Part I. Multithreaded MPI for Hybrid MPI+Threads Model

Problem in hybrid MPI + Threads
- Thread Funneled or Serialized mode
- Multiple threads parallelize computation
- Only one thread issues MPI calls
- Most cores are IDLE during MPI Calls

Solution: MT-MPI — Sharing Idle Threads with Application inside MPI

OpenMP Runtime Extension
- Expose number of IDLE threads
- MPI Internal Parallelism
  1. Derived datatype processing
  2. SHM communication
  3. InfiniBand network
- Nested parallelism
  - Creates new Pthreads and offloads scheduling to OS
  - Threads overrunning.

Part II. Process-based Asynchronous Progress Model for MPI RMA

Problem in MPI RMA-based PGAS
- Suitable for applications with large memory requirements. (i.e., NWChem)

Traditional ASYNC Progress
1. Thread-based Approach
   - Per-MPI process background thread
   - Waste half cores or oversubscription
   - Multithreading overhead in MPI
2. Interrupt-based Approach
   - Per-operation hardware interrupts
   - Overhead of frequent interrupts

Solution: Casper — Process-based Asynchronous Progress Model

Core concept
- #cores is rapidly growing
- Not all of the cores are always keeping busy
- Dedicate small & user-specified number of cores to ghost processes
- Ghost process intercepts all RMA operations to the user processes
- Improve ASYNC progress for SW-handled operations without affecting HW-handled RMA
- No multithreading or interrupt overhead
- Flexible core deployment
- Portable PMPI® redirection

Correctness and Performance challenges
- Ensuring correctness
  1. Lock permission for shared ghost processes
  2. Managing multiple ghost processes
  3. Self lock consistency
  4. Multiple simultaneous epochs
- Ensuring performance
  1. Memory locality

Reference

Figure: Execution Time Speedup

Table: Execution Time Speedup

Graph: Execution Time Speedup

Figure: Execution Time Speedup