HPE Message Passing Interface (MPI) User Guide

Abstract

This publication describes how to use the HPE Message Passing Interface (MPI) 1.4 and MPT 2.20, which facilitate parallel programming and support the MPI standard.

Part Number: 007-3773-036
Published: June 2019
Edition: 1

Contents

About the HPE Message Passing Interface (MPI) software

The Message Passing Interface (MPI) standard supports C and Fortran programs with a library and supporting commands. MPI operates through a technique known as **message passing**. Message passing uses library calls to request data delivery in the following ways:

- From one process to another
- Between groups of processes

MPI also supports parallel file I/O and remote memory access (RMA).

The MPI from HPE software supports the MPI standard. MPI from HPE facilitates parallel programming. This publication describes MPI from HPE 1.4, which supports the MPI 3.1 standard.

MPI from HPE supports the OpenSHMEM standard. The OpenSHMEM standard describes a low-latency library that supports RMA on symmetric memory in parallel environments. The OpenSHMEM programming model is a partitioned global address space (PGAS) programming model. The model presents distributed processes with symmetric arrays that are accessible through `PUT` and `GET` operations from other processes. MPI from HPE supports OpenSHMEM version 1.4. The SGI SHMEM programming model is the basis for the OpenSHMEM™ programming model specification. The Open Source Software Solutions multivendor working group is developing the OpenSHMEM programming model.

The following significant features make MPI from HPE the preferred implementation:

- Data transfer optimizations for the HPE Superdome Flex Grid technology, where available, including single-copy data transfer.
- Multirail InfiniBand support, which takes full advantage of the multiple InfiniBand fabrics available on HPE SGI 8600 and SGI ICE systems.
- Optimized MPI remote memory access (RMA) one-sided commands.
- Support for multiple application binary interfaces (ABIs), including MPICH and Open MPI.

HPE support for MPI and OpenSHMEM is built on top of the Message Passing Toolkit (MPT). MPT is a high-performance communications middleware software product. On some platforms, MPT uses Array Services to launch applications. MPT is optimized for large-scale, high-performance cluster computing.

The following table describes compatibility between MPI from HPE 1.4 and other products.

Technology	Notes
Red Hat Enterprise Linux (RHEL) operating system SLES operating system CentOS operating system	RHEL 7.X and RHEL 6.X. SLES 15, SLES 12 SPX, and SLES 11 SPX. CentOS 7.X CentOS 6.X. For information about specific operating system support on cluster platforms, see the HPE Performance Cluster Manager documentation.
Fortran 2008	Supports Fortran 2008.
Computing platforms	Cluster platforms and HPE Superdome Flex Grid platforms. For a list of supported platforms, see the HPE Message Passing Interface (MPI) release notes.
Multirail InfiniBand (IB)	
Multirail Intel Omni-Path Architecture (OPA)	Supported on systems running IFS 10.2 or later.
TCP/IP communication	
Mellanox Fabric Collective Accelerator (FCA) 3.x / HCOLL	
NVIDIA GPUDirect remote direct memory access (RDMA) over IB and over OPA NVIDIA GPUDirect RDMA	For IB, requires Mellanox Open Fabrics Enterprise Distribution (OFED). No support for MPI RMA passive windows.
Checkpoint-restart (CPR), supported through Berkeley Lab checkpoint restart (BLCR).	Supports jobs running over shared memory, InfiniBand, and TCP/IP. No support for CPR when using the following: • OpenSHMEM • MPI remote memory access (RMA) passive windows • MPI Spawn • Process management interface (PMI), which is commonly used by the simple Linux utility for resource management (Slurm)

Table Continued

Technology	Notes
Third-party debugging and profiling tools: • DDT from Allinea Software • Intel VTune Amplifier • TotalView for HPC from Rogue Wave Software • Tuning and Analysis Utilities (TAU) • Vampir	Contact HPE for information about additional debugging and profiling tools.
Process management interfaces (PMIs), specifically PMIx and PMI-2.	Supported when running under Slurm.
Third-party workload managers: • Altair PBS Professional • Slurm • UNIVA Grid Engine • IBM LSF • Moab HPC Suite and Torque	

The MPI from HPE documentation uses certain terms in the following specific ways:

- **Cluster**, **cluster system**.

  For example, the following are all HPE cluster systems:
  - HPE Apollo systems
  - HPE SGI 8600 systems
  - SGI ICE systems
  - SGI Rackable systems

- **HPE Superdome Flex Grid technology system**.

  HPE Superdome Flex Grid systems support a single Linux image of thousands of processors. This image is distributed over many sockets and many hub application-specific integrated circuits (ASICs). The HPE Superdome Flex Grid technology was formerly known as **NUMAlink technology**.

  For example, the following are all HPE Superdome Flex Grid systems:
  - HPE Superdome Flex Server platforms
  - HPE Integrity MC990 X systems
  - SGI UV systems

Configuring the Message Passing Toolkit (MPT)

When you install the MPI from HPE software, you also install MPT. Before you can run any MPI from HPE programs, configure the MPT software. The procedures in this chapter explain how to configure MPT.

High-performance computers often host several released versions of MPT. This environment provides users with the flexibility to develop and run MPI from HPE programs. If your site installs multiple MPT versions, use the configuration instructions in this chapter to accommodate these multiple versions.

The following list shows where you can find information about how to install and configure MPT:

- If you have a cluster that runs the HPE Performance Cluster Manager, see the following:

  HPE Performance Cluster Manager Installation Guide
- If you have a cluster that runs HPE SGI Management center, see the following:

  HPE SGI Management Suite Installation and Configuration Guide
- If you have a cluster that runs the HPE Insight Cluster Manager Utility, see the following:

  Installing the Array Services software on clusters that use the HPE Insight Cluster Management Utility (CMU) on page 9
- If you have an HPE Flex Grid system, there is no cluster manager. See the following:

  Configuring MPT on an HPE Superdome Flex Grid system on page 11

Installing the Array Services software on clusters that use the HPE Insight Cluster Management Utility (CMU)

Before you can run MPI from HPE programs, install the Array Services software on systems that do not use Slurm. If Slurm is installed on the cluster, do not install Array Services. Systems with Slurm do not use the Array Services software.

The HPE factory installs Array Services on many cluster systems, but not all. Verify whether the Array Services software is installed on your system before you attempt an installation.

The installation process uses the `arrayconfig` command. The `arrayconfig` command creates the following files on the compute service node to which you are logged in:

- `/etc/array/arrayd.conf`
- `/etc/array/arrayd.auth`

NOTE: On ARM architecture systems, such as Apollo 70 systems, make sure that the appropriate `*.aarch64.rpm` package is installed.

The following topics explain how to install and configure Array Services:

- **Installing and configuring Array Services on clusters that use the HPE Insight Cluster Manager Utility (CMU)** on page 10
- **Array Services security options** on page 10
- **Array Services configuration examples** on page 11

Installing and configuring Array Services on clusters that use the HPE Insight Cluster Manager Utility (CMU)

The following procedure explains how to install Array Services on systems that use CMU.

Procedure

1. Log into the admin node.
2. Write the list of cluster nodes to a temporary file:

   ```
   $ cmu_show_nodes > nodes.txt
   ```

3. (Conditional) Edit the node list to include only the compute nodes that you want to include in the array.

   Complete this step if you do not want to include all the cluster nodes in the array.
4. Copy the list of cluster nodes to a directory of your choosing on one of the compute nodes.

   For example:

   ```
   $ scp nodes.txt first_compute_node:/somewhere
   ```

   The preceding command copies the `nodes.txt` file to the `somewhere` directory on `first_compute_node`.
5. Use the `ssh` command to log into the compute node to which you wrote the node list.

   For example:

   ```
   $ ssh first_compute_node
   ```

6. Change to the directory in which the node list resides.

   For example:

   ```
   $ cd somewhere
   ```

7. Use the `arrayconfig` command to configure the nodes into an array.

   The command format is as follows:

   ```
   /usr/sbin/arrayconfig -a arrayname -f -m -A method node node ...
   ```

   The command parameters are as follows:

   - For *arrayname*, specify a name for the array. The default is `default`.
   - For *method*, specify `munge`, `none`, or `simple`. For information about the security levels, see the following:

     Array Services security options on page 10
   - For *node*, specify the nodes to include. For example, `/tmp/nodelist` or a list of individual node names.

