SGI Message-Passing Status and Plans

Karl Feind

kaf@sgi.com

SGI Parallel Communication Team

SUMMIT 2001

43rd CUG Conference on High Performance Computing & Visualization

MPT Themes

- Performance
- Platforms and Interconnects
- Standards

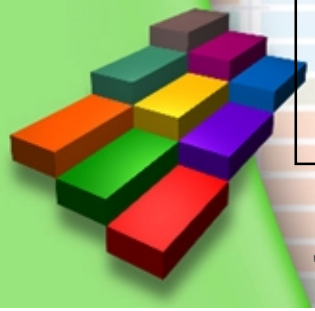
Performance Features in MPT

- Low latency and high bandwidth.
- Fetchop-assisted fast message queuing
- Fast fetchop tree barriers
- MPI and SHMEM one-sided communication
- Large SSI support
- Automatic NUMA placement
- Optimized MPI collectives
- Internal MPI statistics reporting
- Integration with PCP
- Single copy send/recv transfers
- Runtime MPI tuning

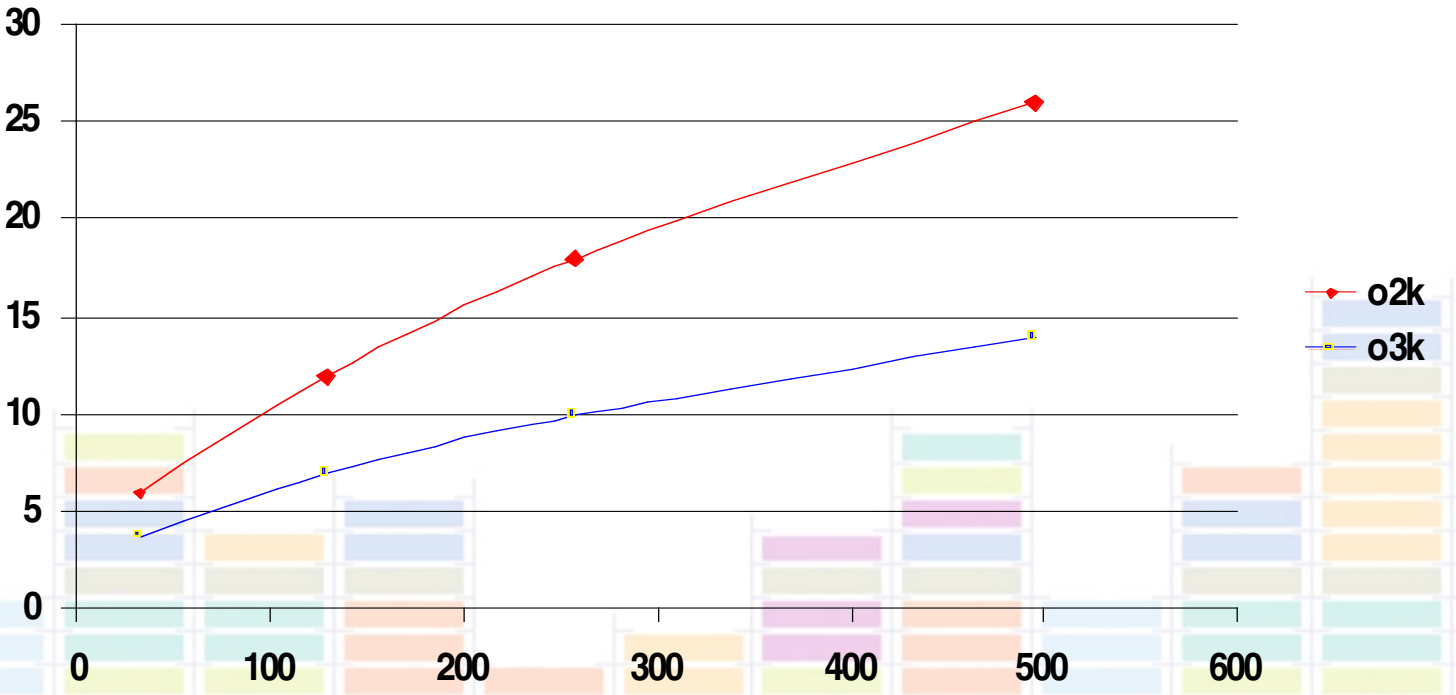
Communication on NUMAlink

MPI Performance on 400 MHz Origin 3000)

send/recv latency	5.5 usec
put/get latency	1 usec
Peak bandwidth (bcopy)	280 Mbytes/sec
Peak bandwidth (BTE)	600 Mbytes/sec
All communicate peak bandwidth per transfer	170 Mbytes/sec



