MPT and Cluster Software Update for IRIX(TM), UNICOS/mk(TM), and UNICOS(TM) Systems

Karl Feind
Silicon Graphics/Cray Research
655F Lone Oak Drive
Eagan, MN 55121
kaf@cray.com
http://reality.sgi.com/kaf_craypark

ABSTRACT:

SGI/Cray Message Passing Toolkit (MPT) and cluster software continues evolving to add both high availability and scalability functions. This paper briefly summarizes recent and planned changes with the MPT, IRIS FailSafe(TM), and IRISconsole(TM) products. The majority of the talk centers around MPI and SHMEM in MPT 1.2.1. In addition to review of new features, performance of MPI on Origin and SHMEM on Origin and CRAY T3E(TM) will be overviewed.

KEYWORDS:

Message Passing Toolkit, IRIS FailSafe, IRISconsole

Introduction

The Message Passing Toolkit (MPT) and Cluster Products Group supports a variety of software products that support parallel programming and applications support for highly parallel and clustered Silicon Graphics and Cray computer systems. The MPT product is comprised of MPI, SHMEM, and PVM message passing software, used by highly parallel applications to support applications launch and inter-process communication.

IRIS FailSafe provides high-availability (HA) support for applications running on Silicon Graphics servers through automatic fail-over of applications from one server node to another. In the event of a failure, in combination with a RAID or mirrored disk configuration, an IRIS FailSafe cluster provides resilience from any single point of failure and acts as insurance against unplanned outages.

IRISconsole is the third major product supported by our group. The IRISconsole multi-server management system manages multi-server clusters from a single workstation with an easy-to-use interface. Servers are managed and monitored, and their activity logging is supported. IRISconsole performs intelligent actions based upon this information and also allows remote console access.

IRIS FailSafe Software Releases

The most recent IRIS FailSafe release was version 1.2 released in 1997. This release provided fail-over capability for paired server nodes.

The next planned release will be IRIS FailSafe 2.0. A June 1998 beta release is planned, with general product release planned for later in the year. IRIS FailSafe 2.0 will support more generalized fail-over configurations, with the limit of two nodes increasing to eight nodes. This will provide more flexibility for highly available applications because any nodes in the group of up to eight machines can serve as the fail-over system for one or several of the other nodes and applications in the group.

For more information about IRIS FailSafe see `http://www.sgi.com/Products/software/failsafe.html`.

IRISconsole Software Releases

The current IRISconsole release is version 1.2, which was primarily a maintenance release without many new features. The upcoming 1.3 release will include the following features:

- Support for EtherLite 8 and 16 multiplexers
- Enhanced control over logging and alarms
- Support for partitioned arrays
- Year 2000 compliance

For more information about IRISconsole, see `http://www.sgi.com/products/remanufactured/challenge/ti_irisconsole.html`.

Message Passing Software Releases

We support the MPI, PVM, and SHMEM message passing models. All are released in the MPT software release, with the exception of SHMEM on CRAY T3E systems which is delivered as part of CrayLibs and Programming Environment software releases.

Software Releases Containing Message Passing Software

Message Passing Model	**Hardware Platform**	**Release Package**
MPI	all	MPT
PVM	all	MPT
SHMEM	T3E	Programming Environment
SHMEM	MIPS and PVP	MPT

Message passing software is supported on all Sillcon Graphics and Cray computer systems.

For more information about Message Passing Software, see `http://www.sgi.com/Products/software/mpt.html`.

Message Passing Features Released in 1997

The most significant Message Passing release in 1997 was the June MPT 1.1 release. This was the first time the MPT product was released on IRIX systems. The MPI and PVM message passing models had previously been released independently, and bringing them together with SHMEM in the MPT release package on IRIX systems provided more product consistency across IRIX, CRAY T3E, and CRAY PVP systems.

MPT 1.1

Major features released with MPT 1.1 included:

IRIX MPI support for Origin 2000 NUMA capabilities

> IRIX release 6.4 for Origin 2000 contained NUMA-control capabilities which permit message passing applications to request that processes be distributed across the system. MPI was enhanced to use these capabilities.

SHMEM introduced on IRIX systems.

> The first IRIX version of SHMEM message passing was included in MPT 1.1. This version contained an initial portion of the SHMEM API.

XMPI released on IRIX systems.

> XMPI is an MPI program debugger. This tool was originally developed by the LAM project at Ohio State University.

CrayLibs Releases

The T3E version of the SHMEM library was enhanced in late 1997 in CrayLibs 3.0.1, 3.0.1.2, and 3.0.2 revision and update releases.

SHMEM support for CRAY T3E-900 coherent streams.

> Program start-up code in libsma was enhanced to activate data streams in SHMEM programs when hardware enforced coherency of streams with SHMEM communication.

CRAY T3E-1200 **get** bandwidth increased.

