# Contents

1. **Sun HPC ClusterTools 8.1**  
   **Software Release Notes**  
   - Major New Features  
   - Related Software  
   - Installing the ClusterTools 8.1 Linux Packages  
   - Changes to the Default Paths  
   - Disabling Installation Notification  
   - Mellanox Host Channel Adapter Support  
   - Change in Open MPI Terminology  
   - Outstanding CRs

   **MPI Library is Not Thread-Safe (CR 6474910)**  
   Using `udapl` BTL on Local Zones Fails for MPI Programs (CR 6480399)  
   `udapl` BTL in Open MPI Should Detect That a `udapl` Connection is Not Accessible and Not Just Hang (CR 6497612)  
   MPI Is Not Handling Resource Exhaustion Gracefully (CR 6499679)  
   Request Script Prevents `SUNWompiat` From Propagating to Nonglobal Zone During Zone Creation (CR 6539860)  
   `udapl` BTL Use of Fragment Free Lists Can Potentially Starve a Peer Connection and Prevent Progress (CR 6542966)  
   TotalView: MPI-2 Support Is Not Implemented (CR 6597772)
TotalView: Message Queue for Unexpected Messages is Not Implemented (CR 6597750) 8

Slow Startup Seen on Large SMP (CR 6559928) 9

Bad Shared Memory Performance for Large Message Sizes Seen With Memory on Sun SPARC Enterprise M9000 Server 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

VampirTrace Wrapper Compiler Attempts to Use Incompatible gzip Library (CR 6738216) 11

MPI_Comm_spawn Fails When uDAPL BTL is in Use on Solaris (CR 6742102) 11
Sun HPC ClusterTools 8.1 Software Release Notes

This document describes late-breaking news about the Sun HPC ClusterTools™ 8.1 (ClusterTools 8.1) software. The information is organized into the following sections:

- “Major New Features” on page 1
- “Related Software” on page 2
- “Installing the ClusterTools 8.1 Linux Packages” on page 3
- “Changes to the Default Paths” on page 3
- “Disabling Installation Notification” on page 4
- “Mellanox Host Channel Adapter Support” on page 4
- “Change in Open MPI Terminology” on page 5
- “Outstanding CRs” on page 5

## Major New Features

The major new features of the ClusterTools 8.1 software include:

- Support for the Linux operating system — Sun HPC ClusterTools 8.1 includes support for Red Hat Linux versions 4 and 5 and SuSe Linux versions 9 and 10. See “Related Software” on page 2 for information on the types and versions of Linux that this version of ClusterTools 8.1 supports.

- Enhanced scalability — ClusterTools 8.1 software now supports clusters of 1024 nodes and 4096 processes.

- MPI profiling with VampirTrace on the Solaris OS — The VampirTrace program tracing tool is integrated into Open MPI version 1.4 and subsequent versions. The Open MPI FAQ contains information about VampirTrace, its components, and its documentation: [http://www.open-mpi.org/faq/?category=vampirtrace](http://www.open-mpi.org/faq/?category=vampirtrace).
■ MPI profiling with PERUSE — Sun HPC ClusterTools 8.1 and the upcoming Open MPI 1.3 release support the PERUSE data collection and analysis tool. For more information about PERUSE and current specifications, see http://www.mpi-peruse.org/index.html Chapter 8 of the Sun HPC ClusterTools 8.1 Software User’s Guide contains instructions about how to use mpiperuse profiling and some usage examples.

■ uDAPL BTL multi-rail support — If your cluster contains multiple Infiniband (IB) Host Channel Adapter (HCA) cards or other physical connections into the IB fabric from a single node, Sun HPC ClusterTools 8.1 allows you to select a particular HCA or physical connection when running using the uDAPL BTL.

■ libstlport4 support — The MPI C++ library can now be compiled and linked against either the C++ standard or the STLPORT4 library.

■ Support for the MPI OpenFabrics communications — Sun HPC ClusterTools 8.1 supports performance tuning for Infiniband and iWARP communications in Linux. For more information about OpenFabrics (formerly Open IB), see http://www.open-mpi.org/faq/?category=openfabrics.

■ Open MPI 1.3 now supports the use of MPI_Accumulate with user-defined types.

---

**Related Software**

Sun HPC ClusterTools 8.1 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 8.1 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 4 and 5 (RHEL)

■ SuSe Linux versions 9 and 10 (SLES)

■ Sun™ Studio 10, 11, and 12 C, C++, and Fortran compilers

■ gcc Linux compiler versions 3.3.3, 3.4.6, and 4.1.2

■ Distributed resource management (DRM) frameworks:
  ■ Sun™ Grid Engine Version 6
  ■ Altair PBS Professional 9.2 or Cluster Resources Torque 2.3

■ Debugging software:
  ■ 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. In addition, see the Open MPI FAQ at http://www.open-mpi.org for more information about using TotalView with Open MPI.

- The DDT 2.1.3 debugger from Allinea also supports debugging on SPARC- and AMD-based systems running the Solaris OS.

