MVAPICH Saves the World!

MVAPICH User's Group Meeting

August 26, 2013

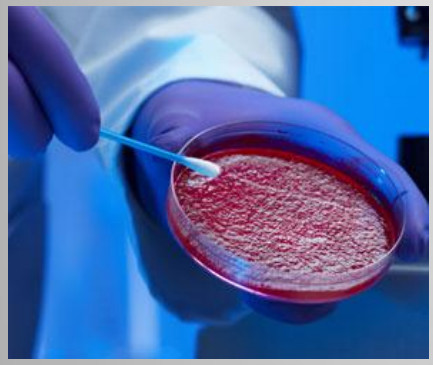
Presented by Adam Moody

Lawrence Livermore National Laboratory

LLNL-PRES-642835

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

LLNL's mission is applying world-class science, technology, and engineering to national & global problems

Bio-Security	Counterterrorism	Defense	Energy
			
Intelligence	**Nonproliferation**	**Science**	**Weapons**
			

LLNL systems by purpose

Capability
Capacity
Visualization
Serial

System	Top500 Rank	Program	Manufacture/ Model	OS	Inter-connect	Nodes	Cores	Memory (GB)	Peak TFLOP/s
Unclassified Network (OCF)									
Vulcan	8	ASC+M&IC+HPCIC	IBM BGQ	RHEL/CNK	5D Torus	24,576	393,216	393,216	5,033.2
Sierra	207	M&IC	Dell	TOSS	IB QDR	1,944	23,328	46,656	243.7
Cab (TLCC2)	79	ASC+M&IC+HPCIC	Appro	TOSS	IB QDR	1,296	20,736	41,472	426.0
Ansel		M&IC	Dell	TOSS	IB QDR	324	3,888	7,776	43.5
RZMerl (TLCC2)		ASC+ICF	Appro	TOSS	IB QDR	162	2,592	5,184	53.9
RZZeus		M&IC	Appro	TOSS	IB DDR	267	2,144	6,408	20.6
Edge	469	M&IC	Appro	TOSS	IB QDR	216	2,592	20,736	239.9
Aztec		M&IC	Dell	TOSS	N/A	96	1,152	4,608	12.9
Herd		M&IC	Appro	TOSS	IB DDR	9	256	1,088	1.6
OCF Totals	Systems	9							6,075.3
Classified Network (SCF)									
Pinot(TLCC2, SNSI)		M&IC	Appro	TOSS	IB QDR	162	2,592	5,184	53.9
Sequoia	3	ASC	IBM BGQ	RHEL/CNK	5D Torus	98,304	1,572,864	1,572,864	20132.7
Dawn	66	ASC	IBM BGP	SLES/CNK	3D Torus	36,864	147,456	147,456	501.4
Zin (TLCC2)	34	ASC	Appro	TOSS	IB QDR	2,916	46,656	93,312	961.1
Juno (TLCC)	295	ASC	Appro	TOSS	IB DDR	1,152	18,432	36,864	162.2
Muir	427	ICF	Dell	TOSS	IB QDR	1,296	15,552	31,104	168.0
Coastal		ICF	Dell	TOSS	IB DDR	1,152	9,216	27,648	88.5
Graph		ASC	Appro	TOSS	IB DDR	576	13,824	72,960	107.5
Inca		ASC	Dell	TOSS	N/A	100	1,216	5,120	13.5
SCF Totals	Systems	9							22,188.8
Combined Totals		18							28,264.1

System	Top500 Rank	Program	Manufacture/ Model	OS	Inter-connect	Nodes	Cores	Memory (GB)	Peak TFLOP/s
Unclassified Network (OCF)									
Vulcan	8	ASC+M&IC+HPCIC	IBM BGQ	RHEL/CNK	5D Torus	24,576	393,216	393,216	5,033.2
Sierra	207	M&IC	Dell	TOSS	IB QDR	1,944	23,328	46,656	243.7
Cab (TLCC2)	79	ASC+M&IC+HPCIC	Appro	TOSS	IB QDR	1,296	20,736	41,472	426.0
Ansel		M&IC	Dell	TOSS	IB QDR	324	3,888	7,776	43.5
RZMerl (TLCC2)		ASC+ICF	Appro	TOSS	IB QDR	162	2,592	5,184	53.9
RZZeus		M&IC	Appro	TOSS	IB DDR	267	2,144	6,408	20.6
Edge	469	M&IC	Appro	TOSS	IB QDR	216	2,592	20,736	239.9
Aztec		M&IC	Dell	TOSS	N/A	96	1,152	4,608	12.9
Herd		M&IC	Appro	TOSS	IB DDR	9	256	1,088	1.6
OCF Totals	Systems	9							6,075.3
Classified Network (SCF)									
Pinot(TLCC2, SNSI)		M&IC	Appro	TOSS	IB QDR	162	2,592	5,184	53.9
Sequoia	3	ASC	IBM BGQ	RHEL/CNK	5D Torus	98,304	1,572,864	1,572,864	20132.7
Dawn	66	ASC	IBM BGP	SLES/CNK	3D Torus	36,864	147,456	147,456	501.4
Zin (TLCC2)	34	ASC	Appro	TOSS	IB QDR	2,916	46,656	93,312	961.1
Juno (TLCC)	295	ASC	Appro	TOSS	IB DDR	1,152	18,432	36,864	162.2
Muir	427	ICF	Dell	TOSS	IB QDR	1,296	15,552	31,104	168.0
Coastal		ICF	Dell	TOSS	IB DDR	1,152	9,216	27,648	88.5
Graph		ASC	Appro	TOSS	IB DDR	576	13,824	72,960	107.5
Inca		ASC	Dell	TOSS	N/A	100	1,216	5,120	13.5
SCF Totals	Systems	9							22,188.8
Combined Totals		18							28,264.1

Why MVAPICH?

- First MPI available for IB
- Reliable and proven
- Still the fastest for many users
- Familiarity with MPICH code base
- Acceptance of feedback and patches
- Good ties and communication with OSU

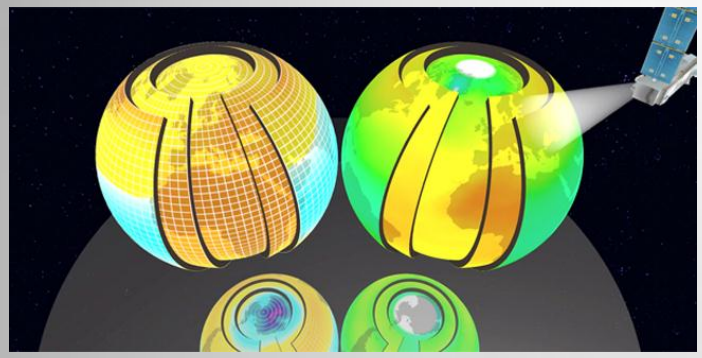
Livermore Computing / OSU: Successful history of collaboration

- Matt Koop
- Hari Subramoni
- Krishna Kandalla
- Raghunath Rajachandrasekar
- Hyperion
- “Collaborative Zone” systems

