Cray MPI for KNL

Peter Mendygral
pjm@cray.com

Legal Disclaimer

Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.

Cray Inc. may make changes to specifications and product descriptions at any time, without notice.

All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Cray hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.

Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, and URIKA. The following are trademarks of Cray Inc.: APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, THREADSTORM, REVEAL. The following system family marks, and associated model number marks, are trademarks of Cray Inc.: CS, CX, XC, XE, XK, XMT, and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis. Other trademarks used in this document are the property of their respective owners.

XC Topology and Aries

- Not covering details of the XC topology and Aries interconnect today

- Please refer to the following document or feel free to ask questions at any time the rest of the workshop

http://www.cray.com/sites/default/files/resources/CrayXCNetwork.pdf

Agenda

- **Introduction to Cray MPICH**
- **KNL/MCDRAM optimizations**
- **Optimizations for Hybrid MPI+OpenMP applications**
  - Hybrid MPI+OpenMP application study

Introduction to Cray MPICH

Brief Introduction to Cray MPICH

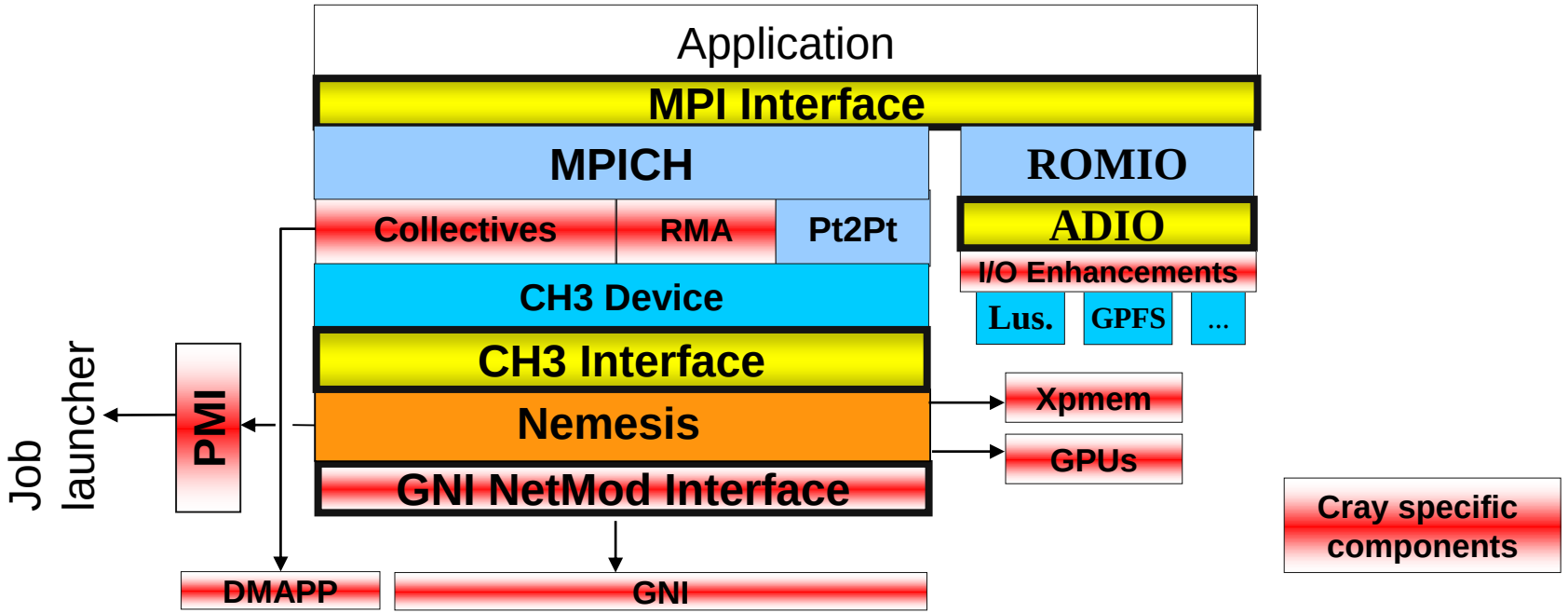
- **Cray MPI compliant with MPI 3.1**
  - Merge to ANL MPICH 3.2 – latest release MPT 7.7.0 (Dec 2017)
- **I/O, collectives, P2P, and one-sided all optimized for XC architecture**
  - SMP aware collectives
  - High performance single-copy on-node communication via xpmem (not necessary to program for shared memory)
- **Highly tunable through environment variables**
  - Defaults should generally be best, but some cases benefit from fine tuning
- **Integrated within the Cray Programming Environment**
  - Compiler drivers manage compile flags and linking automatically
  - Profiling through Cray Perftools

MPI Dynamic Process Management

- **All DPM functionality (except spawn) supported as of MPT 7.7.0**
- **Spawn support coming in Q2 2018**
- **New MPIX function to make MPI_Comm_spawn simpler**
  - MPIX_Comm_rankpool converts an MPI communicator into a pool of compute resources for spawning new MPI jobs
  - Generally called from a separate MPI job
  - Rank pool “name” parameter used to specify pool when spawning new MPI job, i.e., MPI_Info_set(info, “rankpool”, “name_of_pool”)
  - A “timeout” parameter determines the lifetime of the rank pool