   For example, the following command configures the nodes into an array and uses the `-D` parameter to distribute the `arrayd.conf` file:

   ```
   $ arrayconfig -fmD -a default -A simple `cat nodes.txt`
   ```

Array Services security options

The following table explains the Array Services security options.

Security option name	Effect
`simple` (default)	Generates hostname/key pairs by using either the OpenSSL, `rand` command, 64-bit values (if available) or by using `$RANDOM` Bash facilities.
`munge`	Configures additional security provided by MUNGE. The installation process installs `munge` by default but does not configure `munge`.
`none`	You can configure `none` in one of the following ways: • `none` on all nodes, including the login node. The login node can be a compute node that is designated for user logins and/or other services. When you specify `none`, Array Services disables all authentication. **OR** • `none` on the compute nodes and `noremote` on the login node. A security setting of `noremote` prevents remote logins to the array. With the preceding settings, users must run their jobs directly from the compute services nodes. Users cannot submit MPI from HPE jobs remotely. The following topic explains how to configure `noremote` on the login node: **Manually configuring Array Services on multiple hosts** on page 94

On cluster systems, Array Services uses the pluggable authentication mechanism (PAM) to control machine access. Typically, the PAM file requires no editing or other user intervention. The Array Services PAM configuration file is as follows:

```
/etc/pam.d/arrayd
```

On RHEL platforms, MPI from HPE supports SELinux.

More information

(Conditional) Configuring SELinux (HPE Superdome Flex Grid RHEL platforms only) on page 16

Array Services configuration examples

Example 1. To specify that array `myarray` use MUNGE security and include all HPE Apollo compute nodes, enter the following command:

```
# /usr/sbin/arrayconfig -a myarray -f -m -A munge /tmp/nodelist
```

Example 2. To specify that array `yourarray` use no security, include one compute service node, and include all HPE Apollo compute nodes, enter the following command:

```
# /usr/sbin/arrayconfig -a yourarray -f -m -A none n1 /tmp/nodelist
```

Configuring MPT on an HPE Superdome Flex Grid system

The information in the following procedures explains how to configure MPT on an HPE Superdome Flex Grid system:

- **Verifying prerequisites** on page 12
- **(Optional) Installing the MPT software into a nondefault working directory** on page 12
- **Adjusting file resource limits** on page 14

- **Completing the configuration** on page 15
- **(Conditional) Configuring SELinux (HPE Superdome Flex Grid RHEL platforms only)** on page 16

Verifying prerequisites

The following procedure explains how to verify the MPT software installation prerequisites.

Procedure

1. As the `administrator` user, log into the HPE Superdome Flex Grid system.
2. Verify that one of the supported operating systems is installed and configured.

   You can enter the following command to verify your operating system version:

   ```
   # cat /etc/*release
   ```
3. Verify that the MPI from HPE 1.4 release is installed.

   For example:

   ```
   # cat /etc/hpe-mpi-release
   HPE MPI 1.4, Build xxxxxx.rhel76-xxxxxxxxxx
   ```
4. Proceed to one of the following:

   - To configure MPT in a nondefault installation directory, proceed to the following:

     (Optional) Installing the MPT software into a nondefault working directory on page 12

     The preceding topic assumes that you want to maintain more than one released version of the software on your system.
   - To configure MPT in the default installation directory, proceed to the following:

     Adjusting file resource limits on page 14

     The preceding topic assumes that you want only one version of MPT on your system.

(Optional) Installing the MPT software into a nondefault working directory

Complete the procedure in this topic to install MPT into a custom, nondefault working directory. You might want to complete this procedure if you have a nondefault file system.

The RPM utility enables you to create, install, and manage relocatable packages. You can install a matched set of MPT RPMs in either the default location or an alternative location. The default location for installing the MPT RPM is `/opt/hpe/hpc/mpt/mpt-2.`*`rel_level`*. To install the MPT RPM in an alternative location, use the `--relocate` parameter to the `rpm` command. The `--relocate` parameter specifies an alternative base directory for the MPT software installation.

Either `/opt/hpe/hpc/mpt/mpt-2.`*`rel_level`* or both `/opt/hpe/hpc/mpt/mpt-2.`*`rel_level`* and `/usr/share/modules/modulefiles/mpt` can be relocated. The post installation script reconfigures the module file for the new location as long as the *oldpath* argument in the `rpm` command precisely matches the description in the RPM info. The general format for the `rpm` command is as follows:

```
rpm --relocate oldpath=newpath
```

Variable	Specification
oldpath	The current location of the MPT software. If you install the MPT software in an alternative location, the `rpm` command *oldpath* argument must match the relocation listed in the RPM. This match is necessary for the environment module automatic modification feature to be correct.
newpath	The location in which you want to install the MPT software.

Procedure

1. Plan how to avoid problems related to uninstalled RPM dependencies.

   The following are two approaches:

   - Option 1: Did you install from a system that does not run MPT jobs? If yes, it might be appropriate to use the `--nodeps` parameter on the `rpm` command line. This parameter directs the `rpm` command to ignore dependencies.
   - Option 2: Did you install from a system or from cluster nodes upon which MPT jobs run? If yes, enter package manager commands on each cluster node or cluster node image. These commands install the needed prerequisites on all the cluster nodes locally. For example:
     - On SLES platforms, enter the following command:

       ```
       # zypper install cpuset-utils arraysvcs xpmem libbitmask
       ```
     - On RHEL platforms, enter the following command:

       ```
       # yum install cpuset-utils arraysvcs xpmem  libbitmask
       ```

2. Use the `rpm` command to specify an alternative location for the MPT software bundle.

   Example 1. The following example shows how to install MPT in `/usr/local/hpe/mpt/2.20` rather than in `/opt`, which is the default:

   ```
   # rpm -i --relocate /opt/hpe/hpc/mpt/2.20=/usr/local/hpe/mpt/2.20 sgi-mpt-*.x86.rpm
   ```

   Example 2: The following RHEL example shows how to install the modules and the total MPT software bundle into `/usr/local/hpe/mpt/2.20` and `/usr/local/mod/2.20`:

   ```
   # rpm -i --relocate /opt/hpe/hpc/mpt/2.20=/usr/local/hpe/mpt/2.20 \
   --relocate /usr/share/Modules/modulefiles/mpt=/usr/local/mod/2.20 sgi-mpt-*.x86_64.rpm
   ```

   In the preceding RHEL example, the `Modules` directory in the argument to the second `--relocate` parameter begins with an uppercase letter.

   Example 3. The following SLES example shows how to install the modules and the total MPT software bundle to `/usr/local/hpe/mpt/2.20` and `/usr/local/mod/2.20`:

   ```
   # rpm -i --relocate /opt/hpe/hpc/mpt/2.20=/usr/local/hpe/mpt/2.20 \
   --relocate /usr/share/modules/modulefiles/mpt=/usr/local/mod/2.20 sgi-mpt-*.x86_64.rpm
   ```

   In the preceding SLES example, the `modules` directory in the argument to the second `--relocate` parameter begins with a lowercase letter.

   Example 4:

The following example `rpm` command output shows the available relocations:

```
# rpm -qpi sgi-mpt-2.20-sgi*.x86_64.rpm
... Relocations: /opt/hpe/hpc/mpt/2.20 /usr/share/modules/modulefiles/mpt
```

NOTE: In the preceding output, the example shows only the significant message at the end of the output string.

3. Proceed to the following:

   Adjusting file resource limits on page 14

For more information about using the `rpm` command, see the `rpm`(8) manpage.

Adjusting file resource limits

The following procedure explains how to increase resource limits on the number of open files and enforce new security policies.

Procedure

1. Enter the following command to retrieve the number of cores on this computer:

   ```
   # cat /proc/cpuinfo | grep processor | wc -l
   ```

   In the preceding line, the last character is a lowercase `L`, not the number `1`.

   This `cat` command returns the number of cores.

2. Use a text editor to open file `/etc/security/limits.conf`.

3. Add the following line to file `/etc/security/limits.conf`:

   ```
   *     hard    nofile      limit
   ```

   For *limit*, specify an open file limit, for the number of MPI processes per host, based on the guidelines in the following table.

