



# Sun HPC ClusterTools™ 8.2 Software User's Guide

---

Sun Microsystems, Inc.  
[www.sun.com](http://www.sun.com)

Part No. 821-0225-10  
June 2009, Revision A

Submit comments about this document by clicking the Feedback[+] link at: <http://docs.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, Sun Microsystems, the Sun logo, Java, Netra, Solaris, docs.sun.com, Sun HPC ClusterTools and Sun Cluster 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, Sun Microsystems, le logo Sun, Java, Netra, Solaris, docs.sun.com, Sun HPC ClusterTools, et Sun Cluster 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 américaine en matière de contrôle des exportations et peuvent être soumis au droit d'autres pays dans le domaine des exportations et importations. Les utilisations finales, ou utilisateurs finaux, pour des armes nucléaires, des missiles, des armes biologiques et chimiques ou du nucléaire maritime, directement ou indirectement, sont strictement interdites. Les exportations ou réexportations vers des pays sous embargo des Etats-Unis, ou vers des entités figurant sur les listes d'exclusion d'exportation américaines, y compris, mais de manière non exclusive, la liste de personnes qui font objet d'un ordre de ne pas participer, d'une façon directe ou indirecte, aux exportations des produits ou des services qui sont regis par la législation américaine en matière de contrôle des exportations et la liste de ressortissants spécifiquement désignés, 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

---

<b>Preface</b>	<b>xi</b>
<b>1. Introduction to Sun HPC ClusterTools Software</b>	<b>1</b>
Supported Configurations	1
Open Runtime Environment (ORTE)	2
Executing Programs With <code>mpirun</code>	2
Integration With Distributed Resource Management Systems	3
Open MPI Features	3
Debugging With TotalView	4
<b>2. Fundamental Concepts</b>	<b>5</b>
Clusters and Nodes	5
Processes	6
How Programs Are Launched	6
How the Open MPI Environment Is Integrated With Distributed Resource Management Systems	7
Using Sun Grid Engine With ORTE	7
Submitting Jobs Under Sun Grid Engine Integration	8
MCA Parameters	8
How ORTE Works With Zones in the Solaris 10 Operating System	8

<b>3. Before You Begin</b>	<b>11</b>
Prerequisites	11
Command and Man Page Paths	11
Setting Up Your Path	12
▼ To Set Up Your Path for the Solaris OS and the Sun Studio Compiled Linux Version	12
▼ To Set Up Your Path for the GNU Compiled Linux Version	13
Core Files	13
Setting Up a <code>known_hosts</code> File	13
<b>4. Compiling MPI Programs</b>	<b>15</b>
Supported Compilers	15
Using the Compiler Wrappers	16
Using Non-Default Error Handlers	16
Compiling Fortran 90 Programs	17
<b>5. Running Programs With the <code>mpirun</code> Command</b>	<b>19</b>
About the <code>mpirun</code> Command	19
Syntax for the <code>mpirun</code> Command	20
<code>mpirun</code> Options	20
Using Environment Variables With the <code>mpirun</code> Command	21
Using MCA Parameters With the <code>mpirun</code> Command	21
Canceling Send and Receive Operations	22
<code>mpirun</code> Command Examples	22
▼ To Run a Program With Default Settings	23
▼ To Run Multiple Processes	23
▼ To Direct <code>mpirun</code> By Using an Appfile	23
Mapping MPI Processes to Nodes	24
Specifying Available Hosts	24
Specifying Hosts By Using a Hostfile	25

Specifying Hosts By Using the <code>--host</code> Option	26
▼ To Specify Multiple Slots Using the <code>--host</code> Option	26
Excluding Hosts From Scheduling By Using the <code>--host</code> Option	26
Oversubscribing Nodes	27
Scheduling Policies	27
Scheduling By Slot	27
▼ To Specify By-Slot Scheduling	28
Scheduling By Node	29
▼ To Specify By-Node Scheduling	29
Comparing By-Slot to By-Node Scheduling	30
Controlling Input/Output	30
▼ To Redirect Standard I/O	31
Controlling Other Job Attributes	31
▼ To Change the Working Directory	31
▼ To Specify Debugging Output	32
▼ To Display Command Help ( <code>-h</code> )	32
Submitting Jobs Under Sun Grid Engine Integration	36
Defining Parallel Environment (PE) and Queue	36
▼ To Use PE Commands	36
▼ To Use Queue Commands	37
Submitting Jobs in Interactive Mode	38
▼ To Set the Interactive Display	38
▼ To Submit Jobs Interactively	38
▼ To Verify That Sun Grid Engine Is Running	39
▼ To Start an Interactive Session Using <code>qcrsh</code>	39
Using MPI Client/Server Applications	39
▼ To Launch the Client/Server Job	39
Using Name Publishing	40

Troubleshooting Client/Server Jobs	41
For More Information	42
<b>6. Running Programs With <code>mpirun</code> in Distributed Resource Management Systems</b>	<b>43</b>
<code>mpirun</code> Options for Third-Party Resource Manager Integration	43
Checking Your Open MPI Configuration	44
▼ To Check for <code>rsh/ssh</code>	44
▼ To Check for PBS/Torque	44
▼ To Check for Sun Grid Engine	44
Running Parallel Jobs in the PBS Environment	44
▼ To Run an Interactive Job in PBS	45
▼ To Run a Batch Job in PBS	46
Running Parallel Jobs in the Sun Grid Engine Environment	47
Defining Parallel Environment (PE) and Queue	48
▼ To Use PE Commands	48
▼ To Use Queue Commands	49
Submitting Jobs Under Sun Grid Engine Integration	49
▼ To Set the Interactive Display	49
▼ To Submit Jobs in Batch Mode	50
▼ To See a Running Job	51
▼ To Delete a Running Job	51
<code>rsh</code> Limitations	51
Using <code>rsh</code> as the Job Launcher	52
Using Sun Grid Engine as the Job Launcher	52
For More Information	53
<b>7. Using MCA Parameters With <code>mpirun</code></b>	<b>55</b>
About the Modular Component Architecture	56
Open MPI Frameworks	56

OMPI Frameworks	57
ORTE Frameworks	58
OPAL Frameworks	59
The <code>ompi_info</code> Command	59
Command Options	60
Using the <code>ompi_info</code> Command With MCA Parameters	63
▼ To List All MCA Parameters	63
▼ To List All MCA Parameters For a Framework	63
▼ To Display All MCA Parameters For a Selected Component	67
Using MCA Parameters	68
▼ To Set MCA Parameters From the Command Line	69
Using MCA Parameters As Environment Variables	69
▼ To Set MCA Parameters in the <code>sh</code> Shell	69
▼ To Set MCA Parameters in the <code>C</code> Shell	70
▼ To Specify MCA Parameters Using a Text File	71
Including and Excluding Components	72
▼ To Include and Exclude Components Using the Command Line	73
Processor and Memory Affinity	73
Using Processor Affinity	73
Using Memory Affinity	74
▼ To Find Out Whether Memory Affinity Is Supported	74
Running MPI Jobs With Processor and Memory Affinity	75
▼ To Enable Affinity Using the Command Line	75
Using MCA Parameters With Sun Grid Engine	76
Changing the Default Values in MCA Parameters	78
For More Information	79
<b>8. Using the DTrace Utility With Open MPI</b>	<b>81</b>
Checking the <code>mpi.run</code> Privileges	82

▼	To Determine the Correct Privileges on the Cluster	82
	Running DTrace with MPI Programs	83
	Running an MPI Program Under DTrace	84
▼	To Trace a Program Using the <code>mpitrace.d</code> Script	84
▼	To Trace a Parallel Program and Get Separate Trace Files	84
	Attaching DTrace to a Running MPI Program	85
▼	To Attach DTrace to a Running MPI Program	85
	Simple MPI Tracing	85
	Tracking Down Resource Leaks	87
	Using the DTrace <code>mpiperuse</code> Provider	92
	DTrace Support in the ClusterTools Software	92
	Available <code>mpiperuse</code> Probes	92
	Specifying an <code>mpiperuse</code> Probe in a D Script	93
	Available Arguments	94
	How To Use <code>mpiperuse</code> Probes to See Message Queues	94
	<code>mpiperuse</code> Usage Examples	96
▼	To Count the Number of Messages To or From a Host	96
▼	To Count the Number of Messages To or From Specific BTLs	96
▼	To Obtain Distribution Plots of Message Sizes Sent or Received From a Host	97
▼	To Create Distribution Plots of Message Sizes By Communicator, Rank, and Send/Receive	97
<b>A.</b>	<b>Troubleshooting</b>	<b>99</b>
	MPI Messages	99
	Standard Error Classes	99
	MPI I/O Error Handling	101
	Exceeding the File Descriptor Limit	103
	Increasing the Number of Available File Descriptors	103
▼	To View the Hard Limit from the C Shell	103

▼	To View the Hard Limit from the Bourne Shell	104
▼	To Increase the Number of File Descriptors	104
	Setting File Descriptor Limits When Using Sun Grid Engine	105
<b>B.</b>	<b>List of Available MCA Parameters</b>	<b>107</b>
	<b>Index</b>	<b>131</b>



# Preface

---

This manual explains how to use distributed resource management packages for effective resource management and utilization accounting. The following packages work in conjunction with the Open Message-Passing Interface (Open MPI) parallel applications:

- Sun Grid Engine Version 6.1 software
- Altair PBS Professional 9.2 or Torque 2.3

---

## Before You Read This Book

The *Sun HPC ClusterTools™ 8.2 Software Release Notes* includes release note information for the other components in this suite. For information about writing MPI programs, refer to the *Open MPI Software Programming and Reference Guide*. For information about a specific distributed resource management package, refer to the documentation supplied with that package.

For more information about Open MPI and its components, see the Open MPI web site at:

<http://www.open-mpi.org>

---

## Using UNIX Commands

This document might not contain information on basic UNIX® commands and procedures such as shutting down the system, booting the system, and configuring devices.

See one or more of the following for this information:

- Software documentation that you received with your system
- Solaris™ Operating System documentation, which is at

<http://www.sun.com/documentation>

---

# Typographic Conventions

Typeface*	Meaning	Examples
AaBbCc123	The names of commands, files, and directories; on-screen computer output	Edit your <code>.login</code> file. Use <code>ls -a</code> to list all files. % You have mail.
<b>AaBbCc123</b>	What you type, when contrasted with on-screen computer output	% <b>su</b> password:
<i>AaBbCc123</i>	Book titles, new words or terms, words to be emphasized. Replace command-line variables with real names or values.	Read Chapter 6 in the <i>User's Guide</i> . These are called <i>class</i> options. You <i>must</i> be superuser to do this. To delete a file, type <code>rm filename</code> .

\* The settings on your browser might differ from these settings.

---

# Shell Prompts

Shell	Prompt
C shell	<i>machine-name%</i>
C shell superuser	<i>machine-name#</i>
Bourne shell and Korn shell	\$
Bourne shell and Korn shell superuser	#

---

# Related Documentation

This book focuses on Open MPI and assumes familiarity with the *MPI Standard*. The following materials provide useful background about using Open MPI and about the

*MPI Standard.*

<b>Application</b>	<b>Title</b>	<b>Part Number</b>
Sun HPC ClusterTools Software	<i>Sun HPC ClusterTools 8.2 Software Release Notes</i>	821-0223-10
	<i>Sun HPC ClusterTools 8.2 Software Installation Guide</i>	821-0224-10

The Sun HPC ClusterTools documentation is available online at:

<http://www.sun.com/documentation>

For more information about the Sun HPC ClusterTools software, see the related Web site at:

<http://www.sun.com/clustertools>

For more information about Open MPI and its components, see the Open MPI web site at:

<http://www.open-mpi.org>

For more information about Sun Grid Engine software, see the Sun Grid Engine web site at:

<http://www.sun.com/software/gridware>

---

# Documentation, Support, and Training

Sun Function	URL
Documentation	<a href="http://www.sun.com/documentation/">http://www.sun.com/documentation/</a>
Support	<a href="http://www.sun.com/support/">http://www.sun.com/support/</a>
Training	<a href="http://www.sun.com/training/">http://www.sun.com/training/</a>

---

## Third-Party Web Sites

Sun is not responsible for the availability of third-party web sites mentioned in this document. Sun does not endorse and is not responsible or liable for any content, advertising, products, or other materials that are available on or through such sites or resources. Sun will not be responsible or liable for any actual or alleged damage or loss caused by or in connection with the use of or reliance on any such content, goods, or services that are available on or through such sites or resources.

---

## Sun Welcomes Your Comments

Sun is interested in improving its documentation and welcomes your comments and suggestions. Submit comments about this document by clicking the Feedback[+] link at <http://www.sun.com>.

Please include the title and part number of your document with your feedback:

*Sun HPC ClusterTools 8.2 Software User's Guide*, part number 821-0225-10

# Introduction to Sun HPC ClusterTools Software

---

Sun HPC ClusterTools™ 8.2 software is a set of parallel development tools that extend the Sun network computing solutions to high-end distributed-memory applications. This chapter summarizes its required configuration and principal components. It contains the following sections:

- “Supported Configurations” on page 1
- “Open Runtime Environment (ORTE)” on page 2
- “Integration With Distributed Resource Management Systems” on page 3
- “Open MPI Features” on page 3
- “Debugging With TotalView” on page 4

---

## Supported Configurations

Sun HPC ClusterTools 8.2 software can be used in any of the following operating system environments:

- Solaris(TM) 10 Operation System (Solaris 10 OS)
- Red Hat Linux versions 5 (RHEL)
- SuSe Linux versions 10 (SLES)
- CentOS 5.3 Linux
- OpenSolaris 2009.06

Sun HPC ClusterTools 8.2 software supports the following compilers in the Solaris OS and Linux environments:

- **Solaris OS**
  - Sun Studio 10, 11, 12, and 12U1 C, C++, and Fortran

- **Linux**
  - Sun Studio 12U1 C, C++, and Fortran
  - gcc 3.3.3, 3.4.6, 4.1.2
  - Intel 11.0 20081105
  - PGI 7.1-4
  - Pathscale 3.2

Sun HPC ClusterTools 8.2 software can run MPI jobs of up to 4096 processes on as many as 1024 nodes. It also provides support for spawning MPI processes.

The Sun HPC ClusterTools 8.2 software runs on clusters connected by any TCP/IP-capable interconnect, such as high-speed Ethernet, Gigabit Ethernet, Infiniband, and Myrinet MX.

---

## Open Runtime Environment (ORTE)

Sun HPC ClusterTools 8.2 is based on the Open MPI message-passing interface. Open MPI operates using the Open Runtime Environment (ORTE). ORTE starts jobs and provides some status information.

The Open MPI `mpirun` and `mpiexec` commands are actually symbolic links to the `orterun` command. All three commands perform the same function, which is to launch MPI jobs.

ORTE is compatible with a number of other launchers, including `rsh/ssh`, Sun Grid Engine, and PBS Professional/Torque.

Each of ORTE's primary operations is summarized in the sections that follow. Subsequent chapters contain the procedures.

## Executing Programs With `mpirun`

Sun HPC ClusterTools 8.2 software can start both serial and parallel jobs. The syntax and use of `mpirun` are described in [Chapter 5](#).

---

# Integration With Distributed Resource Management Systems

Sun HPC ClusterTools 8.2 software provides integration facilities with two select distributed resource management (DRM) systems. These systems provide proper resource allocation, parallel job control and monitoring, and proper job accounting. They are:

- Sun Grid Engine Version 6.1 and 6.2 software
- Altair PBS Professional 9.2, or Cluster Resources Torque 2.3

---

**Note** – Open MPI itself supports other third-party launchers supported by Open MPI, such as SLURM. However, these launchers are currently not supported in Sun HPC ClusterTools software. To use these other third-party launchers, you must download the Open MPI source, compile, and link with the libraries for the launchers.

---

You can launch parallel jobs directly from these distributed resource management systems. The DRM interacts closely with ORTE for proper resource description and with the multiple processes comprising the requested parallel job.

For a description of the scalable and open architecture of the DRM integration facilities, see [“How the Open MPI Environment Is Integrated With Distributed Resource Management Systems”](#) on page 7. For instructions, see [Chapter 6](#).

---

## Open MPI Features

Open MPI is a highly optimized version of the Message Passing Interface (MPI) communications library. It implements all of the *MPI 1.2 Standard* and the *MPI 2.0 Standard*. Its highlights are:

- Integration with the Open Runtime Environment (ORTE)
- Support for MPI I/O
- Seamless use of different network protocols; for example, code compiled on a Sun HPC cluster that has a fast Ethernet network can be run without change on a cluster that has an Infiniband network
- Multiprotocol support so that MPI picks the fastest available medium for each type of connection (such as shared memory, fast Ethernet, Infiniband)

- Communication via shared memory for fast performance
  - Optimized collectives for multiprocessors and clusters of multiprocessors
  - Full F77, C, and C++ support, and basic F90 support
- 

## Debugging With TotalView

TotalView is a third-party multiprocess debugger from TotalView Technologies (formerly Etnus) that runs on many platforms. Support for using the TotalView debugger on Open MPI applications includes:

- Making Sun HPC ClusterTools software compatible with the TotalView debugger
- Allowing Open MPI jobs to be debugged by TotalView using the Sun Grid Engine or the Portable Batch System (PBS)
- Allowing multiple instantiations of TotalView on a single cluster
- Supporting TotalView in Sun HPC ClusterTools software

Refer to the TotalView documentation at <http://www.totalviewtech.com> for more information about using TotalView.

In addition, the Open MPI Frequently Asked Questions (FAQ) contains information about how to use the TotalView debugger with Open MPI. This information is available at:

<http://www.open-mpi.org/faq/?category=running#run-with-tv>

## Fundamental Concepts

---

This chapter summarizes a few basic concepts that you should understand to get the most out of Sun's HPC ClusterTools software. It contains the following sections:

- ["Clusters and Nodes" on page 5](#)
  - ["Processes" on page 6](#)
  - ["How the Open MPI Environment Is Integrated With Distributed Resource Management Systems" on page 7](#)
  - ["MCA Parameters" on page 8](#)
  - ["How ORTE Works With Zones in the Solaris 10 Operating System" on page 8](#)
- 

### Clusters and Nodes

High performance computing clusters<sup>1</sup> are groups of servers interconnected by any Sun-supported interconnect. Each server in a cluster is called a *node*. A cluster can consist of a single node.

ORTE (Open Run-Time Environment) is the runtime support system for Open MPI that allows users to execute their applications in a distributed clustering environment.

When using ORTE, you can select the cluster and nodes on which your MPI programs will run and how your processes will be distributed among them. For instructions, see [Chapter 5, "Running Programs With the `mpi.run` Command."](#)

---

1. Sun<sup>TM</sup> Cluster is a completely different technology used for high availability (HA) applications.

For more information about how Open MPI allocates computing resources, see the FAQ entitled “Running MPI Jobs” at:

<http://www.open-mpi.org/faq/?category=running>

---

## Processes

Open MPI allows you to control several aspects of job and process execution, such as:

- Number of processes to be launched
- Number of available slots on each node
- Process launcher to be used (such as Sun Grid Engine, PBS, rsh/ssh)
- Mapping processes to nodes

## How Programs Are Launched

The exact instructions vary from one resource manager to another, and are affected by your Open MPI configuration, but they all follow these general guidelines:

1. You can launch the job either interactively or through a script. Instructions for both are provided in [Chapter 5](#) and [Chapter 6](#).
2. You can enter the DRM processing environment (for example, Sun Grid Engine) before launching jobs with `mpirun`.
3. You can reserve resources for the parallel job and set other job control parameters from within the DRM, or use a hosts file to specify the parameters.

For tasks and instructions, see [Chapter 5](#).

---

# How the Open MPI Environment Is Integrated With Distributed Resource Management Systems

As described in [Chapter 1](#), the Open MPI/Sun HPC ClusterTools 8.2 environment provides close integration between ORTE and several different DRM systems, including the following:

- Sun Grid Engine
- Torque
- PBS

The integration process is similar for all DRM systems, with some individual differences. At run time, `mpirun` calls the specified DRM system (launcher), which in turn launches the job.

For information on the ways in which `mpirun` interacts with DRM systems, see [Chapter 5](#). In addition, see the FAQ on running MPI jobs at:

<http://www.open-mpi.org/faq/?category=running>

[Chapter 6](#) provides instructions for script-based and interactive job launching.

## Using Sun Grid Engine With ORTE

HPC sites use batch systems to share resources fairly and accountably, and also to guarantee that a job can obtain the resources it needs to run at maximum efficiency. To properly monitor a job's resource consumption, the batch system must be the agent that launches the job.

Sun Grid Engine, like many other batch systems, cannot launch multiple process jobs (such as MPI applications) on its own. In Sun HPC ClusterTools 8.2, ORTE launches the multiple process jobs and sets up the environment required by Open MPI.

When Sun Grid Engine launches a parallel job in cooperation with ORTE, Sun Grid Engine “owns” the resulting launched processes. Sun Grid Engine monitors the resources for these processes, thereby creating a tightly integrated environment for resource accounting. OpenRTE allows users to execute their parallel applications.

---

**Note** – There is also an open source version of Grid Engine (GE) hosted on <http://www.sunsource.net>. Although the Sun HPC ClusterTools 8.2/Open MPI integration is developed with Sun Grid Engine, this integration should work for the open source Grid Engine as well.

---

## Submitting Jobs Under Sun Grid Engine Integration

To submit jobs under Sun Grid Engine integration in Sun HPC ClusterTools 8.2, you must first create a Sun Grid Engine (SGE) environment using `qsub`, `qsh`, and so on. Instructions about how to set up the parallel environment (PE) and queue in Sun Grid Engine are described in the *Sun HPC ClusterTools 8.2 Software User's Guide*.

There are two ways to submit jobs under Sun Grid Engine integration: interactive mode and batch mode. “[Running Parallel Jobs in the Sun Grid Engine Environment](#)” on page 47 explains how to submit jobs in both modes in the Sun Grid Engine environment.

---

## MCA Parameters

Open MPI provides MCA (Modular Component Architecture) parameters for use with the `mpirun` command. These parameters and their values direct `mpirun` to perform specified functions. To specify an MCA parameter, use the `-mca` flag and the parameter name and value with the `mpirun` command.

For more information about how to use MCA parameters, see [Chapter 7](#).