Cray MPI Software Stack (CH3 device)

Gemini/Aries



Cray MPI Resources

- **Primary user resource for tuning and feature documentation is the manpage**
  - man intro_mpi

  OR
  - man MPI
- **Standard function documentation available as well**
  - E.g., man mpi_isend

Key Environment Variables for XC

- **MPICH_RANK_REORDER_METHOD**
  - Vary your rank placement to optimize communication
  - Can be a quick, low-hassle way to improve performance
  - Use Craypat to produce a specific **MPICH_RANK_ORDER** file to maximize intra-node communication
  - Or, use perf_tools **grid_order** command with your application's grid dimensions to layout MPI ranks in alignment with data grid
  - To use:
    - name your custom rank order file: MPICH_RANK_ORDER
    - export MPICH_RANK_REORDER_METHOD=3

Key Environment Variables for XC

- **MPICH_RANK_REORDER_METHOD (cont.)**
  - A topology and placement aware reordering method is also available
  - Optimizes rank ordering for Cartesian decompositions using the layout of nodes in the job
  - To use:
    - export MPICH_RANK_REORDER_METHOD=4
    - export MPICH_RANK_REORDER_OPTS="—ndims=3 –dims=16,16,8"

Key Environment Variables for XC

- **Use HUGEPAGES**
  - While this is not an MPI env variable, linking and running with hugepages can offer a significant performance improvement for many MPI communication sequences, including MPI collectives and basic MPI_Send/MPI_Recv calls
  - Most important for applications calling MPI_Alltoall[v] or performing point to point operations with a similarly well connected pattern and large data footprint
  - To use HUGEPAGES:
    - **`module load craype-hugepages8M`** (many sizes supported)
    - *<< compile your app >>*
    - **`module load craype-hugepages8M`**
    - *<< run your app >>*

Key Environment Variables for XC

- **MPICH_USE_DMAPP_COLL and hardware supported collectives**
  - Most of MPI's optimizations are enabled by default, but not the DMAPP-optimized features, because…
  - Using DMAPP may have some disadvantages
    - May reduce resources MPICH has available (share with DMAPP)
    - Requires more memory (DMAPP internals)
    - DMAPP does not handle transient network errors
  - These are highly-optimized algorithms which may result in significant performance gains, but user has to request them
  - Supported DMAPP-optimized functions:
    - MPI_Allreduce (4-8 bytes)
    - MPI_Bcast (4 or 8 bytes)
    - MPI_Barrier
  - To use (link with libdmapp):
    - Collective use: export MPICH_USE_DMAPP_COLL=1

Key Environment Variables for XC

- **MPICH GNI environment variables**
  - To optimize inter-node traffic using the Aries interconnect, the following are the most significant env variables to play with (*avoid significant deviations from the default if possible*):
  - **MPICH_GNI_MAX_VSHORT_MSG_SIZE**
    - Controls max message size for E0 mailbox path (Default: varies)
  - **MPICH_GNI_MAX_EAGER_MSG_SIZE**
    - Controls max message size for E1 Eager Path (Default: 8K bytes)
  - **MPICH_GNI_NUM_BUFS**
    - Controls number of 32KB internal buffers for E1 path (Default: 64)
  - **MPICH_GNI_NDREG_MAXSIZE**
    - Controls max message size for R0 Rendezvous Path (Default: 4MB)
  - **MPICH_GNI_RDMA_THRESHOLD**
    - Controls threshold for switching to BTE from FMA (Default: 1K bytes)
- **See the MPI man page for further details**

Key Environment Variables for XC

- **Specific Collective Algorithm Tuning**
  - Different algorithms may be used for different message sizes in collectives (e.g.)
    - Algorithm A might be used for Alltoall for messages < 1K.
    - Algorithm B might be used for messages >= 1K.
  - To optimize a collective, you can modify the cutoff points when different algorithms are used. This may improve performance. A few important ones are:
    - MPICH_ALLGATHER_VSHORT_MSG
    - MPICH_ALLGATHERV_VSHORT_MSG
    - MPICH_GATHERV_SHORT_MSG
    - MPICH_SCATTERV_SHORT_MSG
    - MPICH_GNI_A2A_BLK_SIZE
    - MPICH_GNI_A2A_BTE_THRESHOLD
- **For thread multiple support**
  - export MPICH_MAX_THREAD_SAFETY=multiple
- **See the MPI man page for further details**

KNL Optimizations

MCDRAM on KNL

- **KNL nodes in cache mode are a good starting point for many applications**
  - No extra work for user
  - Typically good performance
- **KNL nodes in flat mode is attractive when**
  - Can benefit from extra memory of DDR+MCDRAM
  - Observing performance variability related to MCDRAM direct map cache
- **Effective use of flat mode requires users understand more about what memory in their application matters and how it is being used**
  - All performance critical memory should reside in MCDRAM if possible