Processes/ host	***limit***
Fewer than 512	3000
Up to 1024	6000
Up to 2048	8192 (default)
4096 or more	21000

   MPI jobs require many file descriptors. On larger systems, you might need to increase the systemwide limit on the number of open files. The default value for the file-limit resource is 8192. For example, the following line is suitable for 512 MPI processes per host:

   ```
   *     hard    nofile  3000
   ```

4. Add the following line to file `/etc/security/limits.conf`:

   ```
   *     hard   memlock  unlimited
   ```

The preceding line increases the resource limit for locked memory.

5. Save and close file `/etc/security/limits.conf`.
6. Use a text editor to open file `/etc/pam.d/login`, which is the Linux pluggable authentication module (PAM) configuration file.
7. Add the following line to file `/etc/pam.d/login`:

   ```
   session    required     /lib/security/pam_limits.so
   ```

8. Save and close the file.
9. (Conditional) Verify resource limits.

   Complete this step on nodes that run Array Services and are installed with the SLES 15 operating system or the SLES12 SP4 operating system.

   Examine the following file:

   ```
   /sys/fs/cgroup/pids/system.slice/array.service/pids.max
   ```

   Ensure that the preceding file contains the value `max`.

   The default `pids.max` value inherited from the operating system can be too small for some system configurations.

10. (Conditional) Update other authentication configuration files as needed.

    Perform this step if your site allows other login methods, such as `ssh` or `rlogin`.

11. Proceed to the following:

    Completing the configuration on page 15

Completing the configuration

The following procedure explains how to complete the MPT configuration.

Procedure

1. Run a test MPI program to make sure that the new software is working as expected.
2. (Conditional) Inform your user community of the location of the new MPT release on this computer.

   Complete this step if you moved the MPT software to a nondefault location.

   In examples, the module files are located in the following directories:

   - On RHEL platforms:

     ```
     /opt/mpt/mpt-2.20/usr/share/Modules/modulefiles/mpt/2.20
     ```

   - On SLES platforms:

     ```
     /opt/mpt/mpt-2.20/usr/share/modules/modulefiles/mpt/2.20
     ```

3. (Conditional) Configure SELinux.

   Proceed to the following:

   (Conditional) Configuring SELinux (HPE Superdome Flex Grid RHEL platforms only) on page 16

(Conditional) Configuring SELinux (HPE Superdome Flex Grid RHEL platforms only)

The procedure in this topic explains how to load the Array Services policy module into SELinux. Before you use SELinux enforcing mode for multihost applications, make sure to complete the procedure in this topic.

After you complete this procedure, if you encounter problems when launching multihost applications, check for SELinux reports in the system log.

For information about how to configure SELinux, see your HPE Superdome Flex Server documentation.

For information about how to fun MPI from HPE with security software, contact your HPE support representative.

Procedure

1. Log into the system as the administrator user.
2. Use the `semodule` command to load the policy module.

   For example, the following command loads the policy module from its default location:

   ```
   $ semodule -i /usr/share/selinux/packages/sgi_arraysvcs.pp.bz2
   ```
3. Use the `fixfiles` command to update the Array Services files to reflect the appropriate security labels.

   For example:

   ```
   $ fixfiles -R sgi-arraysvcs restore
   ```

   The preceding steps show how to use the `semodule` command and the `fixfiles` command to configure SELinux. As an alternative to these commands, use the `arrayconfig -S` command as described on the `arrayconfig`(1M) manpage. For example:

   ```
   $ arrayconfig -fmD -S hostA hostB hostC...
   ```

Getting started with MPI from HPE

This chapter provides procedures for building MPI applications. It provides examples of the use of the `mpirun` command to launch MPI jobs. It also provides procedures for building and running SHMEM applications.

The process of running MPI applications consists of the following procedures:

- **Loading the MPI software module** on page 17
- **Compiling and linking the MPI program** on page 17
- **Launching the MPI application** on page 19

Loading the MPI software module

A program must be able to find the MPT commands and libraries. The paths to these commands and libraries are set in the MPT software module.

For example, to load the MPT module, enter the following command:

```
% module load mpt
```

MPI applications compiled with one implementation of MPI do not necessarily run with another implementation. For example, applications compiled with HPE MPI do not run with Open MPI. On the x86_64 platform, there are two versions of HPE MPI:

- The classic version of HPE MPI. Applications compiled with the classic version of HPE MPI do not run with other MPI implementations. This version requires you to load the `mpt` module.
- The compatibility version of HPE MPI. Applications compiled with the compatibility version of HPE MPI run with derivatives of MPICH. Likewise, applications compiled with derivatives of MPICH run with HPE MPI. These derivatives include Intel MPI, MVAPICH, and Cray MPI. This version of HPE MPI requires you to load the `hmpt` module. On AArch64 (ARM) platforms, for example, on an Apollo 70 systems, HPE MPI applications run with MPICH derivatives. Both the `mpt` and `hmpt` modules load this version.

If your site has a custom MPT installation location, edit the MPT software module to reflect the installation location. Sample MPT module files reside in the following locations:

- `/usr/share/modules/modulefiles/mpt/`*`mpt_rel`*
- `/opt/hpe/hpc/mpt/mpt-`*`mpt_rel`*`/doc/sgi-mpt.`*`rel`*`.template`

To modify the software module, edit the file. Change the value of the `BASEPATH` variable to the base installation location of MPT.

Compiling and linking the MPI program

You can use one of the MPI wrapper compiler commands to run your program, or you can call the compiler directly. The following topics explain these two alternatives:

- **Compiling with the wrapper compilers** on page 17
- **Compiling with the GNU or Intel compilers** on page 18

Compiling with the wrapper compilers

The MPI wrapper compilers automatically incorporate the compiling and linking functions into the compiler command. If possible, use one of the following wrapper compiler commands to run your program:

- `mpif08 -I /install_path/usr/include file.f -L lib_path/usr/lib`
- `mpif90 -I /install_path/usr/include file.f -L lib_path/usr/lib`
- `mpicxx -I /install_path/usr/include file.c -L lib_path/usr/lib`
- `mpicc -I /install_path/usr/include file.c -L lib_path/usr/lib`

The following table explains variables in the preceding commands.

Variable	Specification
install_path	The path to the directory in which the MPT software is installed.
file	The name of your C or Fortran program file name.
lib_path	The path to the library files.

For example:

```
% mpicc -I /tmp/usr/include simple1_mpi.c -L /tmp/usr/lib
```

Compiling with the GNU or Intel compilers

This topic explains how to run an MPI program to call GNU or Intel compilers directly. When the MPT RPM is installed as default, the commands to build an MPI-based application using the `.so` files are as follows:

- To compile using GNU compilers, choose one of the following commands:

  ```
  % g++ -o myprog myprog.C -lmpi++ -lmpi
  % gcc -o myprog myprog.c -lmpi
  ```

- To compile programs with the Intel compilers, choose one of the following commands:

  ```
  % icc -o myprog myprog.c -lmpi     # C - version 8
  % mpif08 simple1_mpi.f             # Fortran 2008 wrapper compiler
  % mpif90 simple1_mpi.f             # Fortran 90 wrapper compiler
  % ifort -o myprog myprog.f -lmpi   # Fortran - version 8
  % mpicc -o myprog myprog.c         # MPI C wrapper compiler
  % mpicxx -o myprog myprog.C        # MPI C++ wrapper compiler
  ```

  NOTE: Use the Intel compiler to compile Fortran 90 programs.

- Insert a `USE MPI` statement near the beginning of each subprogram if you want to accomplish the following:

  - You want to compile the program with the Intel Fortran compiler
  - You want to enable compile-time checking of MPI subroutine calls

  In this situation, use the following compiler command:

  ```
  % ifort -I/usr/include -o myprog myprog.f -lmpi # version 8
  ```

NOTE: The preceding command assumes a default MPT installation. Your system administrator can tell you the location of the module files, when the following conditions are true:

◦ Your site installed more than one MPT version.

◦ Your site installed MPT into a nondefault location.

For a nondefault installation location, replace `/usr/include` with the name of the relocated directory.

