



Sun HPC ClusterTools™ 8.2 Software Release Notes

Sun Microsystems, Inc.
www.sun.com

Part No. 821-0223-11
June 2009, Revision A

Submit comments about this document by clicking the Feedback[+] link at: <http://www.sun.com>

Copyright 2009 Sun Microsystems, Inc., 4150 Network Circle, Santa Clara, California 95054, U.S.A. All rights reserved.

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

This distribution may include materials developed by third parties.

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 Microsystems, the Sun logo, Java, Netra, Solaris, docs.sun.com, Sun HPC ClusterTools, Sun Cluster, and Sun are trademarks or registered trademarks of Sun Microsystems, Inc. or its subsidiaries in the U.S. and other countries.

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

AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices.

Products covered by and information contained in this service manual are controlled by U.S. Export Control laws and may be subject to the export or import laws in other countries. Nuclear, missile, chemical biological weapons or nuclear maritime end uses or end users, whether direct or indirect, are strictly prohibited. Export or reexport to countries subject to U.S. embargo or to entities identified on U.S. export exclusion lists, including, but not limited to, the denied persons and specially designated nationals lists is strictly prohibited.

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 2009 Sun Microsystems, Inc., 4150 Network Circle, Santa Clara, California 95054, Etats-Unis. Tous droits réservés.

Cette distribution peut comprendre des composants développés par des tierces parties.

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 Microsystems, le logo Sun, Java, Netra, Solaris, docs.sun.com, Sun HPC ClusterTools, Sun Cluster, et Sun sont des marques de fabrique ou des marques déposées de Sun Microsystems, Inc. ou ses filiales 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 portant les marques SPARC sont basés sur une architecture développée par Sun Microsystems, Inc.

AMD, Opteron, le logo AMD, et le logo AMD Opteron sont des marques de fabrique ou des marques déposées de Advanced Micro Devices.

Les produits qui font l'objet de ce manuel d'entretien et les informations qu'il contient sont regis par la législation americaine en matiere de controle des exportations et peuvent etre soumis au droit d'autres pays dans le domaine des exportations et importations. Les utilisations finales, ou utilisateurs finaux, pour des armes nucleaires, des missiles, des armes biologiques et chimiques ou du nucleaire maritime, directement ou indirectement, sont strictement interdites. Les exportations ou reexportations vers des pays sous embargo des Etats-Unis, ou vers des entites figurant sur les listes d'exclusion d'exportation americaines, y compris, mais de maniere non exclusive, la liste de personnes qui font objet d'un ordre de ne pas participer, d'une facon directe ou indirecte, aux exportations des produits ou des services qui sont regi par la legislation americaine en matiere de controle des exportations et la liste de ressortissants specifiquement designes, sont rigoureusement interdites.

LA DOCUMENTATION EST FOURNIE "EN L'ETAT" 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 CONTREFACON.



Contents

1. Sun HPC ClusterTools 8.2 Software Release Notes	1
Major New Features	1
Related Software	2
Installing the ClusterTools 8.2 Linux Packages	3
Changes to Installation Method	3
Disabling Installation Notification	4
Mellanox Host Channel Adapter Support	4
Known Issues	5
On Some Linux Variants, Analyzer May Not Show ClusterTools MPI State Profiling Data (CR 6854789)	5
ClusterTools 8.2 Built with Pathscale or PGI Compilers Does Not Support XRC (CR 6852175)	5
MPI Library is Not Thread-Safe (CR 6474910)	6
Using udap1 BTL on Local Zones Fails for MPI Programs (CR 6480399)	6
udap1 BTL in Open MPI Should Detect That a udap1 Connection is Not Accessible and Not Just Hang (CR 6497612)	6
MPI Is Not Handling Resource Exhaustion Gracefully (CR 6499679)	7
Request Script Prevents SUNWompiat From Propagating to Nonglobal Zone During Zone Creation (CR 6539860)	7

udapl BTL Use of Fragment Free Lists Can Potentially Starve a Peer
Connection and Prevent Progress (CR 6542966) 8