Cray MPI support for MCDRAM on KNL

- **Cray MPI offers allocation + hugepage support for MCDRAM on KNL**
  - Must use: MPI_Alloc_mem() or MPI_Win_Allocate()
  - Dependencies: memkind, NUMA libraries and dynamic linking.
    module load cray-memkind
- **Feature controlled with env variables**
  - Users select: Affinity, Policy and PageSize
  - MPICH_ALLOC_MEM_AFFINITY = DDR or MCDRAM
    - DDR = allocate memory on DDR (default)
    - MCDRAM = allocate memory on MDCRAM
  - MPICH_ALLOC_MEM_POLICY = M/ P/ I
    - M = Mandatory: fatal error if allocation fails
    - P = Preferred: fall back to using DDR memory (default)
    - I = Interleaved: Set memory affinity to interleave across MCDRAM NUMA nodes (For SNC* cases)
  - MPICH_ALLOC_MEM_PG_SZ
    - 4K, 2M, 4M, 8M, 16M, 32M, 64M, 128M, 256M, 512M (default 4K)

Cray MPI support for MCDRAM on KNL

- **MPI_Alloc_mem: not restricted to be used only for communication buffers, or MPI's internal buffers.**
  - Can also be used to allocate application's data buffers
- **Cray MPI does not register the memory returned by Alloc_mem**
- **Cray MPI also does not "touch" memory allocated via Alloc_mem()**
  - NUMA Affinity resolved when the memory pages are first touched by the process/threads.
  - Not ideal from a NUMA perspective to have the master thread alone touch the entire buffer right after allocation
- **MPI_Alloc_mem returns page-aligned memory for all page sizes**

Cray MPI support for MCDRAM on KNL Typical use cases

- **When the entire data set fits within MCDRAM, on a KNL nodes in flat mode:**

  aprun –Nx –ny numactl –membind=1 ./a.out

  - Easiest way to utilize hugepages on MCDRAM
  - craype-hugepage module is honored.
  - Allocations (malloc, memalign) on MCDRAM will be backed by hugepages
  - However, all memory allocated on MCDRAM (including MPI's internal memory)
  - Memory available per node limited to % of MCDRAM configured as FLAT memory

- **Alternate solutions needed to utilize hugepage memory on MCDRAM, when the data set per node exceeds 16G**
  - Necessary to identify performance critical buffers
  - Replace memory allocation calls with MPI_Alloc_mem() or MPI_Win_allocate()
  - Use Cray MPI env. vars to control page size, memory policy and memory affinity for allocations

Cray MPI support for MCDRAM on KNL: Typical use cases (Dataset size > 16GB)

- **Flat mode, without numactl options:**
  - malloc(), memalign() will use DDR first
  - Can access MCDRAM via hbw_* or compiler directives.
  - craype-hugepages module honored *only* on DDR
  - hbw_malloc will return memory backed by basepages
  - Memkind can be used to get 2M hugepages on MCDRAM (but not larger)

- **Users need to identify critical buffers and use MPI_Alloc_mem() to allocate hugepages with larger page sizes, and set affinity to MCDRAM**

- **Use following env. vars:**

  MPICH_ALLOC_MEM_AFFINITY=M (or MCDRAM)

  MPICH_ALLOC_MEM_PG_SZ = 16M (16M hugepages)

  MPICH_ALLOC_MEM_POLICY = P (or Preferred)

Cray MPI support for MCDRAM on KNL: Typical use cases (Dataset size > 16GB)

- **Flat mode, with numactl --membind=1**
  - Malloc(),memalign() will use MCDRAM
  - Hugepage allocations via the craype-hugepages module now possible on MCDRAM
  - But, MCDRAM space is limited. Memory scaling issues

- **Users can identify buffers *not* critical to application performance and use MPI_Alloc_mem() to set affinity to DDR**

- **Use following env. vars:**

  MPICH_ALLOC_MEM_AFFINITY=D (or DDR)

  MPICH_ALLOC_MEM_PG_SZ = <as needed, defaults to 4KB base pages>

  MPICH_ALLOC_MEM_POLICY = P (or Preferred)

  Add internal MPI mailbox and SHM collectives buffer affinity variable

Using MPI_Alloc_mem/MPI_Free_mem with a 3DFFT Kernel (Fortran)

```
..
lenr    = nx * ny * mynz
lenc    = MAX((nxd2p1 * ny * mynz),
                (nxd2p1 * nz * myny))
lenm    = MAX((nxd2p1 * ny * mynz),
                (nxd2p1 * nz * myny))
rc_grid = FFTW_ALLOC_REAL(lenr * nvars)
cc_grid = FFTW_ALLOC_COMPLEX(lenc * nvars)
mc_grid = FFTW_ALLOC_COMPLEX(lenm)

CALL C_F_POINTER(rc_grid, rgrid,
                        (/nx, ny, mynz, nvars/))
CALL C_F_POINTER(cc_grid, cgrid,
                        (/lenc, INT(nvars, C_SIZE_T)/))
CALL C_F_POINTER(mc_grid, mgrid, (/lenm/))

..
```

