What's New in the Message Passing Toolkit

Karl Feind

Message-Passing Engineering Team

SGI

41st Cray User Group Conference
Minneapolis, Minnesota

SGI Message-Passing Software

- Message Passing Toolkit (MPT) for IRIX, UNICOS/mk, and UNICOS
- MPI, SHMEM, PVM
- MPT web site:
  - http://www.sgi.com/software/mpt/

Outline

- New features in the past year
  - MPT 1.2.1, 1.3, 1.3.0.1
- Getting high performance
  - High Performance message passing features
  - IRIX scheduling for best performance
- Message-passing future plans

Message-Passing Feature Themes

- Large cluster support
- Performance optimizations
- Usability improvements
- Inter-operability
- More MPI-2 API

MPT 1.2.1 Highlights

- MPI/SHMEM inter-operability on IRIX
- IRIX Miser support
  - -miser option on mpirun restored parent/child relationship
  - Limited Miser support in IRIX 6.5
  - Stable Miser available in IRIX 6.5.4

MPT 1.3 Highlights

- Large cluster support
  - Up to 6144 processes on 48 hosts
  - Multiple Hippi boards per host
- Origin 2000 Performance Improvements
  - MPI_Barrier use of fetch-ops
  - Cluster-aware MPI_Barrier

MPT 1.3 Highlights

- Usability Enhancements for IRIX systems
  - mpirun failure diagnosis
  - Fortran MPI interface checking
    - -auto_use mpi_interface,shmem_interface
- Performance Analysis and Debugging
  - Totalview support for MPI on PVP
  - -stats option on mpirun command for IRIX
  - Totalview support for message queue display on IRIX

