Sun HPC ClusterTools™ 8.2.1c Software

Release Notes

Part No. 821-1317-10
April 2010, Revision A

Please
Recycle

Adobe PostScript

Contents

CHAPTER

Sun HPC ClusterTools 8.2.1c Software Release Notes

This document describes late-breaking news about the Sun HPC ClusterTools™ 8.2.1c (ClusterTools 8.2.1c) software. The information is organized into the following sections:

Major New Features

The following feature has been added to Sun HPC ClusterTools software:

- Failover support for multi-rail IB configurations — Sun HPC ClusterTools 8.2.1c software provides protection against completion errors in multi-rail Infiniband configurations. If a completion error occurs on an active rail, failover software maps out the rail and communication traffic continues on other available rails.

Related Software

Sun HPC ClusterTools 8.2.1c 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 8.2.1c 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.

Note – When TotalView is used to debug applications compiled with the Intel compiler, the stack trace feature is unable to display the full execution stack.

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 8.2.1c 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 8.2.1c software development alias at `ct-feedback@sun.com`.

Known Issues

This section highlights some of the outstanding CRs (Change Requests) for the ClusterTools 8.2.1c 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.

Slow startup for large NPs on single SPARC platform (CR 6898896)

When running connectivity on a single SPARC node with `np=64`, run times of 30 seconds with default versus 10 seconds without the `sparcopl` check.

Workaround: Specify the `--mca memcpy ^sparcopl` option with `mpirun`.

PLPA does not recognize multiple hardware threads per core (CR 6887809)

Running the default Clustertools 8.2.1c on a system with Hyper-Threads (such as Intel Xeon Processor x5570) could cause multiple processes to be bound to the same core, resulting in poor performance.

Workaround: Unless you are an expert user, you may want to avoid binding in this situation. You could use the default behavior or explicitly specify `-bind-to-none`. If you *are* an expert user, you can specify the exact binding behavior you want with rankfiles. See the `mpirun` man page for more information about rankfiles

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 built with Pathscale compiler does not support XRC in OpenIB BTL. (CR 6852175)

The Pathscale and PGI environments in which ClusterTools 8.2.1c was built did not include OFED 1.3.1 or higher. Consequently, XRC support is not available with ClusterTools 8.2.1c built with either of these two compilers.

Workaround: Use ClusterTools 8.2.1c software with Sun Studio or GCC compiled libraries.

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 8.2.1c 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 8.2.1c 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/HPC8.2.1c/bin/mpirun` instead of `/opt/SUNWhpc/bin/mpirun`.
2. Reinstall Sun HPC ClusterTools 8.2.1c software in the non-global zone. This process allows you to activate Sun HPC ClusterTools 8.2.1c 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`

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
```