The National Ignition Facility: World's Highest Energy Laser

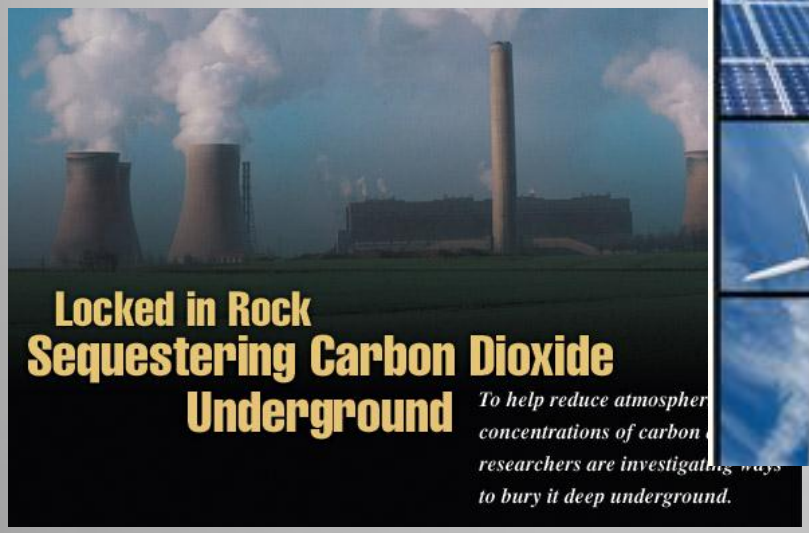
The Hayward Fault: Impact of 7.0 Earthquake

And lots more…



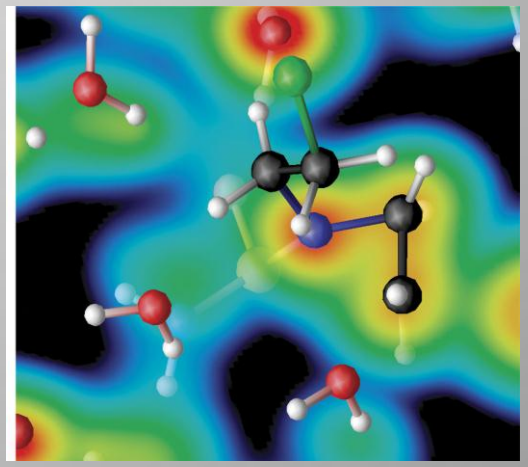
Climate Change

BP Oil Spill



Carbon Sequestration

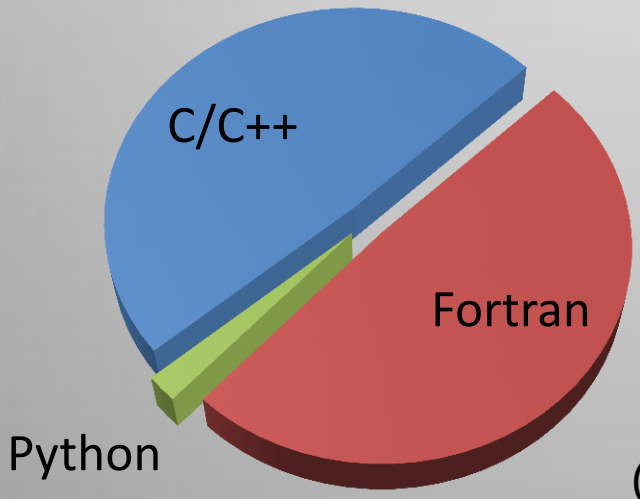
Renewable Energy



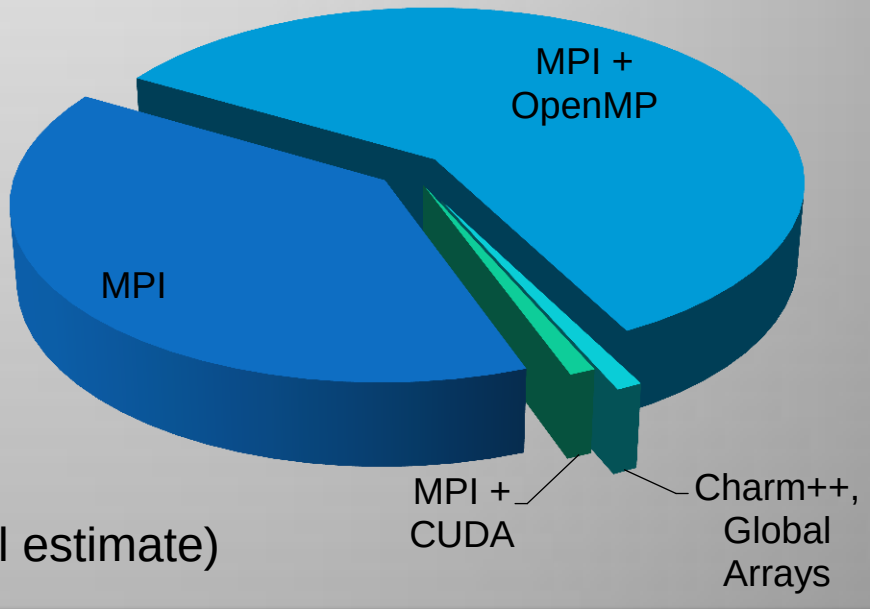
Anti-cancer drugs

Background on LLNL apps

- Many developed at lab
- Users often != the developers
- Users are scientists, but perhaps not computer scientists

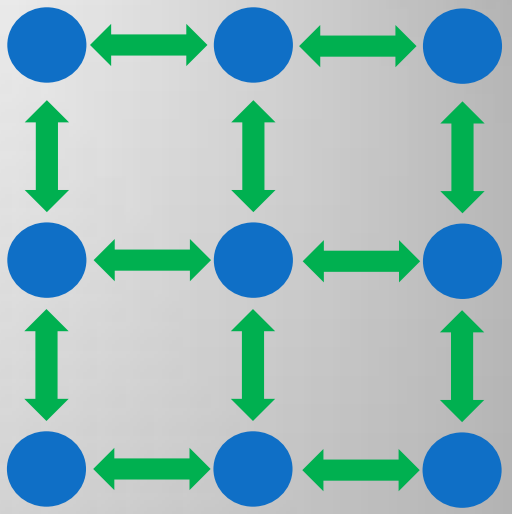


(personal estimate)



LLNL apps mostly use MPI-1.2

- Point-to-point
  - halo exchanges
- Collectives
  - Allreduce(1 double/int) a lot
  - Reduce
  - Barrier
  - Bcast
  - Allgather/v
  - Alltoall/v
- Communicators
  - Create once up front (divide problem space)
  - Create and free throughout run (load balancing)
- MPI-IO (a little bit)
  - Typically via I/O libs like HDF, NetCDF



Excited for shared memory and non-blocking collectives

- Shared memory (RMA)
  - Share large, read-only data, like equation-of-state tables
  - Communicate w/o copies
- Non-blocking collectives
  - Overlap comm / comp
  - Multiple outstanding colls
  - Two-phase barriers
- MPI_THREAD_MULTIPLE
  - As more apps use threads, more requests for (fast) thread-multiple support

```
while (residual > tolerance)
    Compute()
    MPI_Allreduce(residual)
```

```
while (residual > tolerance)
    MPI_Iallreduce(residual)
    Compute()
    MPI_Wait
```

```
MPI_Iallreduce(&req[0])
MPI_Iallreduce(&req[1])
MPI_Ialltoall(&req[2])
MPI_Waitall(3, req)
```