## Installing the ClusterTools 8.1 Linux Packages

Sun HPC ClusterTools 8.1 software supports Red Hat Linux (RHEL) versions 4 and 5 and SuSe Linux (SLES) versions 9 and 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 8.1 documentation, contains the information about how to find and install the Linux RPM that corresponds to the type of Linux you are using. This information is also available in Chapter 6 of the *Sun HPC ClusterTools 8.1 Software Installation Guide*.

For each version of Linux, there are two types of ClusterTools 8.1 RPMs: one built with Sun Studio compilers, and the other built with the GNU compiler `gcc`.

**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 8.1 package (built with the Sun Studio compiler) for SuSe Linux version 9.

```
# rpm --erase ClusterTools
# rpm --install ClusterTools-8.1-23.x86_64-sles9-built-with-sun.rpm
```

## Changes to the Default Paths

In earlier versions of the Sun HPC ClusterTools software, the default directory in which the ClusterTools binaries are installed was `/opt/SUNWhpc/HPC*/bin`, where `*` is the ClusterTools version number.
In Sun HPC ClusterTools 8.1 software, the binaries reside in one of two locations.

- If you installed the Solaris OS or Sun Studio compiled Linux version of the ClusterTools software, its default directory is `/opt/SUNWhpc/HPC8.1/sun/bin`.
- If you installed the GNU compiled Linux version of the ClusterTools software, its default directory is `/opt/SUNWhpc/HPC8.1/gnu/bin`.

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

The Mellanox ConnectX Infiniband Host Channel Adapter (HCA) requires the following software in order to run Sun HPC ClusterTools 8.1:

- Solaris™ 10 5/08 OS, or any subsequent Solaris 10 OS release that supports Sun HPC ClusterTools 8.1 software.

Note — The Infiniband software, Update 2, requires a firmware upgrade to the HCA. To download the firmware update, see the Mellanox Technologies Web site at [http://www.mellanox.com/support/firmware_download.php](http://www.mellanox.com/support/firmware_download.php)

Note — There is a known issue with the Mellanox HCA support on the Solaris OS. For more information, see the Sun HPC ClusterTools Software product page at [http://www.sun.com/clustertools](http://www.sun.com/clustertools)
For more information about Mellanox HCA support, contact the ClusterTools 8.1 software development alias at ct-feedback@sun.com.

Change in Open MPI Terminology

Sun HPC ClusterTools 8.1 is based on Open MPI version 1.3. In that version of Open MPI, the Process launch Subsystem (PLS), a component of the Open Runtime Environment (ORTE), was renamed to Process Launch Module (PLM). This means that all MCA parameters under the ORTE p1s framework were renamed to begin with plm instead.

For more information about the ORTE frameworks, refer to Chapter 7 in the Sun HPC ClusterTools 8.1 Software User’s Guide.

Outstanding CRs

This section highlights some of the outstanding CRs (Change Requests) for the ClusterTools 8.1 software components. A CR might be a defect, or it might be an RFE (request for enhancement).

Note – The heading of each CR description includes the CR bugster number in parentheses, if the description has a corresponding bugster entry.

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 8.1 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 8.1 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_Utilities 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/HPC8.1/bin/mpirun instead of /opt/SUNWhpc/bin/mpirun.

2. Reinstall Sun HPC ClusterTools 8.1 software in the non-global zone. This process allows you to activate Sun HPC ClusterTools 8.1 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.

Workaround: None.

Bad Shared Memory Performance for Large Message Sizes Seen With Memory on Sun SPARC Enterprise M9000 Server

When running ClusterTools 8.1 software with MPI applications on a Sun SPARC Enterprise M9000 server using shared memory for large message sizes, you might observe reduced communication performance.

Workaround: To achieve the best on-node performance, set the MCA parameter \texttt{btl\_sm\_max\_send\_size} to 132000. For example:

\begin{verbatim}
% setenv OMPI_MCA_btl_sm_max_send_size 132000
% mpirun -np 4 ./a.out
\end{verbatim}

You can also set the parameter at the command line as in the following example:

\begin{verbatim}
% mpirun -np 4 --mca btl_sm_max_send_size 132000 ./a.out
\end{verbatim}

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 are compiling MPI programs written in Fortran 90, you must use the -xalias=actual switch. Otherwise, your program could fail.

This condition is due to a known condition 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 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.
VampirTrace Wrapper Compiler Attempts to Use Incompatible gzip Library (CR 6738216)

When using Red Hat or SuSe Linux with the VampirTrace wrapper compiler (mpicxx-vt), you might see an error message similar to the following:

```
usr/lib64/libz.so: file not recognized: File format not recognized
```

This error occurs because Sun Studio does not support Red Hat Linux version 5 or SuSe Linux version 10. For more information about this issue, see the following:

http://developers.sun.com/sunstudio/support/matrix/index.jsp

**Workaround:** Use Red Hat Linux version 4 or SuSe Linux version 9.

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
```