- To compile hybrid MPI/OpenMP applications with the Open64 compiler, specify separate compilation and link command lines. The Open64 version of the OpenMP library requires you to specify `-openmp` on the compiler command line. The `-openmp` parameter, however, interferes with proper linking of MPI libraries. Use the following sequence:

```
% opencc -o myprog.o -openmp -c myprog.c
% opencc -o myprog myprog.o -lopenmp -lmpi
```

Launching the MPI application

You can use either a workload manager or the `mpirun` command to launch an MPI application.

The following topics explain these alternatives:

- **Using a workload manager to launch an MPI application** on page 19
- **Using the mpirun Command to Launch an MPI Application** on page 21
- **Using MPI Spawn Functions to Launch an MPI Application** on page 22

Using a workload manager to launch an MPI application

You can run an MPI job from a workload manager. The workload manager starts the job on the cluster nodes and CPUs that have been allocated to the job. Workload managers include PBS Professional, Torque, or Load Sharing Facility (LSF). For multinode MPI jobs, use commands to communicate the node and CPU selection information to the workload manager. MPT includes one of these commands, `mpiexec_mpt`, and the PBS Professional workload manager includes another such command, `mpiexec`. The following topics describe how to start MPI jobs with specific workload managers:

- **PBS Professional** on page 19
- **Torque** on page 20
- **Simple Linux utility for resource management (Slurm)** on page 20

PBS Professional

You can run MPI applications from job scripts that you submit through workload managers such as the PBS Professional workload manager.

Process and thread pinning onto CPUs is especially important on HPE Superdome Flex Grid systems. When PBS Professional is set up to run each application in a set of dedicated cpusets, process pinning occurs automatically. Process and thread pinning occur when you use `omplace`.

Example 1. To run an MPI application with 512 processes, include the following in the directive file:

```
#PBS -l select=512:ncpus=1
    mpiexec_mpt ./a.out
```

Example 2. To run an MPI application with 512 Processes and four OpenMP threads per process, include the following in the directive file:

```
#PBS -l select=512:ncpus=4
    mpiexec_mpt omplace -nt 4 ./a.out
```

PBS Professional includes an `mpiexec` command that enables you to run MPI from HPE applications. The PBS Professional command does not support the same set of extended options that the `mpiexec_mpt` command supports.

Some third-party debuggers support the `mpiexec_mpt` command. For more information, see the following:

Debugging MPI applications on page 34

For more information about the PBS Professional workload manager, see the following website:

http://www.pbsworks.com/SupportGT.aspx?d=PBS-Professional,-Documentation

Torque

When running Torque, HPE recommends that you use the following `mpiexec_mpt` command to launch MPT jobs:

`mpiexec_mpt [ -n` ***`P`*** **`] ./a.out`**

The *P* argument is optional. By default, the program runs with the original number of processes specified on the job initialization in Torque. To use *P*, specify is the total number of MPI processes in the application. This syntax applies whether running on a single host or a clustered system.

For more information, see the `mpiexec_mpt`(1) manpage.

For more information about using the `mpiexec_mpt` command with parallel debuggers, see the following:

Using the TotalView debugger with MPI programs on page 34

Simple Linux utility for resource management (Slurm)

MPI from HPE is adapted for use with the Slurm workload manager. To use MPI from HPE with Slurm, use the Slurm PMI-2 MPI plug-in or the Slurm PMIx MPI plug-in.

For task running under Slurm, use one of the following CPU pinning methods:

- Use `omplace` to place tasks and threads launched by Slurm.

  For example, the following command places two threads of each of eight instances of `a.out` across CPUs 20-40:

  ```
  srun -N2 -n 8 --cpus-per-task=2 --mpi=pmi2 omplace -c 20-40 ./a.out
  ```

  MPI from HPE does not support the `omplace -s` option.
- Use automatic placement by HPE MPI.

  Placement occurs automatically when the following conditions exist:

  - The application is launched in Slurm using `salloc` or `srun`.
  - Task affinity is not configured or enabled in Slurm. When task affinity is configured and enabled in Slurm, CPU pinning is left to Slurm.
  - `dplace` and/or `omplace` are not used.

For general information about Slurm, see the following website:

http://slurm.schedmd.com

For more information about how to use MPI with Slurm, see the following website:

http://slurm.schedmd.com/mpi_guide.html

Using the `mpirun` Command to Launch an MPI Application

The `mpirun` command starts an MPI application. Use the `mpirun` command when you are not using a resource manager, such as PBS Professional.

For the complete command-line syntax, see the `mpirun`(1) manpage.

The following topics explain how to use the `mpirun` command to launch a variety of applications:

- **Launching a single program on the local host** on page 21
- **Launching a multiple program multiple data (MPMD) application on the local host** on page 21
- **Launching a distributed application** on page 21

Launching a single program on the local host

To run an application on the local host, enter the `mpirun` command with the `-np` argument. Your entry must include the number of processes to run and the name of the MPI executable file.

For example, the following command starts three instances of the `mtest` application, which is passed an argument list (arguments are optional):

```
% mpirun -np 3 mtest 1000 "arg2"
```

Launching a multiple program multiple data (MPMD) application on the local host

You are not required to use a different host in each entry that you specify on the `mpirun` command. You can start a job that has multiple executable files on the same host.

For example, the following command runs one copy of `prog1` and five copies of `prog2` on the local host:

```
% mpirun -np 1 prog1 : -np 5 prog2
```

In the preceding example, both executable files use shared memory.

Launching a distributed application

You can use the `mpirun` command to start a program that consists of any number of executable files and processes. You can distribute the program to any number of hosts. A host is usually a single machine, but it can be any accessible computer running the Array Services software. For a list of the available nodes on systems running Array Services software, enter the following command:

```
% ainfo machines
```

You can list multiple entries on the `mpirun` command line. Each entry can contain an MPI executable file and a combination of hosts and process counts for running it. This capability lets you run different executable files on the same or different hosts, and all run as part of the same MPI application.

The examples show various ways to start an application that consists of multiple MPI executable files on multiple hosts.

Example 1. The following command runs 10 instances of the `a.out` file on `host_a`:

```
% mpirun host_a -np 10 a.out
```

Example 2. The following command launches 10 instances of `fred` on each of three hosts. `fred` has two input arguments.

```
% mpirun host_a, host_b, host_c -np 10 fred arg1 arg2
```

Example 3. The following command launches 10 instances of `fred`, with different numbers of instances on each processor:

```
% mpirun host_a -np 2, host_b -np 3, host_c -np 5 fred arg1 arg2
```

Example 4. The following command launches an MPI application on different hosts with different numbers of processes and executable files:

```
% mpirun host_a 6 a.out : host_b -np 26 b.out
```

Using MPI Spawn Functions to Launch an MPI Application

The following two functions enable the MPI spawn feature:

- `MPI_Comm_spawn`
- `MPI_Comm_spawn_multiple`

For information about how to use the spawn functions, see the following:

- **Specifying the universe size automatically** on page 22
- **Specifying the universe size directly** on page 22
- **Specifying the universe size on the mpirun command** on page 23
- **Specifying host information** on page 23

For more information, see the `mpiexec_mpt`(1) manpage.

Specifying the universe size automatically

You can specify the universe size automatically when you specify the following:

- The `-spawn` parameter on the `mpiexec_mpt` command
- The following MPI process creation functions:
  - `MPI_Comm_spawn`
  - `MPI_Comm_spawn_multiple`

For example, assume that when you submitted a job, you specified 10 processes. The following command starts three instances of the `mtest` MPI application, so the application can spawn seven more:

```
% mpiexec_mpt -spawn -np 3 mtest
```

Specifying the universe size directly

You can specify the universe size directly by setting one of the following shell variables:

- `MPI_UNIVERSE_SIZE`

  Set `MPI_UNIVERSE_SIZE` to the maximum number of processes possible in the universe.

- `MPI_UNIVERSE`

  Set `MPI_UNIVERSE` to control both the universe size and the possible set of hosts that processes can run on.

  This variable specifies hosts upon which processes can be launched. The syntax for this variable is a list of host program specifications without a specified application or argument list. For example:

- `"host_a, host_b"`
- `"host_a 8, host_b 16"`
- `"host_a, host_b 12"`
- `"host_a, host_b -np 16"`

If `MPI_UNIVERSE` is not specified, MPI `spawn` requests place new processes on the local host.

