Sun HPC ClusterTools™ 4 Product Notes

Sun Microsystems, Inc.
901 San Antonio Road
Palo Alto, CA 94303-4900 U.S.A.
650-960-1300

Part No. 816-0647-10
August 2001, Revision A

Send comments about this document to: docfeedback@Sun.com

Please
Recycle

Adobe PostScript

Contents

Sun HPC ClusterTools 4 Product Notes

This document describes late-breaking news about the Sun HPC ClusterTools 4 software. The information is organized into the following sections:

Advisory Notes

Administrators may want to reset `/etc/system` *parameters.*

The installation process for the Sun HPC ClusterTools 4 software sets the following shared memory parameters in `/etc/system` to enable MPI communication using the RSM protocol:

- `shmsys:shminfo_shmseg`
- `shmsys:shminfo_shmmni`
- `shmsys:shminfo_shmmax`

When a cluster is not RSM-enabled, the administrator may want to change these values. See the RSM setup discussion in the *Sun HPC ClusterTools 4 Administrator's Guide* for details.

MPI jobs may time out using PFS on large clusters.

MPI jobs on large clusters that use PFS with the TCP protocol may time out before the PFS I/O daemon can establish the necessary connections or service data requests. If such timeouts occur, the administrator may need to tune the following parameters in `/etc/system`:

- `tcp_ip_abort_cinterval`
- `tcp_time_wait_interval`
- `tcp_rexmit_interval_max`
- `tcp_ip_abort_interval`
- `tcp_conn_req_max_q`
- `tcp_conn_req_max_q0`

For details, see the discussion of system parameters on TCP clusters in the *Sun HPC ClusterTools 4 Administrator's Guide*.

Major New Features

The major new features of the Sun HPC ClusterTools 4 software include:

- Scalability to 2048 processes per MPI job (up to 64 nodes)
- Open MPI transport-layer architecture, available to third parties
- Enhanced CRE security
- Improved error logging and core file handling in CRE
- Prism™ environment support for dynamically spawned MPI processes
- Prism environment performance improvements
- Prism environment support for `dlopen`
- IPv6 support
- Sun S3L Version 4.0
  - Additional solvers and utilities for sparse systems
  - Support for linear programming and optimization
  - Function for calculating equity option pricing
  - Additional transforms (Walsh, sine, cosine)
  - Support for additional ScaLAPACK APIs (Cholesky, QR)
  - Optimizations for UltraSPARC™ III

Related Software

The Sun HPC ClusterTools 4 software works with the following versions of related software:

- Solaris 8 4/01 (maintenance update 4)

  The following Solaris patches can be applied to the Solaris 8 operating environment in lieu of the Solaris maintenance update 4:

  - 109764-02 or greater
  - 109472-04 or greater
- Forte™ 6, Forte 6 update 1, and Forte 6 update 2 Development software
- Load Sharing Facility (LSF) suite from Platform Computing, Version 4.0.1, which requires the "Sun HPC Integration Package" from Platform Computing.

  LSF Version 4.0.1 contains a regression that requires installation of the file `sbatchd4.0_sparc-sol2.Z` to fix. This file is available in the directory `/lsf/support/4.0.1/os/sparc-sol2/` on the server ftp.platform.com.
- Java™ Runtime Environment (JRE) 1.3.0 for using the Sun HPC ClusterTools installation tool, `install_gui`.

  If installation is performed using the command-line interface, this requirement is waived.

Deprecated Commands

- Future releases will not support the use of file descriptors 3 and above in `mprun -I` expressions.
- Future releases will not support the use of `mprun -Mf` to specify rankmaps. Use the `-m` or `-l` options instead.
- Future releases will not support the use of `-J` and `-j` options to the `bsub -sunhpc` command for colocating jobs in the LSF environment.