---

## How ORTE Works With Zones in the Solaris 10 Operating System

The Solaris 10 Operating System (Solaris 10 OS) enables you to create secure, isolated areas within a single instance of the Solaris 10 OS. These areas, called *zones*, provide secure environments for running applications. Applications that execute in one zone cannot monitor or affect activity in another zone. You can create multiple non-global zones to run as virtual instances of the Solaris OS on the same hardware.

The *global zone* is the default zone for the Solaris system. You install Sun HPC ClusterTools software into the global zone. Any non-global zones running under that Solaris system “inherit” that installation. This means that you may install and configure Sun HPC ClusterTools and compile/run/debug your programs in either a global or a non-global zone.

---

**Note** – The non-global zones do not inherit the links set up in the global zone. This means that you must type out the full path to the Sun HPC ClusterTools executables on the command line.

---



## Before You Begin

---

This chapter provides miscellaneous information about the runtime environment that you should know before you begin to use it. It contains the following sections:

- “Prerequisites” on page 11
- “Command and Man Page Paths” on page 11
- “Core Files” on page 13
- “Setting Up a `known_hosts` File” on page 13

---

### Prerequisites

If your program uses Sun HPC ClusterTools components, compile and link it on a cluster that contains the Sun HPC ClusterTools software.

It is strongly suggested that you use the Sun HPC ClusterTools 8.2 compiler wrapper to compile applications. These compiler wrappers add the appropriate compiler and linker flags to the command line and call the underlying compiler and linker for you. Compiler wrappers are available for C, C++, Fortran 77, and Fortran 90.

For more information about compiling MPI applications, see the “Compiling MPI Applications” FAQ at:

<http://www.open-mpi.org/faq/?category=mpi-apps>

---

### Command and Man Page Paths

Sun HPC ClusterTools commands typically reside in the following directories:

- Solaris OS: `/opt/SUNWhpc/HPC8.2/sun/bin`
- Linux:
  - `/opt/SUNWhpc/HPC8.2/gnu/bin` for the GNU compiled version
  - `/opt/SUNWhpc/HPC8.2/sun/bin` for the Sun Studio compiled version
  - `/opt/SUNWhpc/HPC8.2/pgi/bin` for the PGI compiled version
  - `/opt/SUNWhpc/HPC8.2/intel/bin` for the Intel compiled version
  - `/opt/SUNWhpc/HPC8.2/pathscale/bin` for the Pathscale compiled version

You can run the Sun HPC ClusterTools software directly from the directory in which your ClusterTools commands are installed, or you may add the directory to your PATH or set the PATH environment variable.

The man pages for Sun HPC ClusterTools commands reside in the `/opt/SUNWhpc/HPC8.2/<compiler-name>/man`.

---

**Note** – The examples in this manual refer to the default location for the Solaris and Sun Studio compiled Linux binaries. Be sure to use the path name that corresponds to your operating system and ClusterTools version.

---

## Setting Up Your Path

### ▼ To Set Up Your Path for the Solaris OS and the Sun Studio Compiled Linux Version

For example, if you installed the Sun HPC ClusterTools 8.2 software for the Solaris OS in the default location of `/opt/SUNWhpc/HPC8.2/sun/bin`, you would add this location to your PATH as shown in the following example:

```
% setenv PATH /opt/SUNWhpc/HPC8.2/sun/bin:${PATH}
```

The `setenv` command prefixes the PATH on both the local and remote hosts with `/opt/SUNWhpc/HPC8.2/sun/bin`.

## ▼ To Set Up Your Path for the GNU Compiled Linux Version

For example, if you installed the GNU compiled version of the Sun HPC ClusterTools 8.2 software for Linux in the default location of `/opt/SUNWhpc/HPC8.2/gnu/bin`, you would add this location to your `PATH` as shown in the following example:

```
% setenv PATH /opt/SUNWhpc/HPC8.2/gnu/bin:${PATH}
```

The `setenv` command prefixes the `PATH` on both the local and remote hosts with `/opt/SUNWhpc/HPC8.2/gnu/bin`.

---

## Core Files

Core files are produced as they normally are in the Solaris environment. However, if more than one process dumps core in a multiprocess program, the resulting core file may be overwritten in the same directory. Use `coreadm(1M)` to control the naming and placement of core files.

To disable the core dump, use the `limit(1)` command. You can use the following command in the C shell:

```
% limit coredumpsize 0
```

---

## Setting Up a `known_hosts` File

If you are using `ssh` to connect to your remote nodes, you must set up your `~/ .ssh/known_hosts` file to contain the remote nodes' host key, especially if you try to run on a cluster with many nodes for the first time.

Setting up the known\_hosts file avoids having to respond to the following prompts when running mpirun to the remote nodes:

```
% /opt/SUNWhpc/HPC8.2/sun/bin/mpirun -host host04,host05,host06
hostname

The authenticity of host 'host04 (129.148.9.88)' can't be
established.
RSA key fingerprint is
ff:73:0e:91:7b:19:e6:a3:9a:f8:6c:07:0a:ce:1a:f2.
Are you sure you want to continue connecting (yes/no)?
The authenticity of host 'host05 (129.148.9.84)' can't be
established.
RSA key fingerprint is
11:f6:bd:f7:a9:5b:05:ab:73:ee:81:0f:c8:2f:ac:cb.
Are you sure you want to continue connecting (yes/no)?
The authenticity
of host 'host06 (129.148.9.86)' can't be established.
RSA key fingerprint is
c1:24:c4:9c:86:33:25:3b:08:59:12:b5:1b:9d:b2:c6.
Are you sure you want to continue connecting (yes/no)?
Host key verification failed.
yes
Please type 'yes' or 'no': yes
...
```

## Compiling MPI Programs

---

This chapter describes the compilers that Sun HPC ClusterTools Software supports for both the Solaris OS and Linux. In addition, it describes changes you might make in your application code to recompile and run programs developed with a previous version of Sun HPC ClusterTools software in Sun HPC ClusterTools 8.2.

This chapter contains the following topics:

- [“Supported Compilers” on page 15](#)
- [“Using the Compiler Wrappers” on page 16](#)
- [“Using Non-Default Error Handlers” on page 16](#)
- [“Compiling Fortran 90 Programs” on page 17](#)

If you previously compiled your application with the `tmcc` compiler, then you must recompile your applications using the `mpicc` compiler if you want them to be compatible with Sun HPC ClusterTools 8.2 software. The `tmcc` compiler is not supported in Sun HPC ClusterTools 8.2, and there is no backward compatibility.

---

## Supported Compilers

For the Solaris OS, Sun HPC ClusterTools 8.2 software supports Sun Studio 10, 11, 12, and 12 U1 C, C++, and Fortran compilers.

For the Linux OS, the ClusterTools 8.2 software supports

- Sun Studio 12 U1 compilers
- `gcc` Linux 3.3.3, 3.4.6, and 4.1.2 compilers
- Intel 11.0 20081105 compiler
- PGI 7.1-4 compiler

- Pathscale 3.2 compiler

---

## Using the Compiler Wrappers

Sun HPC ClusterTools 8.2 supplies compiler wrappers for you to use instead of directly calling the compilers when compiling applications for use with the Sun HPC ClusterTools 8.2 software. These compiler wrappers do not actually perform the compilation and linking steps themselves, but they add the appropriate compiler and linker flags and call the compiler and linker.

---

**Note** – Using the compiler wrappers is strongly suggested. If you decide not to use them, the Open MPI Web site at <http://www.open-mpi.org> contains instructions about how to compile without using them.

---

The following compiler wrappers are available:

**TABLE 4-1** Compiler Wrappers

Language	Compiler Wrapper
C	<code>mpicc</code>
C++	<code>mpicc</code> , <code>mpicxx</code> , or <code>mpic++</code> (Note: <code>mpicc</code> is for use on case-sensitive file systems only)
Fortran 77	<code>mpif77</code>
Fortran 90	<code>mpif90</code>

For more information about the compiler wrappers, their use, and troubleshooting, see the Open MPI FAQ at:

<http://www.open-mpi.org/faq/?category=mpi-apps>

---

## Using Non-Default Error Handlers

In Open MPI, the non-default error handler does not persist, and the default error handler is used. This causes any call used after `MPI_Finalize` to be aborted.

---

# Compiling Fortran 90 Programs

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.

For more information about this issue, see

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



## Running Programs With the `mpirun` Command

---

This chapter describes the general syntax of the `mpirun` command and lists the command's options. This chapter also shows some of the tasks you can perform with the `mpirun` command. It contains the following sections:

- [“About the `mpirun` Command” on page 19](#)
- [“Syntax for the `mpirun` Command” on page 20](#)
- [“`mpirun` Command Examples” on page 22](#)
- [“Mapping MPI Processes to Nodes” on page 24](#)
- [“Controlling Input/Output” on page 30](#)
- [“Controlling Other Job Attributes” on page 31](#)
- [“Submitting Jobs Under Sun Grid Engine Integration” on page 36](#)
- [“Using MPI Client/Server Applications” on page 39](#)
- [“For More Information” on page 42](#)

---

**Note** – The `mpirun`, `mpiexec`, and `orterun` commands all perform the same function, and they can be used interchangeably. The examples in this manual all use the `mpirun` command.

---

---

## About the `mpirun` Command

The `mpirun` command controls several aspects of program execution in Open MPI. `mpirun` uses the Open Run-Time Environment (ORTE) to launch jobs. If you are running under distributed resource manager software, such as Sun Grid Engine or PBS, ORTE launches the resource manager for you.

If you are using `rsh/ssh` instead of a resource manager, you must use a hostfile or host list to identify the hosts on which the program will be run. When you issue the `mpirun` command, you specify the name of the hostfile or host list on the command line; otherwise, `mpirun` executes all the copies of the program on the local host, in round-robin sequence by CPU slot. For more information about hostfiles and their syntax, see [“Specifying Hosts By Using a Hostfile” on page 25](#).

Both MPI programs and non-MPI programs can use `mpirun` to launch the user processes.

Some example programs are provided in the `/opt/SUNWhpc/HPC8.2/examples` directory for you to try to compile/run as sanity tests.

---

## Syntax for the `mpirun` Command

The following example shows the general single-process syntax for `mpirun`:

```
% mpirun [options] [program-name]
```

For a simple SPMD (Single Process, Multiple Data) job, the typical syntax is:

```
% mpirun -np x program-name
```

For jobs involving multiple instructions, the command syntax appears similar to the following:

```
% mpirun [options] [program-name] : [options2] [program-name2] ...
```

For an MPMD (Multiple Program, Multiple Data) parallel application, the syntax follows this form:

```
% mpirun -np x program1 : -np y program2
```

This command starts  $x$  number of copies of the program `program1`, and then starts  $y$  copies of the program `program2`.

## `mpirun` Options

The *options* control the behavior of the `mpirun` command. They might or might not be followed by arguments.



---

**Caution** – If you do not specify an argument for an option that expects to be followed by an argument (for example, the `--app <filename>` option), that option will read the next option on the command line as an argument. This might result in inconsistent behavior.

---

The procedure “[To Display Command Help \(-h\)](#)” on page 32 lists the options in alphabetical order, with a brief description of each.

## Using Environment Variables With the `mpirun` Command

Use the `-x args` option (where *args* is the environment variable(s) you want to use) to specify any environment variable you want to pass during runtime. The `-x` option exports the variable specified in *args* and sets the value for *args* from the current environment. For example:

```
% mpirun -x LD_LIBRARY_PATH=/opt/SUNWhpc/HPC8.2/lib -np 4 a.out
```

## Using MCA Parameters With the `mpirun` Command

The `mpirun` command uses MCA (Multiple Component Architecture) parameters to pass environment variables. To specify an MCA parameter, use the `-mca` option with the `mpirun` command, and then specify the parameter type, the parameter you want to pass as an environment variable, and the value you want to set. For example:

```
% mpirun --mca mpi_show_handle_leaks 1 -np 4 a.out
```

This sets the MCA parameter `mpi_show_handle_leaks` to the value of 1 before running the program named `a.out` with four processes. In general, the format used on the command line is `--mca parameter_name value`.

---

**Note** – There are multiple ways to specify the values of MCA parameters. This chapter discusses how to use them from the command line with the `mpirun` command. MCA parameters are discussed in more detail in [Chapter 7](#).

---

# Canceling Send and Receive Operations

Open MPI supports the canceling of receive operations. However, the canceling of sends is not supported; therefore, a send will never be successfully canceled.

For more information about canceling send and receive operations, see the `MPI_Cancel(3)` man page.

---

## mpirun Command Examples

The examples in this section show how to use the `mpirun` command options to specify how and where the processes and programs run.

The following table shows the process control options for the `mpirun` command. The procedures that follow the table explain how these options are used and show the syntax for each.

**TABLE 5-1** Program/Process Control Options

Task	mpirun option	Page Number (For More Information)
To run a program with default settings	(no need to specify an option)	18
To run multiple parallel processes	<code>-c</code> or <code>-np &lt;number of processes&gt;</code>	19
To display command help	<code>-h</code> or <code>--help</code>	27
To change the working directory	<code>-wdir</code> or <code>--wdir &lt;directory&gt;</code>	26
To specify the list of hosts on which to invoke processes (also known as the rankmap string)	<code>-host</code> or <code>--host</code> or <code>-H</code>	22
To specify the list of hosts on which to execute the program (also known as the rankmap file)	<code>-hostfile &lt;filename&gt;</code> or <code>--hostfile &lt;filename&gt;</code> or <code>-machinefile &lt;filename&gt;</code> or <code>--machinefile &lt;filename&gt;</code>	21

**TABLE 5-1** (Continued) Program/Process Control Options

Task	mpirun option	Page Number (For More Information)
To start up in debugging mode	-d or --debug or -debugger or --debugger <sequence>	27
To specify verbose output	-v	
To specify multiple executables	-np 2 exe1 : -np 6 exe2	

## ▼ To Run a Program With Default Settings

- To run the program with default settings, enter the command and program name, followed by any required arguments to the program:

```
% mpirun program-name
```

## ▼ To Run Multiple Processes

By default, an MPI program started with `mpirun` runs as one process.

- To run the program as multiple processes, use the `-np` option:

```
% mpirun -np process-count program-name
```

When you request multiple processes, ORTE attempts to start the number of processes you request, regardless of the number of CPUs available to run those processes. For more information, see [“Oversubscribing Nodes” on page 27](#).

## ▼ To Direct mpirun By Using an Appfile

You can use a type of text file (called an appfile) to direct `mpirun`. The appfile specifies the nodes on which to run, the number of processes to launch on each node, and the programs to execute in a parallel application. When you use the `--app` option, `mpirun` takes all its direction from the contents of the appfile and ignores any other nodes or processes specified on the command line.

For example the following shows an appfile called `my_appfile`:

```
# Comments are supported; comments begin with #
# Application context files specify each sub-application in the
# parallel job, one per line. The first sub-application is the 2
# a.out processes:
-np 2 a.out
# The second sub-application is the 2 b.out processes:
-np 2 b.out
```

- **To use the `--app` option with the `mpirun` command, specify the name and path of the appfile on the command line. For example:**

```
% mpirun --app my_appfile
```

This command produces the same results as running `a.out` and `b.out` from the command line.

---

## Mapping MPI Processes to Nodes

When you issue the `mpirun` command from the command line, ORTE reads the number of processes to be launched from the `-np` option, and then determines where the processes will run.

To determine where the processes will run, ORTE uses the following criteria:

- Available hosts (also referred to as nodes), specified by a hostfile or by the `--host` option
- Scheduling policy (round-robin or by-slot)
- Default and maximum numbers of slots available on each host
- ORTE also checks to see whether the current environment/shell runs with any third-party launcher (such as Sun Grid Engine or PBS) to determine where the processes will launch.

## Specifying Available Hosts

You specify the available hosts to Open MPI in three ways:

- Through the batch scheduler in your resource management software. This option is described in detail in [Chapter 6](#).

- By using a hostfile with the `--hostfile` option. The hostfile is a text file that contains the names of hosts, the number of available slots on each host, and the maximum slots on each host.
- By using the `--host` option. Use this option to specify which hosts to include or exclude.

## Specifying Hosts By Using a Hostfile

The hostfile lists each node, the available number of slots, and the maximum number of slots on that node. For example, the following listing shows a simple hostfile:

```
node0
node1 slots=2
node2 slots=4 max_slots=4
node3 slots=4 max_slots=20
```

In this example file, `node0` is a single-processor machine. `node1` has two slots. `node2` and `node3` both have 4 slots, but the values of `slots` and `max_slots` are the same (4) on `node2`. This disallows the processors on `node2` from being oversubscribed. The four slots on `node3` can be oversubscribed, up to a maximum of 20 processes.

When you use this hostfile with the `--nooversubscribe` option (see [“Oversubscribing Nodes” on page 27](#)), `mpirun` assumes that the value of `max_slots` for each node in the hostfile is the same as the value of `slots` for each node. It overrides the values for `max_slots` set in the hostfile.

Open MPI assumes that the maximum number of slots you can specify is equal to infinity, unless explicitly specified. Resource managers also do not specify the maximum number of available slots.

---

**Note** – Open MPI includes a commented default hostfile at `/opt/SUNWhpc/HPC8.2/etc/openmpi-default-hostfile`. Unless you specify a different hostfile at a different location, this is the hostfile that OpenMPI uses. It is empty by default, but you may edit this file to add your list of nodes. See the comments in the hostfile for more information.

---

## Specifying Hosts By Using the `--host` Option

You can use the `--host` option to `mpirun` to specify the hosts you want to use on the command line in a comma-delimited list. For example, the following command directs `mpirun` to run a program called `a.out` on hosts `a`, `b`, and `c`:

```
% mpirun -np 3 --host a,b,c a.out
```

Open MPI assumes that the default number of slots on each host is one, unless you explicitly specify otherwise.

### ▼ To Specify Multiple Slots Using the `--host` Option

To specify multiple slots with the `--host` option for each host repeat the host name on the command line for each slot you want to use. For example:

```
% mpirun -host node1,node1,node2,node2 ...
```

If you are using a resource manager such as Sun Grid Engine or PBS, the resource manager maintains an accurate count of available slots.

## Excluding Hosts From Scheduling By Using the `--host` Option

You can also use the `--host` option in conjunction with a hostfile to exclude any nodes not explicitly specified on the command line. For example, assume that you have the following hostfile called `my_hosts`:

```
a slots=2 max_slots=20
b slots=2 max_slots=20
c slots=2 max_slots=20
d slots=2 max_slots=20
```

Suppose you issue the following command to run program `a.out`:

```
% mpirun -np 1 --hostfile my_hosts --host c a.out
```

This command launches one instance of `a.out` on host `c`, but excludes the other hosts in the hostfile (`a`, `b`, and `d`).

---

**Note** – If you use these two options (`--hostfile` and `--host`) together, make sure that the host(s) you specify using the `--host` option also exist in the hostfile. Otherwise, `mpirun` exits with an error.

---

## Oversubscribing Nodes

If you schedule more processes to run than there are available slots, this is referred to as *oversubscribing*. Oversubscribing a host is not suggested, as it might result in performance degradation.

`mpirun` has a `--nooversubscribe` option. This option implicitly sets the `max_slots` value (maximum number of available slots) to the same value as the `slots` value for each node, as specified in your hostfile. If the number of processes requested is greater than the `slots` value, `mpirun` returns an error and does not execute the command. This option overrides the value set for `max_slots` in your hostfile.

For more information about oversubscribing, see the following URL:

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

## Scheduling Policies

ORTE uses two types of scheduling policies when it determines where processes will run:

- By slot (default). This scheme schedules processes to run on each successive slot on one host. When all those slots are filled, scheduling begins on the next host in the hostfile.
- By node. In this scheme, Open MPI schedules the processes by finding the first available slot on a host, then the first available slot on the next host in the hostfile, and so on, in a round-robin fashion.

## Scheduling By Slot

This is the default scheduling policy for Open MPI. If you do not specify a scheduling policy, this is the policy that is used.

In by-slot scheduling, Open MPI schedules processes on a node until all of its available slots are exhausted (that is, all slots are running processes) before proceeding to the next node. In MPI terms, this means that Open MPI tries to maximize the number of adjacent ranks in MPI\_COMM\_WORLD on the same host without oversubscribing that host.

## ▼ To Specify By-Slot Scheduling

If you want to explicitly specify by-slot scheduling for some reason, there are two ways to do it:

1. Specify the `--byslot` option to `mpirun`. For example, the following command specifies the `--byslot` and `--hostfile` options:

```
% mpirun -np 4 --byslot --hostfile myfile a.out
```

The following example uses the `-host` option:

```
% mpirun -np 4 --byslot -host node0,node0,node1,node1 a.out
```

2. Set the MCA parameter `rmaps_base_schedule_policy` to the value `slot`. For example:

```
% mpirun --mca rmaps_base_schedule_policy slot -np 4 a.out
```

---

**Note** – The examples in this chapter set MCA parameters on the command line. For more information about the ways in which you can set MCA parameters, see [Chapter 7](#). In addition, the Open MPI FAQ contains information about MCA parameters at the following URL:

<http://www.open-mpi.org/faq/?category=tuning#setting-mca-params>

---

The following output example shows the contents of a simple hostfile called `my-hosts` and the results of the `mpirun` command using by-slot scheduling.

```
% cat my-hosts
node0 slots=2 max_slots=20
node1 slots=2 max_slots=20
% mpirun --hostfile my-hosts -np 8 --byslot hello | sort
Hello World I am rank 0 of 8 running on node0
Hello World I am rank 1 of 8 running on node0
Hello World I am rank 2 of 8 running on node1
Hello World I am rank 3 of 8 running on node1
Hello World I am rank 4 of 8 running on node0
Hello World I am rank 5 of 8 running on node0
Hello World I am rank 6 of 8 running on node1
Hello World I am rank 7 of 8 running on node1
```

## Scheduling By Node

In by-node scheduling, Open MPI schedules a single process on each node in a round-robin fashion (looping back to the beginning of the node list as necessary) until all processes have been scheduled. Nodes are skipped once their default slot counts are exhausted.