Specifying the universe size on the `mpirun` command

To specify the universe size on the `mpirun` command, use the `-up` parameter.

Specifying host information

You can pass host information to `MPI_Comm_spawn` or `MPI_Comm_spawn_multiple` in `MPI_info` objects. When you use this method, you can specify hosts that are outside of the list of hosts specified to `MPI_UNIVERSE`. The following table shows are the supported `MPI_info` keys and the values associated with the keys.

Info key	Value string
hostfile	The name of the file that contains the lists of hosts upon which to spawn MPI processes. Use a space character or a tab character to separate multiple host names.
`MPI_SGI_NODELIST`	The list of hosts upon which to spawn MPI processes. Use a space character or a tab character to separate multiple host names.

The following examples call `MPI_Comm_spawn` using the `MPI_SGI_NODELIST` and `hostlist` `MPI_info` objects to specify the hosts on which to launch MPI processes.

Example 1:

```
MPI_Info info;
MPI_Info_create(&info);
MPI_Info_set(info, "MPI_SGI_NODELIST", "host_b host_b");
MPI_Comm_spawn("b.out", MPI_ARGV_NULL, 2, info,...);
```

Example 2:

```
char * list = "host_b host_b";
int fd = open("list.txt", O_WRONLY);
write(fd, list, strlen(list) + 1);
MPI_Info_set(info, "hostfile", "list.txt");
MPI_Comm_spawn("b.out", MPI_ARGV_NULL, 2, info,...);

% export MPI_UNIVERSE="host_a 4"
% export MPI_UNIVERSE_SIZE=20
% mpiexec_mpt -np 3 coupler
```

The preceding lines run two `b.out` processes on `host_b`. If the `coupler` program launches spawn processes without specifying the desired hosts in their `info` argument, they are placed within the defined `MPI_UNIVERSE`. At any single point in time, the sum of the following cannot be greater than `MPI_UNIVERSE_SIZE`:

- The number of starting processes
- The number of processes launched into the hosts in `MPI_UNIVERSE`
- The number of hosts launched onto specified hosts

Compiling and running OpenSHMEM applications

The following procedure explains how to compile and run OpenSHMEM programs in general terms.

Procedure

1. Use one of the OpenSHMEM wrapper compiler commands to run your program or call the compiler directly.

   To use the wrapper compiler, use one of the following commands:

   - `oshcc`
   - `oshCC`
   - `oshfort`

   To compile the OpenSHMEM program directly, use GNU compiler or Intel compiler commands.

   - To compile OpenSHMEM programs with a GNU compiler, choose one of the following commands:
     - `g++ compute.C -lsma -lmpi`
     - `gcc compute.c -lsma -lmpi`
   - To compile OpenSHMEM programs with an Intel compiler, choose one of the following commands:
     - `icc compute.C -lsma -lmpi`
     - `icc compute.c -lsma -lmpi`
     - `ifort compute.f -lsma -lmpi`

2. Use the `mpirun` command or the `mpiexec_mpt` command to launch the OpenSHMEM application.

   When you are not using a resource manager, such as PBS Professional or Slurm, set the `-np` option on the command to request the desired number of processes to launch. The `NPES` variable has no effect on OpenSHMEM programs.

   The OpenSHMEM programming model supports single-host OpenSHMEM applications.

   For more information, see the `intro_shmem`(3) manpage.

Building MPI Fortran modules

The `/opt/hpe/hpc/mpt/mpt-2.19/fortran_module_generator` directory contains source code you can use to build your own Fortran modules. This topic explains how to use environment variables and the `make` command to generate the modules.

Procedure

1. Set the environment variables you need for the compiler and (optionally) for any additional options you need.

The following table explains the environment variables.

Variable	Content
`FC`	This environment variable is required. Set this variable to the compiler command used to generate the modules.
`FCFLAGS`	This environment variable is optional. If you do not set this environment variable, Intel compiler support is assumed, and the system tags buffers with `!DEC$ ATTRIBUTES NO_ARG_CHECK`. If you set this environment variable, set it to one of the following values: • `USE_GCC_FORTRAN`, which tags buffers with the following: `!GCC$ NO_ARG_CHECK` • `USE_PGI_FORTRAN`, which tags buffers with the following: `!DIR$ IGNORE_TKR` • `USE_TS29113_FORTRAN`, which tags buffers with the following: `TYPE(*), DIMENSION(..)` If your compiler supports TS 29113, set `FCFLAGS` to this value.

Example 1. If you have an Intel compiler that supports TS 29113, specify the following environment variables:

```
FC=ifort FCFLAGS="-DUSE_TS29113_FORTRAN" make
```

Example 2. If you have a GNU Fortran compiler, specify the following environment variables:

```
FC=gfortran FCFLAGS="-DUSE_GCC_FORTRAN" make
```

Example 3. If you have a GNU Fortran compiler that supports TS 29113, specify the following environment variables:

```
FC=gfortran FCFLAGS="-DUSE_TS29113_FORTRAN" make
```

2. Choose an installation path, and run the following command:

   ```
   DESTDIR=install_path make install
   ```

   For *install_path*, specify the path to the generated `*.mod` files.

   For example:

   ```
   % make install DESTDIR=/opt/hpe/hpc/mpt/mpt-2.19/include/custom_dir
   ```

3. Set the `MPI_CUSTOM_FORTRAN_MODULES_PATH` environment variable to the directory path you set with the `make` command.
4. Set the `MPIF90_F90` environment variable and the `MPIF08_F08` environment variable to the compiler that you used to generate the Fortran modules.
5. Set the `LD_LIBRARY_PATH` environment variable to include `MPI_CUSTOM_FORTRAN_MODULES_PATH`.

   Make sure that `MPI_CUSTOM_FORTRAN_MODULES_PATH` appears before the MPT `LD_LIBRARY_PATH` environment variable. This ordering avoids loading the default `libmpi_f08`.

For example, the following is a common idiom:

```
LD_LIBRARY_PATH=$MPI_CUSTOM_FORTRAN_MODULES_PATH:$LD_LIBRARY_PATH
```

If you use the preceding idiom, make sure that the idiom is processed after the MPT module is loaded.

The MPI from HPE compiler helpers automatically detect the path to the MPI from HPE Fortran modules you built. You, however, must specify the compiler you are using. If you do not use `mpif90` or `mpif08`, use the following compiler options:

- For Fortran 2008, use the following:
    - `-I$MPI_CUSTOM_FORTRAN_MODULES_PATH`
    - `-LMPI_CUSTOM_FORTRAN_MODULES_PATH`
    - `-lmpi_f08`
- For Fortran 2003, use `-I$MPI_CUSTOM_FORTRAN_MODULES_PATH`

6. (Optional) Update the following environment variables in the module files for your compilers and for your users:

- `MPI_CUSTOM_FORTRAN_MODULES_PATH`
- `MPIF90_F90`
- `MPIF08_F08`
- `LD_LIBRARY_PATH`

Using huge pages

Huge pages optimize MPI application performance. The `MPI_HUGEPAGE_HEAP_SPACE` environment variable defines the minimum amount of heap space each MPI process can allocate using huge pages. If set to a positive number, `libmpi` verifies that enough `hugetlbfs` overcommit resources are available at program start-up to satisfy that amount on all MPI processes. The heap uses all available `hugetlbfs` space, even beyond the specified minimum amount. A value of 0 disables this check and disables the allocation of heap variables on huge pages. Values can be followed by K, M, G, or T to denote scaling by 1024, 1024^2, 1024^3, or 1024^4.

For more information about the `MPI_HUGEPAGE_HEAP_SPACE` environment variable, see the `mpi`(1) manpage.

For more information about how to configure huge pages for MPI applications, see the `mpt_hugepage_config`(1) manpage.

The following procedure explains how to configure system settings for huge pages.