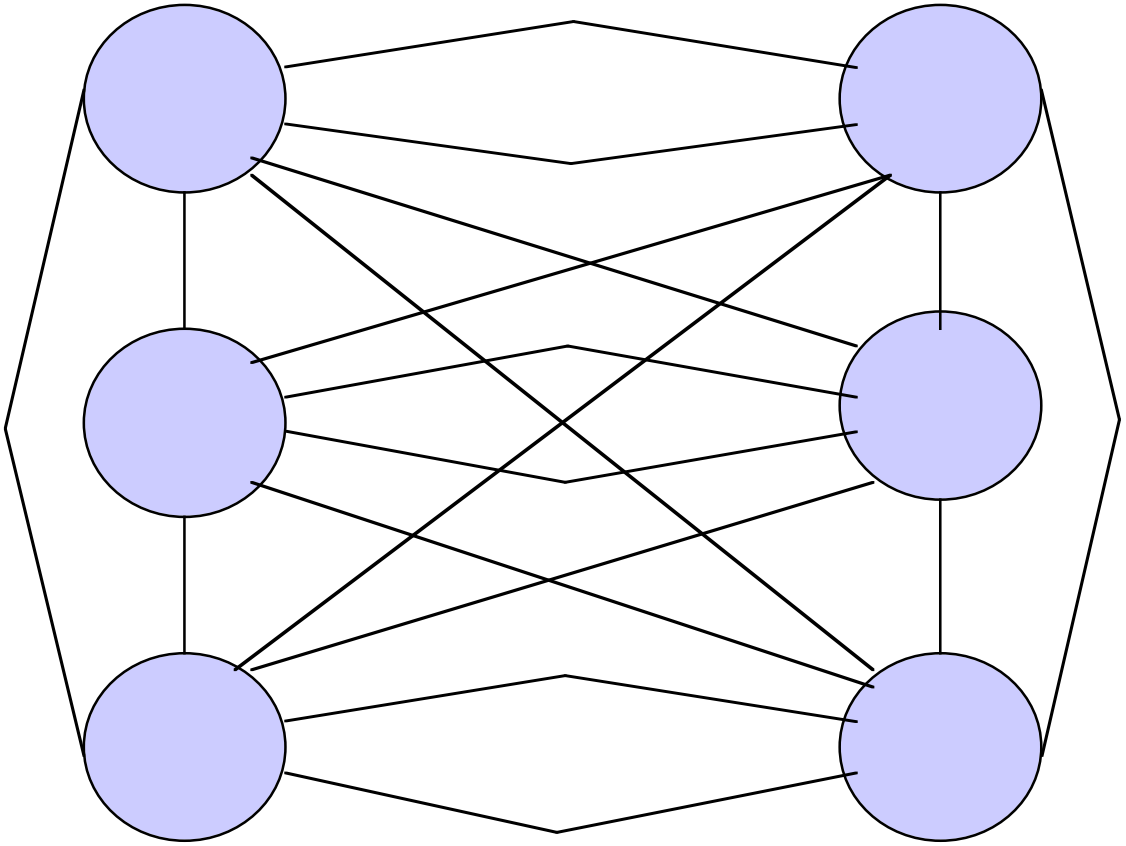
MPT 1.3 Highlights

- MPI-2 API
  - MPI I/O on IRIX and T3E
  - Thread-safe MPI on IRIX
- Features in MPT 1.3.0.1
  - SHMEM support for CRAY SV1 cache and memory system support
  - MPI_Alltoall optimizations on IRIX

High Performance Message Passing

- More parallelism
  - MPI support for larger Origin 2000 clusters
  - Message passing optimizations for large Origin 2000 systems with up to 256 processors.
- Faster collective communication
  - single host
  - multi-host

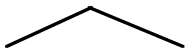
ASCI Cluster Configuration



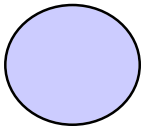
16 way HIPPI-800 switches and the SPPM benchmark led to this topology

There are 36 separate networks with 576 HIPPI adapters.

Inside each 8 host cluster, the connectivity is 12. Going outside a cluster, connectivity drops to 4 or 2.



= 32 connections + 2 HIPPI switches



= 8 x128 O2000

High Performance Message Passing

- MPI_Barrier optimizations on IRIX
  - in-host fetch-op algorithm
  - multi-host cluster-aware algorithm