> The CRAY T3E-1200 system had enhanced router chips which provided higher peak bandwidth in SHMEM **get** operations. The SHMEM **get** algorithm was enhanced to obtain the peak bandwidth.

`SHMEM_GROUP_CREATE_STRIDED added.`

> This function provides a method to declare a group of SHMEM processes (PEs) which will access a global file.

Message Passing Features Released in 1998

There were two all-platform MPT software releases in early 1998 and a third release planned for June 1998. Releases 1.2, 1.2.0.2, and 1.2.1 provided features chiefly for IRIX platforms with a few features that enhanced T3E and PVP message passing.

MPT 1.2

The following features were added in January 1998 in the MPT 1.2 release.

IRIX and PVP MPI common-sourced.

> The MPT 1.2 (MPI 3.1) IRIX version of the MPI library was ported to CRAY PVP systems and merged with the prior PVP MPI implementation. This common sourcing improved maintainability and brought some new features to the PVP platforms while preserving shared memory message passing optimizations on PVP systems.

Enhanced MPI support for multi-host applications.

> The shared memory communication method was enabled for same-host MPI processes when the application as a whole is running on multiple hosts. Previously, the shared memory communication mode was available on PVP systems only when the application ran entirely on one host.

`MPI_ENVIRONMENT` environment variable.

> `mpirun` and array services set this environment variable for the benefit of `.cshrc` or `.profile` start-up files.

IRIX MPI support for 64 CPUs per host.

Stdin/stdout processing in MPI jobs.

> On PVP and IRIX systems, the `mpirun` offers the `-p` option to add optional prefixes to line written by MPI processes to stdin or stdout.

IRIX MPI support for third party products.

> MPI was modified to allow these third party products to be developed for use with MPI: LSF from Platform Computing, Dolphin TotalView from Dolphin Interconnect Solutions, Inc., and ROMIO from Argonne National Laboratory.

MPI and PVM support for CRAY T3E-900 coherent streams.

> When coherent streams are detected, MPI and PVM enable a faster message passing algorithm.

More SHMEM interfaces implemented on IRIX systems.

Additions included shmalloc, shfree, and a set of SHMEM collective communication routines.

SHMEM supports user-control environment variables on IRIX systems.

Support for the following environment variables were added SMA_DSM_OFF, SMA_DSM_VERBOSE, SMA_DSM_PPM, PAGESIZE_DATA, SMA_SYMMETRIC_SIZE, and SMA_VERSION.

IRIX PVM shared arena enhancements.

Shared arena support stabilized and PVM_RSH environment variable.

MPT 1.2.0.2

The following features were added in March 1998 in the MPT 1.2.0.2 release.

F90 SHMEM interface checking made available on IRIX systems.

The "-auto_use shmem_interface" option with version 7.2.1 or higher of the f90 command activates compile-time argument checking of SHMEM function calls.

shmem_barrier_all performance improved on Origin 2000 systems.

A more effective fetch-op algorithm was used.

MPT 1.2.1

The following features are added in the MPT 1.2.1 release.

Checkpoint/restart capability for IRIX MPI.

MPI checkpoint/restart capability is available for MPI applications running within a single host and consisting of a single executable.

MPI Miser support.

Miser is a queuing tool which can dedicate a set of CPUs on an Origin 2000 system to a particular message passing job.

Support MPI on 16 x 128 Origin 2000 clusters.

T3E MPI support for ROMIO.

MPI_Type_get_contents, MPI_Type_get_envelope and MPI-2 error codes are added.

SHMEM **iput/iget** functions added on PVP.

IRIX SHMEM implementation expanded.

`shmem_set_lock` and `shmem_clear_lock` are added.

MPI and SHMEM interoperability on IRIX.

MPI programs can now use SHMEM message passing to communicate with MPI processes on the same host. SHMEM PE numbers are equal to MPI process rank in `MPI_COMM_WORLD`.

Message Passing Library Performance

One of the requirements of the message passing libraries on Silicon Graphics and Cray systems is to provide inter-process communication with lowest possible latency and highest possible bandwidth. This fundamental requirement provides a motivation to continue adding performance enhancements to the message passing libraries.

The enhanced bandwidth for point-to-point communication on T3E-1200 systems was achieved by enhancing SHMEM **get** algorithm. The following graph shows the effective bandwidth obtained over varying transfer sizes on the CRAY T3E and the CRAY T3E-1200.

