current Status
  - v1.4.x series
  - v1.5 series
- Select organization project updates
  - U. Tennessee Knoxville, Cisco, INRIA Bordeaux
- The road to MPI-3

Open MPI Is…

- Evolution of several prior MPI's
- Open source project and community
  - Production quality
  - Vendor-friendly
  - Research- and academic-friendly
- All of MPI-1 / MPI-2

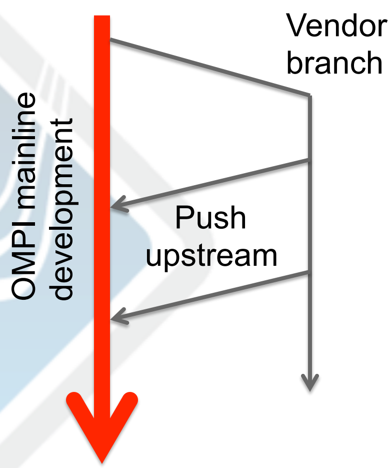


Members, Contributors, Partners



Shout Out to Vendors

- Distributions:
  - Oracle Message Passing Toolkit
  - Bull MPI
  - Voltaire Fabric Collective Accelerator
  - Mellanox Collective Offload
- Pushing most non-proprietary work back upstream



Shout Out to Packagers

- Debian and *BSD packagers quite active
  - Sending patches upstream to us
  - Testing, iterating, helping us fix portability issues
- GNU Autotools maintainers
  - Also send patches to us
- Many, many thanks for your efforts!

Version 1.4 series

George Bosilca

Open MPI 1.4

- First release December 2009
- Current release 1.4.3
- Release Managers:
  - Brad Benton (IBM)
  - George Bosilca (UTK)
- Open MPI has two concurrent release series:
  - "Super stable": for production users (even minor)
  - "Feature driven": not that bleeding edge (odd minor)
  - Trunk for everybody else …

Open MPI 1.4

- Somewhat boring!
  - “Stable” != “Sexy” / “interesting” / etc.
  - But “stable” === “good”
- Let’s discuss the feature list for the v.1 4 series…

1.4 Series Feature List

- Native Windows support
- [Improved] online and offline documentation
- Condensed error messages
- ABI compatibility between versions
  - As long as the MPI doesn’t change your linked applications will run independent on the Open MPI version available (starting with 1.3.2)
- The notifier framework

1.4 Series Feature List

- Thread safety
  - PML OB1 is thread safe
- MPI_THREAD_MULTIPLE
  - Support included for more devices
  - Only the point-to-point and collective support have been tested
- Fixed race conditions with newer compilers / platforms in the shared memory BTL
  - Difficult to hand write assembly code for all compilers

1.4 Series Feature List

- Processor affinity
- Various Fortran fixes
  - Error handlers
  - Array conversion to C
- Improved support for job schedulers
- Full support for Vampir Trace
- Improved singleton regarding support for dynamic processes
- Wrapper compilers (mpicc & friends)

1.4 Series Feature List

- Fault Tolerance
  - Coordinated checkpoint/restart
  - Uncoordinated checkpoint/restart
    - Improved Message Logging (under 5% overhead).
  - Support BLCR and self
  - Able to handle real process migration (i.e. change the network during the migration)
    - MX, IB, TCP, SM, self

Low Level Devices (BTL) Status

Network	Dynamic Processes	Threading support
Self		
Shared Memory		
TCP		
Myrinet (MX)*		
Myrinet (GM)		
Infiniband (openib)		
Infiniband (ofud)		
Elan		
Sicortex		
Portals*		
uDAPL		
SCTP		

•Fully MPI 2.1 compliant: all BTL devices support MPI 1 (pt-to-pt) and MPI 2 (RDMA) communications
•All devices support PERUSE, additional features from the upcoming MPI 3.0 Tools working group
•Table on left shows BTL dynamic / threading status

NOTE: MTL components (*) do not support threading
•Use BTL equiv. (if available)
•MX, Portals, PSM

