Global Arrays Roadmap

Prepared by: Daniel Chavarría
Sriram Krishnamoorthy
Manoj Krishnan
Bruce Palmer
Abhinav Vishnu

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

Roadmap

- Scalability to 200k+ processes
- Thread safety - support for multithreaded execution
- Portability
- Maintainability and Usability
  - User support
- Fault Tolerance
- Data Decomposition and Load balancing
- Support for Hybrid Platforms
- Performance tools for GA/ARMCI
- More GA-based applications

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Scalability

- Scalability to 200k+ processes
- Inter-node and intra-node communication
  - Scaling on manycore, multicore architectures
  - Optimized remote memory access (RMA)
    - Put, get, …
- GA metadata management
  - Reduce memory overhead in GA

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Support thread-based execution models running on multi-core platforms

- Support thread-based execution models running on multi-core platforms
  - OpenMP, pthreads, qthreads, others
- Allow multiple threads executing on a single node to execute GA function calls to any other node
- Move towards a more multi-core friendly execution model
  - More sharing of memory and other resources compared to pure process-based models

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Portability

- Leadership class petaflop machines
  - Cray XTs
  - IBM Bluegene/P
- Multicore Linux Clusters
  - e.g. Chinook – Infiniband Linux cluster
- Future architectures/networks
  - Cray Gemini & Aries, IBM PERCS
  - Hybrid systems
- Maintenance and Usability
  - Modularization of ARMCI protocol layer
  - User support

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

GA Extreme Scale Challenges

- Fault Tolerance
  - Fault resilient GA
  - Fault tolerance support to GA applications
- Data Decomposition and Load balancing
  - Flexible and topology-aware data decomposition
  - Dynamic load balancing
    - Task counters, work queues, work stealing, ..
  - Communication optimization
- Support for Hybrid Platforms
  - Extend the functionality of GA to GPU-accelerated clusters
- Performance tools for GA/ARMCI

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

MPI-3

- The MPI-2 one-sided API – MPI_Get, MPI_Put, etc
  - Proposed eleven years ago
  - Too restrictive for actual application use
- MPI-3 RMA
  - Developing alternative design to the MPI-2 one-sided standard
  - Inspired by

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Short Term Goal – GA 5.0

- Scalability on leadership class machines
  - Cray XT5, Bluegene/P, large-scale Linux clusters
- Usability
  - GA Install and setup
  - Revamping GA webpage and documentation
- Autoconf/automake build mechanism
- Remove memory allocator (MA) dependency
- Thread-safety (coarse)

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Summary

- The idea has proven very successful
  - Efficient on a wide range of architectures
    - Core operations tuned for high performance
  - Library substantially extended but all original (1994) APIs preserved
  - Increasing number of application areas
- Supported and portable tool that works in real applications

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Source Code and More Information

- Version 4.3 available
- Homepage at http://www.emsl.pnl.gov/docs/global/
- Platforms (32 and 64 bit)
  - IBM BlueGene/L, BlueGene/P, LAPI-based systems
  - Cray XTs
  - Linux clusters with InfiniBand, Ethernet, Myrinet, or Quadrics
  - Other systems (SGI Altix, Fujitsu, NEC, Windows)

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965