TotalView: MPI-2 Support Is Not Implemented (CR 6597772) 8

TotalView: Message Queue for Unexpected Messages is Not Implemented (CR
6597750) 9

Slow Startup Seen on Large SMP (CR 6559928) 9

MPI Programs Running on Solaris with uDAPL and IB Network With Hermon
HCA Can Fail (CR 6735630) 9

-xalias=actual is Needed When Compiling Fortran 90 Programs (CR
6735316) 10

DDT Message Queue Hangs When Debugging 64-Bit Programs (CR
6741546) 10

MPI_Comm_spawn Fails When uDAPL BTL is in Use on Solaris (CR
6742102) 11

Sun HPC ClusterTools 8.2 Software Release Notes

This document describes late-breaking news about the Sun HPC ClusterTools™ 8.2 (ClusterTools 8.2) software. The information is organized into the following sections:

- “Major New Features” on page 1
- “Related Software” on page 2
- “Installing the ClusterTools 8.2 Linux Packages” on page 3
- “Changes to Installation Method” on page 3
- “Disabling Installation Notification” on page 4
- “Mellanox Host Channel Adapter Support” on page 4
- “Known Issues” on page 5

Major New Features

The major new features of the ClusterTools 8.2 software include:

- Enhanced shared memory pool sizing — With ClusterTools 8.2 software, a job that has two or more processes on a node will automatically establish a global shared memory pool at startup that is appropriately sized for the job. No special parameters are needed to start very large jobs (e.g., 512 processes on a node)
- Enhanced scalability — Shared memory message latency is low even when there are many processes (e.g., 512) on the same node.
- Added suspend and resume capability -- MPI jobs can be suspended and resumed without loss of state. For a fuller description of this feature, go to the Open MPI FAQ at:
<http://www.open-mpi.org/faq/?category=running#suspend-resume>
- Added support for QDR IB HCA technology — ClusterTools 8.2 software supports Quad Data Rate (QDR) Infiniband (IB) Host Channel Adaptors (HCAs).

- Added support for relocatable installations — ClusterTools 8.2 software can be installed in a non-standard location if the environment variable `OPAL_PREFIX` is set accordingly.
 - Support for automatic path migration (with Linux openib btl)
-

Related Software

Sun HPC ClusterTools 8.2 software works with the following versions of related software:

- Solaris™ 10 11/06 OS, or any subsequent Solaris 10 OS release that supports Sun HPC ClusterTools 8.2 software. If you are running Solaris 10 11/06 OS, you must install patch 125792-01 (for systems with SPARC® processors) or patch 125793-01 (for systems with AMD™ processors), plus any patches that those patches require.
- Red Hat Linux versions 5 (RHEL)
- SuSe Linux versions 10 (SLES)
- CentOS 5.3 Linux
- OpenSolaris 2009.06
- Sun™ Studio 10, 11, 12, and 12 U1 C, C++, and Fortran compilers
- gcc Linux compiler versions 3.3.3, 3.4.6, and 4.1.2
- Intel 11.0 20081105 compiler
- PGI 7.1-4 compiler
- Pathscale 3.2 compiler
- Distributed resource management (DRM) frameworks:
 - Sun™ Grid Engine Version 6.2
 - Altair PBS Professional 9.2 or Cluster Resources Torque 2.3
- Development Tools:
 - The TotalView 8.4.1 debugger from TotalView Technologies (formerly Etnus) supports debugging MPI applications on SPARC and AMD systems running the Solaris OS. Compatibility limitations might exist. See the TotalView Web site at www.totalviewtech.com for a list of compilers that TotalView supports. For more information about using TotalView with OPEN MPI see the Open MPI FAQ at:
<http://www.open-mpi.org/faq>
 - The DDT 2.1.3 debugger from Allinea also supports debugging on SPARC- and AMD-based systems running the Solaris OS.

- The Sun Studio 12 U1 Performance Analyzer supports performance analysis of MPI applications on SPARC and x86/x64 systems.

