Sun HPC ClusterTools™ 5 Software Release Notes

Sun Microsystems, Inc.
4150 Network Circle
Santa Clara, CA 95054 U.S.A.
650-960-1300

Part No. 817-0081-10
February 2003, Revision A

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

Copyright 2003 Sun Microsystems, Inc., 4150 Network Circle, Santa Clara, California 95054, U.S.A. All rights reserved.

Sun Microsystems, Inc. has intellectual property rights relating to technology embodied in the product that is described in this document. In particular, and without limitation, these intellectual property rights may include one or more of the U.S. patents listed at http: / / www.sun.com / patents and one or more additional patents or pending patent applications in the U.S. and in other countries.

This document and the product to which it pertains are distributed under licenses restricting their use, copying, distribution, and decompilation. No part of the product or of this document may be reproduced in any form by any means without prior written authorization of Sun and its licensors, if any.

Third-party software, including font technology, is copyrighted and licensed from Sun suppliers.

Parts of the product may be derived from Berkeley BSD systems, licensed from the University of California. UNIX is a registered trademark in the U.S. and in other countries, exclusively licensed through X / Open Company, Ltd.

Sun, Sun Microsystems, the Sun logo, AnswerBook2, docs.sun.com, Solaris, Sun HPC ClusterTools, Prism, Forte, Sun Performance Library, and UltraSPARC are trademarks or registered trademarks of Sun Microsystems, Inc. in the U.S. and in other countries.

All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. in the U.S. and in other countries. Products bearing SPARC trademarks are based upon an architecture developed by Sun Microsystems, Inc.

The OPEN LOOK and Sun™ Graphical User Interface was developed by Sun Microsystems, Inc. for its users and licensees. Sun acknowledges the pioneering efforts of Xerox in researching and developing the concept of visual or graphical user interfaces for the computer industry. Sun holds a non-exclusive license from Xerox to the Xerox Graphical User Interface, which license also covers Sun's licensees who implement OPEN LOOK GUIs and otherwise comply with Sun's written license agreements.

U.S. Government Rights—Commercial use. Government users are subject to the Sun Microsystems, Inc. standard license agreement and applicable provisions of the FAR and its supplements.

DOCUMENTATION IS PROVIDED "AS IS" AND ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT SUCH DISCLAIMERS ARE HELD TO BE LEGALLY INVALID.

Copyright 2003 Sun Microsystems, Inc., 4150 Network Circle, Santa Clara, California 95054, Etats-Unis. Tous droits réservés.

Sun Microsystems, Inc. a les droits de propriété intellectuels relatants à la technologie incorporée dans le produit qui est décrit dans ce document. En particulier, et sans la limitation, ces droits de propriété intellectuels peuvent inclure un ou plus des brevets américains énumérés à http: / / www.sun.com / patents et un ou les brevets plus supplémentaires ou les applications de brevet en attente dans les Etats-Unis et dans les autres pays.

Ce produit ou document est protégé par un copyright et distribué avec des licences qui en restreignent l'utilisation, la copie, la distribution, et la décompilation. Aucune partie de ce produit ou document ne peut être reproduite sous aucune forme, parquelque moyen que ce soit, sans l'autorisation préalable et écrite de Sun et de ses bailleurs de licence, s'il y ena.

Le logiciel détenu par des tiers, et qui comprend la technologie relative aux polices de caractères, est protégé par un copyright et licencié par des fournisseurs de Sun.

Des parties de ce produit pourront être dérivées des systèmes Berkeley BSD licenciés par l'Université de Californie. UNIX est une marque déposée aux Etats-Unis et dans d'autres pays et licenciée exclusivement par X / Open Company, Ltd.

Sun, Sun Microsystems, le logo Sun, AnswerBook2, docs.sun.com, Solaris, Sun HPC ClusterTools, Prism, Forte, Sun Performance Library, and UltraSPARC sont des marques de fabrique ou des marques déposées de Sun Microsystems, Inc. aux Etats-Unis et dans d'autres pays.

Toutes les marques SPARC sont utilisées sous licence et sont des marques de fabrique ou des marques déposées de SPARC International, Inc. aux Etats-Unis et dans d'autres pays. Les produits protant les marques SPARC sont basés sur une architecture développée par Sun Microsystems, Inc.

L'interface d'utilisation graphique OPEN LOOK et Sun™ a été développée par Sun Microsystems, Inc. pour ses utilisateurs et licenciés. Sun reconnaît les efforts de pionniers de Xerox pour la recherche et le développment du concept des interfaces d'utilisation visuelle ou graphique pour l'industrie de l'informatique. Sun détient une license non exclusive do Xerox sur l'interface d'utilisation graphique Xerox, cette licence couvrant également les licenciées de Sun qui mettent en place l'interface d 'utilisation graphique OPEN LOOK et qui en outre se conforment aux licences écrites de Sun.