Version 1.5 series

Jeff Squyres

Version 1.5(.0)

- First release in new feature series
  - Released Oct 10, 2010
- Release managers
  - Jeff Squyres, Cisco
  - Rainer Keller, HLRS

Major Features (so far)

- Linux KNEM support
- Broke ABI
- Revamped run-time support
- Some (but not all) MPI-2.2 support
- Scalability enhancements
- Dynamic process improvements
- Portability updates
  - BSD, Catamount, Windows, OS X, Solaris
- Millions of other little improvements, updates, and bug fixes

1.5.x Roadmap

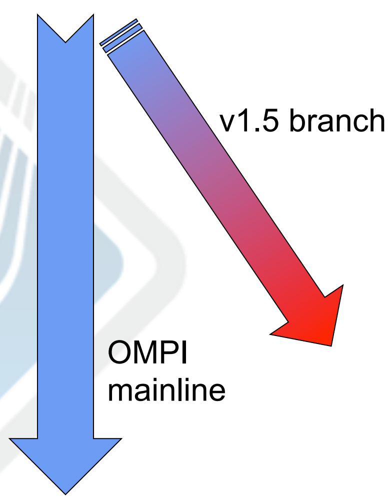
- 1.5.1 to be released “soon”
  - Minor bug fixes against 1.5.0
  - No real new features
  - Too stressful to do a correct SC release
- Expected in December 2010

1.5.x Roadmap

- 1.5.2
  - ROMIO refresh
  - Hwloc hardware affinity (back-end)
  - Linux UMMU notify (possible)
  - “Better” process affinity (more in later slides)
- 1.5.3
  - “It depends”

1.5.x Roadmap

- v1.5(.0) took a looong time to release
- Development mainline has diverged greatly from v1.5 branch
  - There are many, many new features available on the mainline
  - Probably will not come over to v1.5 branch
- Still deciding what to do



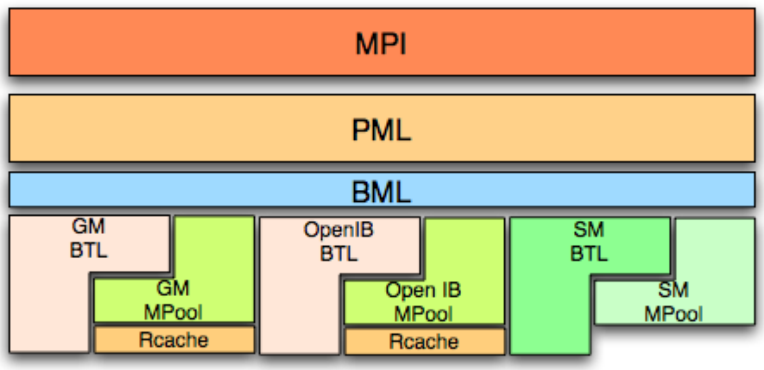
U. Tennessee Updates

George Bosilca

MPI compliance

- MPI 2.1 compliant
- MPI 2.2 – missing parts:
  - MPI_Exscan add MPI_IN_PLACE
  - Fortran/C datatypes
    - MPI_LONG_LONG_INT, MPI_LONG_LONG (as synonym), MPI_UNSIGNED_LONG_LONG, MPI_SIGNED_CHAR, and MPI_WCHAR are now officially supported
    - MPI_(U)INT {8,16,32,64 }_T, MPI_AINT, MPI_OFFSET, MPI_C_BOOL, MPI_C_COMPLEX, MPI_C_FLOAT_COMPLEX, MPI_C_DOUBLE_COMPLEX, and MPI_C_LONG_DOUBLE_COMPLEX
  - MPI_Dist_graph_*

Reminder: OMPI Internals



KNEM & Hwloc

- Create multiple shared memory BTL based on the process distribution inside a "fat" node
    - Each BTL with its own configuration parameters
- Use the RMA interface of KNEM to adapt the collective communications to the underlying topology
    - Tremendous improvements in performance (between 50 and 94%)