### ▼ To Specify By-Node Scheduling

There are two ways to specify by-node scheduling:

- Specify the `--bynode` option to `mpirun`. For example:

```
% mpirun -np 4 --bynode --hostfile my-hosts a.out
```

- Set the MCA parameter `rmaps_base_schedule_policy` to the value `node`. For example:

```
% mpirun --mca rmaps_base_schedule_policy node -np 4 a.out
```

The following output example shows the contents of the same hostfile used in the previous example and the results of the `mpirun` command using by-node scheduling.

```
% cat my-hosts
node0 slots=2 max_slots=20
node1 slots=2 max_slots=20
% mpirun --hostfile my-hosts -np 8 --bynode hello | sort
Hello World I am rank 0 of 8 running on node0
Hello World I am rank 1 of 8 running on node1
Hello World I am rank 2 of 8 running on node0
Hello World I am rank 3 of 8 running on node1
Hello World I am rank 4 of 8 running on node0
Hello World I am rank 5 of 8 running on node1
Hello World I am rank 6 of 8 running on node0
Hello World I am rank 7 of 8 running on node1
```

## Comparing By-Slot to By-Node Scheduling

In the examples in this section, `node0` and `node1` each have two slots. The diagrams show the differences in scheduling between the two methods.

By-slot scheduling for the two nodes can be represented as follows:

node0	node1
0	2
1	3
4	6
5	7

By-node scheduling for the same two nodes can be represented this way:

node0	node1
0	1
2	3
4	5
6	7

---

## Controlling Input/Output

Open MPI directs UNIX standard input to `/dev/null` on all processes except the rank 0 process of `MPI_COMM_WORLD`. The `MPI_COMM_WORLD` rank 0 process inherits standard input from `mpirun`. The node from which you invoke `mpirun`

need not be the same as the node where the MPI\_COMM\_WORLD rank 0 process resides. Open MPI handles the redirection of the `mpirun` standard input to the rank 0 process.

Open MPI directs UNIX standard output and standard error from remote nodes to the node that invoked `mpirun`, and then prints the information from the remote nodes on the standard output/error of `mpirun`. Local processes inherit the standard output/error of `mpirun` and transfer to it directly.

## ▼ To Redirect Standard I/O

To redirect standard I/O for Open MPI applications, use the typical shell redirection procedure on `mpirun`. For example:

```
% mpirun -np 2 my_app < my_input > my_output
```

In this example, only the MPI\_COMM\_WORLD rank 0 process will receive the stream from `my_input` on `stdin`. The `stdin` on all the other nodes will be tied to `/dev/null`. However, the `stdout` from all nodes will be collected into the `my_output` file.

---

## Controlling Other Job Attributes

To Perform This Task	Use This Option
To change the working directory	<code>-wdir</code> or <code>--wdir</code>
To display debugging output	<code>-d</code>
To display command help	<code>-h</code>

## ▼ To Change the Working Directory

Use the `-wdir` or `--wdir` option to specify the path of an alternative working directory to be used by the processes spawned when you run your program:

```
% mpirun --wdir working-directory program-name
```

Setting a path with `--wdir` does not affect where the runtime environment looks for executables. If you do not specify `--wdir`, the default is the current working directory. For example:

```
% mpirun --wdir /home/mystuff/bin a.out
```

The syntax above changes the working directory for `a.out` to `/home/mystuff/bin`.

## ▼ To Specify Debugging Output

Use this syntax to specify debugging output. For example:

```
% mpirun -d a.out
```

The `-d` option shows the user-level debugging output for all of the ORTE modules used with `mpirun`. To see more information from a particular module, you can set additional MCA debugging parameters. The availability of the additional debugging information depends on how the module of interest is implemented.

For more information on MCA parameters, see [Chapter 7](#). For more information about whether a module provides additional verbose or debug mode, run the `ompi_info` command on that module.

## ▼ To Display Command Help (-h)

To display a list of `mpirun` options, use the `-h` option (alone). The following example shows the output from `mpirun -h`:

```
% ./mpirun -h
mpirun (Open MPI) 1.3.3r21324-ct8.2-b09j-r40

Usage: mpirun [OPTION]... [PROGRAM]...
Start the given program using Open RTE

  -am <arg0>           Aggregate MCA parameter set file list
  --app <arg0>         Provide an appfile; ignore all other command line
                       options
  -bynode|--bynode     Whether to allocate/map processes round-robin by
                       node
  -byslot|--byslot     Whether to allocate/map processes round-robin by
```

```

slot (the default)
-c|-np|--np <arg0>      Number of processes to run
  -cf|--cartofile <arg0>
                        Provide a cartography file
-d|-debug-devel|--debug-devel
                        Enable debugging of OpenRTE
  -debug|--debug        Invoke the user-level debugger indicated by the
                        orte_base_user_debugger MCA parameter
  -debug-daemons|--debug-daemons
                        Enable debugging of any OpenRTE daemons used by
                        this application
  -debug-daemons-file|--debug-daemons-file
                        Enable debugging of any OpenRTE daemons used by
                        this application, storing output in files
  -debugger|--debugger <arg0>
                        Sequence of debuggers to search for when "--debug"
                        is used
  -default-hostfile|--default-hostfile <arg0>
                        Provide a default hostfile
  -display-allocation|--display-allocation
                        Display the allocation being used by this job
  -display-devel-allocation|--display-devel-allocation
                        Display a detailed list (mostly intended for
                        developers) of the allocation being used by this
                        job
  -display-devel-map|--display-devel-map
                        Display a detailed process map (mostly intended for
                        developers) just before launch
  -display-map|--display-map
                        Display the process map just before launch
  -do-not-launch|--do-not-launch
                        Perform all necessary operations to prepare to
                        launch the application, but do not actually launch
                        it
  -do-not-resolve|--do-not-resolve
                        Do not attempt to resolve interfaces
  -gmca|--gmca <arg0> <arg1>
                        Pass global MCA parameters that are applicable to
                        all contexts (arg0 is the parameter name; arg1 is
                        the parameter value)
-h|--help              This help message
-H|-host|--host <arg0> List of hosts to invoke processes on
  --hetero              Indicates that multiple app_contexts are being
                        provided that are a mix of 32/64 bit binaries
  -hostfile|--hostfile <arg0>
                        Provide a hostfile
  -launch-agent|--launch-agent <arg0>
                        Command used to start processes on remote nodes
                        (default: orted)

```

```

-leave-session-attached|--leave-session-attached
    Enable debugging of OpenRTE
-loadbalance|--loadbalance
    Balance total number of procs across all allocated
    nodes
-machinefile|--machinefile <arg0>
    Provide a hostfile
-mca|--mca <arg0> <arg1>
    Pass context-specific MCA parameters; they are
    considered global if --gmca is not used and only
    one context is specified (arg0 is the parameter
    name; arg1 is the parameter value)
-n|--n <arg0>
    Number of processes to run
-nolocal|--nolocal
    Do not run any MPI applications on the local node
-nooversubscribe|--nooversubscribe
    Nodes are not to be oversubscribed, even if the
    system supports such operation
--noprefix
    Disable automatic --prefix behavior
-npernode|--npernode <arg0>
    Launch n processes per node on all allocated nodes
-ompi-server|--ompi-server <arg0>
    Specify the URI of the Open MPI server, or the name
    of the file (specified as file:filename) that
    contains that info
-output-filename|--output-filename <arg0>
    Redirect output from application processes into
    filename.rank.
-path|--path <arg0>
    PATH to be used to look for executables to start
    processes
-pernode|--pernode
    Launch one process per available node on the
    specified number of nodes [no -np => use all
    allocated nodes]
--prefix <arg0>
    Prefix where Open MPI is installed on remote nodes
--preload-files <arg0>
    Preload the comma separated list of files to the
    remote machines current working directory before
    starting the remote process.
--preload-files-dest-dir <arg0>
    The destination directory to use in conjunction
    with --preload-files. By default the absolute and
    relative paths provided by --preload-files are
    used.
-q|--quiet
    Suppress helpful messages
-report-pid|--report-pid
    Print out pid
-rf|--rankfile <arg0>
    Provide a rankfile file
-s|--preload-binary
    Preload the binary on the remote machine before
    starting the remote process.

```

```

-server-wait-time|--server-wait-time <arg0>
    Time in seconds to wait for ompi-server (default:
    10 sec)
-slot-list|--slot-list <arg0>
    List of processor IDs to bind MPI processes to
    (e.g., used in conjunction with rank files)
-stdin|--stdin <arg0>
    Specify procs to receive stdin [rank, none]
    (default: 0, indicating rank 0)
-tag-output|--tag-output
    Tag all output with [job,rank]
-timestamp-output|--timestamp-output
    Timestamp all application process output
-tmpdir|--tmpdir <arg0>
    Set the root for the session directory tree for
    orterun ONLY
-tv|--tv
    Deprecated backwards compatibility flag; synonym
    for "--debug"
-v|--verbose
    Be verbose
-V|--version
    Print version and exit
-wait-for-server|--wait-for-server
    If ompi-server is not already running, wait until
    it is detected (default: false)
-wd|--wd <arg0>
    Synonym for --wdir
-wdir|--wdir <arg0>
    Set the working directory of the started processes
-x <arg0>
    Export an environment variable, optionally
    specifying a value (e.g., "-x foo" exports the
    environment variable foo and takes its value from
    the current environment; "-x foo=bar" exports the
    environment variable name foo and sets its value to
    "bar" in the started processes)
-xml|--xml
    Provide all output in XML format
-xterm|--xterm <arg0>
    Create a new xterm window and display output from
    the specified ranks there

```

Report bugs to <http://www.open-mpi.org/community/help/>

---

# Submitting Jobs Under Sun Grid Engine Integration

There are two ways to submit jobs under Sun Grid Engine integration: interactive mode and batch mode. The instructions in this chapter describe how to submit jobs interactively. For information about how to submit jobs in batch mode, see [Chapter 6](#).

## Defining Parallel Environment (PE) and Queue

A PE needs to be defined for all the queues in the Sun Grid Engine cluster to be used as ORTE nodes. Each ORTE node should be installed as an Sun Grid Engine execution host. To allow the ORTE to submit a job from any ORTE node, configure each ORTE node as a submit host in Sun Grid Engine.

Each execution host must be configured with a default queue. In addition, the default queue set must have the same number of slots as the number of processors on the hosts.

### ▼ To Use PE Commands

- To display a list of available PEs (parallel environments), type the following:

```
% qconf -spl  
make
```

- To define a new PE, you must have Sun Grid Engine manager or operator privileges. Use a text editor to modify a template for the PE. The following example creates a PE named `orte`.

```
% qconf -ap orte
```

- To modify an existing PE, use this command to invoke the default editor:

```
% qconf -mp orte
```

- To show a particular PE that has been defined, type this command:

```
% qconf -sp orte
pe_name      orte
slots       8
user_lists   NONE
xuser_lists  NONE
start_proc_args /bin/true
stop_proc_args /bin/true
allocation_rule $round_robin
control_slaves TRUE
job_is_first_task FALSE
urgency_slots min
```

The value NONE in `user_lists` and `xuser_lists` mean enable everybody and exclude nobody.

The value of `control_slaves` must be TRUE; otherwise, `qcrsh` exits with an error message.

The value of `job_is_first_task` must be FALSE or the job launcher consumes a slot. In other words, `mpirun` itself will count as one of the slots and the job will fail, because only `n-1` processes will start.

## ▼ To Use Queue Commands

- To show all the defined queues, type the following command:

```
% qconf -sql
all.q
```

The queue `all.q` is set up by default in Sun Grid Engine.

- To configure the `orte` PE from the example in the previous section to the existing queue, type the following:

```
% qconf -mattr queue pe_list "orte" all.q
```

You must have Sun Grid Engine manager or operator privileges to use this command.

# Submitting Jobs in Interactive Mode

## ▼ To Set the Interactive Display

Before you submit a job, you must have your `DISPLAY` environment variable set so that the interactive window will appear on your desktop, if you have not already done so.

For example, if you are working in the C shell, type the following command:

```
% setenv DISPLAY desktop:0.0
```

## ▼ To Submit Jobs Interactively

1. Use the `source` command to set the Sun Grid Engine environment variables from a file:

```
mynode4% source /opt/sge/default/common/settings.csh
```

2. Use the `qsh` command to start the interactive X Windows session, and specify the parallel environment (in this example, ORTE) and the number of slots to use:

```
mynode4% qsh -pe orte 2
waiting for interactive job to be scheduled...
Your interactive job 324 has been successfully scheduled.
```

3. On a different node in the cluster, use the `cd` command to switch to the directory where your executable is located.

```
mynode5% cd /workspace/joeuser/ompi/trunk/builds/sparc32-g/bin
```

4. Issue the `mpirun` command.

```
mynode5% /opt/SUNWhpc/HPC8.2/sun/bin/mpirun -np 4 hostname
```

In the above example, Sun Grid Engine starts the user executable `hostname` with 4 processes on the two Sun Grid Engine assigned slots. The following example shows the output from the `mpirun` command with the specified options.

```
mynode5% /opt/SUNWhpc/HPC8.2/sun/bin/mpirun -np 4 --hostname mynode5
mynode5
```

## ▼ To Verify That Sun Grid Engine Is Running

The following is not required for normal operation, but if you want to verify that Sun Grid Engine is being used, add `--mca ras_gridengine_verbose` to the `mpirun` command line. For example:

```
% ./mpirun -np 4 -mca ras_gridengine_verbose 100 hostname
[mynode6:04234] ras:gridengine: JOB_ID: 28
[mynode6:04234] ras:gridengine: mynode6: PE_HOSTFILE shows slots=2
[mynode6:04234] ras:gridengine: mynode7: PE_HOSTFILE shows slots=2
mynode6
mynode6
mynode7
mynode7
%
```

## ▼ To Start an Interactive Session Using `qrsh`

An alternate way to start an interactive session is by using `qrsh` instead of `qsh`. For example:

```
% qrsh -V -pe orte 8 mpirun -np 4 -byslot hostname
```

---

# Using MPI Client/Server Applications

The instructions in this section explain how to get best results when starting Open MPI client/server applications.

## ▼ To Launch the Client/Server Job

1. Type the following command to launch the server application. Substitute the name of your MPI job's universe for `univ1`:

```
% ./mpirun -np 1 --universe univ1 t_accept
```

2. Type the following command to launch the client application, substituting the name of your MPI job's universe for `univ1`:

```
% ./mpirun -np 4 --universe univ1 t_connect
```

If the client and server jobs span more than 1 node, the first job (that is, the server job) must specify on the mpirun command line all the nodes that will be used. Specifying the node names allocates the specified hosts from the entire universe of server and client jobs.

For example, if the server runs on node0 and the client job runs on node1 only, the command to launch the server must specify both nodes (using the `-host node0,node1` flag) even it uses only one process on node0.

Assuming that the persistent daemon is started on node0, the command to launch the server would look like this:

```
node0% ./mpirun -np 1 --universe univ1 -host node0,node1 t_accept
```

The command to launch the client is:

```
node0% ./mpirun -np 4 --universe univ1 -host node1 t_connect
```

## Using Name Publishing

If you are planning on using name publishing, you must perform some additional tasks. You need to start up an ompi-server process on your server so that both the clients and servers can exchange information using that server.

For information about how to start the ompi-server process, type the following command on your server:

```
% man ompi-server
```

# Troubleshooting Client/Server Jobs

If the MPI client/server job fails to start, you might see error messages similar to this:

```
node0% ./orted --persistent --seed --scope public --universe univ4
--debug
[node0:21760] procdir: (null)
[node0:21760] jobdir: (null)
[node0:21760] unidir:
/tmp/openmpi-sessions-joeuser@node0_0/univ4
[node0:21760] top: openmpi-sessions-joeuser@node0_0
[node0:21760] tmp: /tmp
[node0:21760] orte_init: could not contact the specified
universe name univ4
[node0:21760] [NO-NAME] ORTE_ERROR_LOG: Unreachable in file
/opt/SUNWhpc/HPC8.2/sun/bin/orted/runtime/orte_init_stagel.c
at line 221
```

These messages indicate that there is residual data left in the `/tmp` directory. This can happen if a previous client/server job has already run from the same node.

To empty the `/tmp` directory, use the `orte-clean` utility. For more information about `orte-clean`, see the `orte-clean` man page.

You might also need to run `orte-clean` if you see error messages similar to the following:

```
node0% ./orted --persistent --seed --scope public --universe univ4 --debug
[node0:21760] procdir: (null)
[node0:21760] jobdir: (null)
[node0:21760] unidir:
/tmp/openmpi-sessions-joeuser@node0_0/univ4
[node0:21760] top: openmpi-sessions-joeuser@node0_0
[node0:21760] tmp: /tmp
[node0:21760] orte_init: could not contact the specified
universe name univ4
[node0:21760] [NO-NAME] ORTE_ERROR_LOG: Unreachable in file
/opt/SUNWhpc/HPC8.2/sun/bin/orted/runtime/orte_init_stagel.c
at line 221
-----
It looks like orte_init failed for some reason; your parallel process is likely
to abort. There are many reasons that a parallel process can fail during
orte_init; some of which are due to configuration or environment problems. This
failure appears to be an internal failure; here's some additional information
(which may only be relevant to an Open MPI developer):
    orte_sds_base_contact_universe failed
```

```
--> Returned value -12 instead of ORTE_SUCCESS
```

```
-----  
[node0:21760] [NO-NAME] ORTE_ERROR_LOG: Unreachable in file  
/opt/SUNWhpc/HPC8.2/sun/bin/orted/runtime/orte_system_init.c  
at line 42
```

```
[node0:21760] [NO-NAME] ORTE_ERROR_LOG: Unreachable in file  
/opt/SUNWhpc/HPC8.2/sun/bin/orte/runtime/orte_init.c  
at line 52
```

```
Open RTE was unable to initialize properly. The error occurred while attempting  
to orte_init(). Returned value -12 instead of ORTE_SUCCESS.
```

---

## For More Information

For more information about the `mpirun` command and its options, see the following:

- [Chapter 7, “Using MCA Parameters With `mpirun`” on page 55](#)
- the `mpirun(3)` man page
- Open MPI FAQ at <http://www.open-mpi.org>

# Running Programs With `mpirun` in Distributed Resource Management Systems

---

This chapter describes the options to the `mpirun` command that are used for distributed resource management, and provides instructions for each resource manager. It contains the following sections:

- [“`mpirun` Options for Third-Party Resource Manager Integration”](#) on page 43
- [“Running Parallel Jobs in the PBS Environment”](#) on page 44
- [“Running Parallel Jobs in the Sun Grid Engine Environment”](#) on page 47

---

## `mpirun` Options for Third-Party Resource Manager Integration

ORTE is compatible with a number of other launchers, including `rsh/ssh`, Sun Grid Engine, and PBS.

---

**Note** – Open MPI itself supports other third-party launchers supported by Open MPI, such as SLURM and Torque. However, these launchers are currently not supported in Sun HPC ClusterTools software. To use these other third-party launchers, you must download the Open MPI source, compile, and link with the libraries for the launchers.

---

## Checking Your Open MPI Configuration

To see whether your Open MPI installation has been configured for use with the third-party resource manager you want to use, issue the `mpi_info` command and pipe the output to `grep`. The following examples show how to use `mpi_info` to check for the desired third-party resource manager.

### ▼ To Check for `rsh/ssh`

To see whether your Open MPI installation has been configured to use the `rsh/ssh` launcher:

```
% mpi_info | grep rsh  
MCA plm: rsh (MCA v2.0, API v2.0, Component v1.3)
```

### ▼ To Check for PBS/Torque

To see whether your Open MPI installation has been configured to use the PBS/Torque launcher:

```
% mpi_info | grep tm  
MCA ras: tm (MCA v2.0, API v2.0, Component v1.3)  
MCA plm: tm (MCA v2.0, API v2.0, Component v1.3)
```

### ▼ To Check for Sun Grid Engine

To see whether your Open MPI installation has been configured to use Sun Grid Engine:

```
% mpi_info | grep gridengine  
MCA ras: gridengine (MCA v2.0, API v2.0, Component v1.3)
```

---

## Running Parallel Jobs in the PBS Environment

If your Open MPI environment is set up to include PBS, Open MPI automatically detects when `mpirun` is running within PBS, and will execute properly.

First reserve the number of resources by invoking the `qsub` command with the `-l` option. The `-l` option specifies the number of nodes and the number of processes per node. For example, this command sequence reserves four nodes with four processes per node for the job `myjob.sh`:

```
% qsub -l nodes=4:ppn=4 myjob.sh
```

When you enter the PBS environment, you can launch an individual job or a series of jobs with `mpirun`. The `mpirun` command launches the job using the nodes and processes information from PBS. The resource information is accessed using the `tm` calls provided by PBS; hence, `tm` is the name used to identify the module in ORTE. The job ranks are children of PBS, not ORTE.

You can run an ORTE job within the PBS environment in two different ways: interactive and scripted.

## ▼ To Run an Interactive Job in PBS

1. Enter the PBS environment interactively with the `-I` option to `qsub`, and use the `-l` option to reserve resources for the job.

Here is an example.

```
% qsub -l nodes=2:ppn=2 -I
```

The command sequence shown above enters the PBS environment and reserves one node called `mynode` with two processes for the job. Here is the output:

```
qsub: waiting for job 20.mynode to start
qsub: job 20.mynode ready
Sun Microsystems Inc. SunOS 5.10 Generic June 2006
pbs%
```

2. Launch the `mpirun` command.

Here is an example that launches the `hostname` command with a verbose output:

```
pbs% /opt/SUNWhpc/HPC8.2/sun/bin/mpirun -np 4 -mca plm_tm_verbose 1 hostname
```

The output shows the hostname program being run on ranks r0 and r1:

```
% /opt/SUNWhpc/HPC8.2/sun/bin/mpirun -np 4 -mca plm_tm_verbose 1
hostname
[hostname1:09064] plm:tm: launching on node mynode1
[hostname2:09064] plm:tm: launching on node mynode2
hostname2
hostname1
hostname2
hostname1
```

## ▼ To Run a Batch Job in PBS

### 1. Write a script that calls mpirun.

In the following examples, the script is called `myjob.csh`. The system is called `mynode`. Here is an example of the script.

```
#!/bin/csh

/opt/SUNWhpc/HPC8.2/sun/bin/mpirun -np 2 -mca plm_tm_verbose 1
hostname
```

### 2. Enter the PBS environment and use the `-l` option to `qsub` to reserve resources for the job.

Here is an example of how to use the `-l` option with the `qsub` command.

```
% qsub -l nodes=2:ppn=2 myjob.csh
```

This command enters the PBS environment and reserves one node with two processes for the job that will be launched by the script named `myjob.csh`.

