SiliconGraphics

CUG

CRAY USER GROUP Incorporated

MPT/Cluster Software Update

Karl Feind

MPT and Cluster Software Group
Silicon Graphics

kaf@cray.com
http://reality.sgi.com/kaf_craypark

STUTTGART

Outline

- Introduction
- IRIS Failsafe recent and future releases
- IRISconsole recent and future releases
- Message passing recent releases
- Message passing performance data
- Message passing roadmap

The MPT/Cluster Team

- Products
  - Message Passing Software (MPI, PVM, SHMEM)
  - IRIS FailSafe Agents
  - IrisConsole
- Staff
  - 4 message passing engineers
  - 2 IRIS FailSafe engineers (+3 in other groups)
  - 1 IrisConsole engineer

IRIS FailSafe Product

- Provide fail-over between a team of servers (nodes)
- Highly available (HA) services
- See `http://www.sgi.com/Products/software/failsafe.html`

IRIS Failsafe Software Releases

- FailSafe 1.2
  - Released in 1997
  - Supports 2-way redundancy
  - Dual-active and active-standby modes
- Failsafe 2.0
  - July 1998 beta release planned
  - up to 8 node configurations supported
  - Separate resource groups can have different fail-over policies

IRISconsole Product

- System console support for clusters
- Consoles for multiple servers on same workstation
- Provides remote console access
- Provides system logging for multiple servers
- See `http://www.sgi.com/products/remanufactured/challenge/ti_irisconsole.html`

IRISconsole Software Releases

- IrisConsole 1.2
  - Bugfix/maintenance release

IrisConsole Software Releases

- IrisConsole 1.3
  - June 1998 release planned
  - IRIX 6.5 client support
  - Support EtherLite 8 and 16 multiplexers
  - Enhanced control over logging and alarms
  - Support partitioned arrays
  - Support for up to 64 servers
  - IRIX 6.5 client support
  - Year 2000 compliance

Message Passing Toolkit (MPT)

- Facilitate highly (massively) parallel programming
- MPI, PVM, and SHMEM
- Support standard and optimized message-passing
- Available on T3E, Origin2000/IRIX, and PVP systems
- See `http://www.sgi.com/Products/software/mpt.html`

1997 Software Releases

- MPT 1.1
  - Released in June 1997
  - First IRIX MPT release
  - Supported on IRIX 6.2 and 6.4
- CrayLibs 3.0.1, 3.0.1.2, 3.0.2
  - T3E release package containing SHMEM

1997 Message Passing Features

2Q97 MPT 1.1

-First IRIX MPT release.

-Origin2000 NUMA enhancements in MPI

-Origin2000 intro of SHMEM

-Origin2000 intro of XMPI

1997 Message Passing Features

3Q97	CrayLibs 3.0.1	-T3E-900 SHMEM support activates streams by default.
4Q97	CrayLibs 3.0.1.2	T3E –1200 SHMEM get bandwidth increased
	CrayLibs 3.0.2	T3E SHMEM gets shmem_group_create_strided ()

1998 Software Releases

- MPT 1.2
  - Released January 1998
  - Significant feature release
- MPT 1.2.0.2
  - Released March 1998
  - Bugfix release with some features
- MPT 1.2.1
  - Released June 1998
  - Feature release

1998 Message Passing Features

1Q98 MPT 1.2 -IRIX MPI source code used on PVP

-PVP MPI optimize intra- and inter-host communication in same application

-MPI_ENVIRONMENT variable

-IRIX MPI 64 cpu support

-IRIX/PVP MPI stdout/stderr output options

-IRIX MPI LSF support

1998 Message Passing Features

1Q98 **MPT 1. 2** (continued)

-IRIX MPI Dolphin TotalView support

-IRIX MPI ROMIO support

-T3E-900 MPI streams support
-mpirun syntax convergence

-IRIX SHMEM SC API
expansion (shmalloc, collective)

-IRIX SHMEM job overhead
reduction

1998 Message Passing Features

1Q98 MPT 1.2 (continued)