Procedure

1. Enter the following command to make sure that the current MPT software release module is installed:

```
sys:~ # module load mpt
```

2. Log in as the administrator user, and enter the following command to configure the system settings for huge pages:

```
sys:~ # mpt_hugepage_config -u
Updating system configuration

System config file:         /proc/sys/vm/nr_overcommit_hugepages
```

```
Huge Pages Allowed:         28974 pages (56 GB)  90% of memory
Huge Page Size:             2048 KB
Huge TLB FS Directory:      /etc/mpt/hugepage_mpt
```

3. Enter the following command to retrieve the current system configuration:

   ```
   sys:~ #  mpt_hugepage_config -v
   Reading current system configuration

   System config file:         /proc/sys/vm/nr_overcommit_hugepages
   Huge Pages Allowed:         28974 pages (56 GB)  90% of memory
   Huge Page Size:             2048 KB
   Huge TLB FS Directory:      /etc/mpt/hugepage_mpt   (exists)
   ```

4. When running your MPT program, make sure the `MPI_HUGEPAGE_HEAP_SPACE` environment variable is set to `1`.

   This setting activates the new `libmpi` huge page heap. When you use the `malloc` function to allocate memory, the memory is allocated on huge pages.

5. Log in as the administrator user, and enter the following command to clear the system configuration settings:

   ```
   sys:~ #  mpt_hugepage_config -e
   Removing MPT huge page configuration
   ```

6. To verify that the MPT huge page configuration has been cleared, enter the following command to retrieve the system configuration again:

   ```
   uv44-sys:~ # mpt_hugepage_config -v
   Reading current system configuration

   System config file:         /proc/sys/vm/nr_overcommit_hugepages
   Huge Pages Allowed:         0 pages (0 KB)  0% of memory
   Huge Page Size:             2048 KB
   Huge TLB FS Directory:      /etc/mpt/hugepage_mpt   (does not exist)
   ```

Using MPI from HPE with NVIDIA GPUs

If your system includes NVIDIA GPUs, MPI from HPE supports the following features:

- The data buffers in graphics processing unit (GPU) memory can be the source or target of data movement by MPI or OpenSHMEM functions. To enable this feature, set `MPI_USE_CUDA=true`.
- If your program sends GPU data to other hosts over Mellanox InfiniBand connections, you can also use the GPUDirect RDMA feature.

For more information, see the documentation from NVIDIA and Mellanox about GPUDirect RDMA.

Programming with MPI from HPE

Portability is one of the main advantages MPI has over vendor-specific message passing software. Nonetheless, the MPI Standard offers sufficient flexibility for general variations in vendor implementations. In addition, there are often vendor-specific programming recommendations for optimal use of the MPI library. The topics in this chapter explain how to develop or port MPI applications to HPE systems.

Job termination and error handling

The following topics describe MPI from HPE behavior upon typical job termination, error handling, and the characteristics of atypical job termination:

- **MPI_Abort** on page 28
- **Error handling** on page 28
- **MPI_Finalize and connect processes** on page 29

`MPI_Abort`

In the MPI from HPE implementation, a call to `MPI_Abort` has the following effect:

- The MPI job terminates, regardless of the communicator argument used.
- The error code value is returned as the exit status of the `mpirun` command.
- A stack traceback is displayed that shows where the program called `MPI_Abort`.

Error handling

The MPI Standard describes MPI error handling. Although almost all MPI functions return an error status, an error handler is invoked before returning from the function. If the function has an associated communicator, the error handler associated with that communicator is invoked. Otherwise, the error handler associated with `MPI_COMM_WORLD` is invoked.

The MPI from HPE implementation provides the following predefined error handlers:

- `MPI_ERRORS_ARE_FATAL`. When called, causes the program to abort on all executing processes. The effect is the same as if `MPI_Abort` were called by the process that invoked the handler.
- `MPI_ERRORS_RETURN`. This handler has no effect.

By default, the `MPI_ERRORS_ARE_FATAL` error handler is associated with `MPI_COMM_WORLD` and any communicators derived from it. To handle error statuses returned from MPI calls, make one of the following associations:

- Associate the `MPI_ERRORS_RETURN` handler with `MPI_COMM_WORLD` near the beginning of the application.

  OR
- Associate another user-defined handler with `MPI_COMM_WORLD` near the beginning of the application.

`MPI_Finalize` and connect processes

When using MPI from HPE, all pending communications involving an MPI process must be complete before the process calls `MPI_Finalize`. If there are any pending `send` or `recv` requests that are unmatched or not completed, the application hangs in `MPI_Finalize`. For more information, see the MPI Standard.

If the application uses the MPI remote memory access (RMA) spawn functionality described in the MPI RMA standard, there are additional considerations. In the HPE implementation, all MPI processes are connected. The MPI RMA standard defines what is meant by connected processes. When the MPI RMA spawn functionality is used, `MPI_Finalize` is collective over all connected processes. Thus all MPI processes, both launched on the command line, or subsequently spawned, synchronize in `MPI_Finalize`.

Signals

In the MPI implementation, MPI processes are Linux processes. As such, the general rule regarding signal handling applies as it would to ordinary Linux processes.

Certain signals can be propagated from the `mpirun` process to other processes in the MPI job. This propagation can occur as follows:

- When the signal belongs to the same process group on a single host.
- When the signal belongs to processes that run across multiple hosts in a cluster.

These signals are as follows:

- `SIGURG`
- `SIGUSR1`
- `SIGINT`
- `SIGTERM`

To use this feature, the MPI program must have a signal handler that catches the signal. When the signal is sent to the `mpirun` process ID, the `mpirun` process catches the signal and propagates it to all MPI processes.

Buffering

Most MPI implementations use buffering for overall performance reasons, and some programs depend on it. However, do not assume that there is any message buffering between processes. The MPI Standard does not mandate a buffering strategy.

If sent messages are not buffered, each process hangs in the initial call, waiting for an `MPI_Recv` call to take the message. For example, the following table illustrates a simple sequence of MPI operations that cannot work unless messages are buffered:

Table 1: Outline of improper dependence on buffering

Process 1	Process 2
`MPI_Send(2,....)`	`MPI_Send(1,....)`
`MPI_Recv(2,....)`	`MPI_Recv(1,....)`

Because most MPI implementations buffer messages to some degree, a program like the one in the preceding table does not usually hang. The `MPI_Send` calls return after putting the messages into buffer space, and the `MPI_Recv` calls get the messages. Nevertheless, programs that include the logic shown in the preceding table are invalid according to the MPI Standard. If your program requires this sequence of MPI calls, employ one of the following buffer MPI send calls:

- `MPI_Bsend`
- `MPI_Ibsend`

By default, the MPI from HPE uses buffering under most circumstances. Short messages (64 or fewer bytes) are always buffered. Longer messages are also buffered, although under certain circumstances, buffering can be avoided. For performance reasons, it is sometimes desirable to avoid buffering. For further information on unbuffered message delivery, see the following:

Programming optimizations on page 31

Multithreaded programming

MPI from HPE supports a hybrid programming model. In the programming model MPI handles one level of parallelism in an application and POSIX threads or OpenMP processes are used to handle another level. When mixing OpenMP with MPI, for performance reasons, it is better to consider invoking MPI functions only outside parallel regions or only from within master regions. When used in this manner, it is not necessary to initialize MPI for thread safety. You can use `MPI_Init` to initialize MPI. However, to safely invoke MPI functions from any OpenMP process or when using POSIX threads, use `MPI_Init_thread` to initialize MPI.

When using `MPI_Thread_init()` with the threading level `MPI_THREAD_MULTIPLE`, link your program as follows:

- If you use the compiler wrappers for MPI or SHMEM, use the `-mt` option on the command line.
- If you want to call the compilers directly, use the `-lmpi_mt` parameter instead of the `-lmpi` parameter on the compiler command line.

For more information about compiling and linking MPI programs, see the `mpi`(1) manpage.

Interoperability with the OpenSHMEM programming model

The following list shows actions to take when you mix OpenSHMEM and MPI message passing in the same program:

- Start with an MPI program that calls `MPI_Init` (or `MPI_Init_thread()`) and `MPI_Finalize`. Next, add OpenSHMEM calls. Remember that the processing environment (PE) numbers are equal to the MPI rank numbers in `MPI_COMM_WORLD`.
- Make sure that the program calls the `shmem_init()` and `shmem_finalize()` library routines. This practice is similar to how you include calls to `MPI_Init()` (or `MPI_Init_thread()`) and `MPI_Finalize`.
- Link the application with both the OpenSHMEM and MPI libraries.