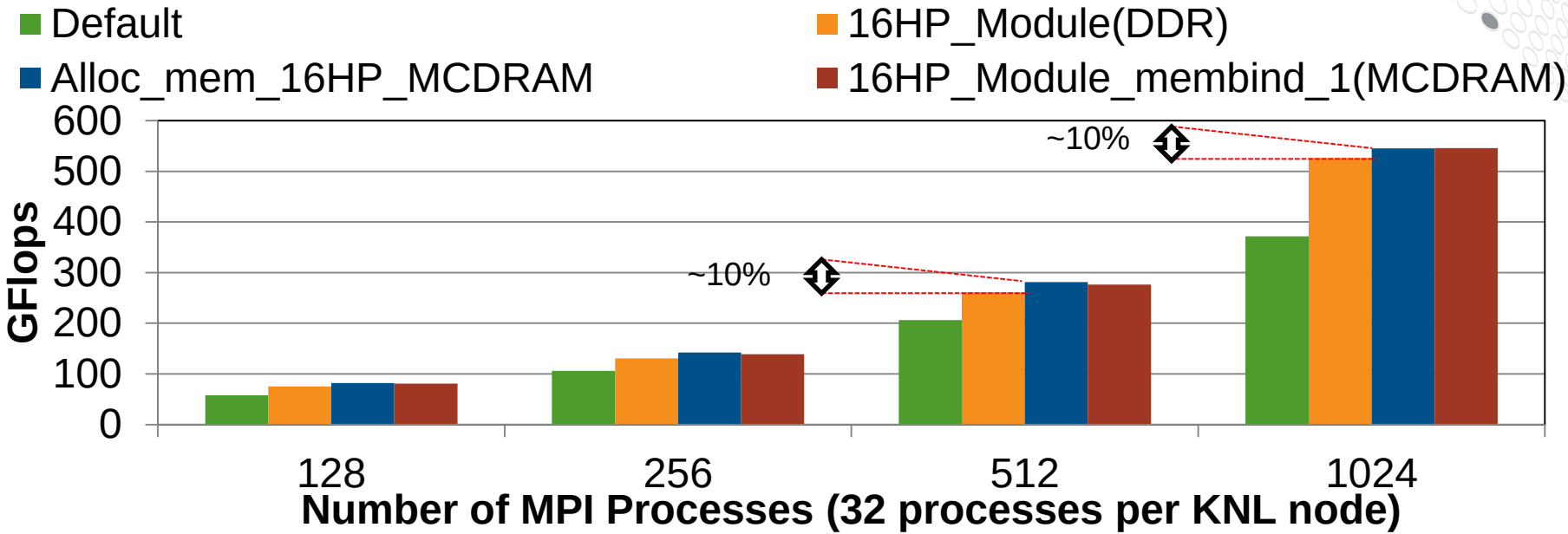
```
lenr    = nx * ny * mynz
lenc    = MAX((nxd2p1 * ny * mynz),
                (nxd2p1 * nz * myny))
lenm    = MAX((nxd2p1 * ny * mynz),
                (nxd2p1 * nz * myny))
info    = MPI_INFO_NULL
CALL MPI_ALLOC_MEM(lenr * nvars * 8_8, info,
                                rc_grid, ierr)
CALL MPI_ALLOC_MEM(lenc * nvars * 16_8, info,
                                cc_grid, ierr)
CALL MPI_ALLOC_MEM(lenm * 16_8, info,
                                mc_grid, ierr)

CALL C_F_POINTER(rc_grid, rgrid,
                        (/nx, ny, mynz, nvars/))
CALL C_F_POINTER(cc_grid, cgrid,
                        (/lenc, INT(nvars, C_SIZE_T)/))
CALL C_F_POINTER(mc_grid, mgrid, (/lenm/))
..
```

Using MPI_Alloc_mem/MPI_Free_mem with a 3DFFT Kernel (Fortran)

```
CALL FFTW_FREE(rc_grid)
CALL FFTW_FREE(cc_grid)
CALL FFTW_FREE(mc_grid)
```

```
CALL MPI_FREE_MEM(rgrid, ierr)
CALL MPI_FREE_MEM(cgrid, ierr)
CALL MPI_FREE_MEM(mgrid, ierr)
```

MCDRAM Experiments with a 3DFFT Kernel



3DFFT Weak scaling (Data Grid: 1024, 1024, 1024)

MPI_Alloc_mem with hugepages offers same performance as using membind=1

Hugepages on MCDRAM performs better than DDR with the same hugepage size

Using MPI_Alloc_mem can help cases where the entire data set does not fit within MCDRAM

Multi-threaded MPI Support and Optimizations

MPI Thread Multiple Support in Cray MPI

- **Thread multiple support for**
  - point to point operations (optimized global lock)
  - Collectives (optimized global lock)
  - MPI-RMA (thread hot)
- **All supported in default library**
- **export MPICH_MAX_THREAD_SAFETY=multiple**
- **Global lock optimization on by default (N/A for MPI-RMA)**
  - 50% better 8B latency than pthread_mutex() (OSU latency_mt, 32 threads per node, Broadwell)
  - export MPICH_OPT_THREAD_SYNC=0 falls back to pthread_mutex()