Here is the output to the script `myjob.csh`.

```
% more myjob.csh.*
::::::::::::
myjob.csh.e2365
::::::::::::
::::::::::::
myjob.csh.o2365
::::::::::::
Warning: no access to tty (Bad file number).
Thus no job control in this shell.
Sun Microsystems Inc.   SunOS 5.10           Generic January 2005
hostname5
hostname4
hostname5
hostname4
```

After the job finishes, it generates two output files:

- `name_of_job.ejob_id` is the file that shows the error outputs (for example, `myjob.csh.e2365` in the above example).
- `name_of_job.ojob_id` is the file that shows the standard outputs (for example, `myjob.csh.o2365` in the above example).

As you can see, the `pbsrun` command calls `mpirun`, which forks into two calls of the `hostname` program, one for each node.

---

## Running Parallel Jobs in the Sun Grid Engine Environment

Sun Grid Engine 6.1 is the supported version of Sun Grid Engine for Sun HPC ClusterTools 8.2.

Before you can run parallel jobs, make sure that you have defined the parallel environment and queue before running the job.

## Defining Parallel Environment (PE) and Queue

A PE needs to be defined for all the queues in the Sun Grid Engine cluster to be used as ORTE nodes. Each ORTE node should be installed as a Sun Grid Engine execution host. To allow the ORTE to submit a job from any ORTE node, configure each ORTE node as a submit host in Sun Grid Engine.

Each execution host must be configured with a default queue. In addition, the default queue set must have the same number of slots as the number of processors on the hosts.

### ▼ To Use PE Commands

- To display a list of available PEs (parallel environments), type the following:

```
% qconf -spl  
make
```

- To define a new PE, you must have Sun Grid Engine manager or operator privileges. Use a text editor to modify a template for the PE. The following example creates a PE named `orte`.

```
% qconf -ap orte
```

- To modify an existing PE, use this command to invoke the default editor:

```
% qconf -mp orte
```

- To show a particular PE that has been defined, type this command:

```
% qconf -sp orte  
pe_name      orte  
slots        8  
user_lists   NONE  
xuser_lists  NONE  
start_proc_args /bin/true  
stop_proc_args /bin/true  
allocation_rule $round_robin  
control_slaves TRUE  
job_is_first_task FALSE  
urgency_slots  min
```

The value `NONE` in `user_lists` and `xuser_lists` mean enable everybody and exclude nobody.

The value of `control_slaves` must be `TRUE`; otherwise, `qcrsh` exits with an error message.

The value of `job_is_first_task` must be `FALSE` or the job launcher consumes a slot. In other words, `mpirun` itself will count as one of the slots and the job will fail, because only `n-1` processes will start.

## ▼ To Use Queue Commands

- To show all the defined queues, type the following command:

```
% qconf -spl  
all.q
```

The queue `all.q` is set up by default in Sun Grid Engine.

- To configure the `orte` PE from the example in the previous section to the existing queue, type the following:

```
% qconf -mattr queue pe_list "orte" all.q
```

You must have Sun Grid Engine manager or operator privileges to use this command.

## Submitting Jobs Under Sun Grid Engine Integration

There are two ways to submit jobs under Sun Grid Engine integration: interactive mode and batch mode. The instructions in this section describe how to submit jobs in batch mode. For information about how to use interactive mode, see [Chapter 5](#).

## ▼ To Set the Interactive Display

Before you submit a job, you must have your `DISPLAY` environment variable set so that the interactive window will appear on your desktop, if you have not already done so.

For example, if you are working in the C shell, type the following command:

```
% setenv DISPLAY desktop:0.0
```

## ▼ To Submit Jobs in Batch Mode

---

**Note** – Before you can use the parallel environment, make sure that you have set it up before running the job. See [“Defining Parallel Environment \(PE\) and Queue” on page 48](#) for more information.

---

1. **Create the script. In this example, `mpirun` is embedded within a script to `qsub`.**

```
mynode4% cat SGE.csh
#!/usr/bin/csh

# set PATH: including location of MPI program to be run
setenv PATH /opt/SUNWhpc/HPC8.2/examples/connectivity:${PATH}

mpirun -np 4 -mca ras_gridengine_verbose 100 connectivity.sparc -v
```

---

**Note** – The `--mca ras_gridengine_verbose 100` setting is used in this example only to show that Sun Grid Engine is being used. This would not be needed for normal operation.

---

2. **Next, source the Sun Grid Engine environment variables from a `settings.csh` file where `$SGE_ROOT` is set to `/opt/sge`:**

```
% source $SGE_ROOT/default/common/settings.csh
```

3. **To start the batch (or scripted) job, specify the parallel environment, slot number and the user executable.**

```
% qsub -pe orte 2 sge.csh
your job 305 ("sge.csh") has been submitted
```

Since this is submitted as a batch job, you would not expect to see output at the terminal. If no indication is given for where the output should go, Sun Grid Engine redirects to your home directory and creates `<job_name>.<job_number>`.

The job creates the output files. The file name with the format `name_of_job.ojob_id` contains the standard output. The file name with the format `name_of_job.ejob_id` contains the error output. If the job executes normally, the error output files will be empty.

The following example lists the files produced by a job called `sgc.csh` with the job ID number 866:

```
% ls -rlt ~ | tail
-rw-r--r-- 1 joeuser  mygroup      0 Jan 16 16:42 sgc.csh.po866
-rw-r--r-- 1 joeuser  mygroup      0 Jan 16 16:42 sgc.csh.pe866
-rw-r--r-- 1 joeuser  mygroup      0 Jan 16 16:42 sgc.csh.e866
-rw-r--r-- 1 joeuser  mygroup    194 Jan 16 16:42 sgc.csh.o866
```

By default, the output files are located in your home directory, but you can use Sun Grid Engine software to change the location of the files, if desired.

---

**Note** – In most cases, you do not need to change the values set in the gridengine MCA parameters. If you run into difficulty and want to change the values for debugging purposes, the option is available. For more information about MCA parameters, see [Chapter 7](#).

---

## ▼ To See a Running Job

- Type the following command:

```
% qstat -f
```

## ▼ To Delete a Running Job

- Type the following command:

```
% qdel job-number
```

where *job-number* is the number of the job you want to delete.

For more information about Sun Grid Engine commands, refer to the Sun Grid Engine documentation.

## rsh Limitations

---

**Note** – This issue affects both `rsh` and the Sun Grid Engine program `qrsh`. `qrsh` uses `rsh` to launch jobs.

---

If you are using `rsh` or `qrsh` as the job launcher on a large cluster with hundreds of nodes, `rsh` might show the following error messages when launching jobs on the remote nodes:

```
rcmd: socket: Cannot assign requested address
rcmd: socket: Cannot assign requested address
rcmd: socket: Cannot assign requested address
[node0:00749] ERROR: A daemon on node m2187 failed to start as
expected.
[node0:00749] ERROR: There may be more information available from
[node0:00749] ERROR: the 'qstat -t' command on the Grid Engine
tasks.
[node0:00749] ERROR: If the problem persists, please restart the
[node0:00749] ERROR: Grid Engine PE job
```

This indicates that `rsh` is running out of sockets when launching the job from the head node.

## Using `rsh` as the Job Launcher

If you are using `rsh` as your job launcher, use `ssh` instead. Add the following to your command line:

```
-mca plm_rsh_agent ssh
```

## Using Sun Grid Engine as the Job Launcher

If you are using Sun Grid Engine version 6.1 or earlier as your job launcher, you can modify the Sun Grid Engine configuration to allow Sun Grid Engine to use `ssh` instead of `rsh` to launch tasks on the remote nodes. The following web site describes how to perform this workaround:

[http://gridengine.sunsource.net/howto/qrsh\\_qlogin\\_ssh.html](http://gridengine.sunsource.net/howto/qrsh_qlogin_ssh.html)

Note that this workaround does not properly track resource usage, nor does it allow proper job accounting. Sun Grid Engine tracks resource usage by attaching an extra groupid when launching tasks as a user of the remote connection.

Sun Grid Engine version 6.2 fixes this issue by not using `rsh` to start jobs on remote nodes. Instead Sun Grid Engine version 6.2 makes use of a native Interactive Job Support (IJS), which removes any dependencies on `rsh`, `ssh`, or `telnet`. It is recommended that you upgrade to the latest available version of Sun Grid Engine.

---

## For More Information

For more information about using the `mpirun` command to perform batch processing, see the following:

- `mpirun(3)` man page
- Running MPI Jobs section of the Open MPI FAQ at <http://www.open-mpi.org/faq/?category=running#plm-available>



## Using MCA Parameters With `mpirun`

---

Open MPI uses Modular Component Architecture (MCA) parameters to provide a way to tune your runtime environment. Each parameter corresponds to a specific function. You change the value of the parameter in order to change the function. [Appendix B](#) contains the complete list of MCA parameters.

Developing an Open MPI application that uses MCA parameters poses a number of advantages. Developers and administrators can customize the Open MPI environment to suit the specific needs of hardware or the operating environment. For example, a system administrator might use MCA parameters to optimize an Open MPI installation on a network so that users only need to run with the default values to obtain the best performance.

This chapter contains the following topics:

- [“About the Modular Component Architecture” on page 56](#)
- [“Open MPI Frameworks” on page 56](#)
- [“The `ompi\_info` Command” on page 59](#)
- [“Using MCA Parameters” on page 68](#)
- [“For More Information” on page 79](#)

In order to understand how MCA parameters fit within Open MPI, you must understand how the Modular Component Architecture is constructed.

---

# About the Modular Component Architecture

The Modular Component Architecture (MCA) is the backbone for much of Open MPI's functionality. It is a series of *frameworks*, *components*, and *modules* that are assembled at runtime to create an MPI implementation.

An MCA *framework* manages a specific Open MPI task (such as process launching for ORTE). Each MCA framework supports a single component type, but can support multiple versions of that type. The framework uses the services from the MCA base functionality to find and/or load components.

An MCA *component* is an implementation of a framework's interface. It is a standalone collection of code that can be bundled into a plug-in that can be inserted into the Open MPI code base, either at runtime and/or at compile time.

An MCA *module* is an instance of a component. For example, if a node running an Open MPI application has multiple Ethernet NICs, the Open MPI application will contain one TCP MPI point-to-point component, but two TCP point-to-point modules.

For more information about the Open MPI Modular Component Architecture, see the Open MPI FAQ on runtime tuning at:

<http://www.open-mpi.org/faq/?category=tuning>

---

## Open MPI Frameworks

There are three types of frameworks in Open MPI:

- In the MPI layer (OMPI)
- In the run-time layer (ORTE)
- In the operating system/platform layer (OPAL)

You might think of these frameworks as ways to group MCA parameters by function. For example, the OMPI `bt1` framework controls the functions in the byte transfer layer, or BTL (point-to-point byte movement) in the network. All of the MCA parameters that are grouped under `bt1` affect the BTL layer.

In addition to the parameters that are grouped under the individual frameworks, there are top-level MCA parameters that affect the frameworks themselves and specify values to your Open MPI installation.

**TABLE 7-1** Top-Level MCA Parameters

Parameter Group	Description
mca	Specify paths or functions for MCA parameters
mpi	Specify MPI behavior at runtime
orte	Specify debugging functions and components for ORTE
opal	Specify stack trace information

To view the available top-level parameters in each group, type the following command:

```
% mpi_info --param groupname groupname
```

where *groupname* stands for the parameter group you want to view. For example, to view the available MPI parameters, you would type:

```
% mpi_info --param mpi mpi
```

## OMPI Frameworks

The following table lists the frameworks in the MPI layer.

**TABLE 7-2** OMPI Frameworks

Framework	Description
allocator	Memory allocator
bml	BTL management layer (managing multiple devices)
btl	Byte transfer layer (point-to-point byte movement)
coll	MPI collective algorithms
io	MPI-2 I/O functionality
mpool	Memory pool management
mtl	Messaging transport layer
osc	One-sided communication

**TABLE 7-2** OMPI Frameworks (*Continued*)

Framework	Description
pml	Point-to-point management layer (fragmenting, reassembly, top-layer protocols, etc.)
rcache	Memory registration management
topo	MPI topology information

Currently, there is no simple way to get a list of the available components in a framework. You can use the `grep` command to search for components. For example, the following command searches for a list of components in the `bt1` framework:

```
% ompi_info | grep bt1
```

## ORTE Frameworks

The following table lists the ORTE frameworks.

**TABLE 7-3** ORTE Frameworks

Framework	Description
errmgr	Error manager
gpr	General purpose registry
iof	I/O forwarding
ns	Name server
oob	Out-of-band communication
p1m	Process launch module (was <code>p1s</code> ).
ras	Resource allocation subsystem
rds	Resource discovery subsystem
rmaps	Resource mapping subsystem
rmgr	Resource manager (upper meta layer for all other Resource frameworks)
rml	Remote messaging layer (routing of OOB messages)
schema	Name schemas
sds	Startup discovery services
soh	State of health

# OPAL Frameworks

The following table lists the OPAL frameworks.

**TABLE 7-4** OPAL Frameworks

Framework	Description
backtrace	Stack trace framework for debugging
maffinity	Memory affinity
memory	Memory hooks
paffinity	Processor affinity
timer	High-resolution timers

A complete list of MCA parameters, grouped under each of these frameworks, appears in [Appendix B](#).

---

## The `ompi_info` Command

The `ompi_info` command returns information about your Sun HPC ClusterTools/Open MPI installation. When you issue the command without any modifiers, `ompi_info` returns the following information:

- Revision information for Open MPI, ORTE, and OPAL
- Installed compilers
- Architecture of the node on which Open MPI is installed
- Version information for the installed frameworks

# Command Options

The `mpi_info` command has the following options:

**TABLE 7-5** Options for the `mpi_info` Command

Option	Description
<code>-a</code> or <code>--all</code>	Shows all configuration options and MCA parameters
<code>--arch</code>	Shows the architecture on which this installation of Open MPI was compiled
<code>-c</code> or <code>--config</code>	Shows configuration options
<code>-gmca</code> or <code>--gmca param-name value</code>	Passes global MCA parameters that apply to all contexts. <i>param-name</i> is the parameter name; <i>value</i> is the value of the parameter
<code>-h</code> or <code>--help</code>	Shows the <code>mpi_info</code> help message
<code>--hostname</code>	Shows the name of the host on which Open MPI was configured and built
<code>--internal</code>	Shows internal MCA parameters (not meant to be modified by users)
<code>-mca</code> or <code>--mca param-name value</code>	Passes context-specific MCA parameters; they are considered global if <code>--gmca</code> is not used. <i>param-name</i> is the name of the parameter; <i>value</i> is the value for that parameter.
<code>--param arg1 arg2</code>	Shows MCA parameters. <i>arg1</i> can be a specific framework name or <code>all</code> . <i>arg2</i> can be a specific parameter name or <code>all</code> .
<code>--parsable</code> or <code>--parseable</code>	Displays output in parsable format
<code>--path pathname</code>	Shows the paths with which Open MPI was configured.
<code>--pretty</code>	Displays output in “prettyprint” format (default)
<code>-v</code> or <code>--version arg0 arg1</code>	Shows version of Open MPI or a component. <i>arg0</i> can be the name of a specific framework or <code>all</code> . <i>arg1</i> can be the name of a specific component or <code>all</code> .

The output from the `mpi_info` command appears similar to the following:

```
% mpi_info
      Open MPI: 1.2r13978-ct7b027r1708
      Open MPI SVN revision: 0
```

```

Open RTE: 1.2r13978-ct7b027r1708
Open RTE SVN revision: 0
    OPAL: 1.2r13978-ct7b027r1708
    OPAL SVN revision: 0
        Prefix: /opt/SUNWhpc/HPC8.2
Configured architecture: i386-pc-solaris2.10
    Configured by: root
    Configured on: Thu Mar  8 16:47:40 EST 2008
    Configure host: burpen-csx10-0
        Built by: root
        Built on: Thu Mar  8 17:04:51 EST 2008
        Built host: burpen-csx10-0
        C bindings: yes
        C++ bindings: yes
        Fortran77 bindings: yes (all)
        Fortran90 bindings: yes
Fortran90 bindings size: trivial
    C compiler: cc
    C compiler absolute: /ws/ompi-tools/SUNWspro/SOS12/bin/cc
    C++ compiler: CC
    C++ compiler absolute: /ws/ompi-tools/SUNWspro/SOS12/bin/CC
    Fortran77 compiler: f77
Fortran77 compiler abs: /ws/ompi-tools/SUNWspro/SOS12/bin/f77
    Fortran90 compiler: f95
Fortran90 compiler abs: /ws/ompi-tools/SUNWspro/SOS12/bin/f95
    C profiling: yes
    C++ profiling: yes
    Fortran77 profiling: yes
    Fortran90 profiling: yes
    C++ exceptions: yes
    Thread support: no
    Internal debug support: no
    MPI parameter check: runtime
    Memory profiling support: no
    Memory debugging support: no
        libltdl support: yes
    Heterogeneous support: yes
    mpirun default --prefix: yes
        MCA backtrace: printstack (MCA v1.0, API v1.0, Component v1.2)
        MCA paffinity: solaris (MCA v1.0, API v1.0, Component v1.2)
        MCA maffinity: first_use (MCA v1.0, API v1.0, Component v1.2)
        MCA timer: solaris (MCA v1.0, API v1.0, Component v1.2)
        MCA allocator: basic (MCA v1.0, API v1.0, Component v1.0)
        MCA allocator: bucket (MCA v1.0, API v1.0, Component v1.0)
            MCA coll: basic (MCA v1.0, API v1.0, Component v1.2)
            MCA coll: self (MCA v1.0, API v1.0, Component v1.2)
            MCA coll: sm (MCA v1.0, API v1.0, Component v1.2)
            MCA coll: tuned (MCA v1.0, API v1.0, Component v1.2)
            MCA io: romio (MCA v1.0, API v1.0, Component v1.2)

```

```

MCA mpool: sm (MCA v1.0, API v1.0, Component v1.2)
MCA mpool: udapl (MCA v1.0, API v1.0, Component v1.2)
  MCA pml: cm (MCA v1.0, API v1.0, Component v1.2)
  MCA pml: obl (MCA v1.0, API v1.0, Component v1.2)
  MCA bml: r2 (MCA v1.0, API v1.0, Component v1.2)
MCA rcache: rb (MCA v1.0, API v1.0, Component v1.2)
MCA rcache: vma (MCA v1.0, API v1.0, Component v1.2)
  MCA btl: self (MCA v1.0, API v1.0.1, Component v1.2)
  MCA btl: sm (MCA v1.0, API v1.0.1, Component v1.2)
  MCA btl: tcp (MCA v1.0, API v1.0.1, Component v1.0)
  MCA btl: udapl (MCA v1.0, API v1.0, Component v1.2)
  MCA topo: unity (MCA v1.0, API v1.0, Component v1.2)
  MCA osc: pt2pt (MCA v1.0, API v1.0, Component v1.2)
MCA errmgr: hnp (MCA v1.0, API v1.3, Component v1.2)
MCA errmgr: orted (MCA v1.0, API v1.3, Component v1.2)
MCA errmgr: proxy (MCA v1.0, API v1.3, Component v1.2)
  MCA gpr: null (MCA v1.0, API v1.0, Component v1.2)
  MCA gpr: proxy (MCA v1.0, API v1.0, Component v1.2)
  MCA gpr: replica (MCA v1.0, API v1.0, Component v1.2)
  MCA iof: proxy (MCA v1.0, API v1.0, Component v1.2)
  MCA iof: svc (MCA v1.0, API v1.0, Component v1.2)
  MCA ns: proxy (MCA v1.0, API v2.0, Component v1.2)
  MCA ns: replica (MCA v1.0, API v2.0, Component v1.2)
MCA oob: tcp (MCA v1.0, API v1.0, Component v1.0)
MCA ras: dash_host (MCA v1.0, API v1.3, Component v1.2)
MCA ras: gridengine (MCA v1.0, API v1.3, Component v1.2)
MCA ras: localhost (MCA v1.0, API v1.3, Component v1.2)
MCA ras: tm (MCA v1.0, API v1.3, Component v1.2)
MCA rds: hostfile (MCA v1.0, API v1.3, Component v1.2)
MCA rds: proxy (MCA v1.0, API v1.3, Component v1.2)
MCA rds: resfile (MCA v1.0, API v1.3, Component v1.2)
MCA rmaps: round_robin (MCA v1.0, API v1.3, Component v1.2)
MCA rmgr: proxy (MCA v1.0, API v2.0, Component v1.2)
MCA rmgr: urm (MCA v1.0, API v2.0, Component v1.2)
MCA rml: oob (MCA v1.0, API v1.0, Component v1.2)
MCA plm: gridengine (MCA v1.0, API v1.3, Component v1.2)
MCA plm: proxy (MCA v1.0, API v1.3, Component v1.2)
MCA plm: rsh (MCA v1.0, API v1.3, Component v1.2)
MCA plm: tm (MCA v1.0, API v1.3, Component v1.2)
MCA sds: env (MCA v1.0, API v1.0, Component v1.2)
MCA sds: pipe (MCA v1.0, API v1.0, Component v1.2)
MCA sds: seed (MCA v1.0, API v1.0, Component v1.2)
MCA sds: singleton (MCA v1.0, API v1.0, Component v1.2)

```

# Using the `mpi_info` Command With MCA Parameters

The `mpi_info` command can list the parameters for a given component, all the parameters for a specific framework, or all parameters. The `mpi_info` output for most parameters contains a description of the parameter. The output for any parameter shows the current value of that parameter.

## ▼ To List All MCA Parameters

- Type the following command at the system prompt:

```
% mpi_info --param all all
```

The output from `mpi_info` lists all of the installed frameworks, their MCA parameters, and their current values. To see the complete list of MCA parameters, see [“List of Available MCA Parameters” on page 107](#).