MPI_T and MPI_Comm_set_info

- API for “environment variables”
- Determine what MPI internals app uses
- Tune MPI per library communicator

```
**************************************************
get_num=23
(UI) posted_recvq_length: 0
(UI) unexpected_recvq_length: 0
<ERROR> UNKNOWN DATATYPE: posted_recvq_match_attempts
<ERROR> UNKNOWN DATATYPE: unexpected_recvq_match_atte
(D) time_failed_matching_postedq: 0.000000
(D) time_matching_unexpectedq: 0.000000
<ERROR> UNKNOWN DATATYPE: unexpected_recvq_buffer_siz
(ULL) mem_allocated_current: 820089
(ULL) mem_allocated_max: 820089
(ULL) mem_allocated_min: 820089
(ULL) coll_bcast_smp: 4
(ULL) coll_bcast_binomial: 4
(ULL) coll_bcast_scatter_doubling_allgather: 0
(ULL) coll_bcast_scatter_ring_allgather: 0
(ULL) coll_bcast_mv2_binomial: 0
(ULL) coll_bcast_mv2_scatter_doubling_allgather: 0
(ULL) coll_bcast_mv2_scatter_ring_allgather: 0
(ULL) coll_bcast_mv2_scatter_ring_allgather_shm: 0
(ULL) coll_bcast_mv2_shmem: 38510
(ULL) coll_bcast_mv2_knomial_internode: 29455
(ULL) coll_bcast_mv2_knomial_intranode: 0
(ULL) coll_bcast_mv2_mcast_internode: 0
(ULL) coll_bcast_mv2_pipelined: 2514
**************************************************
```

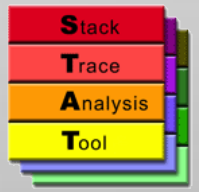
```
================================
Control Variables
================================

Found 24 control variables
Found 24 control variables with verbosity <= D/A-9

Variable                              VRB    Type    Bind     Scope     Value
-------------------------------------------------------------------------------
ALLTOALL_SHORT_MSG_SIZE               U/B-1  INT     n/a      LOCAL     256
ALLTOALL_MEDIUM_MSG_SIZE              U/B-1  INT     n/a      LOCAL     32768
ALLTOALL_THROTTLE                     U/B-1  INT     n/a      LOCAL     4
REDSCAT_COMMUTATIVE_LONG_MSG_SIZE     U/B-1  INT     n/a      LOCAL     524288
BCAST_MIN_PROCS                       U/B-1  INT     n/a      LOCAL     8
BCAST_SHORT_MSG_SIZE                  U/B-1  INT     n/a      LOCAL     12288
BCAST_LONG_MSG_SIZE                   U/B-1  INT     n/a      LOCAL     524288
ALLGATHER_SHORT_MSG_SIZE              U/B-1  INT     n/a      LOCAL     81920
ALLGATHER_LONG_MSG_SIZE               U/B-1  INT     n/a      LOCAL     524288
REDUCE_SHORT_MSG_SIZE                 U/B-1  INT     n/a      LOCAL     2048
ALLREDUCE_SHORT_MSG_SIZE              U/B-1  INT     n/a      LOCAL     2048
GATHER_VSMALL_MSG_SIZE                U/B-1  INT     n/a      LOCAL     1024
GATHER_INTER_SHORT_MSG_SIZE           U/B-1  INT     n/a      LOCAL     2048
GATHERV_INTER_SSEND_MIN_PROCS         U/B-1  INT     n/a      LOCAL     1024
SCATTER_INTER_SHORT_MSG_SIZE          U/B-1  INT     n/a      LOCAL     2048
ALLGATHERV_PIPELINE_MSG_SIZE          U/B-1  INT     n/a      LOCAL     32768
COMM_SPLIT_USE_QSORT                  U/B-1  INT     n/a      LOCAL     1
RMA_ACC_IMMED                         U/B-1  INT     n/a      LOCAL     1
RMA_NREQUEST_THRESHOLD                U/B-1  INT     n/a      LOCAL     4000
RMA_NREQUEST_NEW_THRESHOLD            U/B-1  INT     n/a      LOCAL     128
RMA_LOCK_IMMED                        U/B-1  INT     n/a      LOCAL     0
RMA_MERGE_LOCK_OP_UNLOCK              U/B-1  INT     n/a      LOCAL     1
```

What's it like supporting MPI at LLNL?

- Consulting
  - How do I accomplish X in MPI?
  - Why doesn't my code scale?
- Debugging
  - Identifying and fixing application bugs, sometimes MPI
  - Faulty hardware: NICs, links, switches, CPUs
  - Hangs, hangs, and more hangs
- Research & Development
  - Hangs → STAT
  - Node failures → SCR
  - Slow job launch → SPINDLE



Time consuming to install releases

- Procedure
  - Identify which patches accepted, port those that aren't
  - Study new features, identify build options
  - Fight build steps
  - Install MPI somewhere for testing
  - Simple tests: hello world, mpiBench, OSU benchmarks
  - Compare performance and memory usage to last version
  - Run memory debugging tools
  - Ask friendly users to test
  - Ask experts of other supported software to test
  - Notify all users of new MPI
  - Replace default MPI
- Hangs ups
  - Bad if MPI is slower or uses more memory, esp. if new features don't help
  - Complicated if need recompile
  - Schedule one-on-one time with VIP users

MPI reliability is critical, performance secondary

- New features and optimizations introduce bugs
- Different apps hit different bugs, but if a few users hit different bugs, can build impression that MPI is flaky
- Most apps not bottlenecked by MPI anyway, so LLNL sacrifices performance in trade for fewer bugs
- Build MPI with -O2 instead of –O3 from past compiler bugs
- MPI runtime error checks help
  - User finds solution to his own problem, faster fix
  - Relieves support load from computer center
  - What's broken with your MPI? → What did I do wrong?
- Disabled:
  - registration cache, header caching, XRC, multicast

Verbose error messages and environment variables save time

- User hits error, passes log file to LC Hotline
- Helpful, but need more than just MPI rank:
  - Hostname, timestamp, return code, errno & strerror, file & line number
  - Info of remote side if communication
- Use environment variables to quickly isolate problems
  - Turn off whole portions, e.g., shared memory
  - Then drill down, e.g, shm pt2pt vs. shm collectives
  - Further, shm bcast 1 vs shm bcast 2
  - Much easier than rebuilding MPI to disable things
- Variables to change resource limits like pool sizes
  - Some bugs fatal only after a long time, e.g., slow leak of request objects
  - Reduce pool size to hit error faster (for debugging)