Threading

- Playground for asynchronous progress
  - Design ready for TCP
  - Implementation underway
- MPI_THREAD_MULTIPLE
  - Minimize the overhead
  - More safety for "non-sexy" parts of MPI
  - Performance tuning

Fault Tolerance

- Fault Tolerance
  - Network agnostic
  - BFO (network fail over) [Oracle]
  - Coordinated [IU]
  - Process migration / automatic recovery, debugging [IU]
  - Uncoordinated + Message Logging [UTK]
  - Similar with FT-MPI approach
    - Or try to stay in sync with the MPI Forum

Cisco Updates

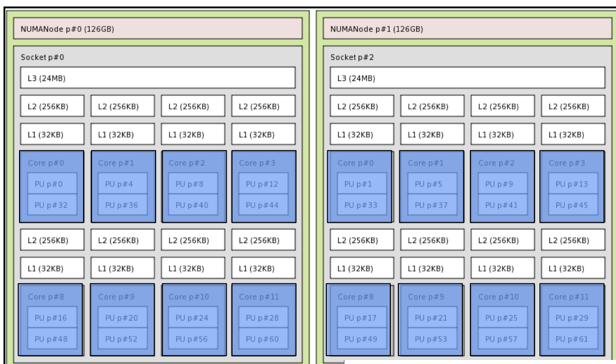
Jeff Squyres

Cisco Open MPI Work

- Community Open MPI development
  - Mainline development
  - Prototype / test various ethernet interfaces
- Research into next-generation core routers
  - Highly fault tolerant embedded systems
  - Mainly using underpinnings of OMPI (ORTE)
  - Also developing the Open Resilient Cluster Manager (ORCM)

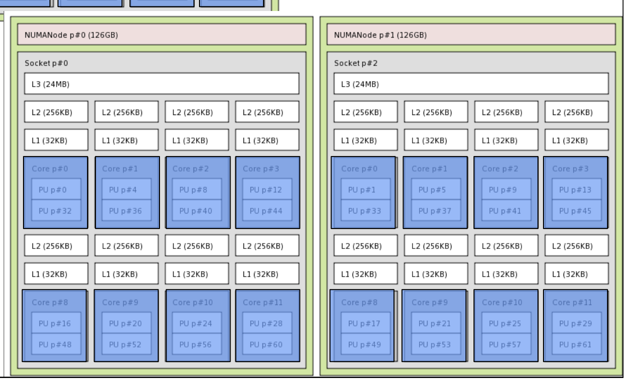
Processor Affinity

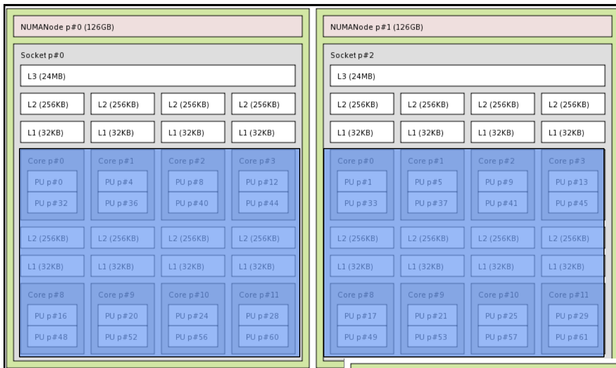
- Core counts are rising
  - "mpi_paffinity_alone" is good, but not enough
- Users are asking for powerful, flexible affinity controls
  - Bind processes to an entire sockets
  - Bind processes to half the cores in a socket
  - Bind processes to a NUMA locality
  - …etc.
- Joint work with Oracle, ORNL