-IRIX SHMEM **SMA_*** environment variables

-IRIX PVM shared arena fixes

-IRIX PVM **PVM_RSH**

-IRIX: Optional alternate location installation via Modules

1998 Message Passing Features

1Q98 MPT 1.2.0.2 -IRIX SHMEM f90 interfaces

-enhanced fetch-op barrier

-IRIX MPI support for 128 cpus per host

1998 Message Passing Features

2Q98 MPT 1.2.1

-IRIX MPI intra-host CPR

-IRIX MPI miser support

-IRIX MPI 16x128 support

-T3E MPI ROMIO hooks

-PVP SHMEM iget/iput

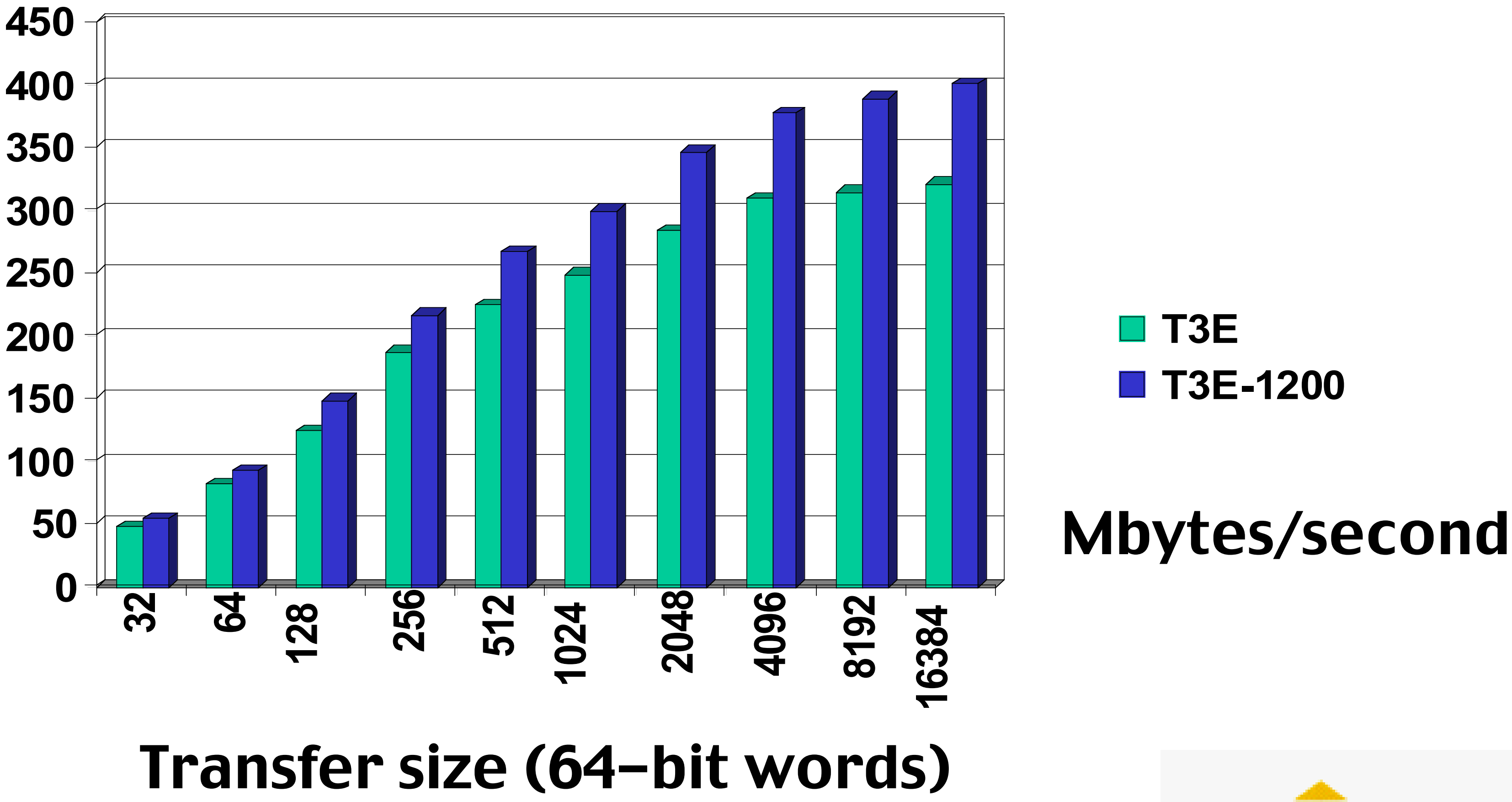
-IRIX SHMEM API expanded

-MPI/SHMEM interoperability

SHMEM Bandwidth on T3E

SiliconGraphics

CRAY USER GROUP Incorporated

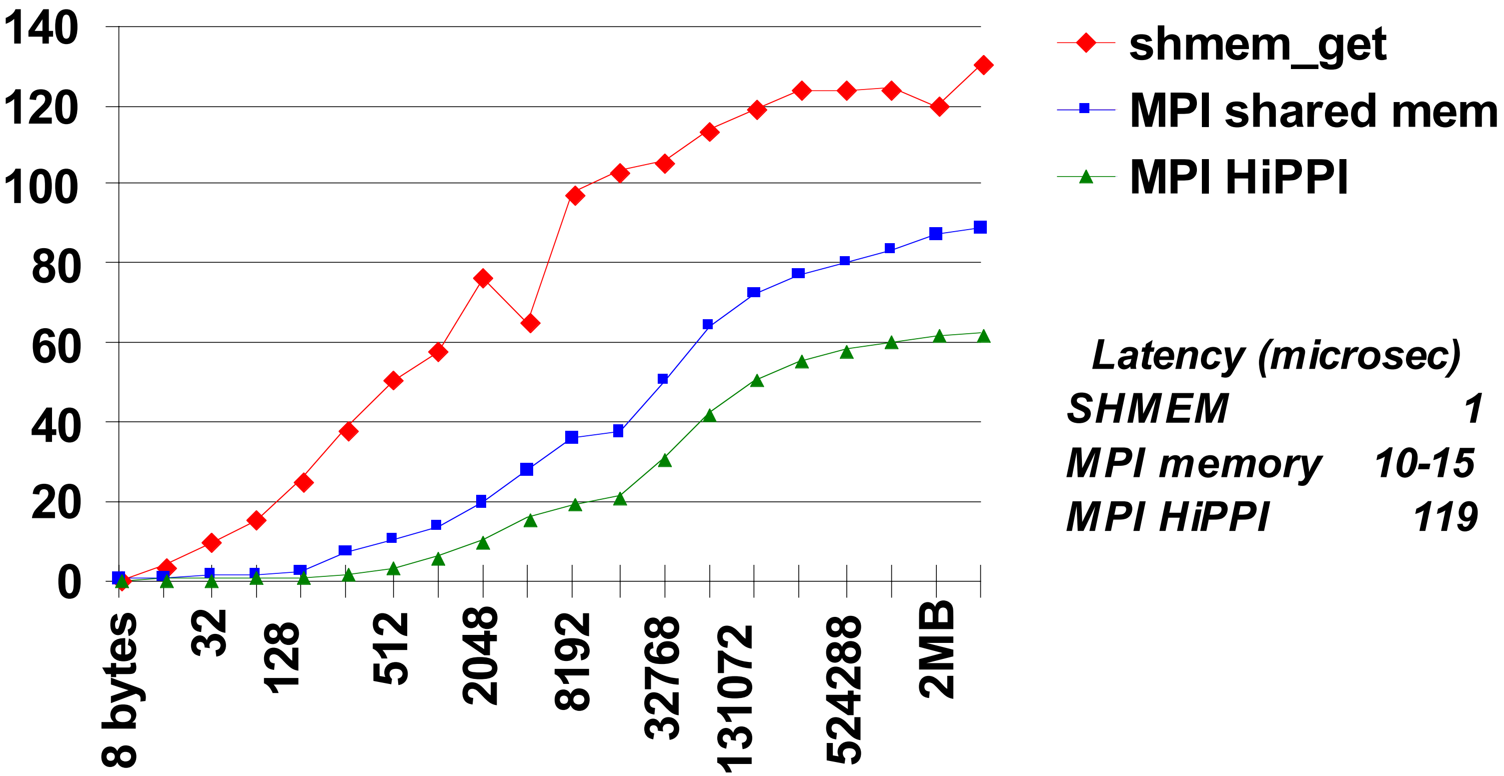


STUTTGART

Message Passing Bandwidth with MPT 1.2

SiliconGraphics

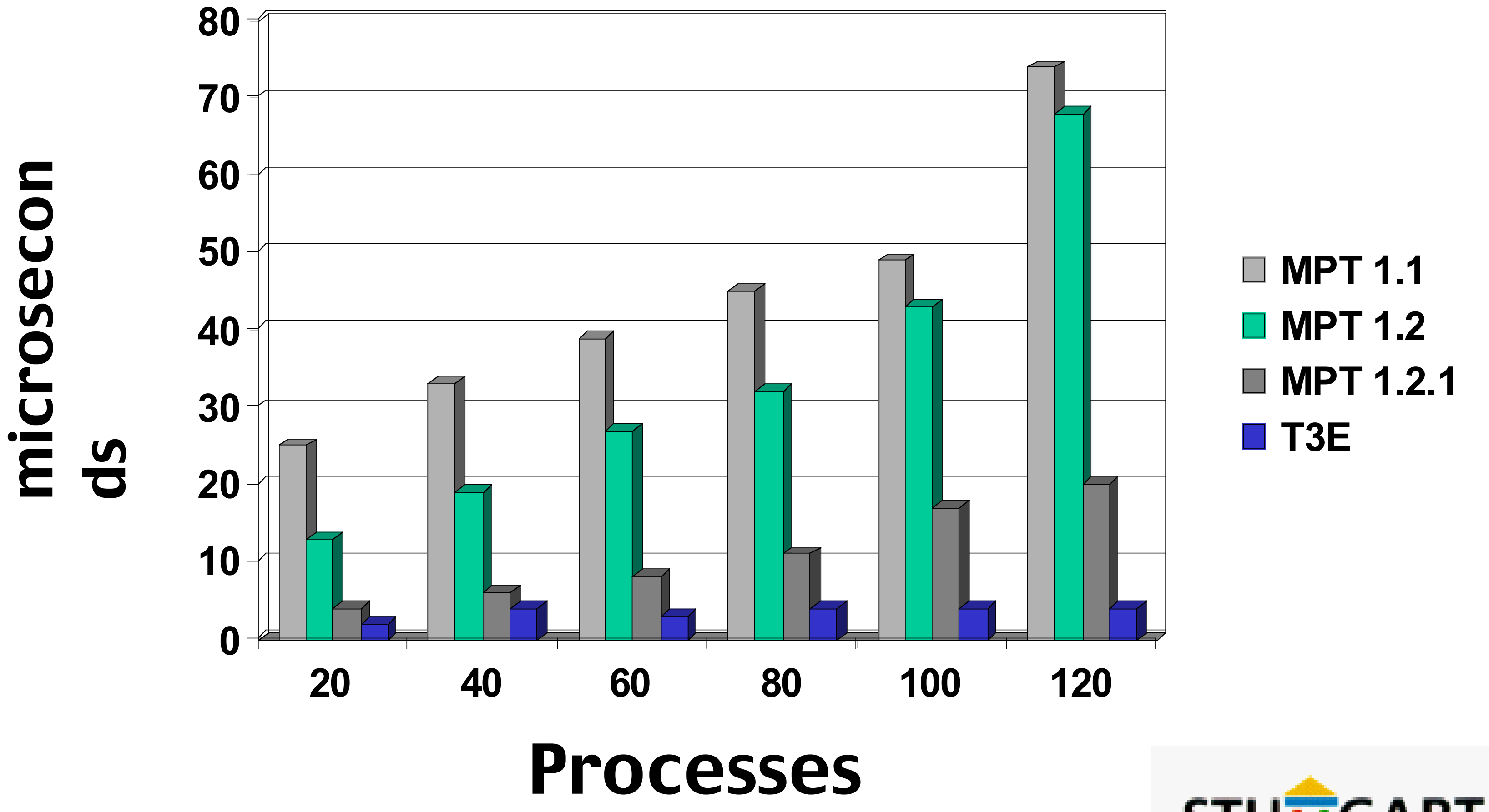
CRAY USER GROUP Incorporated



Latency (microsec)	