System	Top500 Rank	Program	Manufacture/ Model	OS	Inter-connect	Nodes	Cores	Memory (GB)	Peak TFLOP/s
Unclassified Network (OCF)									
Vulcan	8	ASC+M&IC+HPCIC	IBM BGQ	RHEL/CNK	5D Torus	24,576	393,216	393,216	5,033.2
Sierra	207	M&IC	Dell	TOSS	IB QDR	1,944	23,328	46,656	243.7
Cab (TLCC2)	79	ASC+M&IC+HPCIC	Appro	TOSS	IB QDR	1,296	20,736	41,472	426.0
Ansel		M&IC	Dell	TOSS	IB QDR	324	3,888	7,776	43.5
RZMerl (TLCC2)		ASC+ICF	Appro	TOSS	IB QDR	162	2,592	5,184	53.9
RZZeus		M&IC	Appro	TOSS	IB DDR	267	2,144	6,408	20.6
Edge	469	M&IC	Appro	TOSS	IB QDR	216	2,592	20,736	239.9
Aztec		M&IC	Dell	TOSS	N/A	96	1,152	4,608	12.9
Herd		M&IC	Appro	TOSS	IB DDR	9	256	1,088	1.6
OCF Totals	Systems	9							6,075.3
Classified Network (SCF)									
Pinot(TLCC2, SNSI)		M&IC	Appro	TOSS	IB QDR	162	2,592	5,184	53.9
Sequoia	3	ASC	IBM BGQ	RHEL/CNK	5D Torus	98,304	1,572,864	1,572,864	20132.7
Dawn	66	ASC	IBM BGP	SLES/CNK	3D Torus	36,864	147,456	147,456	501.4
Zin (TLCC2)	34	ASC	Appro	TOSS	IB QDR	2,916	46,656	93,312	961.1
Juno (TLCC)	295	ASC	Appro	TOSS	IB DDR	1,152	18,432	36,864	162.2
Muir	427	ICF	Dell	TOSS	IB QDR	1,296	15,552	31,104	168.0
Coastal		ICF	Dell	TOSS	IB DDR	1,152	9,216	27,648	88.5
Graph		ASC	Appro	TOSS	IB DDR	576	13,824	72,960	107.5
Inca		ASC	Dell	TOSS	N/A	100	1,216	5,120	13.5
SCF Totals	Systems	9							22,188.8
Combined Totals		18							28,264.1

Different Platform Types

Qlogic
Mellanox
Mellanox + NVIDIA
Shared Mem

System	Top500 Rank	Program	Manufacture/ Model	OS	Inter-connect	Nodes	Cores	Memory (GB)	Peak TFLOP/s
Unclassified Network (OCF)									
Vulcan	8	ASC+M&IC+HPCIC	IBM BGQ	RHEL/CNK	5D Torus	24,576	393,216	393,216	5,033.2
Sierra	207	M&IC	Dell	TOSS	IB QDR	1,944	23,328	46,656	243.7
Cab (TLCC2)	79	ASC+M&IC+HPCIC	Appro	TOSS	IB QDR	1,296	20,736	41,472	426.0
Ansel		M&IC	Dell	TOSS	IB QDR	324	3,888	7,776	43.5
RZMerl (TLCC2)		ASC+ICF	Appro	TOSS	IB QDR	162	2,592	5,184	53.9
RZZeus		M&IC	Appro	TOSS	IB DDR	267	2,144	6,408	20.6
Edge	469	M&IC	Appro	TOSS	IB QDR	216	2,592	20,736	239.9
Aztec		M&IC	Dell	TOSS	N/A	96	1,152	4,608	12.9
Herd		M&IC	Appro	TOSS	IB DDR	9	256	1,088	1.6
OCF Totals	Systems	9							6,075.3
Classified Network (SCF)									
Pinot(TLCC2, SNSI)		M&IC	Appro	TOSS	IB QDR	162	2,592	5,184	53.9
Sequoia	3	ASC	IBM BGQ	RHEL/CNK	5D Torus	98,304	1,572,864	1,572,864	20132.7
Dawn	66	ASC	IBM BGP	SLES/CNK	3D Torus	36,864	147,456	147,456	501.4
Zin (TLCC2)	34	ASC	Appro	TOSS	IB QDR	2,916	46,656	93,312	961.1
Juno (TLCC)	295	ASC	Appro	TOSS	IB DDR	1,152	18,432	36,864	162.2
Muir	427	ICF	Dell	TOSS	IB QDR	1,296	15,552	31,104	168.0
Coastal		ICF	Dell	TOSS	IB DDR	1,152	9,216	27,648	88.5
Graph		ASC	Appro	TOSS	IB DDR	576	13,824	72,960	107.5
Inca		ASC	Dell	TOSS	N/A	100	1,216	5,120	13.5
SCF Totals	Systems	9							22,188.8
Combined Totals		18							28,264.1

Building MPI: A Nightmare of Permutations

- Multiple compilers
  - GNU, Intel, PGI
  - several versions of each
- Multiple MPI implementations
  - MVAPICH, MVAPICH2, Open MPI
  - 2-3 versions of each
  - normal + debug
- Multiple system types

MPI	Open MPI	MVAPICH2
Compilers	3	3
x MPI Versions	3	3
x (Normal + Debug)	2	2
x Platforms	1	4
= Total	18	72 !!!