When running the application across a cluster using OpenSHMEM and OpenSHMEM functions, some processes might not be able to communicate with other processes. Use the `shmem_pe_accessible` and `shmem_addr_accessible` functions to determine whether an OpenSHMEM call can access data residing in another process.

The OpenSHMEM model functions only with respect to `MPI_COMM_WORLD`. You cannot use `shmem_pe_accessible` and `shmem_addr_accessible` to exchange data between MPI processes that are connected through MPI intercommunicators returned from MPI spawn-related functions.

For more information about the OpenSHMEM programming model, see the `intro_shmem`(3) manpage.

Miscellaneous MPI from HPE features

The following other characteristics of the MPI from HPE implementation might interest you:

- `stdin`/`stdout`/`stderr`.

  In this implementation, `stdin` is enabled for only the process that is rank 0 in the first `MPI_COMM_WORLD`. You can locate such processes on the same host as `mpirun` or on a host that is different from `mpirun`.

  The `stdout` and `stderr` results are enabled for all MPI processes in the job, whether started by `mpirun` or started by one of the MPI spawn functions.

- `MPI_Get_processor_name`

  The `MPI_Get_processor_name` function returns the Internet host name of the computer upon which the MPI process that started this subroutine is running.

Programming optimizations

You might need to modify your MPI application to use the MPI from HPE optimization features.

The following topics describe how to use the optimized features of MPI from HPE:

- **Using MPI point-to-point communication routines** on page 31
- **Using MPI collective communication routines** on page 32
- **Using MPI_Pack and MPI_Unpack** on page 32
- **Avoiding derived data types** on page 32
- **About wildcards** on page 32
- **Avoiding message buffering - single copy methods** on page 33
- **Managing memory placement** on page 33

Using MPI point-to-point communication routines

MPI provides a number of different routines for point-to-point communication. The most efficient ones in terms of latency and bandwidth are the blocking and nonblocking `send`/`receive` functions, which are as follows:

- `MPI_Send`
- `MPI_Isend`
- `MPI_Recv`
- `MPI_Irecv`

Unless required for application semantics, avoid the synchronous send calls, which are as follows:

- `MPI_Ssend`
- `MPI_Issend`

Also avoid the buffered send calls, which double the amount of memory copying on the sender side. These calls are as follows:

- `MPI_Bsend`
- `MPI_Ibsend`

This implementation treats the ready-send routines, `MPI_Rsend` and `MPI_Irsend`, as standard `MPI_Send` and `MPI_Isend` routines. Persistent requests do not offer any performance advantage over standard requests in this implementation.

Using MPI collective communication routines

Frequently, MPI collective calls are layered on top of point-to-point primitive calls. For small process counts, this practice can be reasonably effective. However, for higher process counts of 32 processes or more, or for clusters, this approach can be less efficient. For this reason, a number of the MPI library collective operations have been optimized to use more complex algorithms.

The MPI from HPE collectives have been optimized for use with clusters. In these cases, steps are taken to reduce the number of messages using the relatively slower interconnect between hosts.

Some of the collective operations have been optimized for use with shared memory. The `MPI_Alltoall` routines also use special techniques to avoid message buffering when using shared memory.

For more information, see the following:

Avoiding message buffering - single copy methods on page 33

Using `MPI_Pack` and `MPI_Unpack`

`MPI_Pack` and `MPI_Unpack` are useful for porting parallel virtual machine (PVM) codes to MPI. However, `MPI_Pack` and `MPI_Unpack` double the amount of data to be copied by both the sender and receiver. Generally, either restructure your data or use derived data types to avoid using these functions. Note, however, that use of derived data types can lead to decreased performance in certain cases.

Avoiding derived data types

Avoid derived data types when possible. Generally, there is no performance gain when you use derived data types with MPI from HPE. For example, use of derived data types might disable certain types of optimizations such as unbuffered or single copy data transfer.

About wildcards

The use of wildcards (`MPI_ANY_SOURCE`, `MPI_ANY_TAG`) involves searching multiple queues for messages. This search is not significant for small process counts, but for large process counts, the cost increases quickly.

MPT can make certain optimizations if the application does not make calls to variations of `MPI_Recv()` with `MPI_ANY_SOURCE`. When `MPI_WILDCARDS=false` is in effect, MPT assumes that the application does not contain receive calls with rank wildcards. This assumption enables MPT to make some bandwidth optimizations in its Intel Omni-Path Architecture code. MPT supports the `MPI_WILDCARDS` environment variable only on systems that include the Intel Omni-Path Architecture. For information about more environment variables that MPT supports on the Intel Omni-Path Architecture, see the following:

Tuning for running applications over the Intel Omni-Path interconnect on page 48

Avoiding message buffering - single copy methods

One of the most significant optimizations for bandwidth-sensitive applications in the MPI library is single-copy optimization. Single-copy optimization avoids using shared memory buffers. However, as discussed in **Buffering** on page 29, some incorrectly coded applications might hang because of buffering assumptions. For this reason, single-copy optimization is not enabled by default for `MPI_Send`, but you can use the `MPI_BUFFER_MAX` environment variable to enable this optimization at run time. The following guidelines show how to increase the opportunity for use of the unbuffered pathway:

- The MPI data type on the send side must be a contiguous type.
- The sender and receiver MPI processes must reside on the same host.
- The sender data must be globally accessible by the receiver. MPI from HPE allows data allocated from the static region (common blocks), the private heap, and the stack region to be globally accessible. In addition, the following memory is globally accessible:
    - Memory allocated through the `MPI_Alloc_mem` function
    - Memory allocated through the SHMEM symmetric heap accessed through the `shpalloc` or `shmalloc` functions.

To enable the unbuffered, single-copy method, set the run-time environment variables described in the following:

Avoiding message buffering - enabling single copy on page 40

Managing memory placement

For single-process and small multiprocess applications, the HPE Superdome Flex Grid architecture behaves similarly to flat memory architectures. For more highly parallel applications, memory placement becomes important. MPI takes placement into consideration when it lays out shared memory data structures and the address spaces of the individual MPI processes. In general, do not try to manage memory placement explicitly.

To control the placement of the application at run time, however, see the following:

Run-time tuning on page 39

More programming model considerations

A number of additional programming options might be worth consideration when developing MPI applications. For example, using the SHMEM programming model can improve performance in the latency-sensitive sections of an application. Usually, this enhancement requires replacing MPI `send`/`recv` calls with `shmem_put`/`shmem_get` and `shmem_barrier` calls. The SHMEM programming model can deliver lower latencies for short messages than traditional MPI calls. As an alternative to `shmem_get`/`shmem_put` calls, you might consider the MPI remote memory access (RMA) `MPI_Put`/ `MPI_Get` functions. These functions provide almost the same performance as the SHMEM calls, while providing a greater degree of portability.

Alternately, consider exploiting the shared memory architecture by implementing the following:

- Handling one or more levels of parallelism with OpenMP.
- Using MPI to handle with the coarser-grained levels of parallelism.

There are special placement considerations to be aware of when running hybrid MPI/OpenMP applications. For more information, see the following:

Run-time tuning on page 39

Debugging MPI applications

MPI routine argument checking

Debugging MPI applications can be more challenging than debugging sequential applications. By default, the MPI from HPE does not check the arguments to some performance-critical MPI routines, such as most of the point-to-point and collective communication routines. You can force MPI from HPE to always check the input arguments to MPI functions by setting the `MPI_CHECK_ARGS` environment variable. However, setting this variable might result in some degradation in application performance. HPE recommends that you do not set `MPI_CHECK_ARGS` unless you are debugging.

Using the TotalView debugger with MPI programs

The Message Passing Toolkit (MPT) `mpiexec_mpt` command supports the debugging of MPI applications with the TotalView Debugger.

Example 1. To run an MPT MPI job with the TotalView Debugger with a workload manager, such as PBS Professional or Torque, enter the following and select a parallel environment of `MPT` in TotalView:

```
% totalview a.out
```

Example 2. To run an MPT MPI job with TotalView without a workload manager, enter the following and select a parallel environment of `none` in TotalView:

```
% totalview mpirun -a hostA, hostB -np 4 a.out
```

NOTE: TotalView does not operate with MPI processes started by the `MPI_Comm_spawn` or `MPI_Comm_spawn_multiple` functions.