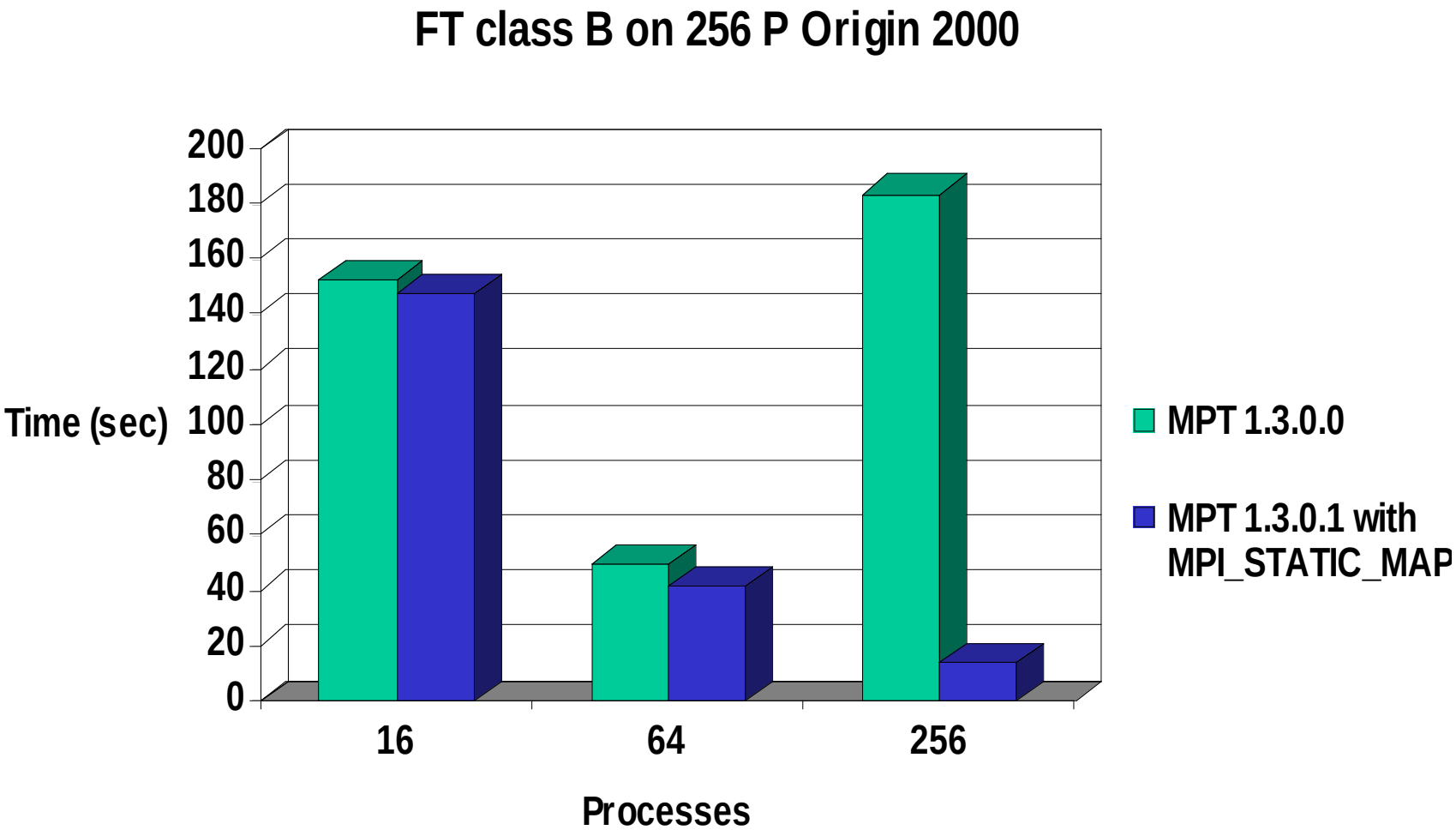
MPI_Barrier performance (microseconds)

Hosts x Processes	*Pre-optimization*	*Optimized*	*Improvement*
1 x 64	*3140*	*10*	*300 x*
1 x 128	*24000*	*26*	*1000 x*
2 x 4	*670*	*174*	*4 x*
4 x 16	*26000*	*994*	*26 x*