MPI-RMA Thread Hot Communication in Cray MPI

- "Thread hot" means high performance thread multiple support
- Design Objectives
  - Contention free progress and completion
  - High bandwidth and high message rate
  - Independent progress – thread(s) flush outstanding traffic, other threads make uninterrupted progress
  - Dynamic mapping between threads and network resources
  - Locks needed only if the number of threads exceed the number of network resources
- MPI-3 RMA
  - Epoch calls (Win_complete, Win_fence) are thread-safe, but not intended to be thread hot
  - All other RMA calls (including request-based operations) are thread hot
  - Multiple threads doing Passive Synchronization operations likely to perform best

Multi-threading Approaches with Cray MPI

- **Easy way to hit the ground running on a KNL – MPI only mode**
  - Works quite well in our experience
  - Scaling to more than 2-8 threads most likely requires a different application design approach
- **“Bottom-Up” OpenMP development approach is very common**
  - Likely will not offer best performance and thread scaling
- **“Top-Down” SPMD model is more appealing for KNL**
  - Increases the scope of code executed by OpenMP
  - Allows for better load balancing and overall compute scaling on KNL
  - Leads to multiple threads calling MPI concurrently
  - In this model, performance is limited by the level of support offered by MPI for multi-threaded communication
  - MPI implementations must offer “thread hot” communication capabilities to improve communication performance for highly threaded use cases on KNL

MPI+OpenMP Models

“bottom up”

```
! Keep OpenMP within a “compute” loop

DO WHILE (t .LT. tend)

   DO patch = 1, npatches

      CALL update_patch()

      …CALL MPI…

   END DO

END DO

SUBROUTINE update_patch()

   !$OMP PARALLEL DO
   DO i = 1, nx
   …do work…
   END DO

END SUBROUTINE
```

“SPMD”

```
! Move OpenMP near the top of the call stack

!#OMP PARALLEL
DO WHILE (t .LT. tend)

   !#OMP DO
   DO patch = 1, npatches

      CALL update_patch()

      …CALL MPI…

   END DO

END DO
```

High-level OpenMP

- **Benefits of high-level SPMD OpenMP**
  - Application more closely mimics completely independent processes
    - Less likely to be in the same portion of code at the same time
    - Bandwidth competition may decrease
    - Amdahl's law
  - Threads are less coupled => infrequent thread synchronization
  - Less likely to have issues with memory conflicts between threads
  - Potentially simpler to implement
    - Large reduction in the amount of OpenMP directives
    - Very little variable scoping needed as most everything is shared => reduced memory footprint
  - Easier to make use of all cores on node (e.g., 68) that can be hard to use for domain decomposition reasons
  - Effective way to manage hardware induced imbalance and algorithmic load imbalance

High-level OpenMP

- **Challenges for high-level SPMD OpenMP**
  - Requires full understanding of data dependencies and potential for race conditions
  - Best performance requires new approach to MPI
    - Goal should be to remove any thread synchronization you can
    - Serializing MPI will limit the benefit and scalability of SPMD
  - SPMD is a data centric model
    - Present work as independent units
    - Let threads work on that set in any order with a non-static schedule
  - MPI work should be treated the same as compute if possible
    - Independent units of communication to be worked on in any order with non-static schedule

Wombat Project Hybrid Application Study

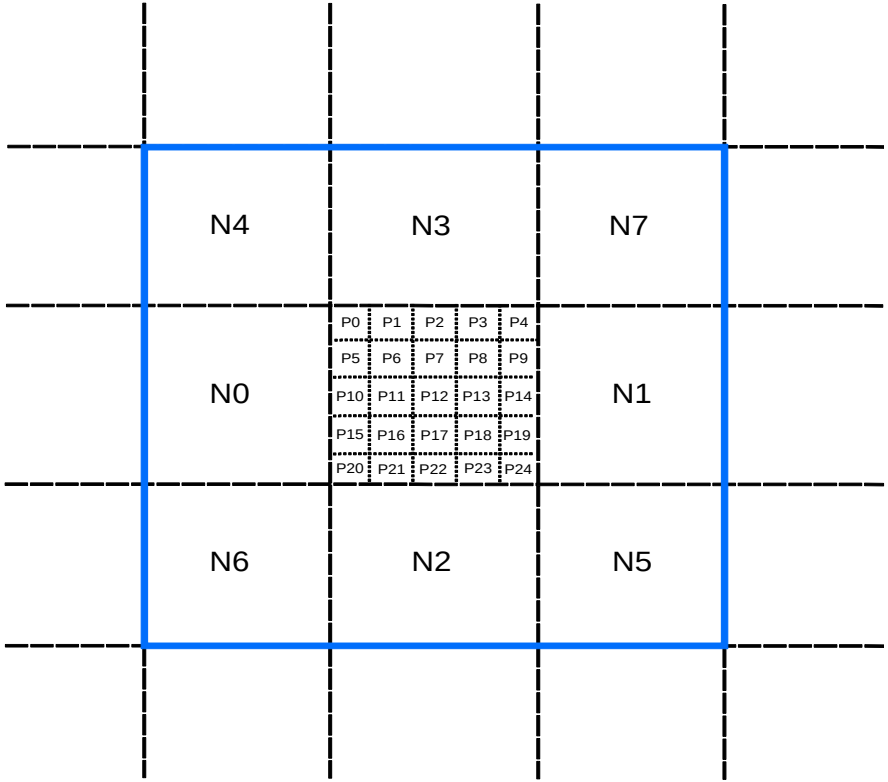
- **Wombat is being developed through a collaboration between Cray Programming Environments and the University of Minnesota Institute for Astrophysics**
  - Peter Mendygral (Cray) is lead developer
  - Tom Jones (UofM) supervises graduate students contributing to the code and drives scientific goals
  - Other contributors/users: Dongsu Ryu (UNIST), Julius Donnert (INAF)
- **Scientific goal is to study turbulence in astrophysical fluids over cosmological scales**
  - MHD + dark matter
- **Application goal is to develop a code capable of achieving the science goals on the latest HPC architectures**