Outstanding Bugs

This section highlights some of the outstanding bugs for the following Sun HPC ClusterTools 4 software components:

Installation

Bug - When one node is removed from an NFS cluster, the `SUNWtnfv` package is removed from the remaining nodes. [4474365]

The recovery action for this is to install the `SUNWtnfv` package from the CD-ROM with `pkgadd` on all remaining NFS client nodes. For example:

```
# pkgadd -d /cdrom/hpc_4_0_ct/Product SUNWtnfv
```

Note that only the Prism environment requires `SUNWtnfv`. Consequently, it needs to be installed only on the node operating as Prism host node.

Bug - Removal of Sun HPC ClusterTools software fails when `hpc_remove` is invoked from `/opt/SUNWhpc/HPC4.0/bin/`. [4475377]

To prevent this from happening, initiate the removal process from the CD-ROM instead.

When `hpc_remove` is initiated from the `../HPC4.0/bin/` directory, it will stop when it encounters a `SUNtnfv` package. This will result in one of the following conditions, depending on whether the software is installed on a single node or on multiple nodes:

- Single node – The `SUNtnfv` package will be the only remaining package. All other Sun HPC ClusterTools packages will have been removed. Use `pkgrm` to remove the `SUNtnfv` package.
- Multiple nodes – The `SUNtnfv` package will remain on the node where the removal process began and no packages will have been removed from the other nodes. Use `hpc_remove –f` from the CD-ROM to forcibly remove all remaining Sun HPC ClusterTools packages.

Bug - When installing ClusterTools 4 on a cluster larger than eight nodes already containing ClusterTools 3.1 and LSF, only the first group of nodes receive a copy of the `hpc.conf` file. [4475946]

Ordinarily, when HPC ClusterTools software is installed on clusters with more than eight nodes, the installation process proceeds in eight-node groups (referred to as *chunking*). However, if the nodes already contain HPC ClusterTools 3.1 and LSF, the `hpc.conf` file does not get copied to the second group of nodes.

Please note that this problem also arises when you use the `–parallel:`*n* flag on the `install_gui` command line.

The recovery for this problem is to manually copy the *INSTALL_LOC*`/SUNWhpc/HPC4.0/conf/hpc.conf` file from any node in the first group to each subsequent group of nodes and then restart LSF.

Bug - Removing nodes from an NFS cluster can also remove software from the NFS server [4477388]

To work around this restriction and remove nodes safely, do the following steps.

1. **On each node to be removed, execute these commands.**

```
# /etc/init.d/sunhpc.spind stop

# /etc/init.d/sunhpc.cre_node stop

# pkgrm SUNWpfsx SUNWpfsrt SUNWmpirt SUNWcrert
```

2. **On the CRE master node, execute this command.**

```
# mpadmin –c "node delete removed_node"
```

3. **Update the** `hpc_config` **file, deleting the names of removed nodes from the** `NODES` **list.**

4. **Restart the software.**

```
# hpc_reconfigure –a 4.0 –c hpc_config
```

Bug - When using `telnet` to remove nodes from an NFS cluster in which the NFS server is not part of the cluster, the NFS server can hang. [4461127]

To avoid this problem, use `rsh` to remove NFS client nodes.

Or, do the following steps to bypass the problem:

1. **Deactivate ClusterTools 4 software on the running cluster.**

```
# hpc_reconfigure -d 4.0 -c config_dir
```

2. **Edit the** `hpc_config` **file, deleting the desired subset of nodes from the** `NODES` **list.**

3. **Activate ClusterTools 4 on the running cluster.**

```
# hpc_reconfigure -a 4.0 -c config_dir
```

This action restarts the cluster without the removed nodes.

4. **Use** `mpadmin` **to update the partition descriptions to reflect node removal. For information about** `mpadmin`**, refer to the** *Sun HPC ClusterTools 4 Administrator's Guide***.**

Bug - Re-installation after repeated activation and deactivation can cause corruption of the `SUNWpfsx` package. [4474081]

Re-installation can corrupt already-installed packages and leave the installation in a state where activation no longer is possible.

