

MPI + X (OpenSHMEM?)

Michael A. Raymond
SGI MPT Team Leader

Agenda

- Models of +X
- Why
- Data model
- What to do about the Symheap?
- Accelerators
- Fault Tolerance
- Development Requirements

Models of +X

- Each section has its own programming model
- Intra-node
 - E.g. OpenMP, OpenACC

Why +X?

- Communication topography matters
- Endpoints are expensive
- MPI is the de facto inter-node model
- “Easy” way to parallelize things within a node

How is OpenSHMEM different?

- Low overhead
- Symmetric heap
- Different mental model
 - Message passing vs. direct access

Why intra-node OpenSHMEM?

- Simpler code paths
- SIMD-like?
- More NUMA-aware than OpenMP
- Better CPU – accelerator interaction?

Why not OpenSHMEM?

- Maybe it doesn't go far enough
- Hard to convert existing codes?
- Harder than OpenMP conversion?
- The symmetric heap

The Symmetric Heap

- Per-thread heaps at same address?
- MPI interaction?
- Separate process & thread symheaps?
- Interaction with “communicator” heaps?
- Dynamic creation and destruction of threads?

Heterogeneous Nodes

- Intel Phi / “Smart” GPU / Tiler / FPGA
- OpenSHMEM would work well between them and the CPU
- Keep type-specific SHMEM calls?

Cache Coherency / Flushing

- Part of the SHMEM standard designed for the Cray T3D
- Today, InfiniBand AMOs are not cache coherent
- Similar problems with accelerators?

Fault Tolerance

- Fault tolerance is a lot easier for MPI
- No one has solved restarting AMOs
- What if you restart on a different size node?
- Can I hot swap an accelerator card?

Development Requirements

- Special OpenSHMEM library
- Thread protection?
- VM tricks?
 - Symheap
 - Heterogeneous nodes
- Symheap aware MPI

sgi