## ▼ To List All MCA Parameters For a Framework

- Type the following command at the system prompt:

```
% mpi_info --param btl all
```

In this example, the command lists all of the available MCA parameters for the `btl` framework. The output from `mpi_info` looks similar to the following:

```
MCA btl: parameter "btl_base_debug" (current value: "0")
        If btl_base_debug is 1 standard debug is output, if >
        1 verbose debug is output
MCA btl: parameter "btl" (current value: <none>)
        Default selection set of components for the btl
        framework (<none> means "use all components that can
        be found")
MCA btl: parameter "btl_base_verbose" (current value: "0")
        Verbosity level for the btl framework (0 = no
        verbosity)
MCA btl: parameter "btl_self_free_list_num" (current value:
        "0")
        Number of fragments by default
MCA btl: parameter "btl_self_free_list_max" (current value:
        "-1")
        Maximum number of fragments
```

```

MCA btl: parameter "btl_self_free_list_inc" (current value:
    "32")
    Increment by this number of fragments
MCA btl: parameter "btl_self_eager_limit" (current value:
    "131072")
    Eager size fragmeng (before the rendez-vous ptoocol)
MCA btl: parameter "btl_self_min_send_size" (current value:
    "262144")
    Minimum fragment size after the rendez-vous
MCA btl: parameter "btl_self_max_send_size" (current value:
    "262144")
    Maximum fragment size after the rendez-vous
MCA btl: parameter "btl_self_min_rdma_size" (current value:
    "2147483647")
    Maximum fragment size for the RDMA transfer
MCA btl: parameter "btl_self_max_rdma_size" (current value:
    "2147483647")
    Maximum fragment size for the RDMA transfer
MCA btl: parameter "btl_self_exclusivity" (current value:
    "65536")
    Device exclusivity
MCA btl: parameter "btl_self_flags" (current value: "10")
    Active behavior flags
MCA btl: parameter "btl_self_priority" (current value: "0")
MCA btl: parameter "btl_sm_free_list_num" (current value: "8")
MCA btl: parameter "btl_sm_free_list_max" (current value:
    "-1")
MCA btl: parameter "btl_sm_free_list_inc" (current value:
    "64")
MCA btl: parameter "btl_sm_exclusivity" (current value:
    "65535")
MCA btl: parameter "btl_sm_latency" (current value: "100")
MCA btl: parameter "btl_sm_max_procs" (current value: "-1")
MCA btl: parameter "btl_sm_sm_extra_procs" (current value:
    "2")
MCA btl: parameter "btl_sm_mpool" (current value: "sm")
MCA btl: parameter "btl_sm_eager_limit" (current value:
    "4096")
MCA btl: parameter "btl_sm_max_frag_size" (current value:
    "32768")
MCA btl: parameter "btl_sm_size_of_cb_queue" (current value:
    "128")
MCA btl: parameter "btl_sm_cb_lazy_free_freq" (current value:
    "120")
MCA btl: parameter "btl_sm_priority" (current value: "0")
MCA btl: parameter "btl_tcp_if_include" (current value:
    <none>)
MCA btl: parameter "btl_tcp_if_exclude" (current value: "lo")
MCA btl: parameter "btl_tcp_free_list_num" (current value:

```

```

"8")
MCA btl: parameter "btl_tcp_free_list_max" (current value:
"-1")
MCA btl: parameter "btl_tcp_free_list_inc" (current value:
"32")
MCA btl: parameter "btl_tcp_sndbuf" (current value: "131072")
MCA btl: parameter "btl_tcp_rcvbuf" (current value: "131072")
MCA btl: parameter "btl_tcp_endpoint_cache" (current value:
"30720")
MCA btl: parameter "btl_tcp_exclusivity" (current value: "0")
MCA btl: parameter "btl_tcp_eager_limit" (current value:
"65536")
MCA btl: parameter "btl_tcp_min_send_size" (current value:
"65536")
MCA btl: parameter "btl_tcp_max_send_size" (current value:
"131072")
MCA btl: parameter "btl_tcp_min_rdma_size" (current value:
"131072")
MCA btl: parameter "btl_tcp_max_rdma_size" (current value:
"2147483647")
MCA btl: parameter "btl_tcp_flags" (current value: "122")
MCA btl: parameter "btl_tcp_priority" (current value: "0")
MCA btl: parameter "btl_udapl_free_list_num" (current value:
"8")
Initial size of free lists (must be >= 1).
MCA btl: parameter "btl_udapl_free_list_max" (current value:
"-1")
Maximum size of free lists (-1 = infinite, otherwise
must be >= 1).
MCA btl: parameter "btl_udapl_free_list_inc" (current value:
"8")
Increment size of free lists (must be >= 1).
MCA btl: parameter "btl_udapl_mpool" (current value: "udapl")
Name of the memory pool to be used.
MCA btl: parameter "btl_udapl_max_modules" (current value:
"8")
Maximum number of supported HCAs.
MCA btl: parameter "btl_udapl_num_recvs" (current value: "8")
Total number of receive buffers to keep posted per
endpoint (must be >= 1).
MCA btl: parameter "btl_udapl_num_sends" (current value: "7")
Maximum number of sends to post on an endpoint (must
be >= 1).
MCA btl: parameter "btl_udapl_sr_win" (current value: "4")
Window size at which point an explicit credit message
will be generated (must be >= 1).
MCA btl: parameter "btl_udapl_eager_rdma_num" (current value:
"32")
Number of RDMA buffers to allocate for small messages

```

(must be  $\geq 1$ ).

MCA btl: parameter "btl\_udapl\_max\_eager\_rdma\_peers" (current value: "16")  
Maximum number of peers allowed to use RDMA for short messages (independently RDMA will still be used for large messages, (must be  $\geq 0$ ; if zero then RDMA will not be used for short messages).

MCA btl: parameter "btl\_udapl\_eager\_rdma\_win" (current value: "28")  
Window size at which point an explicit credit message will be generated (must be  $\geq 1$ ).

MCA btl: parameter "btl\_udapl\_timeout" (current value: "10000000")  
Connection timeout, in microseconds.

MCA btl: parameter "btl\_udapl\_conn\_priv\_data" (current value: "1")  
Use connect private data to establish connections (not supported by all uDAPL implementations).

MCA btl: parameter "btl\_udapl\_async\_events" (current value: "1000000000")  
The asynchronous event queue will only be checked after entering progress this number of times.

MCA btl: parameter "btl\_udapl\_buffer\_alignment" (current value: "256")  
Preferred communication buffer alignment, in bytes (must be  $\geq 1$ ).

MCA btl: parameter "btl\_udapl\_evd\_qlen" (current value: "256")  
The event dispatcher queue length is a function of the number of connections as well as the maximum number of outstanding data transfer operations.

MCA btl: parameter "btl\_udapl\_max\_request\_dtos" (current value: "44")  
Maximum number of outstanding submitted sends and rdma operations per endpoint, (see Section 6.6.6 of uDAPL Spec.).

MCA btl: parameter "btl\_udapl\_max\_recv\_dtos" (current value: "8")  
Maximum number of outstanding submitted receive operations per endpoint, (see Section 6.6.6 of uDAPL Spec.).

MCA btl: parameter "btl\_udapl\_exclusivity" (current value: "1014")  
uDAPL BTL exclusivity (must be  $\geq 0$ ).

MCA btl: parameter "btl\_udapl\_eager\_limit" (current value: "8192")  
Eager send limit, in bytes (must be  $\geq 1$ ).

MCA btl: parameter "btl\_udapl\_min\_send\_size" (current value: "16384")  
Minimum send size, in bytes (must be  $\geq 1$ ).

MCA btl: parameter "btl\_udapl\_max\_send\_size" (current value: "65536")

```

Maximum send size, in bytes (must be >= 1).
MCA btl: parameter "btl_udapl_min_rdma_size" (current value:
"524288")
Minimum RDMA size, in bytes (must be >= 1).
MCA btl: parameter "btl_udapl_max_rdma_size" (current value:
"131072")
Maximum RDMA size, in bytes (must be >= 1).
MCA btl: parameter "btl_udapl_flags" (current value: "2")
BTL flags, added together: PUT=2 (cannot be 0).
MCA btl: parameter "btl_udapl_bandwidth" (current value:
"225")
Approximate maximum bandwidth of network (must be >=
1).
MCA btl: parameter "btl_udapl_priority" (current value: "0")
MCA btl: parameter "btl_base_include" (current value: <none>)
MCA btl: parameter "btl_base_exclude" (current value: <none>)
MCA btl: parameter "btl_base_warn_component_unused" (current
value: "0")
This parameter is used to turn on warning messages
when certain NICs are not used

```

## ▼ To Display All MCA Parameters For a Selected Component

- Type the following command at the system prompt:

```
% mpi_info --param btl tcp
```

The `mpi_info` output looks similar to the following:

```

MCA btl: parameter "btl_base_debug" (current value: "0")
    If btl_base_debug is 1 standard debug is output, if >
    1 verbose debug is output
MCA btl: parameter "btl" (current value: <none>)
    Default selection set of components for the btl
    framework (<none> means "use all components that can
    be found")
MCA btl: parameter "btl_base_verbose" (current value: "0")
    Verbosity level for the btl framework (0 = no
    verbosity)
MCA btl: parameter "btl_tcp_if_include" (current value:
<none>)
MCA btl: parameter "btl_tcp_if_exclude" (current value: "lo")
MCA btl: parameter "btl_tcp_free_list_num" (current value:

```

```

"8")
MCA btl: parameter "btl_tcp_free_list_max" (current value:
"-1")
MCA btl: parameter "btl_tcp_free_list_inc" (current value:
"32")
MCA btl: parameter "btl_tcp_sndbuf" (current value: "131072")
MCA btl: parameter "btl_tcp_rcvbuf" (current value: "131072")
MCA btl: parameter "btl_tcp_endpoint_cache" (current value:
"30720")
MCA btl: parameter "btl_tcp_exclusivity" (current value: "0")
MCA btl: parameter "btl_tcp_eager_limit" (current value:
"65536")
MCA btl: parameter "btl_tcp_min_send_size" (current value:
"65536")
MCA btl: parameter "btl_tcp_max_send_size" (current value:
"131072")
MCA btl: parameter "btl_tcp_min_rdma_size" (current value:
"131072")
MCA btl: parameter "btl_tcp_max_rdma_size" (current value:
"2147483647")
MCA btl: parameter "btl_tcp_flags" (current value: "122")
MCA btl: parameter "btl_tcp_priority" (current value: "0")
MCA btl: parameter "btl_base_warn_component_unused" (current
value: "0")
This parameter is used to turn on warning messages
when certain NICs are not used

```

## Using MCA Parameters

There are three ways to use MCA parameters with Open MPI:

1. Setting the parameter from the command line using the `mpirun --mca` command. This method assumes the highest precedence; values set for parameters using this method override any other values specified for the same parameter.
2. Using the parameter as an environment variable. Values for parameters set in this fashion assume the next highest priority.
3. Setting the parameter values in a text file. Parameter values specified using this method have the lowest priority.

## ▼ To Set MCA Parameters From the Command Line

- Type the following command at the system prompt:

```
% mpirun --mca param-name value
```

In this example, *param-name* stands for the name of the MCA parameter you want to set, and *value* stands for the new value you want to specify for the parameter. For example, the following command sets the value of the `mpi_show_handle_leaks` parameter to 1 for the specified job:

```
% mpirun --mca mpi_show_handle_leaks 1 -np 4 a.out
```

This sets the value of MCA parameter `mpi_show_handle_leaks` to 1 before running the program `a.out` with four processes.

## Using MCA Parameters As Environment Variables

As with other types of environment variables, the syntax for setting MCA parameters as environment variables varies with the type of command shell.

## ▼ To Set MCA Parameters in the `sh` Shell

1. Type the following command at the prompt:

```
% OMPI_MCA_param-name=value
```

where *param-name* is the name of the MCA parameter you want to set, and *value* is the desired value for the parameter. For example, the following command sets the `mpi_show_handle_leaks` parameter to 1:

```
% OMPI_MCA_mpi_show_handle_leaks=1
```

2. Type the following command:

```
% export OMPI_MCA_param-name
```

For example, an export command using the parameter used in the previous step would look like this:

```
% export OMPI_MCA_mpi_show_handle_leaks
```

3. Issue the `mpirun` command with the desired options. For example:

```
% mpirun -np 4 a.out
```

## ▼ To Set MCA Parameters in the C Shell

1. Use the `setenv` command to set the MCA parameter.

```
% setenv OMPI_MCA_param-name value
```

where *param-name* is the name of the MCA parameter you want to set, and *value* is the desired value for the parameter. The following example shows how to set the `mpi_show_handle_leaks` parameter to 1.

```
% setenv OMPI_MCA_mpi_show_handle_leaks 1
```

2. Issue the `mpirun` command for the program (in this example, `a.out`).

```
% mpirun -np 4 a.out
```

## ▼ To Specify MCA Parameters Using a Text File

1. Create a text file, specifying each parameter/value pair on a separate line. Comments are allowed. For example:

```
# This is a comment
# Set the same MCA parameter as in previous examples
mpi_show_handle_leaks = 1

# Default to rsh always
plm_rsh_agent = rsh

mpi_preconnect_all = 1
mpi_param_check = 0
#
# udapl parameters - comment or uncomment as needed
#
#btl = self,tcp,sm
#btl = self,udapl,sm
btl = ^tcp
```

2. Name the file `mca-params.conf` and save it.

You can save the file either to your home directory under `$HOME/.openmpi/mca-params.conf`, where the parameter values in the file will only affect your jobs, or you can save it to `/opt/SUNWhpc/HPC8.2/lib/openmpi-mca-params.conf`, where the parameter values in the file affect all users.

The following example shows the output from the `ompi_info` command for `mca_param_files`.

```
% ompi_info --param mca mca_param_files
MCA mca: parameter "mca_param_files" (current value:
"/home/joeuser/.openmpi/mca-params.conf:
/opt/SUNWhpc/HPC8.2/etc/openmpi-mca-params.conf")
Path for MCA configuration files containing default parameter values
MCA mca: parameter "mca_component_path" (current value:
"/opt/SUNWhpc/HPC8.2/lib/openmpi:/home/joeuser/.openmpi/components")
Path where to look for Open MPI and ORTE components
MCA mca: parameter "mca_verbose" (current value: <none>)
Top-level verbosity parameter
MCA mca: parameter "mca_component_show_load_errors" (current value: "1")
Whether to show errors for components that failed to load or not
MCA mca: parameter "mca_component_disable_dlopen" (current value: "0")
Whether to attempt to disable opening dynamic components or not
```

The MCA parameter `mca_param_files` specifies a colon-delimited path of files to search for MCA parameters. Files to the left of the colon have lower precedence; files to the right of the colon have higher precedence. At runtime, `mpirun` searches the following two files in order when the `mca_param_files` parameter is set:

1. `$HOME/.openmpi/mca-params.conf`: The user-supplied set of values takes the highest precedence.
2. `$prefix/etc/openmpi-mca-params.conf`: The system-supplied set of values has a lower precedence.

In the above example, Open MPI first searches `/home/joeuser/.openmpi/mca-params.conf` for MCA parameters, and then searches `/opt/SUNWhpc/HPC8.2/etc/openmpi-mca-params.conf`. If a parameter appears in both locations, the value set in the second file (the file to the right of the colon) is used.

## Including and Excluding Components

Each MCA framework has a top-level MCA parameter that you can use to select which components are to be used at runtime. In other words, there is an MCA parameter of the same name as each MCA framework (for example, `btl`) that you can use to include or exclude components from a given run.

You can use top-level parameters in the same way you would use other MCA parameters (for example, you can set them from the command line, as environment variables, or in text files).

For example, the `btl` MCA parameter is used to control which byte transfer layer (BTL) components are used with `mpirun`. The value for the `btl` parameter is a list of components separated by commas, with the optional prefix `^` (caret symbol).

---

**Note** – Do not mix “include” instructions with “exclude” instructions in the same command; otherwise, `mpirun` returns an error.

---

## ▼ To Include and Exclude Components Using the Command Line

- Type the following command at the system prompt:

```
% mpirun --mca framework comp1, comp2 ^comp3
```

In this example, the components *comp1* and *comp2* are included for the framework specified by `--mca framework`. Component *comp3* is excluded, since it is preceded by the ^ (caret) symbol.

For example, the following command excludes the `tcp` and `openib` components from the BTL framework, and implicitly includes all the other components:

```
% mpirun --mca btl ^tcp,openib ...
```

The use of the caret followed by the ellipsis in the command means “Perform the opposite action with the rest of the components.” When the `mpirun --mca` command specifies components to be excluded, the caret followed by the ellipsis implicitly includes the rest of the components in that framework. When the `mpirun --mca` command specifically includes components, the caret followed by the ellipsis means “and exclude the components not specified.”

For example, the following command includes only the `self`, `sm`, and `gm` components of `btl` and implicitly excludes the rest:

```
% mpirun --mca btl self,sm,gm ...
```

## Processor and Memory Affinity

### Using Processor Affinity

The term *processor affinity* refers to the state where the operating system allows only that process to run on a specific processor. On multi-processor machines, this can help improve performance by not allowing the operating system to move processes between processors. This can eliminate the “jitter” from performance characteristics due to the OS moving processes, which means that performance characteristics should be consistent among multiple runs. This approach can dramatically improve performance.

---

**Note** – Processor affinity should not be used when a node is over-subscribed (that is, when more processes are launched than there are processors). This can lead to a serious degradation in performance (even more than simply oversubscribing the node). Open MPI usually detects this situation and automatically disables the use of processor affinity (and displays run-time warnings to this effect). For more information about oversubscribing nodes, see [“Oversubscribing Nodes” on page 27](#).

---

## Using Memory Affinity

Memory affinity is only relevant for Non-Uniform Memory Access (NUMA) machines, such as many models of multi-processor Opteron™ machines. In a NUMA architecture, memory is physically distributed throughout the machine, even though it is virtually treated as a single address space. That is, memory may be physically local to one or more processors; therefore, the memory is remote to other processors. This means that some memory can be accessed more quickly by a process than other memory.

Open MPI supports general and specific memory affinity, which means that it generally tries to allocate all memory local to the processor that asked for it. When shared memory is used for communication, Open MPI uses memory affinity to make certain pages local to specific processes in order to minimize memory network/bus traffic.

### ▼ To Find Out Whether Memory Affinity Is Supported

Open MPI supports memory affinity on a variety of systems.

- **To find out which systems are supported, type the `ompi_info` command and look for maffinity components to see if your system is supported. For example:**

```
% ompi_info | grep maffinity
MCA maffinity: first_use (MCA v1.0, API v1.0, Component v1.2)
```

---

**Note** – Memory affinity support is enabled only when processor affinity is enabled. This is because processes might allocate local memory and then move to a different processor, and the second processor might be remote from the memory that the process just allocated. This negates the purpose of specifying memory affinity.

---

## Running MPI Jobs With Processor and Memory Affinity

If your system supports processor and memory affinity as shown using the `ompi_info` command, you can explicitly tell Open MPI to use affinity when running MPI jobs.

---

**Note** – Processor and memory affinity function only on multi-processor machines.

---

Currently, Open MPI only offers coarse-grained controls for processor affinity. For this reason, you can obtain the best results if processes in an Open MPI job using processor affinity are the only intensive processes running on the nodes being used for the job. Since most schedulers do not provide information on which processors should be used for specific processes, Open MPI assumes that its processes are "alone" on the node. Open MPI then exclusively claims CPUs, starting with the first one.

This means that if two processor-affinity-enabled jobs are running on the same node, they will both attempt to claim the first processor(s) on the node, resulting in CPU thrashing (and severely degraded performance).

---

**Note** – When running with processor affinity, all processors must be operational. Otherwise, processor affinity will not function because all the processors must be accessed in sequence.

---

### ▼ To Enable Affinity Using the Command Line

- To enable processor (and potentially memory) affinity, set the MCA parameter `opal_paffinity_alone` to 1.

For example, the following command enables processor affinity while running the program `a.out` on four processors:

```
% mpirun --mca opal_paffinity_alone 1 -np 4 a.out
```

The command shown in this example assumes that this job is running on a single 4-processor machine or two 2-processor machines. Setting `opal_paffinity_alone` tells Open MPI to bind each process to a specific processor. If memory affinity is supported, Open MPI also attempts to use memory affinity for this job.

You set values for `opal_paffinity_alone` in the same way you set other MCA parameters. For more information about setting MCA parameters, see [“Using MCA Parameters” on page 68](#).

---

**Note** – Open MPI automatically disables processor affinity on any node that is oversubscribed (that is, where more Open MPI processes are launched in a single job on a node than it has processors) and returns warning messages. However, you may use processor affinity with degraded performance mode if the nodes are not oversubscribed.

---

## Using MCA Parameters With Sun Grid Engine

The `ras_gridengine` parameters enable you to specify the output from the Open MPI RAS (Resource Allocation Subsystem). The `rsh` PLM (Process Launch Module) contains the `gridengine` parameters.

The following example shows the `mpirun` command with a specified MCA parameter.

```
% mpirun -np 4 -mca plm_gridengine_debug 100 connectivity.sparc -v
```

The following table shows the available MCA parameters and their default values.