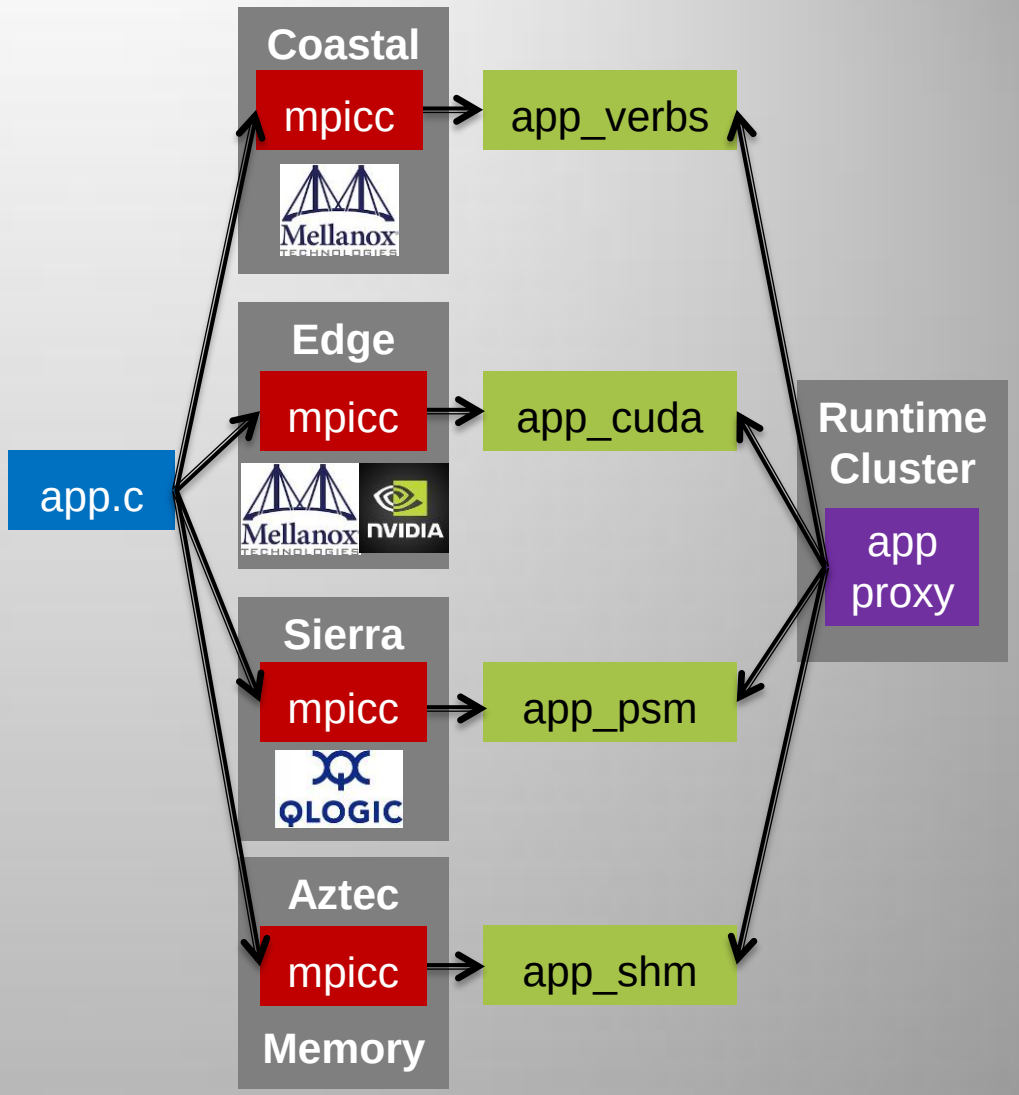
mpicc complicates build & run

- MPI wrappers link network libs

  >>: mpicc -show

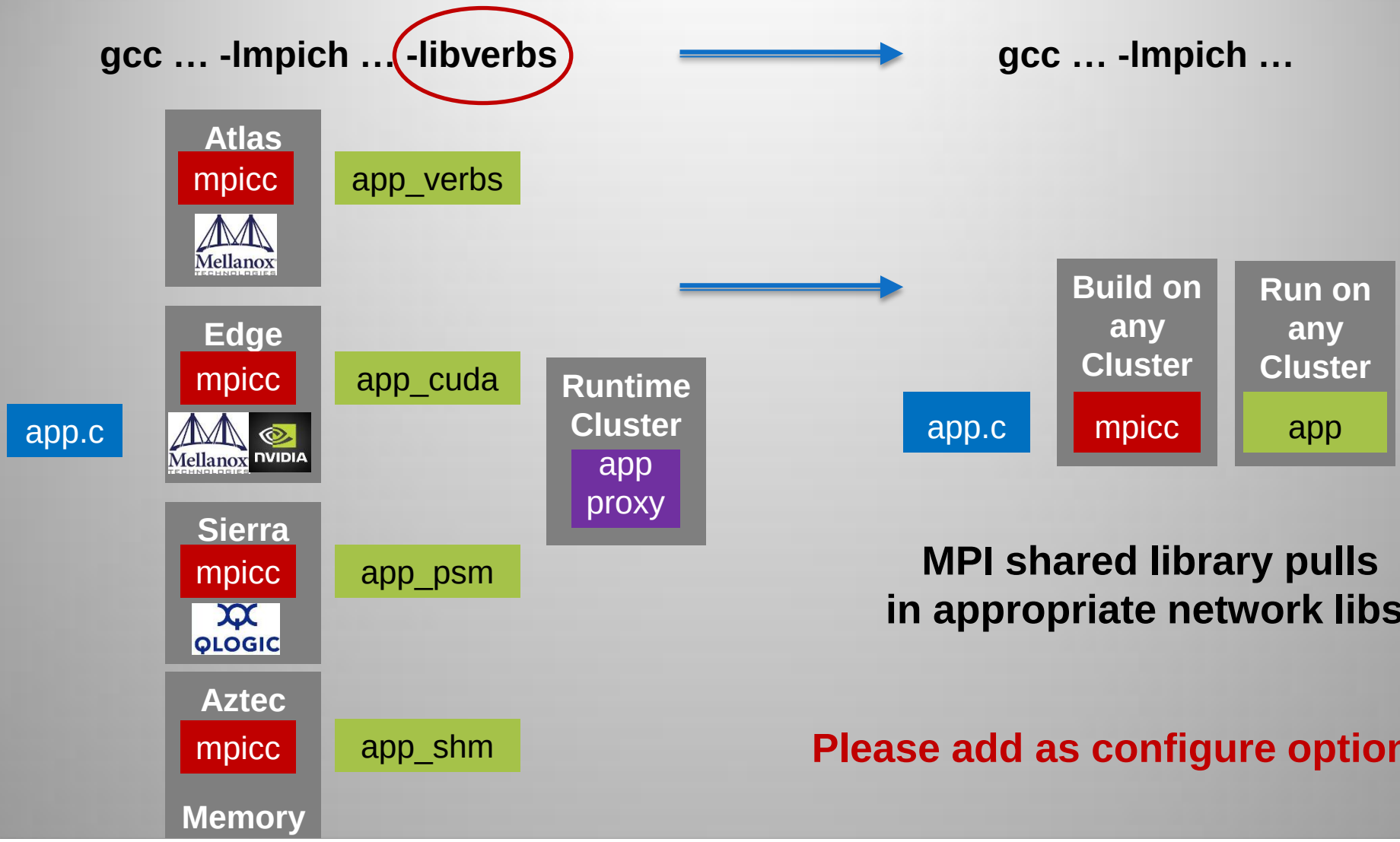
  gcc … -lmpich ... -libverbs

- LLNL has 4 network types
  - 4 different application binaries
  - 4 versions of every lib the app uses
  - Built on 4 different clusters
  - Documentation and/or scripting so user gets the right binary
- Cumbersome and error prone for app developers and their users



libmpich.so > 4*(libmpich + libverbs)

gcc … -lmpich … -libverbs → gcc … -lmpich …



MPI shared library pulls in appropriate network libs.

Please add as configure option!