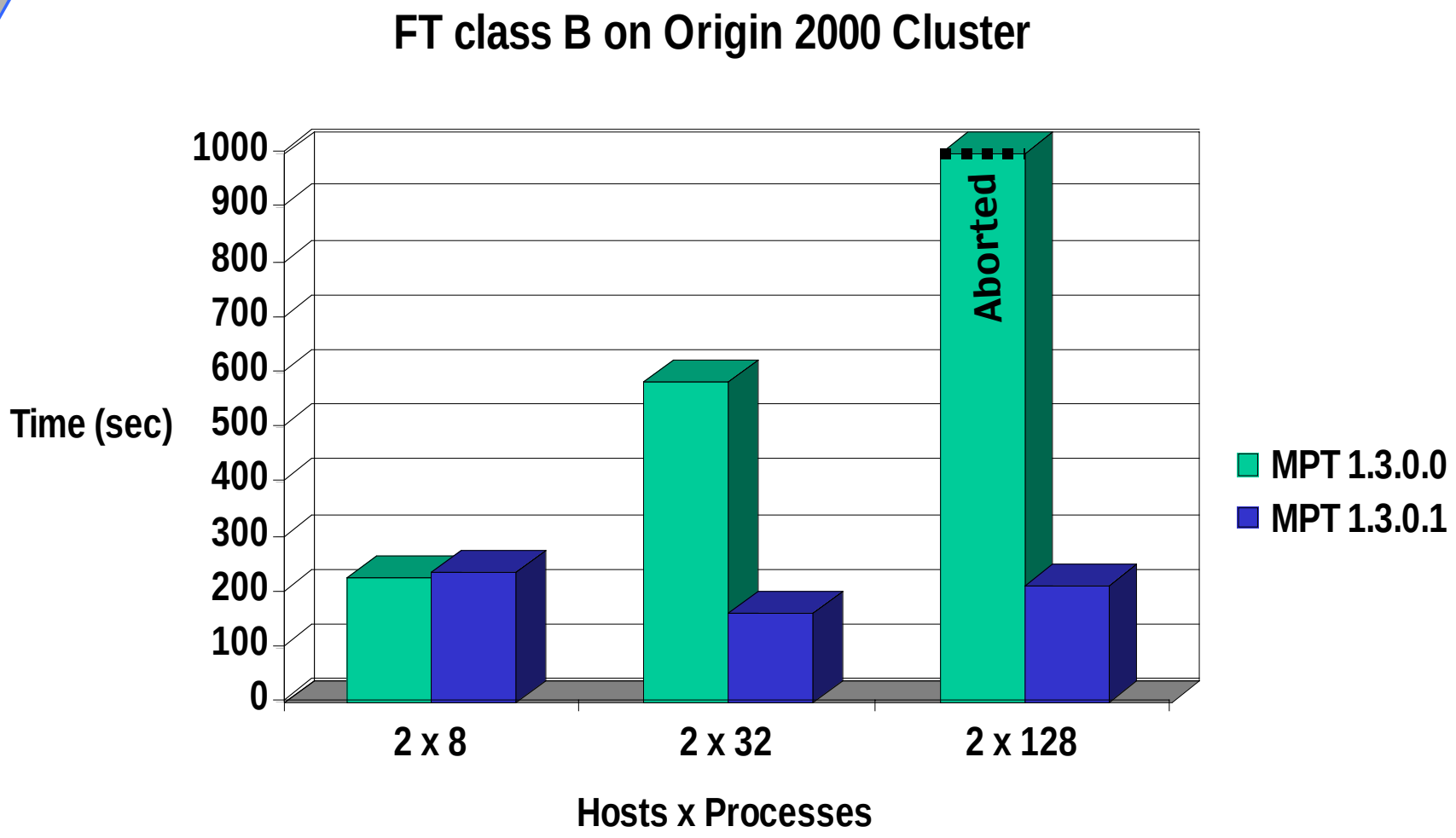
High Performance Message Passing

- **IRIX MPI_Alltoall optimizations**
  - **MPI buffer bypass for in-host optimization**
  - **Cluster-aware recursive algorithm for multi-host optimization**

NAS Parallel FT Execution Time



NAS Parallel FT Execution Time



Scheduling Parallel Jobs on IRIX

- Synchronized execution is key
  - SPMD programs are loosely or tightly synchronized
  - Waiting processes consume memory and CPU
- Use Miser in IRIX 6.5.4
  - “static” qualifier on miser_submit
  - queue repacking policy
  - stability

Scheduling Parallel Jobs on IRIX

- Miser job submission
  - MPI
    - miser_submit –q default –o c=64,m=4g,t=10h,static \
    - mpirun –miser –np 64 a.out
  - SHMEM
    - setenv NPES 64
    - miser_submit –q default –o c=64,m=4g,t=10h,static a.out

Message-Passing Future Plans

- MPT 1.4, late 1999
  - HIPPI resiliency support for large Origin clusters
  - MPI support for CRAY SV1
  - MPI-2 C++ bindings on IRIX
  - MPI collectives optimization on IRIX
  - GSN support infrastructure
  - Improved cleanup of aborted MPI jobs on IRIX
- MPT 1.5, 2000
  - MPI-2 one-sided communication on IRIX
  - MPI-2 I/O enhancements on IRIX
  - MPI support for Origin clustering with GSN
  - MPI/LSF inter-operability enhancements