**TABLE 7-6** MCA Parameters For Use With Sun Grid Engine Integration

MCA Parameter	Default Value	Function
<code>ras_gridengine_debug</code>	0	Enable debugging output for the <code>gridengine ras</code> component
<code>ras_gridengine_verbose</code>	0	Enable verbose output for the <code>gridengine ras</code> component
<code>ras_gridengine_show_jobid</code>	0	Show the <code>JOB_ID</code> of the Grid Engine job
<code>ras_gridengine_priority</code>	100	Priority of the <code>gridengine ras</code> component
<code>plm_base_reuse_daemons</code>	0	Specifies whether to reuse daemons to launch dynamically spawned processes
<code>plm_gridengine_debug</code>	0	Enable debugging of <code>gridengine plm</code> component
<code>plm_gridengine_verbose</code>	0	Enable verbose output of the <code>gridengine qrsh -inherit</code> command
<code>plm_gridengine_priority</code>	100	Priority of the <code>gridengine plm</code> component
<code>plm_gridengine_orted</code>	<code>orted</code>	The command name that the <code>gridengine plm</code> component will invoke for the ORTE daemon

To view a list of the RAS parameters from the command line, use the `mpi_info` command. The following example shows how to specify the RAS parameters and the output from the `mpi_info` command.

```
% mpi_info -param ras gridengine
MCA ras: parameter "ras" (current value: <none>)
          Default selection set of components for the ras
          framework (<none> means "use all components that can
          be found")
MCA ras: parameter "ras_gridengine_debug" (current value: "0")
          Enable debugging output for the gridengine ras
          component
MCA ras: parameter "ras_gridengine_priority" (current value:
"100")
          Priority of the gridengine ras component
MCA ras: parameter "ras_gridengine_verbose" (current value:
"0")
          Enable verbose output for the gridengine ras
          component
MCA ras: parameter "ras_gridengine_show_jobid" (current value:
"0")
          Show the JOB_ID of the Grid Engine job
```

This example shows the output from the `ompi_info` command when the PLM parameters are specified:

```
% ompi_info -param plm gridengine
MCA plm: parameter "plm_base_reuse_daemons" (current value:
        "0")
        If nonzero, reuse daemons to launch dynamically
        spawned processes.  If zero, do not reuse daemons
        (default)
MCA plm: parameter "plm" (current value: <none>)
        Default selection set of components for the plm
        framework (<none> means "use all components that can
        be found")
MCA plm: parameter "plm_base_verbose" (current value: "0")
        Verbosity level for the plm framework (0 = no
        verbosity)
MCA plm: parameter "plm_gridengine_debug" (current value: "0")
        Enable debugging of gridengine plm component
MCA plm: parameter "plm_gridengine_verbose" (current value:
        "0")
        Enable verbose output of the gridengine qrsh -inherit
        command
MCA plm: parameter "plm_gridengine_priority" (current value:
        "100")
        Priority of the gridengine plm component
MCA plm: parameter "plm_gridengine_orted" (current value:
        "orted")
        The command name that the gridengine plm component
        will invoke for the ORTE daemon
```

## Changing the Default Values in MCA Parameters

---

**Note** – In most cases, you do not need to change the default values in the gridengine MCA parameters. If you encounter a difficulty and want to change the values for debugging purposes, the options are available.

---

There are options available in the MCA PLM and RAS components and modules to allow changes of the default values.

For more information about how to change the values in MCA parameters, see the General Run-time Tuning FAQ on the Open MPi Web site at:

<http://www.open-mpi.org/faq/?category=tuning#setting-mca-params>

---

## For More Information

For more information about the Modular Component Architecture and MCA parameters, refer to the following sources:

- [Chapter 5, “Running Programs With the `mpirun` Command” on page 19](#)
- [Appendix B, “List of Available MCA Parameters” on page 107](#)
- Open MPI FAQ about runtime tuning: <http://www.open-mpi.org/faq/?category=tuning>
- The `ompi_info` man page
- The `ompi_info --help` command



# Using the DTrace Utility With Open MPI

---

This chapter describes how to use the Solaris™ Dynamic Tracing (DTrace) utility with Open MPI. DTrace is a comprehensive dynamic tracing utility that you can use to monitor the behavior of applications programs as well as the operating system itself. You can use DTrace on live production systems to understand those systems' behavior and to track down any problems that might be occurring.

The D language is the programming language used to create the source code for DTrace programs.

The content of this chapter assumes knowledge of the D language and how to use DTrace.

The following topics are covered in this chapter:

- [“Checking the `mpirun` Privileges” on page 82](#)
- [“Running DTrace with MPI Programs” on page 83](#)
- [“Tracking Down Resource Leaks” on page 87](#)
- [“Using the DTrace `mpiperuse` Provider” on page 92](#)

For more information about the D language and DTrace, refer to the *Solaris Dynamic Tracing Guide* (Part Number 817-6223). This guide is part of the Solaris 10 OS Software Developer Collection.

Solaris 10 OS documentation can be found on the web at the following location:

<http://www.sun.com/documentation>

Follow these links to the *Solaris Dynamic Tracing Guide*:

Solaris Operating Systems -> Solaris 10 -> Solaris 10 Software Developer Collection

---

**Note** – The programs and script mentioned in the sections that follow are located at:

`/opt/SUNWhpc/examples/mpi/dtrace`

---

## Checking the `mpirun` Privileges

Before you run a program under DTrace, you need to make sure that you have the correct `mpirun` privileges.

In order to run the script under `mpirun`, make sure that you have `dtrace_proc` and `dtrace_user` privileges. Otherwise, DTrace will return the following error because it does not have sufficient privileges:

```
dtrace: failed to initialize dtrace: DTrace requires additional
privileges
```

### ▼ To Determine the Correct Privileges on the Cluster

To determine whether you have the appropriate privileges on the entire cluster, perform the following steps:

1. Use your favorite text editor to create the following shell script, called `mpppriv.sh`:

```
#!/bin/sh
# mpppriv.sh - run ppriv under a shell so you can get the privileges
# of the process that mpirun creates
ppriv $$
```

2. Type the following command, replacing `host1` and `host2` with the names of hosts in your cluster:

```
% mpirun -np 2 --host host1,host2 mpppriv.sh
```

If the output of `ppriv` shows that the E privilege set has the `dtrace` privileges, then you will be able to run `dtrace` under `mpirun` (see the following two examples). Otherwise, you must adjust your system to get `dtrace` access.

The following example shows the output from `ppriv` when the privileges have not been set:

```
% ppriv $$
4084:  -csh
flags = <none>
E:   basic
I:   basic
P:   basic
L:   all
```

This example shows `ppriv` output when the privileges have been set:

```
% ppriv $$
2075:  tcsh
flags = <none>
E:basic,dtrace_proc,dtrace_user
I:basic,dtrace_proc,dtrace_user
P:basic,dtrace_proc,dtrace_user
L:   all
```

---

**Note** – To update your privileges, ask your system administrator to add the `dtrace_user` and `dtrace_proc` privileges to your account in the `/etc/user_attr` file.

---

After the privileges have been changed, you can use the `ppriv` command to view the changed privileges.

---

## Running DTrace with MPI Programs

There are two ways to use dynamic tracing with MPI programs:

- Run the MPI program directly under DTrace
- Attach DTrace to a running MPI program

# Running an MPI Program Under DTrace

For illustration purposes, assume you have a program named `mpiapp`.

## ▼ To Trace a Program Using the `mpitrace.d` Script

- Type the following command:

```
% mpirun -np 4 dtrace -s mpitrace.d -c mpiapp
```

The advantage of tracing an MPI program in this way is that all the processes in the job will be traced from the beginning. This method is probably most useful in doing performance measurements, when you need to start at the beginning of an application and you need all the processes in a job to participate in collecting data.

This approach also has some disadvantages. One disadvantage of running a program like the one in the above example is that all the tracing output for all four processes is directed to standard output (`stdout`). One way around this problem is to create a script similar to the script in the following section:

## ▼ To Trace a Parallel Program and Get Separate Trace Files

1. Create a shell script (called `partrace.sh` in this example) similar to the following:

```
#!/bin/sh
# partrace.sh - a helper script to dtrace Open MPI jobs from the
# start of the job.
dtrace -s $1 -c $2 -o $2.$OMPI_COMM_WORLD_RANK.trace
```

2. Type the following command to run the `partrace.sh` shell script:

```
% mpirun -np 4 partrace.sh mpitrace.d mpiapp
```

This will run `mpiapp` under `dtrace` using the `mpitrace.d` script. The script saves the trace output for each process in a job under a separate file name, based on the program name and rank of the process. Note that subsequent runs will append the data into the existing trace files.

---

**Note** – The status of the `OMPI_COMM_WORLD_RANK.trace` variable is unstable and subject to change. Use this variable with caution.

---

# Attaching DTrace to a Running MPI Program

The second way to use `dtrace` with Open MPI is to attach `dtrace` to a running MPI program.

## ▼ To Attach DTrace to a Running MPI Program

Perform the following procedure:

1. **Log in to the node in which you are interested.**
2. **Type commands similar to the following command to get the process ID (PID) of the running program on the node of interest.**

```
% prstat 0 1 | grep mpiapp
24768 joeuser      526M 3492K sleep  59    0    0:00:08 0.1% mpiapp/1
24770 joeuser      518M 3228K sleep  59    0    0:00:08 0.1% mpiapp/1
```

3. **Decide which rank you want to use to attach `dtrace`.**  
The lower PID number is usually the lower rank on the node.
4. **Type the following command to attach to the rank 1 process (identified by its process ID, which is 24770 in the example) and run the DTrace script `mpitrace.d`:**

```
% dtrace -p 24770 -s mpitrace.d
```

## Simple MPI Tracing

DTrace enables you to easily trace programs. When used in conjunction with MPI and the more than 200 functions defined in the MPI standard, DTrace provides an easy way to determine which functions might be in error during the debugging process, or those functions that might be of interest. After you determine the function showing the error, it is easy to locate the desired job, process, and rank on which to run your scripts. As demonstrated above, DTrace allows you to perform these determinations while the program is running.

Although the MPI standard provides the MPI profiling interface, using DTrace does provide a number of advantages. The advantages of using DTrace include the following:

- The PMPI interface requires you to restart a job every time you make changes to the interposing library.

- DTrace allows you to define probes that let you capture tracing information on MPI without having to code the specific details for each function you want to capture.
- The DTrace scripting language D has several built-in functions that help in debugging problematic programs.

The following example shows a simple script that traces the entry and exit into all the MPI API calls.

```
mpitrace.d:
pid$target:libmpi:MPI_*:entry
{
    printf("Entered %s...", probefunc);
}

pid$target:libmpi:MPI_*:return
{
    printf("exiting, return value = %d\n", arg1);
}
```

When you use this example script to attach DTrace to a job that performs send and recv operations, the output looks similar to the following:

```
% dtrace -q -p 24770 -s mpitrace.d
Entered MPI_Send...exiting, return value = 0
Entered MPI_Recv...exiting, return value = 0
Entered MPI_Send...exiting, return value = 0
Entered MPI_Recv...exiting, return value = 0
Entered MPI_Send...exiting, return value = 0 ...
```

You can easily modify the `mpitrace.d` script to include an argument list. The resulting output resembles `truss` output. For example:

```
mpitruss.d:
pid$target:libmpi:MPI_Send:entry,
pid$target:libmpi:MPI_*send:entry,
pid$target:libmpi:MPI_Recv:entry,
pid$target:libmpi:MPI_*recv:entry
{
printf("%s(0x%x, %d, 0x%x, %d, %d, 0x%x)",probefunc, arg0, arg1,
arg2, arg3, arg4, arg5);
}
pid$target:libmpi:MPI_Send:return,
pid$target:libmpi:MPI_*send:return,
pid$target:libmpi:MPI_Recv:return,
pid$target:libmpi:MPI_*recv:return
{
printf("\t\t = %d\n", arg1);
}
```

The `mpitruss.d` script shows how you can specify wildcard names to match the functions. Both probes will match all send and receive type function calls in the MPI library. The first probe shows the usage of the built-in `arg` variables to print out the `arglist` of the function being traced.

Take care when wildcarding the entrypoint and the formatting argument output, because you could end up printing either too many arguments, or not enough arguments, for certain functions. For example, in the above case, the `MPI_Irecv` and `MPI_Isend` functions will not have their Request handle parameters printed out.

The following example shows a sample output of the `mpitruss.d` script:

```
% dtrace -q -p 24770 -s mpitruss.d
MPI_Send(0x80470b0, 1, 0x8060f48, 0, 1, 0x8060d48) = 0
MPI_Recv(0x80470a8, 1, 0x8060f48, 0, 0, 0x8060d48) = 0
MPI_Send(0x80470b0, 1, 0x8060f48, 0, 1, 0x8060d48) = 0
MPI_Recv(0x80470a8, 1, 0x8060f48, 0, 0, 0x8060d48) = 0 ...
```

---

## Tracking Down Resource Leaks

One of the biggest issues with programming is the unintentional leaking of resources (such as memory). With MPI, tracking and repairing resource leaks can be somewhat more challenging because the objects being leaked are in the middleware, and thus are not easily detected by the use of memory checkers.

DTrace helps with debugging such problems using variables, the profile provider, and a callstack function. The `mpicommcheck.d` script (shown in the example below) probes for all the MPI communicator calls that allocate and deallocate communicators, and keeps track of the stack each time the function is called. Every 10 seconds the script dumps out the current count of MPI communicator calls and the total calls for the allocation and deallocation of communicators. When the `dtrace` session ends (usually by pressing Ctrl-C, if you attached to a running MPI program), the script will print out the totals and all the different stack traces, as well as the number of times those stack traces were reached.

In order to perform these tasks, the script uses DTrace features such as variables, associative arrays, built-in functions (`count`, `ustack`) and the predefined variable `probefunc`.

The following example shows the `mpicommcheck.d` script.

```
mpicommcheck.d:
BEGIN
{
    allocations = 0;
    deallocations = 0;
    prcnt = 0;
}

pid$target:libmpi:MPI_Comm_create:entry,
pid$target:libmpi:MPI_Comm_dup:entry,
pid$target:libmpi:MPI_Comm_split:entry
{
    ++allocations;
    @counts[probefunc] = count();
    @stacks[ustack()] = count();
}

pid$target:libmpi:MPI_Comm_free:entry
{
    ++deallocations;
    @counts[probefunc] = count();
    @stacks[ustack()] = count();
}

profile:::tick-1sec
/++prcnt > 10/
{
    printf("=====");
    printa(@counts);
    printf("Communicator Allocations = %d \n", allocations);
    printf("Communicator Deallocations = %d\n", deallocations);
    prcnt = 0;
}

END
{
    printf("Communicator Allocations = %d, Communicator
Deallocations = %d\n",
    allocations, deallocations);
}
```

This script attaches `dtrace` to a suspect section of code in your program (that is, a section of code that might contain a resource leak). If, during the process of running the script, you see that the printed totals for allocations and deallocations are starting to steadily diverge, you might have a resource leak. Depending on how your

program is designed, it might take some time and observation of the allocation/deallocation totals in order to definitively determine that the code contains a resource leak. Once you do determine that a resource leak is definitely occurring, you can press Ctrl-C to break out of the `dtrace` session. Next, using the stack traces dumped, you can try to determine where the issue might be occurring.

The following example shows code containing a resource leak, and the output that is displayed using the `mpicommcheck.d` script.

The sample MPI program containing the resource leak is called `mpicommleak`. This program performs three `MPI_Comm_dup` operations and two `MPI_Comm_free` operations. The program thus “leaks” one communicator operation with each iteration of a loop.

When you attach `dtrace` to `mpicommleak` using the `mpicommcheck.d` script above, you will see a 10-second periodic output. This output shows that the count of the allocated communicators is growing faster than the count of deallocations.

When you finally end the `dtrace` session by pressing Ctrl-C, the session will have output a total of five stack traces, showing the distinct three `MPI_Comm_dup` and two `MPI_Comm_free` call stacks, as well as the number of times each call stack was encountered.

For example:

```
% prstat 0 1 | grep mpicommlleak
 24952 joeuser    518M 3212K sleep  59    0    0:00:01 1.8% mpicommlleak/1
 24950 joeuser    518M 3212K sleep  59    0    0:00:00 0.2% mpicommlleak/1
% dtrace -q -p 24952 -s mpicommlcheck.d
=====
  MPI_Comm_free                4
  MPI_Comm_dup                  6
Communicator Allocations = 6
Communicator Deallocations = 4
=====
  MPI_Comm_free                8
  MPI_Comm_dup                  12
Communicator Allocations = 12
Communicator Deallocations = 8
=====
  MPI_Comm_free                12
  MPI_Comm_dup                  18
Communicator Allocations = 18
Communicator Deallocations = 12
^C
Communicator Allocations = 21, Communicator Deallocations = 14

libmpi.so.0.0.0`MPI_Comm_free
  mpicommlleak`deallocate_comms+0x19
  mpicommlleak`main+0x6d
  mpicommlleak`0x805081a
  7

  libmpi.so.0.0.0`MPI_Comm_free
  mpicommlleak`deallocate_comms+0x26
  mpicommlleak`main+0x6d
  mpicommlleak`0x805081a
  7

  libmpi.so.0.0.0`MPI_Comm_dup
  mpicommlleak`allocate_comms+0x1e
  mpicommlleak`main+0x5b
  mpicommlleak`0x805081a
  7

  libmpi.so.0.0.0`MPI_Comm_dup
  mpicommlleak`allocate_comms+0x30
  mpicommlleak`main+0x5b
  mpicommlleak`0x805081a
  7

  libmpi.so.0.0.0`MPI_Comm_dup
  mpicommlleak`allocate_comms+0x42
  mpicommlleak`main+0x5b
  mpicommlleak`0x805081a
  7
```

---

# Using the DTrace `mpiperuse` Provider

PERUSE is an MPI interface that allows you to obtain detailed information about the performance and interactions of processes, software, and MPI. PERUSE provides a greater level of detail about process performance than does the standard MPI profiling interface (PMPI).

For more information about PERUSE and the current PERUSE specification, see:

<http://www.mpi-peruse.org>

Open MPI includes a DTrace provider named `mpiperuse`. This provider enables you to configure Open MPI to support DTrace probes into the Open MPI shared library `libmpi`.

## DTrace Support in the ClusterTools Software

In Sun HPC ClusterTools 8.2 software, there are preconfigured executables and libraries with the `mpiperuse` provider probes built in. They are located in the `/opt/SUNWhpc/HPC8.2/sun/instrument` directory. Use the wrappers and utilities located in this directory to access the `mpiperuse` provider.

---

**Note** – No recompilation is necessary in order to use the `mpiperuse` provider. Just run the application to be DTraced using `/opt/SUNWhpc/HPC8.2/sun/instrument/bin/mpirun`.

---

## Available `mpiperuse` Probes

The DTrace `mpiperuse` probes expose the events specified in the current PERUSE specification. These events track the life cycle of requests within the MPI library. For more information about this life cycle and the actual events provided by PERUSE, see Section 4 of the PERUSE Specification.

Sections 4.3.1 and 4.4 of the PERUSE Specification list and describe the individual events exposed by PERUSE.

The `mpiperuse` provider makes these events available to DTrace. The probe names correspond to the event names listed in Sections 4.3.1 and 4.4 of the PERUSE specification. For each event, the corresponding probe name is similar, except that

the leading PERUSE is removed, the probe name is all lowercase, and underscores are replaced with hyphens. For example, the probe for PERUSE\_COMM\_MSG\_ARRIVED is comm-msg-arrived.

All of the probes are classified under the `mpiperuse` provider. This means that to find the probe names, you would look under the `mpiperuse` name. It also means that when you make a DTrace statement, you can include a wildcard for all probes simply by using the `mpiperuse` classification.

## Specifying an `mpiperuse` Probe in a D Script

In the D scripting language, specifying an `mpiperuse` provider takes the following form:

```
mpiperuse$target::probe-name
```

where *probe-name* is the name of the `mpiperuse` probe you want to use.

For example, to specify a probe to capture a PERUSE\_COMM\_REQ\_ACTIVATE event, add the following line to a D script:

```
mpiperuse$target::comm-req-activate
```

This alerts DTrace that you want to use the `mpiperuse` provider to capture the PERUSE\_COMM\_REQ\_ACTIVATE event. In this example, the optional object and function fields in the probe description are omitted. This directs DTrace to find all occurrences of the `comm-req-activate` probes in the MPI library and its plugins instead of a specific probe. This is necessary because certain probes can happen in multiple places in the MPI library.

For more information about the D language and its syntax, refer to the *Solaris Dynamic Tracing Guide* (Part Number 817-6223). This guide is part of the Solaris 10 OS Software Developer Collection.

# Available Arguments

All of the `mpiperuse` probes receive the following arguments:

**TABLE 8-1** Available `mpiperuse` Arguments

<code>args[0] = mpiconninfo_t *i</code>	This provides a basic source and destination for the request and which protocol is expected to be used for the transfer. This typedef is defined in <code>/usr/lib/dtrace/mpi.d</code> .
<code>args[1] = uintptr_t uid</code>	This is the PERUSE unique id for the request that fired the probe (as defined by the PERUSE specifications). For OMPI this is the address of the actual request.
<code>args[2] = uint_t op</code>	This value indicates whether the probe is for a <code>send == 0</code> or <code>recv == 1</code> request.
<code>args[3] = mpicomm_spec_t *cs</code>	This structure is defined in <code>/usr/lib/dtrace/mpi.d</code> and mimics the <code>spec</code> structure, as defined on page 22 of the PERUSE specification.

## How To Use `mpiperuse` Probes to See Message Queues

To use the `mpiperuse` provider, make reference to the appropriate `mpiperuse` provider probes and arguments in a DTrace script, as you would for any other provider (such as the `pid` provider).

The procedure for running scripts with `mpiperuse` probes follows the same steps as those shown in [“Running an MPI Program Under DTrace” on page 84](#) and [“Attaching DTrace to a Running MPI Program” on page 85](#), except that you must edit the `partrace.sh` script before you run it.

Change `partrace.sh` to include a `-Z` switch after the `dtrace` command, as shown in the following example.

```
#!/bin/sh
# partrace.sh - a helper script to dtrace Open MPI jobs from the
# start of the job.
dtrace -Z -s $1 -c $2 -o $2.$OMPI_COMM_WORLD_RANK.trace
```

This change allows probes that do not exist at initial load time to be used in a script (that is, the probes are in plugins that have not been `dlopened`).

The following example shows how to use the `mpiperuse` probes when running a DTrace script. Use the example script provided in `/opt/SUNWhpc/HPC8.2/sun/examples/dtrace/mpistat.d`

## 1. Compile and run a script against a program.

In this example, the script file is called `dtest.c`. Substitute the name and path of your script for `dtest.c`.

```
% /opt/SUNWhpc/HPC8.2/sun/instrument/bin/mpicc
~myhomedir/scraps/usdt/examples/dtest.c -o dtest
% /opt/SUNWhpc/HPC8.2/sun/instrument/bin/mpirun -np 2 dtest
Initing MPI...
Initing MPI...
Do communications...
Do communications...
attach to pid 13371 to test tracing.
```

## 2. In another window, type the following command:

```
% dtrace -q -p 13371 -s /opt/SUNWhpc/HPC8.2/sun/examples/dtrace/mpistat.d
input(Total) Q-sizes      Q-Matches      output
bytes active posted unexp posted unexp      bytes active
  0      0      0      0      0      0      0      0
  0      0      0      0      0      0      0      0
  0      0      0      0      0      0      0      0
  0      0      0      0      0      0      0      0
  0      0      0      0      1      0      5      0
  0      0      0      0      1      0      5      0
  0      0      0      0      1      0      5      0
  0      0      0      0      1      0      5      0
  0      0      0      0      1      0      5      0
  0      0      0      0      1      0      5      0
  0      0      0      0      1      0      5      0
  0      0      0      0      2      0     10      0
  0      0      0      0      2      0     10      0
  0      0      0      0      2      0     10      0
  0      0      0      0      2      0     10      0
  0      0      0      0      2      0     10      0
  0      0      0      0      2      0     10      0
  0      0      0      0      2      0     10      0
  0      0      0      0      2      0     10      0
  0      0      0      0      3      0     15      0
  0      0      0      0      3      0     15      0
  0      0      0      0      3      0     15      0
  0      0      0      0      3      0     15      0
  0      0      0      0      3      0     15      0
  0      0      0      0      3      0     15      0
  0      0      0      0      3      0     15      0
```

## mpiperuse Usage Examples

The examples in this section show how to perform the described DTrace operations from the command line.