Installing the ClusterTools 8.2 Linux Packages

Sun HPC ClusterTools 8.2 software supports Red Hat Linux (RHEL) version 5 and SuSe Linux (SLES) version 10. The Linux packages are delivered in RPM format.

The `Quick_Installation_Guide.txt` file, included in the binary packages and with the Sun HPC ClusterTools 8.2 documentation, contains the information about how to find and install the Linux RPM.

For each version of Linux, there are five types of ClusterTools 8.2 RPMs:

- `ClusterTools_sun`
- `ClusterTools_gnu`
- `ClusterTools_intel`
- `ClusterTools_pgi`
- `ClusterTools_pathscale`

Note – You must install the RPM packages individually on each Linux node in your cluster. To facilitate the process, you might want to use a parallel SSH tool.

The following example shows the commands you would type to install the ClusterTools 8.2 package (built with the Sun Studio compiler) for SuSe Linux version 9.

```
# rpm --erase ClusterTools
# rpm --install ClusterTools_*-8.2-*x86_64.rpm
```

Changes to Installation Method

ClusterTools 8.2 is available in two formats: RPM and tarball. RPM provides the system administrator with metadata on all the files that comprise the package (e.g., whether a file belonging to the ClusterTools 8.2 packages has been modified or removed). For more information on RPM, visit:

<http://www.rpm.org/max-rpm>

Alternatively, tarball files can be used to install ClusterTools 8.2. For example:

```
$ tar xvfz ClusterTools_sun-8.2-*.x86_64.tar.gz
bin/
bin/opal_wrapper
bin/orte-clean
bin/orte-iof
...
```

If ClusterTools 8.2 is installed in a non-standard location (i.e., anywhere but `/opt/SUNWhpc/HPC8.2/compiler_name`) you must set the environment variable `OPAL_PREFIX`. For additional information how to set `OPAL_PREFIX`, see

<http://www.open-mpi.org/faq/?category=building#installdirs>

Disabling Installation Notification

To improve ClusterTools, Sun collects anonymous information about your cluster during installation. If you want to turn this feature off, use the `-w` option with `ctinstall`.

The communication between `ctinstall` and Sun works only if the Sun HPC ClusterTools software installation process completes successfully. It does not work if the installation fails for any reason.

Mellanox Host Channel Adapter Support

Sun HPC ClusterTools 8.2 software requires the Solaris OS to have the latest Infiniband updates to support use of the Mellanox ConnectX IB HCA.

This download is available at:

<http://www.sun.com/download/index.jsp?cat=Hardware%20Drivers&tab=3&subcat=InfiniBand>

For more information about Mellanox HCA support, contact the ClusterTools 8.2 software development alias at `ct-feedback@sun.com`.

Known Issues

This section highlights some of the outstanding CRs (Change Requests) for the ClusterTools 8.2 software components. A CR might be a defect, or it might be an RFE (request for enhancement).

Each CR has an identifying number assigned to it. To avoid ambiguity when inquiring about a CR, include its CR number in any communications. The heading for each CR description includes the associated CR number.

On Some Linux Variants, Analyzer May Not Show ClusterTools MPI State Profiling Data (CR 6854789)

Analyzer experiments may not contain ClusterTools MPI State profiling data on some Linux systems when the application is compiled with GNU or Intel compilers. This issue is exhibited on the Linux variants RHEL 5.3 and CentOS 5.3.

Workaround: Supply the option `-wl, --enable-new-dtags` to ClusterTools `mpi*` link commands. This flag causes the compiled executable to define `RUNPATH` in addition to `RPATH`, allowing ClusterTools MPI State libraries to be enabled via the `LD_LIBRARY_PATH` environment variable.

ClusterTools 8.2 Built with Pathscale or PGI Compilers Does Not Support XRC (CR 6852175)

The Pathscale and PGI environments in which ClusterTools 8.2 was built did not include OFED 1.3.1 or higher. Consequently, XRC support is not available with ClusterTools 8.2 built with either of these two compilers.

Workaround: Use ClusterTools 8.2 software that was built with a Sun Studio or GCC compiler.

