MPICH2

Performance and Portability

MPICH2 is a **high-performance** and **widely portable** implementation of the Message Passing Interface (MPI) standard (both MPI-1 and MPI-2).

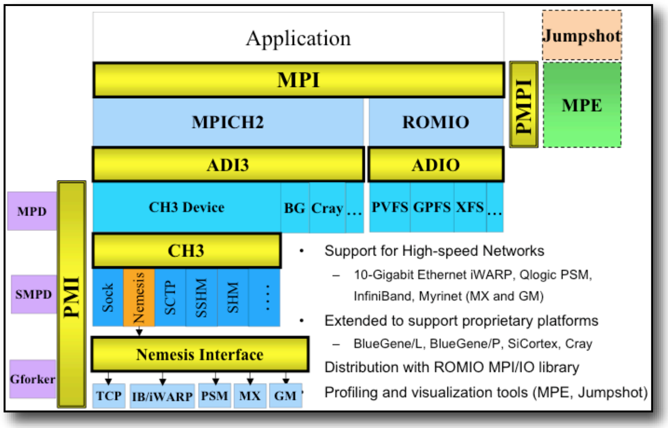
MPICH2 won the R&D 100 award in 2005.

Goals of MPICH2

1. Provide an MPI implementation that efficiently supports different computation and communication platforms including commodity clusters *(desktop systems, shared-memory systems, multi-core architectures)*, high-speed networks *(10 Gigabit Ethernet, InfiniBand, Myrinet, Quadrics)* and proprietary high-end computing systems *(Blue Gene, Cray, SiCortex)*.

2. Enable cutting-edge research in MPI through an **easy-to-extend** modular framework for other derived implementations.

MPICH2 Architecture



- Support for High-speed Networks
  - 10-Gigabit Ethernet iWARP, Qlogic PSM, InfiniBand, Myrinet (MX and GM)
- Extended to support proprietary platforms
  - BlueGene/L, BlueGene/P, SiCortex, Cray

Distribution with ROMIO MPI/IO library

Profiling and visualization tools (MPE, Jumpshot)

MPICH2 implements both MPI-1 and MPI-2 standards, including support for:

✓ Dynamic process management
✓ Remote-memory access (one-sided)
✓ Parallel I/O
✓ C, C++, Fortran (77 & 90) language bindings
✓ Singleton init
✓ Thread safety at MPI_THREAD_MULTIPLE level.

Visit us at
http://www.mcs.anl.gov/mpi/mpich2

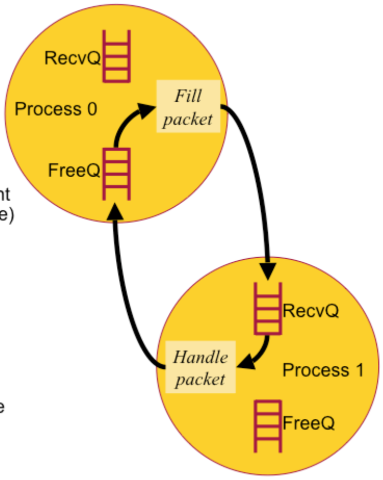
NEMESIS – A high performance communication channel for MPICH2

NEMESIS is a scalable, high-performance, shared-memory, multi-network communications subsystem within MPICH2. Nemesis offers low-latency, high-bandwidth communication, particularly for intra-node communication.

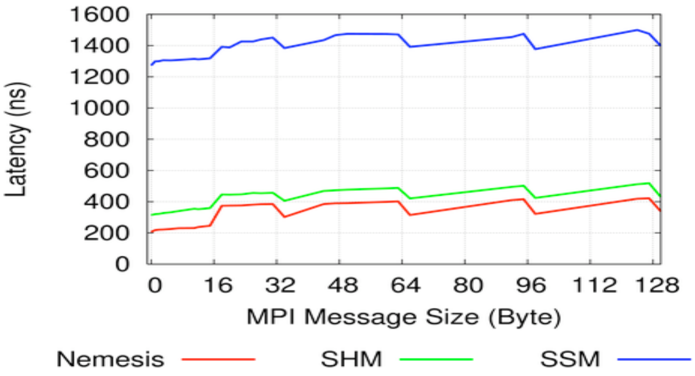
Extensive work went into tuning Nemesis for high-performance communication. This includes the usage of shared memory and lock-free algorithms as well as optimized memory copy routines. Nemesis allows MPI processes to communicate over multiple communication paths. So, for example, a process can communicate with another process on the same node over shared memory, with another process on a different node over InfiniBand or Myrinet. This eliminates the need to choose a single lowest common denominator communications protocol.

