

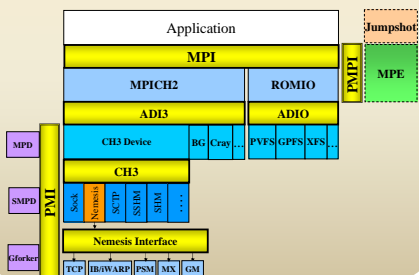
# MPI Research and MPICH2

Pavan Balaji, Darius Buntinas, Anthony Chan, Dave Goodell,  
Jayesh Krishna, Rusty Lusk, Rob Ross, Rajeev Thakur  
Argonne National Laboratory  
William Gropp, University of Illinois at Urbana-Champaign

## Thrust Areas

- MPI implementation research, embodied in MPICH2
- General MPI research, including tools, test suites, programming models

## MPICH2 Architecture



## Major Implementations Derived from MPICH2

- IBM for BG/L and BG/P
- Cray for XT-3/4
- Intel
- Microsoft
- SiCortex
- Myricom
- Myricom
- Ohio State (MVA-PICH2)

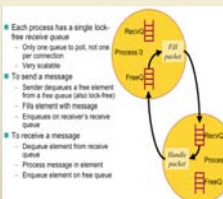
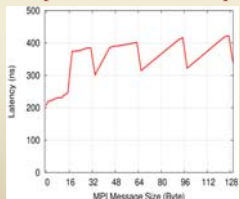


## Research Areas

- Fast, low-latency communication
- Efficient support for multicore architectures
- Scalability to more than 100,000 processes
- Efficient support for multithreading
- One-sided communication
- Derived datatypes
- Process management
- Collective communication
- Scalable parallel I/O
- Fault tolerance (CIFTS, MPICH-V)
- Extensions to MPI in MPI-3
- Performance and Correctness Tools
- Formal verification of MPI programs and implementations

## Nemesis

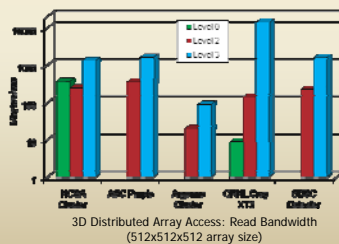
- High-performance communication channel
- Supports shared memory and multiple networks
- Optimized to minimize function calls, instructions, cache misses, etc.
- Uses assembly-level atomic operations to implement lock-free queues



207 nanoseconds MPI latency over shared memory. 2.6 GHz Intel Clovertown machine

## ROMIO

- High-performance MPI-IO implementation
- Implemented on top of numerous parallel file systems (PVFS, GPFS, Lustre)
- Efficient support for collective I/O and noncontiguous I/O
- Forms substrate for many high-level I/O libraries: parallel HDF-5, PnetCDF



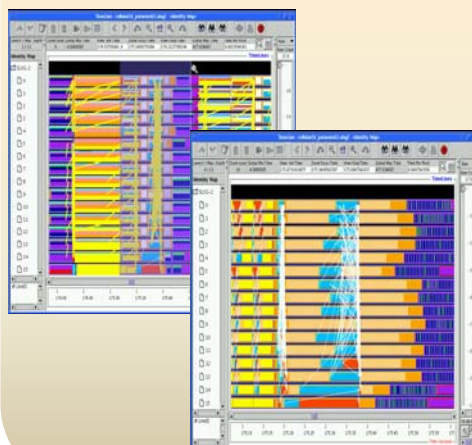
## Fine-Grain Threads

- Collaboration with IBM Watson for next-generation Blue Gene
- Needed to support high message rates with multiple MPI threads on multicore systems
- Exploring various approaches to thread safety and their performance implications
- Single global lock
- Multiple smaller locks
- Lock-free operations
- Developed performance test suite for MPI calls from multiple threads

## MPE: Multi-Processing Environment

- Included in MPICH2; works with any MPI implementation
- Five main components:
  - Library for logging MPI and user events
  - Tool for visualizing log files: Jumpshot
  - Validation library for MPI collectives and datatypes
  - Tracing library that prints MPI calls to stdout
  - Shared-display parallel X graphics library
- Jumpshot used by TAU, PPX for UPC and AIX's PE benchmarking tool

## Jumpshot Log File Visualizer



## Formal Verification of MPI Programs

- Collaboration with University of Utah
- Two approaches:
  - Generate a model of an MPI program using a modeling language (Promela) and verify it using a model checker (SPIN)
  - Use in-situ model checking without creating a model first
- Found deadlock in a published locking algorithm that uses MPI one-sided communication
- *The authors were unaware of this bug until the model checker found it!*