MPI Library is Not Thread-Safe (CR 6474910)

The Open MPI library does not currently support thread-safe operations. If your applications contain thread-safe operations, they might fail.

Workaround: None.

Using udapl BTL on Local Zones Fails for MPI Programs (CR 6480399)

If you run an MPI program using the udapl BTL in a local (nonglobal) zone in the Solaris OS, your program might fail and display the following error message:

```
Process 0.1.3 is unable to reach 0.1.0 for MPI communication.  
If you specified the use of a BTL component, you may have  
forgotten a component (such as "self") in the list of  
usable components.
```

```
PML add procs failed  
--> Returned "Unreachable" (-12) instead of "Success" (0)
```

```
-----  
*** An error occurred in MPI_Init  
*** before MPI was initialized  
*** MPI_ERRORS_ARE_FATAL (goodbye)
```

Workarounds: Either run the udapl BTL in the Solaris global zone only, or use another interconnect (such as tcp) in the local zone.

udapl BTL in Open MPI Should Detect That a udapl Connection is Not Accessible and Not Just Hang (CR 6497612)

This condition happens when the udapl BTL is not available on one node in a cluster. The Infiniband adapter on the node could be unavailable or misconfigured, or there might not be an Infiniband adapter on the node.

When you run an Open MPI program using the udapl BTL under such conditions, the program might hang or fail, but no error message is displayed. When a similar operation fails under the tcp BTL, the failure results in an error message.

Workaround: Add the following MCA parameter to your command line to exclude the udapl BTL:

```
--mca btl ^udapl
```

For more information about MCA parameters and how to exclude functions at the command line, refer to the *Sun HPC ClusterTools 8.2 Software User's Guide*.

MPI Is Not Handling Resource Exhaustion Gracefully (CR 6499679)

If an MPI job exhausts the resources of the CPUs, the program can fail or show segmentation faults. This might happen when nodes are oversubscribed.

Workaround: Avoid oversubscribing the nodes.

For more information about oversubscribing nodes and the `--nooversubscribe` option, refer to the *Sun HPC ClusterTools 8.2 Software User's Guide*.

Request Script Prevents SUNWompiat From Propagating to Nonglobal Zone During Zone Creation (CR 6539860)

When you set up nonglobal zones in the Solaris OS, the Solaris OS packages propagate from the global zone to the new zones.

However, if you installed Sun HPC ClusterTools software on the system before setting up the zones, `SUNWompiat` (the Open MPI installer package) does not get propagated to the new nonglobal zone. It causes the `Install_Uutilities` directory not to be available on nonglobal zones during new zone creation. This also means that the links to `/opt/SUNWhpc` do not get propagated to the local zone.

Workaround: There are two workarounds for this issue.

1. From the command line, use the full path to the Sun HPC ClusterTools executable you want to use. For example, type `/opt/SUNWhpc/HPC8.2/bin/mpirun` instead of `/opt/SUNWhpc/bin/mpirun`.
2. Reinstall Sun HPC ClusterTools 8.2 software in the non-global zone. This process allows you to activate Sun HPC ClusterTools 8.2 software (thus creating the links to the executables) on nonglobal zones.

udapl BTL Use of Fragment Free Lists Can Potentially Starve a Peer Connection and Prevent Progress (CR 6542966)

When using a peer-to-peer connection with the `udapl` BTL (byte-transfer layer), the `udapl` BTL allocates a free list of fragments. This free list is used for send and receive operations between the peers. The free list does not have a specified maximum size, so a high amount of communication traffic at one peer might increase the size of the free list until it interferes with the ability of the other peers to communicate.

This issue might appear as a memory resource issue to an Open MPI application. This problem has only been observed on large jobs where the number of uDAPL connections exceeds the default value of `btl_udapl_max_eager_rdma_peers`.

Workaround: For example, if an Open MPI application running over uDAPL/IB (Infiniband) reports an out-of-memory error for `alloc` or for privileged memory, and if those two values have already been increased, the following might allow the program to run successfully.