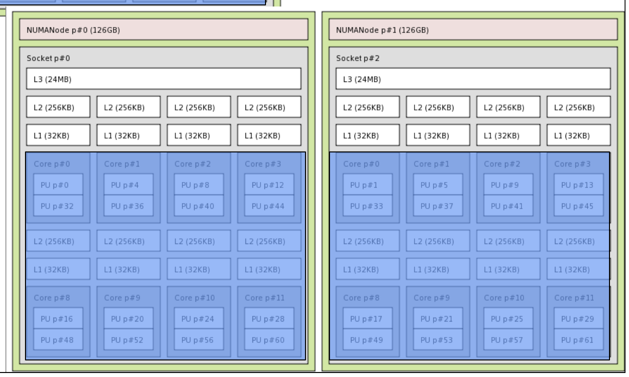


4 sockets, each with 8 cores, each with 2 PUs

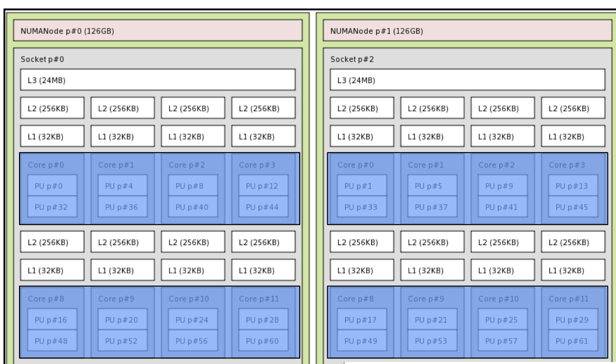
Each core has its own L1 and L2, shared L3 across socket

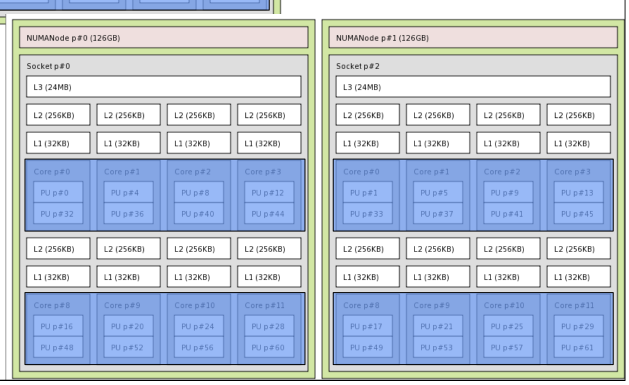






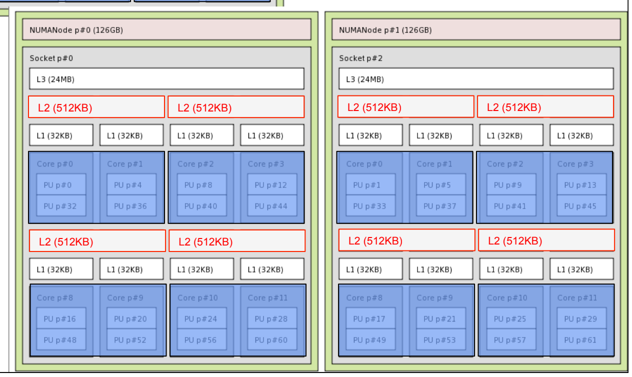
Bind processes to the whole socket



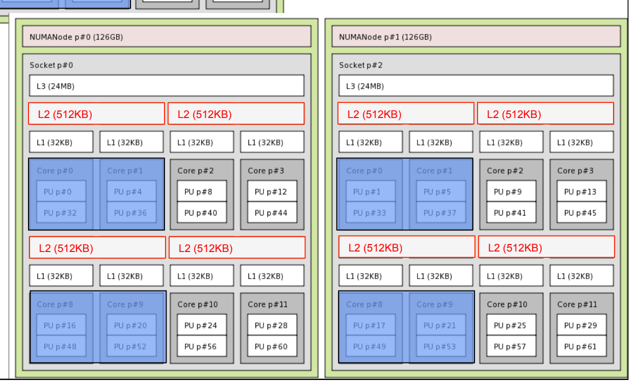


Bind processes to half the socket

Bind processes to an L2 locality



Bind processes to an L2 locality, leaving holes



How to order the resulting processes?

How to order the resulting processes?