LA DOCUMENTATION EST FOURNIE "EN L'ÉTAT" ET TOUTES AUTRES CONDITIONS, DECLARATIONS ET GARANTIES EXPRESSES OU TACITES SONT FORMELLEMENT EXCLUES, DANS LA MESURE AUTORISEE PAR LA LOI APPLICABLE, Y COMPRIS NOTAMMENT TOUTE GARANTIE IMPLICITE RELATIVE A LA QUALITE MARCHANDE, A L'APTITUDE A UNE UTILISATION PARTICULIERE OU A L'ABSENCE DE CONTREFAÇON.

Please
Recycle

Adobe PostScript

Contents

Major New Features 1

Product Migration 2

TNF EOL 2

PFS EOL 2

Transferring Files From HPC ClusterTools 4 Software's Parallel File System 2

Attempting To Use PFS Utilities in an HPC ClusterTools 5 Software Installation 3

Related Software 4

Outstanding Bugs 4

MPI 5

SCSL 7

CRE 8

Performance Issues 9

Sun HPC ClusterTools 5 Software Release Notes

This document describes late-breaking news about the Sun HPC ClusterTools 5 software. The information is organized into the following sections:

Section
Major New Features
Product Migration
Related Software
Outstanding Bugs
Performance Issues

Major New Features

The major new features of the Sun HPC ClusterTools 5 software include:

- Scalability to clusters of up to 256 computational nodes (running up to 2048 processes per MPI job)
- New and improved installation procedures, supplying new graphical and command line user interfaces
- Full one-sided MPI communication features in the Sun MPI Library
- Full MPI-2 compliance in the Sun MPI Library
- Sun CRE integration with distributed resource management frameworks: Sun Grid Engine, Load Sharing Facility from Platform Computing, and PBS from Veridian

- MPProf profiling tool for identifying and correcting performance problems in message-passing programs
- Debugging of Sun MPI programs supported by the TotalView debugger from Etnus.

Product Migration

TNF EOL

TNF (Trace Normal Form) probes and the `tnfview` trace file viewer are no longer actively supported within Sun and have been eliminated in ClusterTools 5 software. An alternative solution for tracing MPI calls in applications is available in the Sun™ ONE Studio 7 (formerly Forte™ Developer) Performance Analyzer.

The Performance Analyzer GUI and the IDE are part of the Sun™ONE Studio 4 Enterprise Edition for Java™. The GUI version of Performance Analyzer now includes a timeline viewer.

Case studies of profiling MPI applications with Performance Analyzer can be found in the Sun HPC ClusterTools Performance Guide.

For information about Sun ONE program performance tools, see the Program Performance Analysis Tools (816-2548-10) manual. See also the `collect(1)`, `collector(1)`, `libcollector(3)`, `analyzer(1)`, and `er_print(1)` man pages and the Performance Analyzer online help.

PFS EOL

The Parallel File System (PFS) is no longer actively supported within Sun and has been eliminated in ClusterTools 5.

Transferring Files From HPC ClusterTools 4 Software's Parallel File System

The procedure for transferring files from PFS to another file system is very straightforward. The following example assumes that PFS is mounted at `/pfs`.

1. **Change directory to the directory above the PFS mount point**

   For example,

   ```
   % cd /
   ```

2. **Archive your files**

   For example,

   ```
   % tar cvf pfs.tar pfs
   ```

3. **Copy your files to your target filesystem**

   Copy your files to the file system you want to use, for example, *ufs*.

   ```
   % cp pfs.tar /ufs/ufs.tar
   ```

4. **Unarchive your files**

   ```
   % cd /ufs
   ```

   Then, reverse the process you used in archiving your files.

   ```
   % tar xvf ufs.tar
   ```

   Your files appear under a subdirectory of /ufs named pfs/

   ```
   % ls
     pfs/
   ```

Note – When migrating from an HPC ClusterTools 4 software installation to an HPC ClusterTools 5 software installation, any PFS-related sections in the HPC ClusterTools 4 software's `hpc.conf` file are automatically commented out in the HPC ClusterTools 5 software's `hpc.conf` file

---

Attempting To Use PFS Utilities in an HPC ClusterTools 5 Software Installation

PFS utilities have no effect in HPC ClusterTools 5. Their use merely generates a warning. For example,

```
Commandname: This command is not supported.
Sun PFS is no longer provided as part of Sun HPC Cluster Tools.
```

Related Software

The Sun HPC ClusterTools 5 software works with the following versions of related software:

- Solaris 8 [2/02] (maintenance update 7) or any subsequent Solaris release that supports Sun HPC ClusterTools 5 software.
- Forte 6 update 2, and Sun ONE Studio 7 Compiler Collection for C, C++, and Fortran compilers (formerly Forte Development 7 software)
- Distributed resource management frameworks operating under integration with Sun CRE:
  - Sun Grid Engine SGE Version 5.3 and Sun Grid Engine Enterprise Edition SGEEE Version 5.3.
  - Load Sharing Facility Version 4.x of Platform Computing.
  - Portable Batch System (PBS) PBS Pro 5.x.x of Veridian.
- Java Runtime Environment (JRE) 1.2.0 (or compatible) for using the Sun HPC ClusterTools installation tool's graphic interface.
- TotalView debugger supports debugging Sun MPI applications. Compatibility limitations may exist, please see www.etnus.com for compilers that TotalView supports.