1. At the command line, add the following MCA parameter to your `mpirun` command:

```
--mca btl_udapl_max_eager_rdma_peers x
```

where `x` is equal to the number of peer uDAPL connections that the Open MPI job will establish.

2. If the setting in [Step 1](#) does not fix the problem, then set the following MCA parameter with the `mpirun` command at the command line:

```
--mca mpi_preconnect_all 1
```

TotalView: MPI-2 Support Is Not Implemented (CR 6597772)

The TotalView debugger might not be able to determine if an `MPI_Comm_spawn` operation has occurred, and might not be able to locate the new processes that the operation creates. This is because the current version of the Open MPI message dumping library (`ompi/debuggers/ompi_dll.c`) does not implement the functions and interfaces for the support of MPI 2 debugging and message dumping.

Workaround: None.

TotalView: Message Queue for Unexpected Messages is Not Implemented (CR 6597750)

The Open MPI DLL for the TotalView debugger does not support handling of unexpected messages. Only pending send and receive queues are supported.

Workaround: None.

Slow Startup Seen on Large SMP (CR 6559928)

On a large SMP (symmetric multiprocessor) with many CPUs, ORTE might take a long time to start up before the MPI job runs. This is a known issue with the MPI layer.

Note – This behavior has improved in the ClusterTools 8.2 release as a result of changes in shared memory use. But the CR continues to be in effect.

Workaround: Reduce `mpool_sm_min_size` and `btl_sm_eager_limit` settings. This may shorten startup time. For more information, see the OMPI FAQ entry at:

<http://www.open-mpi.org/faq/?category=sm#decrease-sm>

MPI Programs Running on Solaris with uDAPL and IB Network With Hermon HCA Can Fail (CR 6735630)

If you are running MPI programs on the Solaris OS using the Mellanox HCA (Hermon) with uDAPL and an Infiniband network, you might experience hangs or segmentation faults if you set the number of processors to a value greater than 6.

Workaround: Set the environment variable `DAPL_MAX_INLINE` to 0, and then include the variable on the command line. For example:

```
% setenv DAPL_MAX_INLINE 0  
% mpirun ... -x DAPL_MAX_INLINE ...
```

Note – Setting this environment variable may have some impact on MPI latency when running using this configuration.

`-xalias=actual` is Needed When Compiling Fortran 90 Programs (CR 6735316)

When you compile MPI programs written in Fortran 90, you must use the `-xalias=actual` switch. Otherwise, your program could fail or give incorrect results.

Note – The `-xalias=actual` switch works only with the Sun Studio Fortran 90 compiler.

This behavior is caused by a known issue in the MPI standard. The standard states that “The MPI Fortran binding is inconsistent with the Fortran 90 standard in several respects.” Specifically, the Fortran 90 compiler could break MPI programs that use non-blocking operations.

However, not using `-xalias=actual` might result in silent failures, in which the program will complete, but return incorrect results.

For more information about this issue, see the MPI-1 standard at the following location:

<http://www-unix.mcs.anl.gov/mpi/mpi-standard/mpi-report-2.0/node19.htm#Node19>

DDT Message Queue Hangs When Debugging 64-Bit Programs (CR 6741546)

When using the Allinea DDT debugger to debug an application compiled in 64-bit mode on a SPARC-based system, the program might not run when loaded into the DDT debugger. In addition, if you try to use the View `->Message Queue` command, the debugger issues a popup dialog box with the message Gathering Data, and never finishes the operation.

Workaround: Set the environment variable `DDT_DONT_GET_RANK` to 1.

MPI_Comm_spawn Fails When uDAPL BTL is in Use on Solaris (CR 6742102)

When using `MPI_Comm_spawn` or other spawn commands in Open MPI, the uDAPL BTL might hang and return timeout messages similar to the following:

```
[btl_udapl_component.c:1051:mca_btl_udapl_component_progress]  
WARNING: connection event not handled :  
DAT_CONNECTION_EVENT_TIMED_OUT
```

Workaround: Use the TCP BTL with the spawn commands instead of the uDAPL BTL. For example:

```
--mca btl self,sm,tcp
```