To recover from this problem, perform these steps:

1. **Use** `pkgadd` **to reinstall** `SUNWpfsx` **from the CDROM on all nodes.**

```
# pkgadd -d /cdrom/hpc_4_0_ct/Product SUNWpfsx
```

2. **Remove all remaining** `SYNC` **files from the configuration directory.**

```
# cd config_dir

# rm SYNC*
```

3. **Proceed to activation or deactivation with the** `hpc_config` **file generated during the re-install process.**

Bug - The activation stage of an NFS installation can fail if the NFS server is external to the target cluster or if the cluster contains a previously installed version of HPC ClusterTools 3.1 software. [4471779]

An HPC ClusterTools 3.1 postinstall script is called when the HPC ClusterTools 4 software is activated, which causes the activation process to fail with a reconfiguration error being reported by all NFS clients.

To recover from this problem, take the following steps:

1. **Issue the mount command with the suggested arguments.**

```
# mount –F nfs –o rw,acdirmax=10,acdirmin=5,acregmax=10, \
acregmin=3,actimeo=10 nfs_server:$INSTALL_LOC_SERVER/SUNWhpc \
opt/SUNWhpc
```

where:

nfs_server is the hostname of the NFS server.

`$INSTALL_LOC_SERVER/SUNWhpc` is the location of the software installed on the NFS server.

`/opt/SUNWhpc` is the mount point where the NFS clients nodes mount the HPC ClusterTools software to be installed on the NFS server.

2. **Go to the directory containing the `hpc_config` file and remove the synchronization files.**

```
# cd config_dir

# rm SYNC*
```

3. **Deactivate the NFS cluster.**

```
# hpc_reconfigure –f –c config_dir –d 4.0
```

4. **Activate the NFS cluster.**

```
# hpc_reconfigure –c config_dir –a 4.0
```

If the `hpc_config` file has a different name, specify the full path in place of *config_dir.*

Bug - Using a terminal concentrator to install the software will cause the installation to fail. [4474730]

Extra information about the terminal concentrator and ports is provided in the installation, which results in a mismatch with the list of nodes contained in the configuration definition.

To avoid this problem, do not install the HPC ClusterTools software with a terminal concentrator.

Bug - HPC clusters using LSF do not start properly on initial install. [4474263]

To work around this bug, run `lsfrestart` on the cluster.

Bug - When running `hpc_cluster_tool_setup` on a node that is not the first node in the `NODES` list, as is specified in the `hpc_config` file, `hpc_cluster_tool_setup` fails to find other nodes in the `NODES` list. [4476600]

Two workarounds are available.

- First workaround:

  Install the Sun HPC ClusterTools 4 Cluster Console Manager software on any desired node by issuing the following commands:

  On nodes running 64-bit Solaris:

```
# pkgadd –d /cdrom/hpc_4_0_ct/Product/Sol_2.7/ SUNWccon \
SUNWscch
```

  On nodes running 32-bit Solaris:

```
# pkgadd –d /cdrom/hpc_4_0_ct/Product/Sol_2.6/ SUNWccon \
SUNWscch
```

- Second workaround:

  The alternative workaround is to edit the `hpc_config` file so that the node on which you want to install Cluster Console Manager is the first of the `NODES` list. Then, rerun `hpc_cluster_tool_setup`. After successfully installing the `SUNWccon` and `SUNWscch` packages, change the `NODES` list back to the original order.

Cluster Runtime Environment (CRE)

Bug - CRE may not survive Kerberos ticket expiration. [4291039]

Kerberos tickets that exceed their `max_life` need to be renewed. If CRE is started with a given ticket, CRE becomes unstable when that ticket's `max_life` expires, regardless of the renewal state of the ticket.

To work around, restart the CRE daemons before the Kerberos ticket expires.

Parallel File System (PFS)

Bug - The PFS I/O daemon does not permit 100% use of the file system's disk space. [4434579]

If a PFS file system becomes 100% full, the I/O daemon and/or the MPI job may hang or crash. The workaround is to monitor the file system's disk space usage (using `df`) and avoid using more than 99% of the space.