MPI 3.0 Functions

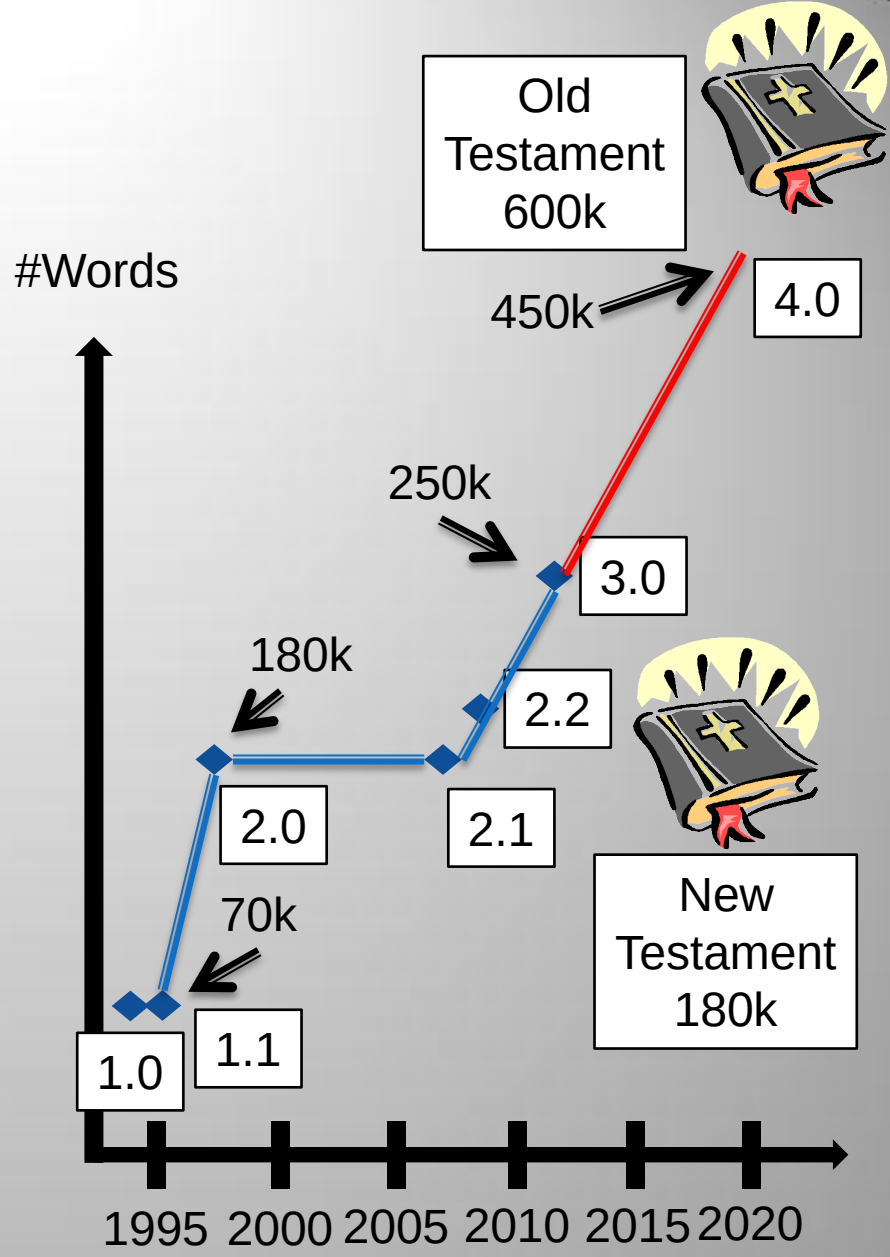
MPI_ABORT
MPI_ACCUMULATE
MPI_ADD_ERROR_CLASS
MPI_ADD_ERROR_CODE
MPI_ADD_ERROR_STRING
MPI_ADDRESS
MPI_ALLGATHER
MPI_ALLGATHERV
MPI_ALLOC_MEM
MPI_ALLOC_MEM_CPTR
MPI_ALLREDUCE
MPI_ALLTOALL
MPI_ALLTOALLV
MPI_ALLTOALLW
MPI_ATTR_DELETE
MPI_ATTR_GET
MPI_ATTR_PUT
MPI_BARRIER
MPI_BCAST
MPI_BSEND
MPI_BSEND_INIT
MPI_BUFFER_ATTACH
MPI_BUFFER_DETACH
MPI_CANCEL
MPI_CART_COORDS
MPI_CART_CREATE
MPI_CART_GET
MPI_CART_MAP
MPI_CART_RANK
MPI_CART_SHIFT
MPI_CART_SUB
MPI_CARTDIM_GET
MPI_CLOSE_PORT
MPI_COMM_ACCEPT
MPI_COMM_C2F
MPI_COMM_CALL_ERRHANDLER
MPI_COMM_COMPARE
MPI_COMM_CONNECT
MPI_COMM_CREATE
MPI_COMM_CREATE_ERRHANDLER
MPI_COMM_CREATE_GROUP
MPI_COMM_CREATE_KEYVAL
MPI_COMM_DELETE_ATTR
MPI_COMM_DISCONNECT
MPI_COMM_DUP
MPI_COMM_DUP_FN
MPI_COMM_DUP_WITH_INFO
MPI_COMM_F2C
MPI_COMM_FREE
MPI_COMM_FREE_KEYVAL
MPI_COMM_GET_ATTR
MPI_COMM_GET_ERRHANDLER
MPI_COMM_GET_INFO
MPI_COMM_GET_NAME
MPI_COMM_GET_PARENT
MPI_COMM_GROUP
MPI_COMM_IDUP
MPI_COMM_JOIN
MPI_COMM_KEYVAL_CREATE
MPI_COMM_NULL_COPY_FN
MPI_COMM_NULL_DELETE_FN
MPI_COMM_RANK
MPI_COMM_REMOTE_GROUP
MPI_COMM_REMOTE_SIZE
MPI_COMM_SET_ATTR
MPI_COMM_SET_ERRHANDLER
MPI_COMM_SET_INFO
MPI_COMM_SET_NAME
MPI_COMM_SIZE
MPI_COMM_SPAWN
MPI_COMM_SPAWN_MULTIPLE
MPI_COMM_SPLIT
MPI_COMM_SPLIT_TYPE
MPI_COMM_TEST_INTER
MPI_COMM_WORLD
MPI_COMPARE_AND_SWAP
MPI_CONVERSION_FN_NULL
MPI_DIMS_CREATE
MPI_DIST_GRAPH_CREATE
MPI_DIST_GRAPH_CREATE_ADJACENT
MPI_DIST_GRAPH_NEIGHBOR_COUNT
MPI_DIST_GRAPH_NEIGHBORS
MPI_DIST_GRAPH_NEIGHBORS_COUNT
MPI_DUP_FN
MPI_ERRHANDLER_C2F
MPI_ERRHANDLER_CREATE
MPI_ERRHANDLER_F2C
MPI_ERRHANDLER_FREE
MPI_ERRHANDLER_GET
MPI_ERRHANDLER_SET
MPI_ERROR_CLASS
MPI_ERROR_STRING
MPI_EXSCAN
MPI_F_SYNC_REG
MPI_FETCH_AND_OP
MPI_FILE_C2F
MPI_FILE_CALL_ERRHANDLER
MPI_FILE_CLOSE
MPI_FILE_CREATE_ERRHANDLER
MPI_FILE_DELETE
MPI_FILE_F2C
MPI_FILE_GET_AMODE
MPI_FILE_GET_ATOMICITY
MPI_FILE_GET_BYTE_OFFSET
MPI_FILE_GET_ERRHANDLER
MPI_FILE_GET_GROUP
MPI_FILE_GET_INFO
MPI_FILE_GET_POSITION
MPI_FILE_GET_POSITION_SHARED
MPI_FILE_GET_SIZE
MPI_FILE_GET_TYPE_EXTENT
MPI_FILE_GET_VIEW
MPI_FILE_IREAD
MPI_FILE_IREAD_AT
MPI_FILE_IREAD_SHARED
MPI_FILE_IWRITE
MPI_FILE_IWRITE_AT
MPI_FILE_IWRITE_SHARED
MPI_FILE_OPEN
MPI_FILE_PREALLOCATE
MPI_FILE_READ
MPI_FILE_READ_ALL
MPI_FILE_READ_ALL_BEGIN
MPI_FILE_READ_ALL_END
MPI_FILE_READ_AT
MPI_FILE_READ_AT_ALL
MPI_FILE_READ_AT_ALL_BEGIN
MPI_FILE_READ_AT_ALL_END
MPI_FILE_READ_ORDERED
MPI_FILE_READ_ORDERED_BEGIN
MPI_FILE_READ_ORDERED_END
MPI_FILE_READ_SHARED
MPI_FILE_SEEK
MPI_FILE_SEEK_SHARED
MPI_FILE_SET_ATOMICITY
MPI_FILE_SET_ERRHANDLER
MPI_FILE_SET_INFO
MPI_FILE_SET_SIZE
MPI_FILE_SET_VIEW
MPI_FILE_SYNC
MPI_FILE_WRITE
MPI_FILE_WRITE_ALL
MPI_FILE_WRITE_ALL_BEGIN
MPI_FILE_WRITE_ALL_END
MPI_FILE_WRITE_ORDERED
MPI_FILE_WRITE_AT
MPI_FILE_WRITE_AT_ALL
MPI_FILE_WRITE_AT_ALL_BEGIN
MPI_FILE_WRITE_AT_ALL_END
MPI_FILE_WRITE_ORDERED
MPI_FILE_WRITE_ORDERED_BEGIN
MPI_FILE_WRITE_ORDERED_END
MPI_FILE_WRITE_SHARED
MPI_FINALIZE
MPI_FINALIZED
MPI_FREE_MEM
MPI_GATHER
MPI_GATHERV
MPI_GET
MPI_GET_ACCUMULATE
MPI_GET_ADDRESS
MPI_GET_COUNT
MPI_GET_ELEMENTS
MPI_GET_ELEMENTS_X
MPI_GET_LIBRARY_VERSION
MPI_GET_PROCESSOR_NAME
MPI_GET_VERSION
MPI_GRAPH_CREATE
MPI_GRAPH_GET
MPI_GRAPH_MAP
MPI_GRAPH_NEIGHBORS
MPI_GRAPH_NEIGHBORS_COUNT
MPI_GRAPHDIMS_GET
MPI_GREQUEST_COMPLETE
MPI_GREQUEST_START
MPI_GROUP_C2F
MPI_GROUP_COMPARE
MPI_GROUP_DIFFERENCE
MPI_GROUP_EXCL
MPI_GROUP_F2C
MPI_GROUP_FREE
MPI_GROUP_INCL
MPI_GROUP_INTERSECTION
MPI_GROUP_RANGE_EXCL
MPI_GROUP_RANGE_INCL
MPI_GROUP_RANK
MPI_GROUP_SIZE
MPI_GROUP_TRANSLATE_RANKS
MPI_GROUP_UNION
MPI_IALLGATHER
MPI_IALLGATHERV
MPI_IALLREDUCE
MPI_IALLTOALL
MPI_IALLTOALLV
MPI_IALLTOALLW
MPI_IBARRIER
MPI_IBCAST
MPI_IBSEND
MPI_IEXSCAN
MPI_IGATHER
MPI_IGATHERV
MPI_IMPROBE
MPI_IMRECV
MPI_INEIGHBOR_ALLGATHER
MPI_INEIGHBOR_ALLGATHERV
MPI_INEIGHBOR_ALLTOALL
MPI_INEIGHBOR_ALLTOALLV
MPI_INEIGHBOR_ALLTOALLW
MPI_INFO_C2F
MPI_INFO_CREATE
MPI_INFO_DELETE
MPI_INFO_DUP
MPI_INFO_F2C