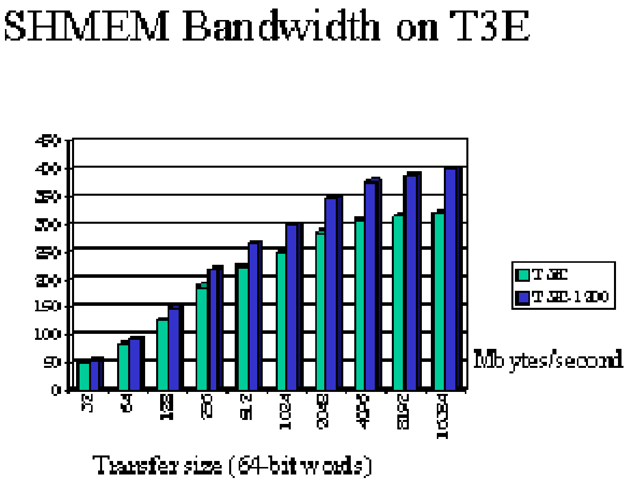

Figure 1: SHMEM_GET effective bandwidth

A common practice in message passing programs that use SHMEM is to use barrier synchronization to separate computational and communication phases. The barrier synchronization must be as fast as possible for these programs to get best performance. The following graph shows how Origin SHMEM barrier synchronization time has been continually improved in the MPT releases the past year

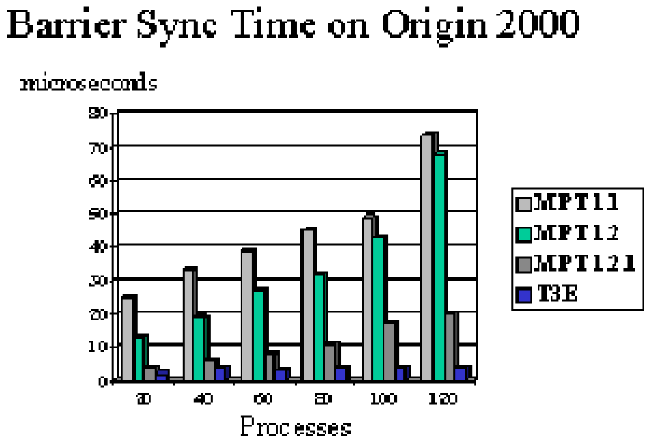

Figure 2: `shmem_barrier_all` synchronization time

The following graphic compares bandwidths and latencies for MPI and SHMEM message passing on Origin 2000 systems.

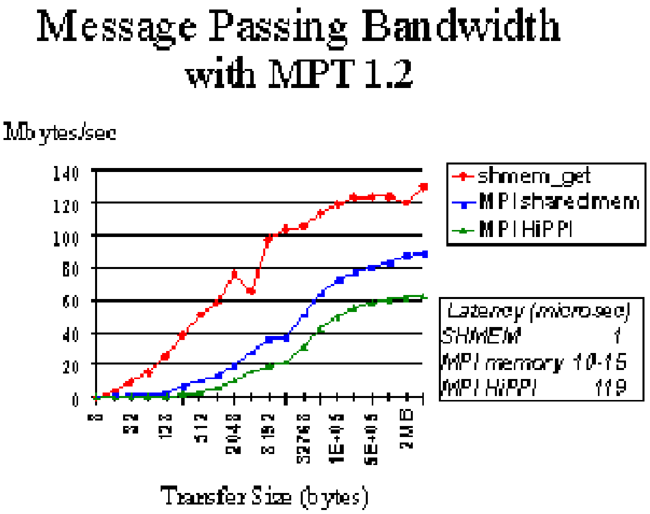

Figure 3: Message Passing Bandwidth on Origin 2000 systems.

MPT Roadmap

The MPT and cluster group will be enhancing message passing software in the coming year to provide many customer-requested features. The chart in this section outlines the planned features for the coming years. These are not commitments. These are target plans that might change.

4Q98	1Q99	3Q99	1Q00
● MPT 1.3 (Nov 98 MR) ● MPI scaling ● MPI latency reduction ● MPI thread-safe phase 1 ● Dolphin TotalView msg queue ● MPI statistics ● T3E ROMIO support	● MPT 1.3.1 (Mar 99 MR) ● T3E MPI-2 one-sided ● MPI cluster CPR phase1 ● Native MPI-2 IO ● MPI error handling ● PMPIO support ● MPI collective performance ● F90 interface blocks ● MPI multi-board messages	● MPT 1.4 (Sep 99 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	● MPT 1.5 (Jan 2000 MR) ● MPI performance and scalability ● MPI-2 dynamic ● T3E MPI-2 bindings ● Additional ST/GSN support ● SGI Roadmap Support

Conclusion

In the past year, MPT and cluster software has been enhanced to better support all Silicon Graphics and Cray hardware platforms. The most significant effort has been made recently for the Origin 2000 system to support increasing CPU counts and to support clustered Origin 2000 configurations. Future enhancements to message passing software will seek to balance the needs of large clustered and non-clustered Silicon Graphics and Cray systems.

Author Biography

Karl Feind is a Core Design Engineer in the MPT and Cluster Products group, where his primary responsibilities are support of Distributed Shared Memory (SHMEM) data passing and Message Passing Interface (MPI). In the past, Karl’s responsibilities have included the Cray Fortran I/O and Flexible File I/O (FFIO) libraries.

Home page: `http://reality.sgi.com/kaf_craypark`
E-mail address: `kaf@cray.com`