Nemesis Intra-node Communication

- Each process has a single lock-free *receive queue*
  - Only one queue to poll, not one per connection
  - Very scalable
- To send a message
  - Sender dequeues a free element from a *free queue* (also lock-free)
  - Fills element with message
  - Enqueues on receiver's receive queue
- To receive a message
  - Dequeue element from receive queue
  - Process message in element
  - Enqueue element on free queue



Nemesis Intra-node Latency



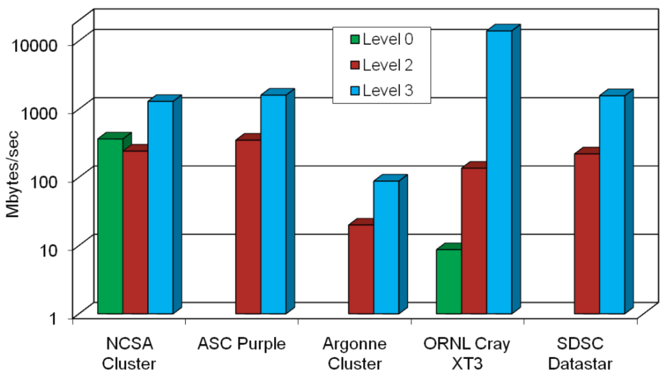
Intra-node performance of Nemesis vs. ssm and shm channels on a dual-core, dual-processor Intel 2.6 GHz Clovertown machine

MPICH2
Performance and Portability

ROMIO - High performance and portable MPI-IO

ROMIO is a high-performance, portable implementation of MPI-IO (part of the MPI-2 standard) that works with any MPI implementation on multiple file systems. It is included as part of MPICH2, MPICH1, vendor MPI implementations, Open MPI and LAM, and supports optimized implementations for PVFS, SGI XFS, PanFS (Panasas), and UFS file systems (the UFS implementation can be used for Lustre and GPFS). ROMIO performs sophisticated optimizations that enable applications to achieve high I/O performance. These include collective I/O, data sieving and I/O aggregation. ROMIO also accepts a number of hints from the user for improving I/O performance, such as file striping and algorithm tuning parameters.

ROMIO Performance



3D Distributed Array Access: Read Bandwidth
(512x512x512 array size)

EVENTS AT SC'07

TUTORIALS

Nov 11th; 8:30am; Loc: A4 **Parallel I/O in Practice**
Nov 12th; 8:30am; Loc: A3 **Advanced MPI**
Nov 12th; 8:30am; Loc: A6 **Designing Systems with InfiniBand and 10-GE iWARP**

MPICH2 BIRDS-OF-A-FEATHER (BoF) Session

Nov 15th; 12:15pm - 1:15pm; Location: A3/A4

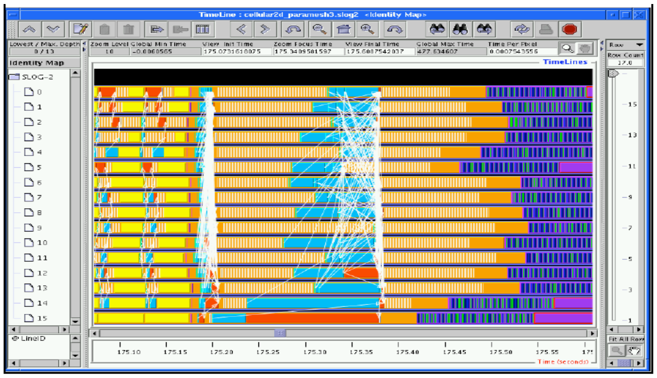
COLLABORATIVE DEMOS (Loc: ANL Booth #551)

Nov 13th, 14th, 15th; 2:00 pm-3:00 pm **CIFTS**
Nov 13th, 14th, 15th; 11:00 am-12:00 noon **PARAMEDIC**

MPE – MPI Parallel Environment

MPE is a suite of performance analysis tools comprising of profiling libraries, utility programs, graphical tools and checking ibraries. MPE is included in the MPICH2, but can be used with any MPI implementation that provides the MPI profiling interface. The MPE profiler generates logs viewable by the integrated Jumpshot visualization tool. The datatype and collective verification library finds argument inconsistency in MPI collective calls. The tracing library records all MPI calls and the animation and X-graphics libraries provide a real-time program animation of the trace using X-window routines.

MPE Jumpshot Tool


TimeLine : cellular2d_paramesh3.slog2 <Identity Map>

Lowest / Max. Depth: 0 / 13
Zoom Level: 10
Global Min Time: -0.0060565
View Init Time: 175.0731618075
Zoom Focus Time: 175.3409581597
View Final Time: 175.6087542037
Global Max Time: 477.634607
Time Per Pixel: 0.0007543556
Row Count: 17.0

FPMPI2

FPMPI2 is a library that takes advantage of the MPI Profiling interface to create a summary of the use of each MPI call. It distinguishes between messages of different sizes within 32 message bins (essentially powers of two) and optionally identifies synchronization time (the time that an MPI call is forced to wait). FPMPI2 may be used with any MPI implementation, and has been run scalably on over 16,000 processes.

PARTNERS

IBM	(MPI BlueGene/L and BlueGene/P)
Cray	(MPI over RedStorm and XT3)
SiCortex	(MPI SiCortex)
Microsoft	(MPICH2-MS)
Intel	(MPICH2-Nemesis)
NetEffect	(MPICH2-iWARP)
Qlogic	(MPICH2-PSM)
Myricom	(MPICH2-MX)
Ohio State Univ.	(MVAPICH and MVAPICH2)
Univ. of British Columbia	(MPICH2/SCTP)