Primary Development Concerns

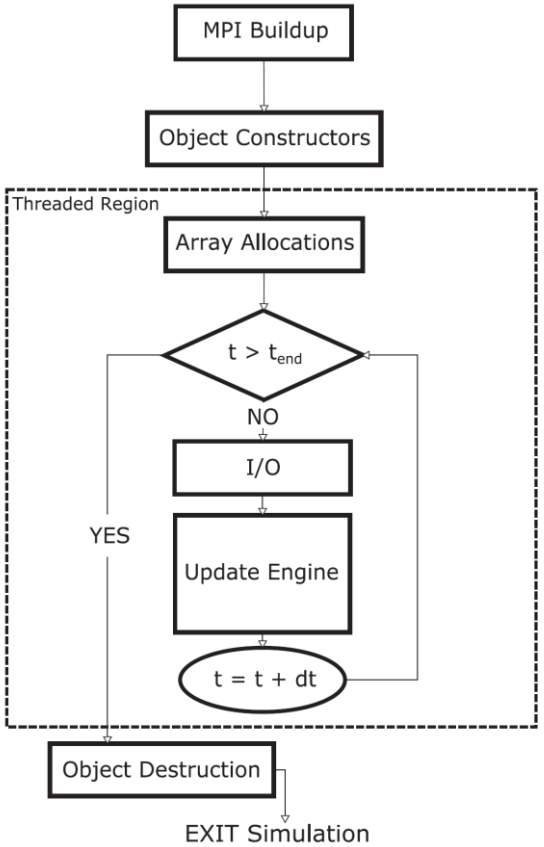
- **Two main issues drove the design of Wombat and explain why other codes have not been sufficient for the science**
  - Scaling to extreme core count required to get to resolutions needed for MHD turbulence
  - Load balancing for SMR/AMR and dark matter particles
- **The approach to these problems in Wombat was**
  - Design to hid communication behind computation as much as possible
  - Wide OpenMP on a node to soften impacts of load imbalances (and hardware imbalances) as much as possible before communicating work between ranks
  - Data structures that reduce AMR/SMR complexity and avoid significant global communication for refined patch tracking

Domain Decomposition

- Domain decomposition is represented in the Fortran data classes and structures
- “Domain” is an array of “Patches” managed by a MPI rank
- “Patch” is a self-contained, self-describing piece of the world grid at some fixed logical location



