

Putting big pegs in small holes: Ameliorating memory limits on Blue Gene/P and related solutions for big data

Jeff Hammond

Leadership Computing Facility
Argonne National Laboratory

24 January 2011



The data problem

The machines keep getting more powerful, but where is the memory?

- Need 450+ MB per MPI rank but my code isn't threaded.
- Need 1+ GB per MPI rank but my code isn't threaded.
- I need hundreds of GB and don't want to use disk.
- I need hundreds of GB and using disk was terrible.

POSIX Shared Memory

POSIX shared memory is a way to share memory between processes.

- POSIX replaces the old Sys5 shared memory API for Linux/Unix.
- Blue Gene/P CNK supports POSIX but not Sys5.
- Every Linux box you use should support POSIX shared memory.

See `man shm_overview` for API.

On Blue Gene/P, the distribution is static.

Using SHM

- 1** Create an MPI communicator per node.
- 2** Allocate SHM using `shm_open` and `ftruncate`.
- 3** Get a pointer to SHM using `mmap`.
- 4** Use SHM properly (i.e. call `msync`).
- 5** Stop complaining about VN mode memory restrictions :-)

`shm_open('x', ...)` is just an alias for `open('/dev/shm/x', ...)`, which is to say, if you understand C file I/O, you are well on your way to understanding POSIX shared memory.

Look at example code

When one node is not enough...

One-sided communication

- 1-sided communication (OSC aka RMA) is not a new concept.
- OSC primitives include Put, Get and Fence, which are a lot like Write, Read and Flush.
- SHMEM, PGAS languages (CAF/UPC), Global Arrays and MPI-2 all provide OSC.
- OSC is a terrible match for Ethernet but maps well to Blue Gene, Cray (Gemini) and Infiniband hardware.
- OSC decouples data movement from synchronization; it *does not* eliminate the need for synchronization.

One-sided programming models

MPI-RMA supports 3 models:

- Bulk Synchronous Parallel (BSP): MPI_Win_fence then Put/Get/Acc then MPI_Win_fence.
- Post-Start-Complete-Wait (PSCW): Pairwise synchronization for halo exchange.
- Passive target (PT): MPI_Win_lock then Put/Get/Acc then MPI_Win_unlock.

PT is similar to GA/ARMCI (GA over ARMCI over MPI-RMA done by Jim Dinan and coworkers).

Look at example code

Final thoughts

- Do not pay attention to the MPI-PGAS wars. Hit nails with hammers and flies with newspaper.
- OSC is closer to the hardware, which is neither good nor bad.
- Breaking apart data movement and synchronization can be useful.
- Supercomputing hardware can do OSC well; the software does not always reflect this.
- I am an OSC fanboy. If you want to use OSC in your app, I can help you.