### ▼ To Count the Number of Messages To or From a Host

- Issue the following DTrace command, substituting the process ID of the process you want to monitor for *pid*:

```
dtrace -p pid -n 'mpiperuse$target:::comm-req-xfer-end { @[args[0]->ci_remote] = count(); }'
```

DTrace returns a result similar to the following. In this example, the process ID is 25428 and the host name is joe-users-host2.

```
% dtrace -p 25428 -n 'mpiperuse$target:::comm-req-xfer-end { @[args[0]->ci_remote] = count(); }'  
dtrace: description 'mpiperuse$target:::comm-req-xfer-end ' matched 17 probes  
^C  
joe-users-host2 recv 3  
joe-users-host2 send 3
```

### ▼ To Count the Number of Messages To or From Specific BTLs

- Issue the following DTrace command, substituting the process ID of the process you want to monitor for *pid*:

```
dtrace -p pid -n 'mpiperuse$target:::comm-req-xfer-end { @[args[0]->ci_protocol] = count(); }'
```

DTrace returns a result similar to the following. In this example, the process ID is 25445.

```
% dtrace -p 25445 -n 'mpiperuse$target:::comm-req-xfer-end { @[args[0]->ci_protocol] = count(); }'  
dtrace: description 'mpiperuse$target:::comm-req-xfer-end ' matched 17 probes  
^C  
  
sm 60
```

## ▼ To Obtain Distribution Plots of Message Sizes Sent or Received From a Host

- Issue the following DTrace command, substituting the process ID of the process you want to monitor for *pid*:

```
dtrace -p pid -n 'mpiperuse$target:::comm-req-xfer-end { @[args[0]->ci_remote] = quantize(args[3]->mcs_count); }'
```

DTrace returns a result similar to the following. In this example, the process ID is 25445.

```
% dtrace -p 25445 -n 'mpiperuse$target:::comm-req-xfer-end { @[args[0]->ci_remote] = quantize(args[3]->mcs_count); }'
dtrace: description 'mpiperuse$target:::comm-req-xfer-end ' matched 17 probes
^C

myhost
value  ----- Distribution ----- count
    2  | 0
    4  | @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@ 4
    8  | 0
```

## ▼ To Create Distribution Plots of Message Sizes By Communicator, Rank, and Send/Receive

- Issue the following DTrace command, substituting the process ID of the process you want to monitor for *pid*:

```
dtrace -p pid -n 'mpiperuse$target:::comm-req-xfer-end { @[args[3]->mcs_comm, args[3]->mcs_peer, args[3]->mcs_op] = quantize(args[3]->mcs_count); }'
```

DTrace returns a result similar to the following. In this example, the process ID is 24937.

```
% dtrace -p 24937 -n 'mpiperuse$target:::comm-req-xfer-end { @[args[3]->mcs_comm, args[3]->mcs_peer, args[3]->mcs_op] = quantize(args[3]->mcs_count); }'
dtrace: description 'mpiperuse$target:::comm-req-xfer-end ' matched 19 probes
^C

134614864      1  recv
value  ----- Distribution ----- count
    2  | 0
    4  | @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@ 9
    8  | 0
```

```

134614864      1 send
value  ----- Distribution ----- count
   2 |                                           0
   4 | @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@ 9
   8 |                                           0

```

# Troubleshooting

---

This appendix describes some common problem situations, resulting error messages, and suggestions for fixing the problems. Open MPI error reporting, including I/O, follows the *MPI-2 Standard*. By default, errors are reported in the form of standard error classes. These classes and their meanings are listed in [TABLE A-1](#) (for non-I/O MPI) and [TABLE A-2](#) (for MPI I/O), and are also available on the MPI man page.

---

## MPI Messages

### Standard Error Classes

Listed below are the error return classes you might encounter in your MPI programs. Error values can also be found in `mpi.h` (for C), `mpif.h` (for Fortran), and `mpi++.h` (for C++).

**TABLE A-1** Open MPI Standard Error Classes

Error Code	Value	Meaning
MPI_SUCCESS	0	Successful return code.
MPI_ERR_BUFFER	1	Invalid buffer pointer.
MPI_ERR_COUNT	2	Invalid count argument.
MPI_ERR_TYPE	3	Invalid datatype argument.
MPI_ERR_TAG	4	Invalid tag argument.
MPI_ERR_COMM	5	Invalid communicator.

**TABLE A-1** Open MPI Standard Error Classes (Continued)

<b>Error Code</b>	<b>Value</b>	<b>Meaning</b>
MPI_ERR_RANK	6	Invalid rank.
MPI_ERR_ROOT	7	Invalid root.
MPI_ERR_GROUP	8	Null group passed to function.
MPI_ERR_OP	9	Invalid operation.
MPI_ERR_TOPOLOGY	10	Invalid topology.
MPI_ERR_DIMS	11	Illegal dimension argument.
MPI_ERR_ARG	12	Invalid argument.
MPI_ERR_UNKNOWN	13	Unknown error.
MPI_ERR_TRUNCATE	14	Message truncated on receive.
MPI_ERR_OTHER	15	Other error; use <code>Error_string</code> .
MPI_ERR_INTERN	16	Internal error code.
MPI_ERR_IN_STATUS	17	Look in status for error value.
MPI_ERR_PENDING	18	Pending request.
MPI_ERR_REQUEST	19	Illegal <code>MPI_Request()</code> handle.
MPI_ERR_KEYVAL	36	Illegal key value.
MPI_ERR_INFO	37	Invalid info object.
MPI_ERR_INFO_KEY	38	Illegal info key.
MPI_ERR_INFO_NOKEY	39	No such key.
MPI_ERR_INFO_VALUE	40	Illegal info value.
MPI_ERR_TIMEDOUT	41	Timed out.
MPI_ERR_SYSRESOURCES	42	Out of resources.
MPI_ERR_SPAWN	45	Error spawning.
MPI_ERR_WIN	46	Invalid window.
MPI_ERR_BASE	47	Invalid base.
MPI_ERR_SIZE	48	Invalid size.
MPI_ERR_DISP	49	Invalid displacement.
MPI_ERR_LOCKTYPE	50	Invalid locktype.

**TABLE A-1** Open MPI Standard Error Classes (Continued)

Error Code	Value	Meaning
MPI_ERR_ASSERT	51	Invalid assert.
MPI_ERR_RMA_CONFLICT	52	Conflicting accesses to window.
MPI_ERR_RMA_SYNC	53	Erroneous RMA synchronization.
MPI_ERR_NO_MEM	54	Memory exhausted.
MPI_ERR_LASTCODE	55	Last error code.

## MPI I/O Error Handling

Open MPI I/O error reporting follows the *MPI-2 Standard*. By default, errors are reported in the form of standard error codes (found in `/opt/SUNWhpc/include/mpi.h`). Error classes and their meanings are listed in [TABLE A-2](#). They can also be found in `mpif.h` (for Fortran) and `mpi.h` (for C).

You can change the default error handler by specifying `MPI_FILE_NULL` as the file handle with the routine `MPI_File_set_errhandler()`, even if no file is currently open. Or, you can use the same routine to change the error handler for a specific file.

**TABLE A-2** Open MPI I/O Error Classes

Error Class	Value	Meaning
MPI_ERR_FILE	20	Bad file handle.
MPI_ERR_NOT_SAME	21	Collective argument not identical on all processes.
MPI_ERR_AMODE	22	Unsupported amode passed to open.
MPI_ERR_UNSUPPORTED_DATAREP	23	Unsupported datarep passed to <code>MPI_File_set_view()</code> .
MPI_ERR_UNSUPPORTED_OPERATION	24	Unsupported operation, such as seeking on a file that supports only sequential access.
MPI_ERR_NO_SUCH_FILE	25	File (or directory) does not exist.
MPI_ERR_FILE_EXISTS	26	File exists.

**TABLE A-2** Open MPI I/O Error Classes (Continued)

<b>Error Class</b>	<b>Value</b>	<b>Meaning</b>
MPI_ERR_BAD_FILE	27	Invalid file name (for example, path name too long).
MPI_ERR_ACCESS	28	Permission denied.
MPI_ERR_NO_SPACE	29	Not enough space.
MPI_ERR_QUOTA	30	Quota exceeded.
MPI_ERR_READ_ONLY	31	Read-only file system.
MPI_ERR_FILE_IN_USE	32	File operation could not be completed, as the file is currently open by some process.
MPI_ERR_DUP_DATAREP	33	Conversion functions could not be registered because a data representation identifier that was already defined was passed to MPI_REGISTER_DATAREP.
MPI_ERR_CONVERSION	34	An error occurred in a user-supplied data-conversion function.
MPI_ERR_IO	35	I/O error.
MPI_ERR_INFO	37	Invalid info object.
MPI_ERR_INFO_KEY	38	Illegal info key.
MPI_ERR_INFO_NOKEY	39	No such key.
MPI_ERR_INFO_VALUE	40	Illegal info value.
MPI_ERR_LASTCODE	55	Last error code.

---

# Exceeding the File Descriptor Limit

If your application attempts to open a file descriptor when the maximum limit of open file descriptors has been reached, the job will fail and display the following message:

```
mynode% mpirun -np 61 hello_c
mpirun noticed that job rank 0 with PID 0 on node burl-ct-v440-1
exited on signal 15 (Terminated).
59 additional processes aborted (not shown)
```

Should this occur, increase the value of the file descriptor hard limit before starting your job again.

The Solaris OS default file descriptor limit is 256. When you start an MPI job, a program called an `orted` (for ORTE daemon) spawns the user processes. For each user process spawned, the `orted` takes up four file descriptors. In addition, the job takes 12 additional file descriptors regardless of the number of processes spawned.

To calculate the number of file descriptors needed to run a certain job, use the following formula:

$$\text{file descriptors} = 12 + 4 * \text{np}$$

where `np` is the number of processes launched.

If the number of file descriptors needed is greater than 256, you must increase the number of available descriptors to a value equal to or greater than the number you calculated. Otherwise, the processes fail and the error message is displayed.

## Increasing the Number of Available File Descriptors

### ▼ To View the Hard Limit from the C Shell

1. Log in to a C shell as superuser.

2. Determine the current hard limit value for your Solaris implementation. Type the following command:

```
# limit -h descriptors
```

## ▼ To View the Hard Limit from the Bourne Shell

1. Log in to a Bourne shell as superuser.
2. Use the `ulimit` function. Type the following command:

```
# ulimit -Hn
```

Each function returns the file descriptor hard limit that was in effect. The new value you set for the number of available file descriptors must be less than or equal to this number. The usual default value for the hard limit in the Solaris OS is 64000 (64K).

## ▼ To Increase the Number of File Descriptors

---

**Note** – You must perform this procedure on each of the nodes on which you plan to run.

---

1. Open the `/etc/system` file in a text editor.
2. Add the following line to the file:

```
set rlim_fd_cur=value
```

where `value` is the new maximum number of file descriptors. For example, the following line added to the `/etc/system` file increases the maximum number of file descriptors to 1024:

```
set rlim_fd_cur=1024
```

3. Save the file and exit the text editor.
4. Reboot the system.

# Setting File Descriptor Limits When Using Sun Grid Engine

If you are using Sun Grid Engine to launch your jobs on very large multi-processor nodes, you might see an error message about exceeding your file descriptor limit, and your jobs might fail. This can happen because Sun Grid Engine cannot set the file descriptor limit in its queue.

There are three ways in which you can adjust the number of available file descriptors when you use Sun Grid Engine:

1. Set the file descriptor limit in your login shell (`.cshrc`, `.tcshrc`, `.bashrc`, and so on).
2. Modify the `/etc/shell` file for each of the nodes on your cluster as described in the previous section, [“To Increase the Number of File Descriptors” on page 104](#). Remember that you must reboot all of the nodes in the cluster once you have finished modifying the files.
3. On a Sun Grid Engine execution host, modify the `$(SGE_ROOT)/default/common/sgeexecd` startup script to increase the file descriptor limit to the same value as the hard limit (as described in the previous section). You must restart the `sgeexecd` daemon on the host. Since this script is shared among the Sun Grid Engine execution hosts in the cluster using NFS, you may make the change on one host, and it will be propagated to the other Sun Grid Engine hosts in the cluster.



## List of Available MCA Parameters

---

This section lists the MCA parameters available in the Solaris OS implementation of Open MPI 1.3. The parameters are grouped under the framework to which they correspond.

To view all MCA parameters available in your implementation of Open MPI, type the following command:

```
% mpi_info --param all all
```

Note that the list of available MCA parameters is different for each operating system and implementation of Open MPI. For more information, see the Open MPI Web site at:

<http://www.open-mpi.org>

---

**Note** – The list of MCA parameters is constantly changing due to ongoing development. See the Open MPI Web site for more information.

---

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
mca	mca_param_files	Path for MCA configuration files containing default parameter values	\$HOME/.openmpi/mca-params.conf:/opt/SUNWhpc/HPC8.2/<compiler-name>/etc/openmpi-mca-params.conf
	mca_component_path	Path where to look for Open MPI and ORTE components	/opt/SUNWhpc/HPC8.2/lib/openmpi:\$HOME/.openmpi/components
	mca_verbose	Top-level verbosity parameter	none
	mca_component_show_load_errors	Whether or not to show errors for components that failed to load	1
	mca_component_disable_dlopen	Whether to attempt to disable opening dynamic components or not	0
mpi	mpi_param_check	Whether you want MPI API parameters checked at runtime or not. Possible values are 0 (no checking) and 1 (perform checking at runtime)	1
	mpi_yield_when_idle	Yield the processor when waiting for MPI communication (for MPI processes, will default to 1 when oversubscribing nodes)	0
	mpi_event_tick_rate	How often to progress TCP communications (-1 = leave default value. 0 = never, otherwise specified in microseconds)	10000
	mpi_show_handle_leaks	Whether or not MPI_FINALIZE shows all MPI handles that were not freed	0

---

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	<code>mpi_no_free_handles</code>	Whether to actually free MPI objects when their handles are freed	0
	<code>mpi_show_mca_params</code>	Whether to show all MCA parameter value during MPI_INIT or not (good for reproducibility of MPI jobs)	0 (don't show)
	<code>mpi_show_mca_params_file</code>	If <code>mpi_show_mca_params</code> is true, setting this string to a valid file name tells Open MPI to dump all the MCA parameter values into a file suitable for reading using the <code>mca_param_files</code> parameter (good for reproducibility of MPI jobs)	none
	<code>opal_paffinity_alone</code>	If nonzero, assume that this job is the only (set of) process(es) running on each node and bind processes to processors, starting with processor ID 0	0
	<code>mpi_keep_peer_hostnames</code>	If nonzero, save the string hostnames of all MPI peer processes (mostly for error / debugging output messages). This can add quite a bit of memory usage to each MPI process.	1
	<code>mpi_abort_delay</code>	If nonzero, print out an identifying message when MPI_ABORT is invoked (hostname, PID of the process that called MPI_ABORT) and delay for that many seconds before exiting (a negative delay value means to never abort). This allows attaching of a debugger before quitting the job.	0

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	<code>mpi_abort_print_stack</code>	If nonzero, print out a stack trace when <code>MPI_ABORT</code> is invoked	0 Note: this is an information parameter. Its value is returned by Open MPI. This value cannot be adjusted manually.
	<code>mpi_preconnect_all</code>	Whether to force MPI processes to create connections/ warmup with <i>all</i> peers during <code>MPI_INIT</code> (vs. making connections lazily -- upon the first MPI traffic between each process/peer pair)	0
	<code>mpi_leave_pinned</code>	Whether to use the "leave pinned" protocol or not. Enabling this setting can help bandwidth performance when repeatedly sending and receiving large messages with the same buffers over RDMA-based networks. Use this parameter with care; setting it to too high a value can lead to resource exhaustion.	0
	<code>mpi_leave_pinned_pipeline</code>	Whether to use the "leave pinned pipeline" protocol or not.	0
orte	<code>orte_base_user_debugger</code>	Sequence of user-level debuggers to search for in <code>orterun</code>	<code>totalview @mpirun@ -a @mpirun_args@ : fxp @mpirun@ -a @mpirun_args@</code>
	<code>orte_debug</code>	Whether or not to enable debugging output for all ORTE components (0 or 1)	0
	<code>orte_no_daemonize</code>	Whether to properly daemonize the ORTE daemons or not	0
	<code>orte_abort_timeout</code>	Time to wait [in seconds] before giving up on aborting an ORTE operation	10

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	orte_timing	Request that critical timing loops be measured	0
opal	opal_signal	If a signal is received, display the stack trace frame	6,10,8,11
backtrace	backtrace	Default selection set of components for the backtrace framework (<none> means "use all components that can be found")	none
	backtrace_base_verbose	Verbosity level for the backtrace framework (0 = no verbosity)	0
	backtrace_printstack_priority		0
memory	memory	Default selection set of components for the memory framework (none means "use all components that can be found")	none
	memory_base_verbose	Verbosity level for the memory framework (0 = no verbosity)	0
paffinity	paffinity	Default selection set of components for the paffinity framework (none means "use all components that can be found")	none
	paffinity_solaris_priority	Priority of the Solaris OS paffinity component	10
maffinity	maffinity	Default selection set of components for the maffinity framework (none means "use all components that can be found")	none
	maffinity_first_use_priority	Priority of the first_use maffinity component	10

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
timer	timer	Default selection set of components for the timer framework (<none> means "use all components that can be found")	none
	timer_base_verbose	Verbosity level for the timer framework (0 = no verbosity)	0
	timer_solaris_priority	Priority of the Solaris OS maffinity component	0
allocator	allocator	Default selection set of components for the allocator framework (none means "use all components that can be found")	none
	allocator_base_verbose	Verbosity level for the allocator framework (0 = no verbosity)	0
	allocator_basic_priority	Priority of the basic allocator component	0
	allocator_bucket_num_buckets	Number of buckets in the allocator bucket component	30
	allocator_bucket_priority	Priority of the allocator bucket component	0
coll	coll	Default selection set of components for the coll framework (none means "use all components that can be found")	none
	coll_base_verbose	Verbosity level for the coll framework (0 = no verbosity)	0
	coll_basic_priority	Priority of the basic coll component	10
	coll_basic_crossover	Minimum number of processes in a communicator before using the logarithmic algorithms	4

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	<code>coll_self_priority</code>	Priority of the self coll component	75
	<code>coll_sm_priority</code>	Priority of the sm coll component	0
	<code>coll_sm_control_size</code>	Length of the control data -- should usually be either the length of a cache line on most multiprocessors, or the size of a page on machines that support direct memory affinity page placement (in bytes)	4096
	<code>coll_sm_bootstrap_filename</code>	File name (in the Open MPI session directory) of the <code>coll_sm</code> component bootstrap rendezvous mmap file	<code>shared_mem_sm_bootstrap</code>
	<code>coll_sm_bootstrap_num_segments</code>	Number of segments in the bootstrap file	8
	<code>coll_sm_fragment_size</code>	Fragment size (in bytes) used for passing data through shared memory (will be rounded up to the nearest <code>control_size</code> size)	8192
	<code>coll_sm_mpool</code>	Name of the mpool component to use	sm
	<code>coll_sm_comm_in_use_flags</code>	Number of "in use" flags, used to mark a message passing area segment as currently being used or not (must be $\geq 2$ and $\leq$ <code>comm_num_segments</code> )	2
	<code>coll_sm_comm_num_segments</code>	Number of segments in each communicator's shared memory message passing area (must be $\geq 2$ , and must be a multiple of <code>comm_in_use_flags</code> )	8
	<code>coll_sm_tree_degree</code>	Degree of the tree for tree-based operations (must be $> 1$ and $\leq$ <code>min(control_size, 255)</code> )	4

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	<code>coll_sm_shared_mem_used_bootstrap</code>	Amount of shared memory used in the shared memory bootstrap area (in bytes)	160 Note: this is an information parameter. Its value is returned by Open MPI. This value cannot be adjusted manually.
	<code>coll_sm_info_num_procs</code>	Number of processes to use for the calculation of the <code>shared_mem_size</code> MCA information parameter (must be => 2)	4
	<code>coll_sm_shared_mem_used_data</code>	Amount of shared memory used in the shared memory data area for <code>info_num_procs</code> processes (in bytes)	548864 Note: this is an information parameter. Its value is returned by Open MPI. This value cannot be adjusted manually.
	<code>coll_tuned_priority</code>	Priority of the tuned coll component	30
	<code>coll_tuned_pre_allocate_memory_comm_size_limit</code>	Size of communicator where you want to stop pre-allocating memory for the fixed internal buffer. The buffer is used for message requests attached to the communicator data segment. For example, if you have a 100,000-node cluster, you might not want to pre-allocate 200,000 request handle slots per communicator instance.	32768
	<code>coll_tuned_use_dynamic_rules</code>	Switch used to decide whether to use static (compiled/if statements) or dynamic (built at runtime) decision function rules	0

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	coll_tuned_init_tree_fanout	Initial fanout used in the tree topologies for each communicator. This is only an initial estimate value. If a tuned collective needs a different fanout for an operation, it builds it dynamically. This parameter is only for the first guess and might save a little time.	4
	coll_tuned_init_chain_fanout	Initial fanout used in the chain (fanout followed by pipeline) topologies for each communicator. This is only an initial estimate. If a tuned collective needs a different fanout for an operation, it builds it dynamically. This parameter is only for the first guess and might save a little time.	4
	coll_tuned_use_dynamic_rules	Switch used to decide whether to use static (compiled/if statements) or dynamic (built at runtime) decision function rules (0 = use static rules)	0
io	io	Default selection set of components for the io framework (none means "use all components that can be found")	none
	io_base_freelist_initial_size	Initial MPI-2 IO request freelist size	16
	io_base_freelist_max_size	Max size of the MPI-2 IO request freelist	64
	io_base_freelist_increment	Increment size of the MPI-2 IO request freelist	16
	io_base_verbose	Verbosity level for the io framework (0 = no verbosity)	0