Flexibility

- Need to represent:
  - Hardware threads, cores, L2 / L3 caches, sockets, boards, nodes
- Need to handle heterogeneous situations
  - E.g., non-uniform socket core count
- Would be nice to handle “offline” units

How To Express?

- Incredibly challenging to represent this on a command line
  - Need to be simple for 95% of users
  - Need to be powerful / flexible for power users
- May introduce “rankfile2” syntax
  - More flexible than current “rankfile” syntax
  - Allow completely arbitrary binding and ordering
- Design discussions are ongoing

INRIA @ Bordeaux

Brice Goglin

KNEM
Kernel-assisted direct-copy for intra-node comms

- Initially developed for MPICH2 (KNEM = Kernel Nemesis)
- Supported in BTL SM since OMPI v1.5
  - Configure with --with-knem=/path/to/knem/install
- Less cache pollution, less memory bandwidth, less CPU usage
  - More initialization work
- Enabled by default after 4kB
  - Direct copy only useful for large messages
    - Tens of kB or more

KNEM details

- Dedicated Linux kernel module
  - Working for any kernel since 2.6.15
- RMA interface
  - Sender creates a memory region, gets a cookie
  - Cookie is passed to another process (using PML)
  - Receiver pulls data from the send region
- Vectorial buffers, asynchronous data transfers, …
- DMA engine support disabled by default
  - Bad on current platforms

KNEM Future

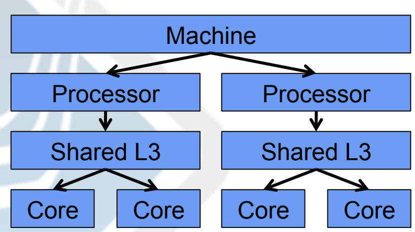
- Use KNEM directly inside collective
  - Use same memory region multiple times, read or write
    - Instead of only using KNEM for point-to-point within collectives
- SSE optimization (not that easy in the kernel)
- Thinking about getting some official support in Linux
  - Christopher Yeoh (IBM) trying to push some basic support
    - Not region based, not vectorial, …
  - Full vectorial and region-based support needs more discussion

Hardware Locality (hwloc)

- Replaces PLPA
- Working towards including in OMPI 1.5.x
  - New paffinity component
- More knowledge of the topology
  - HMT/SMT, shared caches, NUMA, …
- Portable
  - Solaris, AIX, OSF, HP-UX, FreeBSD, Darwin, Windows
    - Topology discovery and binding abilities may vary

Hwloc Details

- Tree of objects
  - Machine, (groups of) NUMA nodes, sockets, caches, cores, threads, ...
  - Ordered logically
- With many attributes
  - Memory size
  - Cache linesize
  - Physical/OS indexes
- XML import/export
  - Stop rereading /proc and /sys over and over



Hwloc Future

- I/O device discovery
  - hwloc 1.1 already knows the affinity of I/O handles
    - Cuda devices, IB devices, MX endpoints, …
  - Working on adding these objects to the main tree
  - May be used for tuning Open MPI components
- NUMA distances

KNEM + hwloc = Your Friends

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

http://runtime.bordeaux.inria.fr/knem

(Google for either of these and you'll find them)

Brice.Goglin@inria.fr

INRIA Booth #2751

The Road to MPI-3

George Bosilca

ICL UT

MPI Forum = Needs Feedback

- MPI Forum BOF yesterday
  - See the slides posted on http://meetings.mpi-forum.org/ (soon)
- PLEASE send your feedback
  - Many of the Forum are implementors
  - Need real world user feedback
- Next face-to-face meeting:
  - Cisco, San Jose, CA, USA, Dec. 6-8, 2010
  - Come join the Forum discussions

MPI-3 Prototyping Work

- MPI-3 has a “freely available implementation” requirement
  - Much work being prototyped in Open MPI
  - Will help speed our final implementation
- Examples
  - Fault tolerance work (Josh Hursey, ORNL)
    - “Crossing the Valley of Death”
  - New Fortran MPI bindings (Craig Rasmussen, LANL)

Come Join Us!

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