Barrier Synchronization Time on O2K and O3K

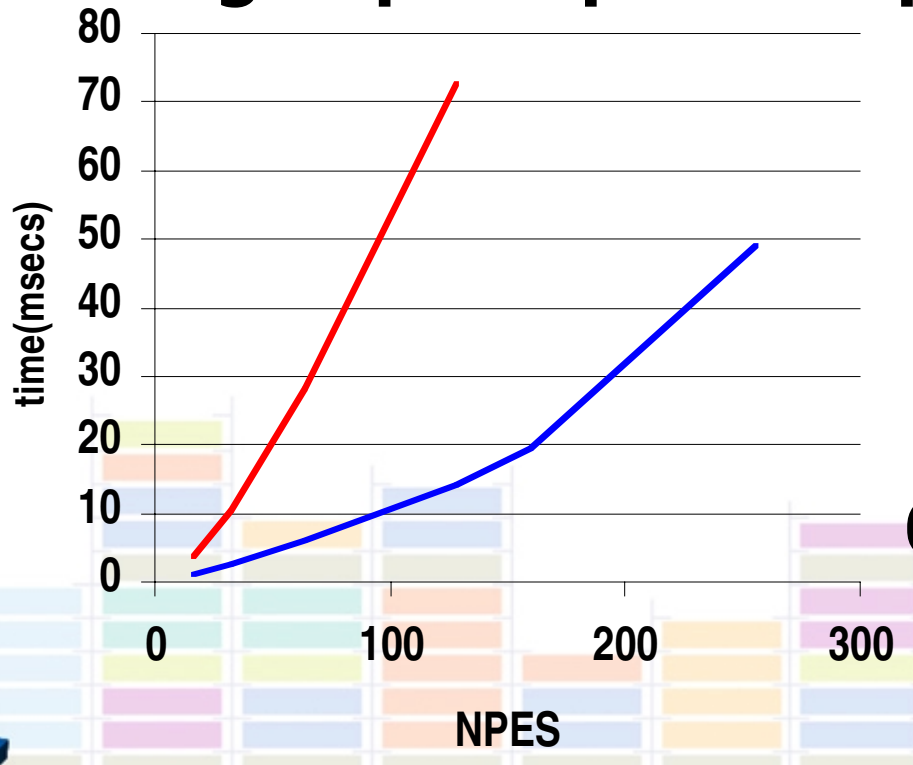


Assign MPI Ranks to Physical CPUs

- Environment Variable Syntax
  - setenv MPI_DSM_CPULIST 0-15
  - setenv MPI_DSM_CPULIST 0,2,4,6
- Also SMA_DSM_CPULIST
- Maps ranks 0 - N onto the physical CPUs in the order specified
- Useful only on quiet systems
- Easier than using dplace command

Single Copy Speed-up

Alltoallv type communication pattern using explicit point to point calls



O2K 300 MHz

Single Copy Send/Receive

- Send buffer must be globally accessible (ABI 64 required)
  - common block
  - symmetric heap (shmalloc/SHPALLOC)
  - global heap (ALLOCATE statement with SMA_GLOBAL_ALLOC set)
- Set MPI_BUFFER_MAX to 2048
- Best used with MPI_Isend or MPI_Bcast
- Best if sender does not immediately wait for send completion.
- Little payoff below 8 Kbyte messages

Reducing Run-Time Variability

- Recommended algorithm for workload manager launch of MPI jobs in SSI:
  - Batch scheduler creates a cpuset
  - Batch scheduler launches mpirun into cpuset
  - MPI job is confiined within cpuset during execution by virtue of fork/exec/cpuset semantics
  - MPI job performs automatic NUMA placement within the cpuset
  - When MPI job completes, the batch scheduler destroys the cpuset.
- NOTE
  - Use exclusive cpusets or restrict interactive use of the system.

Platforms and Interconnects: MPI

- MIPS
  - Single kernel NUMAlink
  - Partitioned NUMAlink (available June 2001)
  - GSN (libst 2.0 work planned in July 2001)
  - Myrinet
  - Sockets
  - HIPPI
- SNIA
  - Single kernel NUMAlink (prototype working)
  - Partitioned NUMAlink (prototyping in late 2001)
  - Myrinet (prototype running on IA64)
  - Sockets (prototype working)

Platforms and Interconnects: SHMEM

- MIPS
  - Single kernel NUMAlink
  - Partitioned NUMAlink (planned Dec 2001)
- SNIA
  - Single kernel NUMAlink (planned Sep 2001)
  - Partitioned NUMAlink (planned Nov 2001)

Platforms and Interconnects: PVM

- MIPS
  - PVM support is retired after MPT 1.6 (2002)
- SNIA
  - SGI will not provide PVM on SNIA

MPI-2 Features Planned

- MPI I/O enhancements: MPI_Wait integration
- MPI-2 datatypes: replacements for deprecated MPI-1 datatypes
- Expanded one-sided communication
- Process spawn

SGI Message-Passing References

- "relnotes mpt" gives information about new features and how to install MPT
- "man mpi" tells about all environment variables
- "man shmem" tells about the SHMEM API
- *MPI Programmer's Manual* viewable with *insight* and on the web at http://techpubs.sgi.com
- MPT web page: http://www.sgi.com/software/mpt/