Prism Environment

Bug - The Prism environment crashes on the command `print` *x* `on cycle` when the pset is `cycle`. [4348951]

To work around this problem, change the scope from the entire `cycle` pset to a specific process in the the pset, before invoking the `print` *x* `on cycle` command.

This example illustrates the workaround:

1. **Start the Prism Environment.**

```
% prism –np 2 program-name
(prism all) stop at 6
(prism all) run
(prism all) pset cycle
(prism 0,1)
```

   The `pset cycle` command makes the Prism prompt appear as `(prism 0,1)` where process ranks 0 and 1 are present (since this is an `np` of 2 run).

2. **Set the current pset to `0`.**

```
(prism 0,1) pset 0
(prism 0)
```

   This command changes the prompt to `(prism 0)`.

3. **You can then print the value of a variable in another window, using the** `on` *windowname* **syntax.**

```
(prism 0) print i on cycle
```

4. **Change the current pset (and prompt) back to `all`:.**

```
(prism 0) pset all
(prism all)
```

   This command returns the session to its original scope. You can then repeat the workaround, examining and printing the values of processes in the `cycle` pset.

5. **Change the current pset (and prompt) back to the cycle group:**

```
(prism all) pset cycle
```

6. **You can then use the** `cycle` **command to move through the** `cycle` **pset:**

```
(prism 0,1) cycle
```

Bug - Running and then immediately interrupting an MPI job in the Prism environment can cause an assertion failure or repeated state warnings. [4400183]

This bug arises in MP mode, when communication of a symbol name between host and node executables of the Prism environment is unable to resolve the symbol uniquely. This occurs when a library contains multiple files with the same name, but with different symbols in each file. If the symbol lookup takes place in the wrong file, the error results. An immediate use of the `interrupt` command finds the program in the runtime linker, which has such files.

The error message reads "`symbol lookup` *symbol_name* `failed`".

At this point, an attempt to rerun the program currently loaded in the Prism environment results in the error:

`unresolved_list == NULL`

To avoid encountering this problem, wait a few seconds before interrupting a program that you have just run.

Bug - The Prism environment issues an unhelpful error message when the TNF tracing data file size limit has been exceeded. [4478713]

The message reads:

```
Loading tracefiles...
Maximum file size reached - some events have been lost.
```

Use the `tnffile` command to increase the size of the file used to hold the TNF tracing data. For information about `tnffile`, refer to the Prism manuals and online help.

Documentation

Bug - Use of the - option is incorrectly documented in the `mprun` man page and is not documented in *Sun HPC ClusterTools 4 User's Guide*. [4471810]

To start a program whose name begins with a dash (for example, `-myprogram`), you must precede the program name with a dash, and include a space between the dashes. The `mprun` man page does not point out the need for the extra space. Here are two examples:

```
% mprun –np 2 myprogram
```

```
% mprun –np 2 – –myprogram
```

The first example uses a program name that does not begin with a dash (`myprogram`). The second example uses a program name that begins with a dash (`-myprogram`), so you must precede the name with an additional dash, and leave a space between the dashes.

Bug - The `mprun` man page and the *Sun HPC ClusterTools 4 User's Guide* do not mention that to launch a program as a different user (`-U` option), you must be superuser. [4471810]

The `mprun -U` option enables you to start a program as a different user than the one currently logged on to the cluster. Here is the syntax:

```
# mprun –U username program-name

# mprun –U userid program-name
```

You can identify the user with either a *username* or *userid*. To do so, however, you must be superuser, unless you are the user whose name or ID you are specifying. The documentation does not point out this requirement.

Bug - The *MPI 5 Reference Manual* does not mention that you can have a maximum of 16 ports open simultaneously. [4475295]

You can open a maximum of 16 ports for client-server communication with the `MPI_Open_port()` function, listed in Appendix A of the *MPI 5 Reference Manual*. If you try to open a 17th port without first closing one of the open ports, you will get an error message. The documentation does not point out this limitation.

Bug - The command `man pfsstop` fails. [4467702]

The `pfsstop` command is described on the same man page as `pfsstart`. Its man page can be accessed by means of `man pfsstart`.