Wombat Driver and Parallel Region

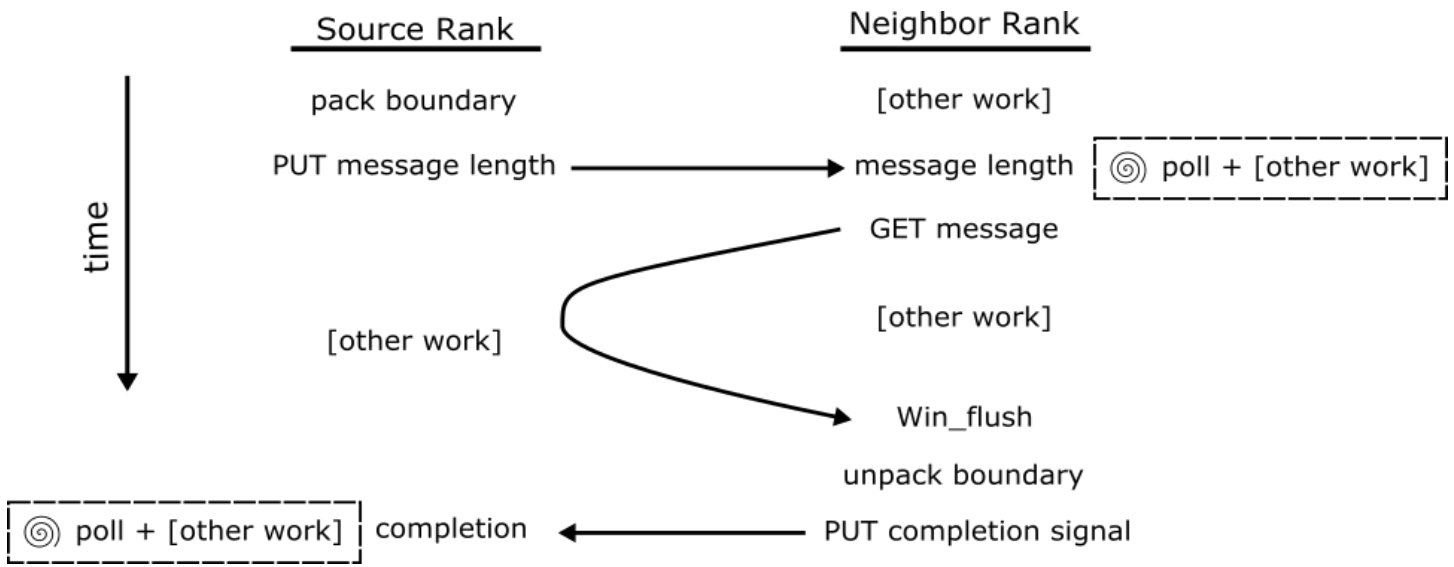


Communication Considerations for Wide Threading

- **If a rank is made much wider with threads, serialization around MPI will limit thread scaling and overall performance**
  - Nearly all MPI libraries implement thread safety with a global lock
  - Cray (and other vendors) addressing this issue
- **Wide OpenMP also means more communication to process per rank**
  - Every Patch now has its own smaller boundaries to communicate
  - Starts tipping the behavior towards the message rate limit
- **Slower serial performance of KNL => maybe look for the lightest weight MPI layer available**
  - MPI-RMA on Cray systems is a thin software layer that achieves similar performance to SHMEM

RMA Boundary Communication Cycle

- **Single passive RMA window used for the duration of the application**
  - No explicit synchronization between ranks
  - RMA semantics make computation/communication overlap simpler to achieve
- **All "solvers" in Wombat utilize a generalized class that implements the communication/computation cycle below**



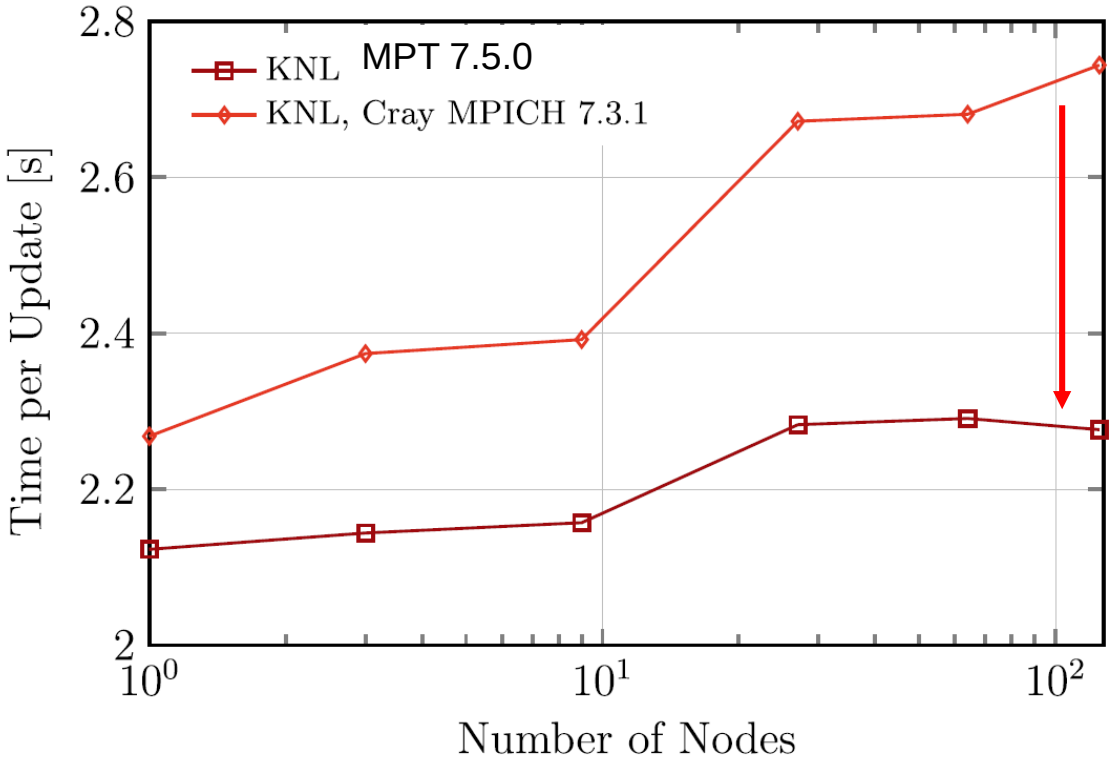
Thread Hot MPI-RMA Synchronization

- **Cray MPICH allows for efficient message completion in a threaded region**
  - Can call completion routines (e.g., MPI_WIN_FLUSH) in a threaded region (not required)
  - Threads collaboratively complete messages issued from all threads
  - Removes the need for any additional thread barriers after MPI_PUT/MPI_GET/etc in threaded region

```
!$OMP DO
DO n = 1, n_neighbors

CALL MPI_PUT(…)

END DO
!$OMP END DO

!$OMP MASTER
CALL  MPI_WIN_FLUSH_ALL()
!$OMP BARRIER
```