MPI_INFO_FREE
MPI_INFO_GET
MPI_INFO_GET_NKEYS
MPI_INFO_GET_NTHKEY
MPI_INFO_GET_VALUELEN
MPI_INFO_SET
MPI_INIT
MPI_INIT_THREAD
MPI_INITIALIZED
MPI_INTERCOMM_CREATE
MPI_INTERCOMM_MERGE
MPI_IPROBE
MPI_IRECV
MPI_IREDUCE
MPI_IREDUCE_SCATTER
MPI_IREDUCE_SCATTER_BLOCK
MPI_IRSEND
MPI_IS_THREAD_MAIN
MPI_ISCAN
MPI_ISCATTER
MPI_ISCATTERV
MPI_ISEND
MPI_ISSEND
MPI_KEYVAL_CREATE
MPI_KEYVAL_FREE
MPI_LOCK_ALL
MPI_LOOKUP_NAME
MPI_MESSAGE_C2F
MPI_MESSAGE_F2C
MPI_MPROBE
MPI_MRECV
MPI_NEIGHBOR_ALLGATHER
MPI_NEIGHBOR_ALLGATHERV
MPI_NEIGHBOR_ALLTOALL
MPI_NEIGHBOR_ALLTOALLV
MPI_NEIGHBOR_ALLTOALLW
MPI_NULL_COPY_FN
MPI_NULL_DELETE_FN
MPI_OP_C2F
MPI_OP_COMMUTATIVE
MPI_OP_CREATE
MPI_OP_F2C
MPI_OP_FREE
MPI_OPEN_PORT
MPI_PACK
MPI_PACK_EXTERNAL
MPI_PACK_EXTERNAL_SIZE
MPI_PACK_SIZE
MPI_PCONTROL
MPI_PROBE
MPI_PUBLISH_NAME
MPI_PUT
MPI_QUERY_THREAD
MPI_RACCUMULATE
MPI_RECV
MPI_RECV_INIT
MPI_REDUCE
MPI_REDUCE_LOCAL
MPI_REDUCE_SCATTER
MPI_REDUCE_SCATTER_BLOCK
MPI_REGISTER_DATAREP
MPI_REQUEST_C2F
MPI_REQUEST_F2C
MPI_REQUEST_FREE
MPI_REQUEST_GET_STATUS
MPI_RGET
MPI_RGET_ACCUMULATE
MPI_RPUT
MPI_RSEND
MPI_RSEND_INIT
MPI_SCAN
MPI_SCATTER
MPI_SCATTERV
MPI_SEND
MPI_SEND_INIT
MPI_SENDRECV
MPI_SENDRECV_REPLACE
MPI_SIZEOF
MPI_SSEND
MPI_SSEND_INIT
MPI_START
MPI_STARTALL
MPI_STATUS_C2F
MPI_STATUS_C2F08
MPI_STATUS_F082C
MPI_STATUS_F082F
MPI_STATUS_F2C
MPI_STATUS_F2F08
MPI_STATUS_SET_CANCELLED
MPI_STATUS_SET_ELEMENTS
MPI_STATUS_SET_ELEMENTS_X
MPI_T_CATEGORY_CHANGED
MPI_T_CATEGORY_GET_CATEGORIES
MPI_T_CATEGORY_GET_CVARS
MPI_T_CATEGORY_GET_INFO
MPI_T_CATEGORY_GET_NUM
MPI_T_CATEGORY_GET_PVARS
MPI_T_CVAR_GET_INFO
MPI_T_CVAR_GET_NUM
MPI_T_CVAR_HANDLE_ALLOC
MPI_T_CVAR_HANDLE_FREE
MPI_T_CVAR_READ
MPI_T_CVAR_WRITE
MPI_T_ENUM_GET_INFO
MPI_T_ENUM_GET_ITEM
MPI_T_FINALIZE
MPI_T_INIT_THREAD
MPI_T_PVAR_GET_INFO
MPI_T_PVAR_GET_NUM
MPI_T_PVAR_HANDLE_ALLOC
MPI_T_PVAR_HANDLE_FREE
MPI_T_PVAR_READ
MPI_T_PVAR_READRESET
MPI_T_PVAR_RESET
MPI_T_PVAR_SESSION_CREATE
MPI_T_PVAR_SESSION_FREE
MPI_T_PVAR_START
MPI_T_PVAR_STOP
MPI_T_PVAR_WRITE
MPI_TEST
MPI_TEST_CANCELLED
MPI_TESTALL
MPI_TESTANY
MPI_TESTSOME
MPI_TOPO_TEST
MPI_TYPE_C2F
MPI_TYPE_COMMIT
MPI_TYPE_CONTIGUOUS
MPI_TYPE_CREATE_DARRAY
MPI_TYPE_CREATE_F90_COMPLEX
MPI_TYPE_CREATE_F90_INTEGER
MPI_TYPE_CREATE_F90_REAL
MPI_TYPE_CREATE_HINDEXED
MPI_TYPE_CREATE_HINDEXED_BLOCK
MPI_TYPE_CREATE_HVECTOR
MPI_TYPE_CREATE_INDEXED_BLOCK
MPI_TYPE_CREATE_KEYVAL
MPI_TYPE_CREATE_RESIZED
MPI_TYPE_CREATE_STRUCT
MPI_TYPE_CREATE_SUBARRAY
MPI_TYPE_DELETE_ATTR
MPI_TYPE_DUP
MPI_TYPE_DUP_FN
MPI_TYPE_EXTENT
MPI_TYPE_F2C
MPI_TYPE_FREE
MPI_TYPE_FREE_KEYVAL
MPI_TYPE_GET_ATTR
MPI_TYPE_GET_CONTENTS
MPI_TYPE_GET_ENVELOPE
MPI_TYPE_GET_EXTENT
MPI_TYPE_GET_EXTENT_X
MPI_TYPE_GET_NAME
MPI_TYPE_GET_TRUE_EXTENT
MPI_TYPE_GET_TRUE_EXTENT_X
MPI_TYPE_HINDEXED
MPI_TYPE_HVECTOR
MPI_TYPE_INDEXED
MPI_TYPE_LB
MPI_TYPE_MATCH_SIZE
MPI_TYPE_NULL_COPY_FN
MPI_TYPE_NULL_DELETE_FN
MPI_TYPE_SET_ATTR
MPI_TYPE_SET_NAME
MPI_TYPE_SIZE
MPI_TYPE_SIZE_X
MPI_TYPE_STRUCT
MPI_TYPE_UB
MPI_TYPE_VECTOR
MPI_UNPACK
MPI_UNPACK_EXTERNAL
MPI_UNPUBLISH_NAME
MPI_WAIT
MPI_WAITALL
MPI_WAITANY
MPI_WAITSOME
MPI_WIN_ALLOC
MPI_WIN_ALLOCATE
MPI_WIN_ALLOCATE_CPTR
MPI_WIN_ALLOCATE_SHARED
MPI_WIN_ALLOCATE_SHARED_CPTR
MPI_WIN_ATTACH
MPI_WIN_C2F
MPI_WIN_CALL_ERRHANDLER
MPI_WIN_COMPLETE
MPI_WIN_CREATE
MPI_WIN_CREATE_DYNAMIC
MPI_WIN_CREATE_ERRHANDLER
MPI_WIN_CREATE_KEYVAL
MPI_WIN_DELETE_ATTR
MPI_WIN_DETACH
MPI_WIN_DUP_FN
MPI_WIN_F2C
MPI_WIN_FENCE
MPI_WIN_FLUSH
MPI_WIN_FLUSH_ALL
MPI_WIN_FLUSH_LOCAL
MPI_WIN_FLUSH_LOCAL_ALL
MPI_WIN_FREE
MPI_WIN_FREE_KEYVAL
MPI_WIN_GET_ATTR
MPI_WIN_GET_ERRHANDLER
MPI_WIN_GET_GROUP
MPI_WIN_GET_INFO
MPI_WIN_GET_NAME
MPI_WIN_LOCK
MPI_WIN_LOCK_ALL
MPI_WIN_NULL_COPY_FN
MPI_WIN_NULL_DELETE_FN
MPI_WIN_POST
MPI_WIN_SET_ATTR
MPI_WIN_SET_ERRHANDLER
MPI_WIN_SET_INFO
MPI_WIN_SET_NAME
MPI_WIN_SHARED_ALLOCATE
MPI_WIN_SHARED_QUERY
MPI_WIN_SHARED_QUERY_CPTR
MPI_WIN_START
MPI_WIN_SYNC
MPI_WIN_TEST
MPI_WIN_UNLOCK
MPI_WIN_UNLOCK_ALL
MPI_WIN_WAIT
MPI_WTICK
MPI_WTIME