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	<code>io_romio_priority</code>	Priority of the <code>io_romio</code> component	10
	<code>io_romio_delete_priority</code>	Delete priority of the <code>io_romio</code> component	10
	<code>io_romio_enable_parallel_optimizations</code>	Enable set of Open MPI-added options to improve collective file I/O performance	0
<code>mpool</code>	<code>mpool</code>	Default selection set of components for the <code>mpool</code> framework (<none> means "use all components that can be found")	none
	<code>mpool_base_verbose</code>	Verbosity level for the <code>mpool</code> framework (0 = no verbosity)	0
	<code>mpool_sm_allocator</code>	Type of shared memory allocator used	bucket
	<code>mpool_sm_max_size</code>	Maximum size of the <code>sm_mpool</code> shared memory file	536870912
	<code>mpool_sm_min_size</code>	Minimum size of the <code>sm_mpool</code> shared memory file	134217728
	<code>mpool_sm_per_peer_size</code>	Size (in bytes) to allocate per local peer in the <code>sm_mpool</code> shared memory file, bounded by <code>min_size</code> and <code>max_size</code>	33554432
	<code>mpool_sm_priority</code>	Priority of the shared memory allocator for <code>mpool</code>	0
	<code>mpool_udapl_priority</code>	Priority of the <code>uDAPL</code> BTL	0
	<code>mpool_base_use_mem_hooks</code>	Use memory hooks for deregistering freed memory	0
	<code>mpool_use_mem_hooks</code>	Deprecated; use <code>mpool_base_use_mem_hook</code>	0
<code>pml</code>	<code>pml</code>	Default selection set of components for the <code>pml</code> framework ( <code>none</code> means "use all components that can be found")	<code>ob1</code>

**TABLE B-1** Available MCA Parameters

<b>Framework</b>	<b>Parameter Name</b>	<b>Description</b>	<b>Default Value</b>
	pml_base_verbose	Verbosity level for the pml framework (0 = no verbosity)	0
	pml_cm_free_list_num	Initial size of request free lists	4
	pml_cm_free_list_max	Maximum size of request free lists	-1
	pml_cm_free_list_inc	Number of elements to add when growing request free lists	64
	pml_cm_priority	CM PML selection priority	1
	pml_dr_free_list_num		4
	pml_dr_free_list_max		-1
	pml_dr_free_list_inc		64
	pml_dr_priority		1
	pml_dr_eager_limit		131072
	pml_dr_send_pipeline_depth		3
	pml_dr_wdog_timer_sec		5
	pml_dr_wdog_timer_usec		0
	pml_dr_wdog_timer_multiplier		1
	pml_dr_wdog_retry_max		1
	pml_dr_ack_timer_sec		10
	pml_dr_ack_timer_usec		0
	pml_dr_ack_timer_multiplier		1
	pml_dr_ack_retry_max		3
	pml_dr_enable_csum		1
	pml_obl_free_list_num		4
	pml_obl_free_list_max		-1
	pml_obl_free_list_inc		64
	pml_obl_priority		1
	pml_obl_eager_limit		131072
	pml_obl_send_pipeline_depth		3
	pml_obl_recv_pipeline_depth		4

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
bml	bml	Default selection set of components for the bml framework ( <i>none</i> means "use all components that can be found")	none
	bml_base_verbose	Verbosity level for the bml framework (0 = no verbosity)	0
	bml_r2_show_unreach_errors	Show error message when procs are unreachable	1
	bml_r2_priority		0
rcache	rcache	Default selection set of components for the rcache framework ( <i>none</i> means "use all components that can be found")	none
	rcache_base_verbose	Verbosity level for the rcache framework (0 = no verbosity)	0
	rcache_rb_priority		0
	rcache_vma_mru_len	The maximum size IN ENTRIES of the MRU (most recently used) rcache list	256
	rcache_vma_mru_size	The maximum size IN BYTES of the MRU (most recently used) rcache list	1073741824
	rcache_vma_priority		0
btl	btl	Default selection set of components for the btl framework ( <i>none</i> means "use all components that can be found")	none
	btl_base_debug	If btl_base_debug is 1 standard debug is output, if > 1 verbose debug is output	0
	btl_base_verbose	Verbosity level for the btl framework (0 = no verbosity)	0
	btl_self_free_list_num	Number of fragments by default	0

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	<code>btl_self_free_list_max</code>	Maximum number of fragments	-1
	<code>btl_self_free_list_inc</code>	Increment by this number of fragments	32
	<code>btl_self_eager_limit</code>	Eager size fragment (before the rendezvous protocol)	131072
	<code>btl_self_min_send_size</code>	Minimum fragment size after the rendezvous	262144
	<code>btl_self_max_send_size</code>	Maximum fragment size after the rendezvous	262144
	<code>btl_self_min_rdma_size</code>	Minimum fragment size for the RDMA transfer	2147483647
	<code>btl_self_max_rdma_size</code>	Maximum fragment size for the RDMA transfer	2147483647
	<code>btl_self_exclusivity</code>	Device exclusivity	65536
	<code>btl_self_flags</code>	Active behavior flags	10
	<code>btl_self_priority</code>	Device priority	0
	<code>btl_sm_free_list_num</code>		8
	<code>btl_sm_free_list_max</code>		-1
	<code>btl_sm_free_list_inc</code>		256
	<code>btl_sm_exclusivity</code>		65535
	<code>btl_sm_latency</code>		100
	<code>btl_sm_max_procs</code>		-1
	<code>btl_sm_sm_extra_procs</code>		2
	<code>btl_sm_mpool</code>		sm
	<code>btl_sm_eager_limit</code>		4096
	<code>btl_sm_max_frag_size</code>		32768
	<code>btl_sm_size_of_cb_queue</code>		128
	<code>btl_sm_cb_lazy_free_freq</code>		120
	<code>btl_sm_priority</code>		0
	<code>btl_tcp_if_include</code>		none
	<code>btl_tcp_if_exclude</code>		lo
	<code>btl_tcp_free_list_num</code>		8

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	btl_tcp_free_list_max		-1
	btl_tcp_free_list_inc		32
	btl_tcp_sndbuf		131072
	btl_tcp_rcvbuf		131072
	btl_tcp_endpoint_cache		30720
	btl_tcp_exclusivity		0
	btl_tcp_eager_limit		65536
	btl_tcp_min_send_size		65536
	btl_tcp_max_send_size		131072
	btl_tcp_min_rdma_size	The user message size at which point the MPI library will begin registering portions of the user's send and receive buffers, and begin using RDMA operations to write directly from user buffer to user buffer. The write operations occur in portions no greater than the value of btl_tcp_max_rdma_size.	131072
	btl_tcp_max_rdma_size	The maximum message size that the MPI library will use to perform the actual RDMA operation.	2147483647
	btl_tcp_flags		10
	btl_tcp_priority		0
	btl_udapl_free_list_num		8
	btl_udapl_free_list_max		-1
	btl_udapl_free_list_inc		8
	btl_udapl_mpool		udapl
	btl_udapl_max_modules		8
	btl_udapl_evd_qlen		32
	btl_udapl_max_request_dtos		18
	btl_udapl_max_recv_dtos		18

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	btl_udapl_num_recvs		8
	btl_udapl_num_sends		8
	btl_udapl_sr_win		4
	btl_udapl_timeout		10000000
	btl_udapl_eager_rdma_num		8
	btl_udapl_max_eager_rdma_peers		16
	btl_udapl_eager_rdma_win		4
	btl_udapl_timeout		10000000
	btl_udapl_eager_rdma_guarantee		0
	btl_udapl_conn_priv_data		1
	btl_udapl_exclusivity		1014
	btl_udapl_eager_limit		32768
	btl_udapl_min_send_size		16384
	btl_udapl_max_send_size		65536
	btl_udapl_min_rdma_size	The user message size at which point the MPI library will begin registering portions of the user's send and receive buffers, and begin using RDMA operations to write directly from user buffer to user buffer. The write operations occur in portions no greater than the value of btl_udapl_max_rdma_size.	524288
	btl_udapl_max_rdma_size	The maximum message size that the MPI library will use to perform the actual RDMA operation.	131072
	btl_udapl_flags		2
	btl_udapl_bandwidth		225
	btl_udapl_priority		0
	btl_base_include		none
	btl_base_exclude		none

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	<code>btl_base_warn_component_unused</code>	This parameter is used to turn on warning messages when certain NICs are not used.	0
mtl	<code>mtl</code>	Default selection set of components for the mtl framework ( <code>none</code> means "use all components that can be found").	<code>none</code>
	<code>mtl_base_verbose</code>	Verbosity level for the mtl framework (0 = no verbosity)	0
topo	<code>topo</code>	Default selection set of components for the topo framework ( <code>none</code> means "use all components that can be found").	<code>none</code>
	<code>topo_base_verbose</code>	Verbosity level for the topo framework (0 = no verbosity).	0
osc	<code>osc</code>	Default selection set of components for the osc framework ( <code>none</code> means "use all components that can be found").	<code>none</code>
	<code>osc_base_verbose</code>	Verbosity level for the osc framework (0 = no verbosity)	0
	<code>osc_pt2pt_no_locks</code>	Enable optimizations available only if <code>MPI_LOCK</code> is not used.	0
	<code>osc_pt2pt_eager_limit</code>	Max size of eagerly sent data	16384
	<code>osc_pt2pt_priority</code>		0
	<code>osc_rdma_fence_sync_method</code>	Specifies how to synchronize fence: <code>reduce_scatter</code> , <code>allreduce</code> , <code>alltoall</code>	<code>reduce_scatter</code>

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	osc_rdma_eager_send	Attempt to start data movement during communication call, instead of at synchronization time. Info key of same name overrides this value, if info key given.	0
	osc_rdma_no_locks	Enable optimizations available only if MPI_LOCK is not used.	0
	osc_rdma_priority		0
errmgr	errmgr	Default selection set of components for the errmgr framework (none means "use all components that can be found")	none
	errmgr_hnp_debug		0
	errmgr_hnp_priority		0
	errmgr_orted_debug		0
	errmgr_orted_priority		0
	errmgr_proxy_debug		0
	errmgr_proxy_priority		0
gpr	gpr	Default selection set of components for the gpr framework (none means "use all components that can be found").	none
	gpr_base_maxsize		2147483647
	gpr_base_blocksize		512
	gpr_null_priority		0
	gpr_proxy_debug		0
	gpr_proxy_priority		0
	gpr_replica_debug		0
	gpr_replica_isolate		0
	gpr_replica_priority		0

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
iof	iof	Default selection set of components for the iof framework (none means "use all components that can be found").	none
	iof_base_window_size		4096
	iof_base_service		0.0.0
	iof_proxy_debug		1
	iof_proxy_priority		0
	iof_svc_debug		1
ns	ns	Default selection set of components for the ns framework (none means "use all components that can be found").	none
	ns_proxy_debug		0
	ns_proxy_maxsize		2147483647
	ns_proxy_blocksize		512
	ns_proxy_priority		0
	ns_replica_debug		0
	ns_replica_isolate		0
	ns_replica_maxsize		2147483647
oob	oob	Default selection set of components for the oob framework (none means "use all components that can be found").	none
	oob_base_verbosity	Verbosity level for the oob framework (0 = no verbosity)	0
	oob_tcp_peer_limit		-1
	oob_tcp_peer_retries		60

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	<code>oob_tcp_debug</code>		0
	<code>oob_tcp_include</code>		none
	<code>oob_tcp_exclude</code>		none
	<code>oob_tcp_sndbuf</code>		131072
	<code>oob_tcp_rcvbuf</code>		131072
	<code>oob_tcp_connect_timeout</code>	connect() timeout in seconds, before trying next interface	10
	<code>oob_tcp_connect_sleep</code>	Enable (1) /Disable (0) random sleep for connection wireup	1
	<code>oob_tcp_listen_mode</code>	Mode for HNP to accept incoming connections. Values are: event, listen_thread	event
	<code>oob_tcp_listen_thread_max_queue</code>	High water mark for queued accepted socket list size	10
	<code>oob_tcp_listen_thread_max_time</code>	Maximum amount of time (in milliseconds) to wait between processing accepted socket list	10
	<code>oob_tcp_accept_spin_count</code>	Number of times to let accept return EWOULDBLOCK before updating accepted socket list	10
	<code>oob_tcp_priority</code>		0
ras	<code>ras</code>	Default selection set of components for the ras framework (none means "use all components that can be found")	none
	<code>ras_dash_host_priority</code>	Selection priority for the dash_host RAS component	5
	<code>ras_gridengine_debug</code>	Enable debugging output for the gridengine RAS component	0

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	ras_gridengine_priority	Priority of the gridengine ras component	100
	ras_gridengine_verbose	Enable verbose output for the gridengine ras component	0
	ras_gridengine_show_jobid	Show the JOB_ID of the Grid Engine job	0
	ras_hostfile_priority	Selection priority for the hostfile RAS component	10
	ras_localhost_priority	Selection priority for the localhost RAS component	0
	ras_tm_priority	Priority of the tm ras component	100
rds	rds	Default selection set of components for the rds framework (none means "use all components that can be found")	none
	rds_hostfile_debug	Toggle debug output for hostfile RDS component. 1=enable; 0=disable.	0
	rds_hostfile_path	Path to the ORTE Host filename.	/opt/SUNWhpc/HPC8.2/etc/openmpi-default-hostfile
	rds_hostfile_priority		0
	rds_proxy_priority		0
	rds_resfile_debug	Toggle debug output for resfile RDS component. 1=enable; 0=disable.	0
	rds_resfile_name	ORTE Resource filename	none
	rds_resfile_priority		0
rmaps	rmaps	Default selection set of components for the rmaps framework (none means "use all components that can be found")	none
	rmaps_base_verbose	Verbosity level for the rmaps framework	0

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	<code>rmaps_base_schedule_policy</code>	Scheduling policy for RMAPS. [slot   node]	slot
	<code>rmaps_base_pernode</code>	Launch one ppn as directed	0
	<code>rmaps_base_n_pernode</code>	Launch n procs/node	-1
	<code>rmaps_base_schedule_local</code>	If nonzero, allow scheduling MPI applications on the same node as mpirun (default). If zero, do not schedule any MPI applications on the same node as mpirun.	1
	<code>rmaps_base_no_oversubscribe</code>	If nonzero, then do not allow oversubscription of nodes. mpirun will return an error if there are not enough nodes to launch all processes without oversubscribing.	0
	<code>rmaps_round_robin_debug</code>	Toggle debug output for Round Robin RMAPS component	1
	<code>rmaps_round_robin_priority</code>	Selection priority for Round Robin RMAPS component	1
rmgr	<code>rmgr</code>	Default selection set of components for the rmgr framework (none means "use all components that can be found")	none
	<code>rmgr_proxy_priority</code>		0
	<code>rmgr_urm_priority</code>		0
rml	<code>rml</code>	Default selection set of components for the rml framework (none means "use all components that can be found")	none
	<code>rml_base_verbose</code>	Verbosity level for the rml framework (0 = no verbosity)	0
	<code>rml_oob_priority</code>		0

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
plm	plm	Default selection set of components for the plm framework (none means "use all components that can be found")	none
	plm_base_reuse_daemons	If nonzero, reuse daemons to launch dynamically spawned processes. If zero, do not reuse daemons (default)	0
	plm_fork_reap	Specifies whether or not to wait to reap all children before finalizing. 1=wait; 0= do not wait	1
	plm_fork_reap_timeout	When killing child processes, first send a SIGTERM, then wait at least this timeout (in seconds), then send a SIGKILL	0
	plm_fork_priority	Priority of this component	1
	plm_fork_debug	Specifies whether or not to enable debugging output. 1=enable; 0= do not enable	0
	plm_gridengine_debug	Enable debugging of gridengine plm component. 1=enable; 0= do not enable	0
	plm_gridengine_verbose	Enable verbose output of the gridengine qrsh -inherit command 1= enable; 0= do not enable	0
	plm_gridengine_priority	Priority of the gridengine plm component	100
	plm_gridengine_orted	The command name that the gridengine plm component will invoke for the ORTE daemon	orted
	plm_proxy_priority		0
	plm_rsh_debug	Enable debugging output for the rsh plm component 1=enable; 0= do not enable	0

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	<code>plm_rsh_num_concurrent</code>	How many <code>plm_rsh_agent</code> instances to invoke concurrently (value must be greater than 0)	128
	<code>plm_rsh_orted</code>	The command name that the <code>rsh</code> <code>plm</code> component will invoke for the ORTE daemon	<code>orted</code>
	<code>plm_rsh_force_rsh</code>	Force the launcher to always use <code>rsh</code> , even for local daemons	0
	<code>plm_rsh_priority</code>	Priority of the <code>rsh</code> <code>plm</code> component	10
	<code>plm_rsh_delay</code>	Delay (in seconds) between invocations of the remote agent, but only used when the <code>debug</code> MCA parameter is set to true, or when the top-level MCA debugging is enabled (otherwise this value is ignored)	1
	<code>plm_rsh_reap</code>	If set to 1, wait for all the processes to complete before exiting. Otherwise, quit immediately without waiting for confirmation that all other processes in the job have completed.	1
	<code>plm_rsh_assume_same_shell</code>	If set to 1, assume that the shell on the remote node is the same as the shell on the local node. Otherwise, probe for what the remote shell.	1
	<code>plm_rsh_agent</code>	The command used to launch executables on remote nodes (typically either " <code>ssh</code> " or " <code>rsh</code> ")	<code>ssh : rsh</code>
	<code>plm_tm_debug</code>	Enable debugging of the TM <code>plm</code> . 1= enable; 0= do not enable	0

**TABLE B-1** Available MCA Parameters

Framework	Parameter Name	Description	Default Value
	plm_tm_verbose	Enable verbose output of the TM plm. 1= enable; 0= do not enable	0
	plm_tm_priority	Default selection priority	75
	plm_tm_orted	Command to use to start proxy orted	orted
	plm_tm_want_path_check	Specify whether the launching process should check for the plm_tm_orted executable in the PATH before launching. Any nonzero value enables this check. Note: the TM API does not give an indication of failure.	1
sds	sds	Default selection set of components for the sds framework (none means "use all components that can be found")	none
	sds_base_verbose	Verbosity level for the sds framework (0 = no verbosity)	0
	sds_env_priority		0
	sds_pipe_priority		0
	sds_seed_priority		0
	sds_singleton_priority		0
soh	soh	Default selection set of components for the soh framework (none means "use all components that can be found")	none

# Index

---

## A

applications  
  migrating, 15  
  recompiling, 15

## B

batch systems, 7

## C

cluster  
  about, 5  
command line interface (CRE), 2  
compilation, 16  
compilers  
  mpicc, 16  
  mpicc compiler, 15  
  mpiCC, mpicxx, mpic++, 16  
  mpif77, 16  
  mpif90, 16  
  wrapper compilers, 16  
compiling  
  using the wrapper compilers, 16  
configurations, supported, 1  
CRE, 2

## D

D language, 81  
default settings  
  how to run a program with, 23  
documentation  
  MPI Reference Manual, xi  
  product notes, xi

Sun documentation on the web, xiii

DTrace, 81  
  advantages over MPI profiling, 85  
  attaching to an MPI process, 85  
  running with MPI programs, 83  
  tracing MPI programs, 85  
  tracking resource leaks, 88  
  using with MPI, 84

dtrace\_proc, 82

dtrace\_user, 82

Dynamic tracing  
  using with MPI, 83

## E

E privilege set  
  dtrace privileges, 83  
error classes, standard, 99  
error classes, Sun MPI I/O, 101  
error handling, MPI I/O, 101  
errors  
  tracing using DTrace, 85  
exceeding the file descriptor limit, 103

## F

File descriptor  
  exceeding the limit, 103

## G

Grid Engine  
  open source version, 8

- H**
- help, how to display, 32
  - How to
    - attach DTrace to an MPI process, 85
    - change the working directory, 31
    - determine which function is returning errors, 85
    - determine your mprun privileges, 82
    - display command help, 32
    - run a program as multiple processes, 23
    - run a program with default settings, 23
    - track down a resource leak, 89
    - use DTrace with an MPI program, 84
    - use DTrace with Sun MPI, 81
- L**
- limit -h, 103
  - linking, 16
- M**
- mapping MPI processes to nodes, 24
  - MCA parameters, 8
    - gridengine parameters, 51
  - messages, MPI, 99
  - MPI
    - attaching DTrace to a process, 85
    - running a program under DTrace, 84
    - Sun MPI, 3
    - tracing programs, 84
    - tracing programs using DTrace, 85
    - tracking resource leaks, 87
  - MPI messages, 99
  - MPI\_COMM\_WORLD
    - inheriting stdin, 30
  - mpicc compiler, 15
  - mpif77, 16
  - mpif90, 16
  - mpirun
    - mca option, 8
  - mprun
    - C, 31
    - default settings, 23
    - determining privileges on the cluster, 82
    - h, 32
    - np, 23
    - privileges for use with DTrace, 82
    - syntax, 20
  - v, 32
- N**
- node
    - mapping MPI processes to, 24
  - nodes
    - about, 5
- O**
- Open MPI
    - and Sun N1 Grid Engine, 7
- P**
- parallel environment (PE), 8
  - process
    - how to run a program as multiple, 23
    - mapping to nodes, 24
- R**
- remote nodes
    - standard output, 31
  - Resource leaks
    - determining using DTrace, 89
    - tracking, 87
- S**
- scalability, 2
  - Solaris Dynamic Tracing utility (DTrace), 81
  - standard error, 31
  - standard input, 30
  - standard output, 31
  - submitting jobs
    - under Sun N1 Grid Engine, 8
  - Sun Grid Engine
    - and Open MPI, 7
  - Sun HPC ClusterTools 6
    - migrating applications, 15
  - Sun N1 Grid Engine
    - gridengine MCA parameters, 51
- T**
- tmcc compiler, 15
  - troubleshooting, 99
- U**
- ulimit -Hn, 104

## **W**

wildcards

    using in tracing scripts, 87

working directory, how to change the, 31