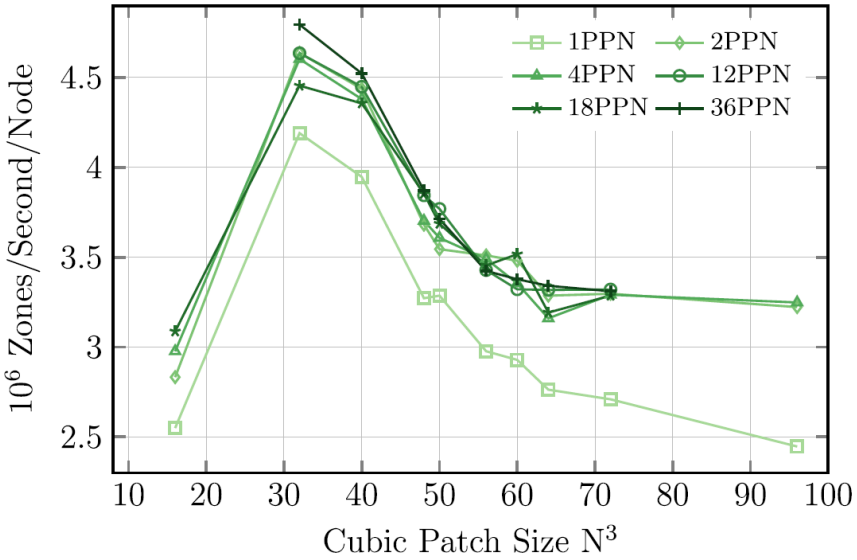
```
!$OMP DO
DO n = 1, n_neighbors

CALL MPI_PUT(…)

END DO
!$OMP END DO NOWAIT

CALL  MPI_WIN_FLUSH_ALL()
!$OMP BARRIER
```

Thread Hot MPI-RMA Speedup

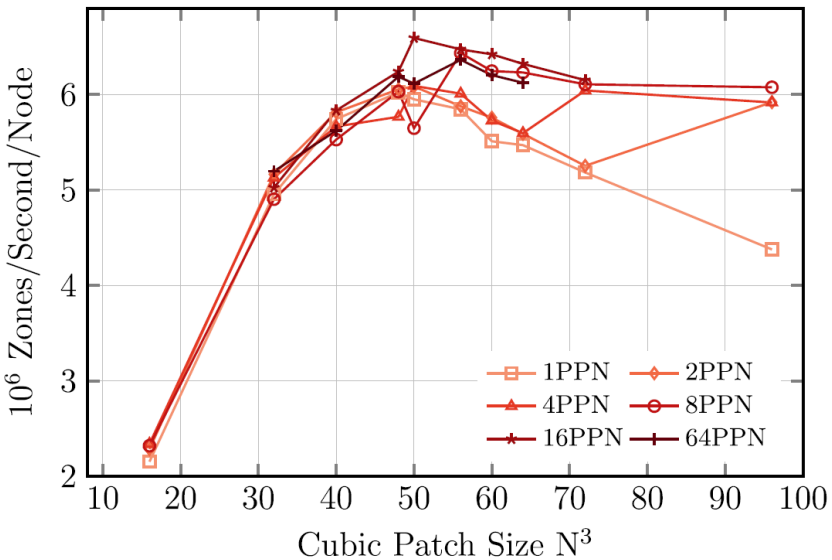


~20% walltime speedup

Dual Socket Broadwell - 36 cores



KNL - 64 cores



- **Measure performance at 27 nodes varying patch size and thread/MPI rank**
- **Fixed grid calculation should not theoretically favor threads or ranks**
  - 9 threads/rank on BDW (nearly optimal performance, good memory footprint)
  - 8 threads/rank on KNL for same reasons
- **~33% higher throughput on KNL over dual socket BDW**

Summary and Recommendations

- Enhancements in Cray MPI to enable users best utilize the MCDRAM technology in flat mode on KNL
- Cray MPI offers Thread-Hot capabilities on Intel Xeon and Intel KNL architectures
- Reviewed design and performance study of Wombat, a high performance astrophysics application that relies on multi-threaded MPI-3 RMA implementation in Cray MPI

- MPI-only works quite well on KNL
  - Threading can be helpful, but unless SPMD with “thread-hot” MPI is used scaling to more than 2-8 threads not recommended
- Using hugepages on MCDRAM can improve large message communication performance
- Thread multiple support in MPI is a key tool on KNL for hybrid applications
- Use asynchronous communication to hide/overlap communication overheads on KNL
- Collectives implemented with user pt2pt is strongly discouraged
  - Especially for alltoall, bcast, and gather
  - Very unlikely pt2pt will perform better
  - If they do, please file a bug with Cray


CRAY
COMPUTE | STORE | ANALYZE
Questions?
Thank You!