SHMEM	*1*
MPI memory	*10-15*
MPI HiPPI	*119*

STUTTGART

Barrier Sync Time on Origin 2000



Message Passing Toolkit Product Roadmap Timeline

2Q98

MPT 1.2.1
(Jun98 MR)
- Single system CPR
- Single system Miser
- MPI/SHMEM interoperability

4Q98

MPT 1.3 (Nov98 MR)
- MPI scaling
- MPI latency reduction
- MPI thread safe phase1
- Dolphin TotalView msg queue
- MPI statistics
- T3E ROMIO support

These are target plans, and not commitments. These plans may change.
For commitments, contact your SGI account representative,
who will in turn contact Peter Rigsbee or Dan Ferber.

Message Passing Toolkit Product Roadmap Timeline

1Q99

MPT 1.3.1 (Mar99 MR)
- T3E MPI-2 one-sided
- MPI cluster CPR phase1
- Native MPI-2 IO
- MPI error handling
- PMPIO support
- MPI collectives performance
- F90 interface blocks
- MPI multi-board messages

MPI scaling
- MPI thread safe phase2
- ASCI cluster
- Supercomputing API in SHMEM

These are target plans, and not commitments. These plans may change.
For commitments, contact your SGI account representative,
who will in turn contact Peter Rigsbee or Dan Ferber.

Message Passing Toolkit Product Roadmap Timeline

3Q99

MPT 1.4 (Sep99 MR)

- MPI performance and scalability
- MPI cluster CPR phase 2
- MPI MPI-2 bindings
- Cluster SHMEM (GSN)
- IRIX MPI-2 one-sided
- MPI over GSN
- MPI for NT

1Q00

MPT 1.5 (Jan2000 MR)

- MPI performance and scalability
- MPI-2 dynamic
- T3E MPI-2 bindings
- Additional ST/GSN support
- SGI Roadmap Support

These are target plans, and not commitments. These plans may change.
For commitments, contact your SGI account representative,
who will in turn contact Peter Rigsbee or Dan Ferber.