History

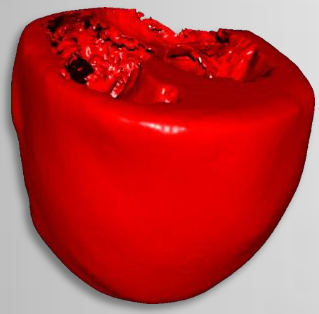
- MPI 1.0 May 1994 – 228 pages
- MPI 1.1 Nov 1995 – 238 pages (128 fns)
- MPI 2.0 Nov 1997 – 608 pages

10 year break

- MPI 2.1 June 2008 – 608 pages
- MPI 2.2 Sep 2009 – 647 pages
- MPI 3.0 Sep 2012 – 852 pages (430 fns)



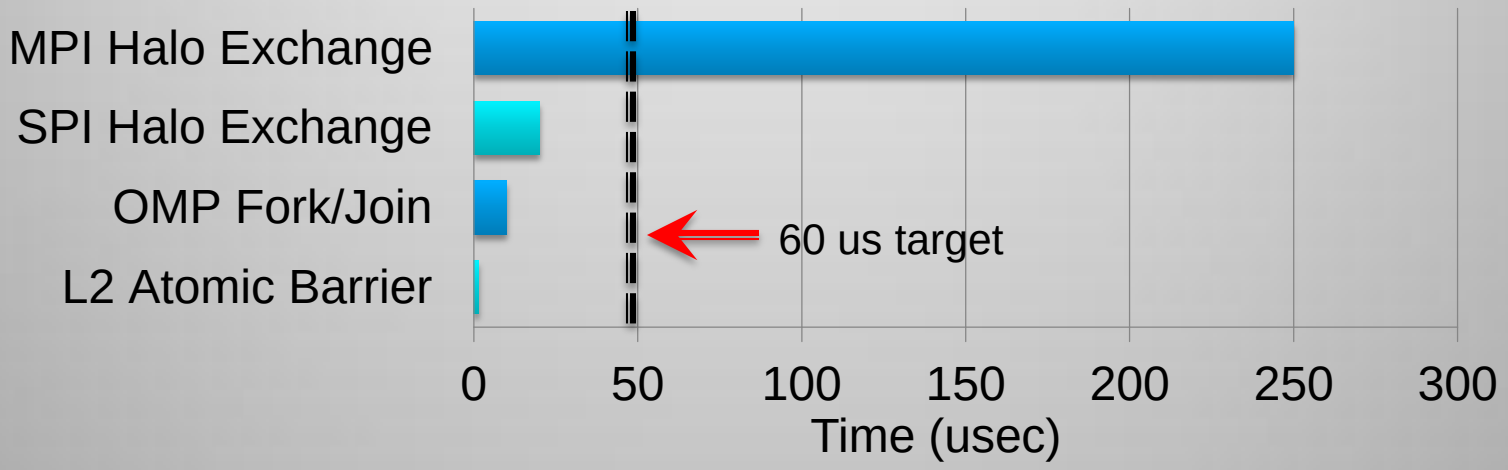
“Large” Overhead of MPI and OpenMP



370 Million Cells 1.6 Million Cores

1600 Flops/cell

60 us per iteration



Message units and L2 atomic operations provided by BG/Q hardware deliver performance even in this demanding regime

LLNL challenges posed to MVAPICH

- Reduce build/run permutations
- Efficient MPI_THREAD_MULTIPLE
- Reduce software overhead
  - instruction count, cache misses, etc
- Reduce memory footprint
- Think billions of processes
  - # procs ~ sizeof(memory)
  - Even MPI_Send is non-trivial!
  - What's MPI look like?