Outstanding Bugs

This section highlights some of the outstanding bugs for the following Sun HPC ClusterTools 5 software components:

Components
"MPI" on page 5
"SCSL" on page 7
"CRE" on page 8

Note – The heading of each bug description includes the bug's Bugtraq number, within brackets.

MPI

Bug - Errors can lead to deadlock when using the `MPI::ERRORS_THROW_EXCEPTIONS` error handler [4425209]

To work around this problem, define and use a new error handler (with `MPI::Comm::Create_errhandler` and `MPI::Comm::Set_errhandler`, respectively) to do some combination of the following:

- Print out an error message
- Spin wait at point of error so a debugger can attach to process
- Dump core

Bug - `MPI_Send` latency increases in the presence of window [4782790]

This problem affects one-sided Sun MPI communications.

To work around this problem,

```
% setenv MPI_RSM_PUTSIZE 0
```

Note – This workaround has the adverse side effect of increasing `MPI_Put` latency.

Bug - Lock files left over in `ufs /tmp` prevent `hpc_rsmd` start on boot [4812693]

When the `hpc_rsmd` starts up, it creates a lockfile to prevent other instances of hpc_rsmd from running concurrently. Subsequent attempts to start `hpc_rsmd` fail when they find `/tmp/.hpc_rsmd_lock`.

When `hpc_rsmd` exits normally, it removes the lock file. If a system with a running `hpc_rsmd` crashes, the lock file is left over in `/tmp`.

On systems with /tmp mounted on volatile file systems this is not a problem since /tmp is wiped clean on each boot. However, if /tmp is mounted on a nonvolatile filesystem such as ufs, the lockfile persists. It can be removed by running

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

Bug - MPI uses too much RSM buffer space at high numbers of processes [4815821]

The default environment variable settings for the amount of RSM buffer space allocated do not scale well with the numbers of processes (np). For Sun Fire 15K clusters with three or more nodes, multiple gigabytes of RSM memory are consumed per node. This can exceed the amount of memory that can be exported by the Sun™ Fire Link driver, and cause the MPI job to fail.

To control this problem, reduce RSM memory consumption using Sun MPI environment variables. The simplest approach is to set MPI_RSM_CPOOLSIZE as shown in the following example,

```
MPI_RSM_CPOOLSIZE=131072
```

An alternative is to set both MPI_RSM_CPOOLSIZE and MPI_RSM_SBPOOLSIZE as follows:

```
MPI_RSM_SBPOOLSIZE=4194304
MPI_RSM_CPOOLSIZE=131072
```

If deadlock results, setting MPI_POLLALL=1 (the default) may help.

You can run an MPI job that requests more RSM buffer memory than is available; perhaps because you have asked for more than the default, or because jobs belonging to other users are currently running and using some of this memory. In this case, your MPI job will wait for memory to become available. It is possible that enough memory will never become available. You must decide whether you have waited too long and terminate the mprun command using Ctrl-C.

SCSL

Bug - SCSL `configure` script needs additional option for PBS Pro [4802380]

When configuring `sunhpc` makefiles for SCSL builds of ClusterTools 5 software, the `configure` script requires the use of a new option if PBS Pro is to be used in close integration with CRE. Specify the PBS Pro installation location as an argument to the `–pbspro` option. For example,

`# ./configure … –pbspro` *PBSPRO_PATH …*

CRE

Bug - Node failure can cause stale job entries [4692994]

If a node crashes while an MPI program is running, CRE does not remove the job entry from its database, so `mpps` continues to show the job indefinitely, often in states such as coring or exiting.

To delete these stale jobs from the database, `su` to `root` and issue this command:

```
# mpkill -C
```

Performance Issues

This section highlights those bugs that have important implications for performance.

Bug - `MPI_Alltoall` with large `SHM_SBPOOLSIZE` [4790032]

The Sun MPI environment variables `MPI_SHM_SBPOOLSIZE` and `MPI_SHM_NUMPOSTBOX` can be tuned to improve performance when MPI processes execute many point-to-point message-passing calls out of step with one another. When all-to-all message passing dominates, however, the default values of these variables can offer significantly better performance.