Using `idb` and `gdb` with MPI programs

The `idb` and `gdb` debuggers are designed for sequential, nonparallel applications. To make them usable for parallel applications, use the `MPI_SLAVE_DEBUG_ATTACH` environment variable.

If you set the `MPI_SLAVE_DEBUG_ATTACH` environment variable to a global rank number, the MPI process sleeps briefly in startup while you use `idb` or `gdb` to attach to the process. A message is printed to the screen, telling you how to use `idb` or `gdb` to attach to the process.

Similarly, to debug the MPI daemon, set `MPI_DAEMON_DEBUG_ATTACH`. This variable puts the daemon to sleep briefly while you attach to it.

Using the DDT debugger with MPI programs

The DDT debugger from Allinea Software is a parallel debugger that supports MPT. You can run DDT in either interactive (online) or batch (offline) mode. In batch mode, DDT can create a text or HTML report that tracks variable values and shows the location of any errors. DDT records the data for program variables across all processes, and DDT logs values in the HTML output files as sparkline charts.

To configure DDT for use with MPI on HPE systems, use the instructions that are posted to the following website:

http://content.allinea.com/downloads/userguide.pdf

Example 1. The following command starts DDT in interactive (online) mode:

```
# ddt -np 4 a.out
```

Example 2. The following command generates a debugging report in HTML format:

```
# ddt -offline my-log.html -np 4 a.out
```

Example 3. Assume that you want to trace variables `x`, `y`, and `my_arr(x,y)` in parallel across all processes. The following command directs DDT to record the values of `x`, `y`, and `my_arr(x,y)` each time it encounters line 147:

```
# ddt -offline my-log.html -trace-at "my-file.f:147,x,y,my_arr(x,y)" -np 4 a.out
```

Example 4. You can specify batch (offline) DDT commands from within a queue submission script. Instead of specifying `mpiexec_mpt -np 4 a.out`, specify the following:

```
# ddt -noqueue -offline my-log.html -trace-at "my-file.f:147,x,y,my_arr(x,y)" \
-np 4 a.out
```

Using Valgrind with MPI programs

Valgrind is a tool that can profile your program and can automatically detect memory management and threading bugs.

Valgrind is not compatible with the memory mapping functionality in MPT. When MPT detects that Valgrind is in use, MPT automatically enables the `MPI_MEMMAP_OFF` environment variable, which disables MPT memory mapping.

Working with other Message Passing Interface (MPI) implementations

The performance boosting tool, `perfboost`, uses a wrapper library to run applications compiled against other MPI implementations under the Message Passing Toolkit (MPT) product on HPE hardware platforms.

NOTE: `perfboost` does not support the MPI C++ API.

Using `perfboost`

The following procedure explains how to use `perfboost` with an MPI from HPE program.

NOTE: On Apollo 70 platforms, MPI from HPE supports `perfboost` for only the Open MPI implementation.

Procedure

1. Load the `perfboost` environment module.

   The module includes the `PERFBOOST_VERBOSE` environment variable.

   If you set the `PERFBOOST_VERBOSE` environment variable, it enables a message when `PerfBoost` activates and also when the MPI application is completed through the `MPI_Finalize()` function. This message indicates that the `perfboost` library is active and also indicates when the MPI application completes through the `libperfboost` wrapper library.

   The MPI environment variables that are documented in the `MPI`(1) manpage are available to `perfboost`. MPI environment variables that are not used by MPT are currently not supported.

   NOTE: Some applications redirect `stderr`. In this case, the verbose messages might not appear in the application output.

2. Enter a command that inserts the `perfboost` command in front of the executable name along with the choice of MPI implementation to emulate.

   In other words, run the executable file with the MPT `mpiexec_mpt` command or the `mpirun` command.

   The following table shows the MPI implementations and corresponding command-line options.

Implementation	Command-line option
Platform MPI 7.1+	`-pmpi`
Intel MPI	`-impi`
Open MPI	`-ompi`
MPICH1	`-mpich`

Table Continued

Implementation	Command-line option
MPICH2, version 2 and later	`-impi`
MVAPICH2, version 2 and later	`-impi`

The following are some examples that use `perfboost`:

```
% module load mpt
% module load perfboost

% mpirun -np 32 perfboost -impi a.out arg1
% mpiexec_mpt perfboost -pmpi b.out arg1
% mpirun host1 32, host2 64 perfboost -impi c.out arg1 arg2
```

MPI supported functions

`perfboost` supports the commonly used elements of the C and Fortran MPI APIs. If a function is not supported, the job aborts and issues an error message. The message shows the name of the missing function. You can contact HPE customer support to schedule a missing function to be added to `perfboost`.

Using Berkeley Lab Checkpoint/Restart (BLCR)

The Message Passing Toolkit (MPT) supports BLCR checkpoint/restart. This checkpoint/restart implementation allows applications to periodically save a copy of their state. If the application crashes or if the job is aborted to free resources for higher-priority jobs, the application can resume from that point.

The following are the implementation limitations:

- BLCR does not set a checkpoint for the state of any data files that the application might be using.
- When using checkpoint/restart, the Message Passing Interface (MPI) does not support certain features, including spawning and one-sided MPI.
- InfiniBand XRC queue pairs are not supported.
- Checkpoint files are often large and require significant disk bandwidth to create in a timely manner.

For more information on BLCR, see the following:

http://crd.lbl.gov/departments/computer-science/CLaSS/research/BLCR/

Installing BLCR

To use checkpoint/restart with MPT, first install BLCR.

Procedure

1. Log in as the administrator user.
2. Install the `blcr-`, `blcr-libs-`, and `blcr-kmp-` RPMs.

   BLCR uses a kernel module that must be built against the specific kernel that the operating system is running. If the kernel module fails to load, rebuild and reinstall. Install the `blcr-` source RPM. In the `blcr.spec` file, set the kernel variable to the name of the current kernel, then rebuild and install the new set of RPMs.
3. Enter the following command to enable BLCR:

   ```
   # chkconfig blcr on
   ```

Using BLCR with MPT

To enable checkpoint/restart within MPT, pass the `-cpr` option to `mpirun` or `mpiexec_mpt`. For example:

```
% mpirun -cpr hostA, hostB -np 8 ./a.out
```

To checkpoint a job, run the `mpt_checkpoint` command on the same host upon which `mpirun` is running. Make sure to pass the `mpt_checkpoint` command the PID of `mpirun` and the name with which you want to prefix all the checkpoint files. For example:

```
% mpt_checkpoint -p 12345 -f my_checkpoint
```

The preceding example command creates a `my_checkpoint.cps` metadata file and a number of `my_checkpoint.*.cpd` files.

To restart the job, pass the name of the `.cps` file to `mpirun`. For example:

```
% mpirun -cpr hostC, hostD -np 8 mpt_restart my_checkpoint.cps
```

You can restart the job on a different set of hosts, but the number of hosts must be the same. In addition, each host must have the same number of ranks as the corresponding host in the original run of the job.

Run-time tuning

Run-time tuning is the process by which you tune the run-time environment to improve the performance of a Message Passing Interface (MPI) message passing application. The methods do not involve application code changes.

The run-time tuning topics are as follows:

Reducing run-time variability

It can be difficult to achieve reproducible timings from run to run when you optimize message passing codes on a large, shared-memory systems. To reduce run-time variability, you can take the following precautions:

- Do not oversubscribe the system. In other words, do not request more CPUs than are available, and do not request more memory than is available. Oversubscribing causes the system to wait unnecessarily for resources to become available, leads to variations in the results, and leads to less than optimal performance.
- Avoid interference from other system activity. The Linux kernel uses more memory on node 0 than on other nodes. Node 0 is also known as **the kernel node**. If your application uses almost all the available memory per processor, the memory for processes assigned to the kernel node can unintentionally spill over to nonlocal memory. By keeping user applications off the kernel node, you can avoid this effect.

  By restricting system daemons to run on the kernel node, you can also deliver an additional percentage of each application CPU to the user program.
- Avoid interference with other applications. Use Linux cgroups or the cpuset software to address this problem. Linux cgroups and the cpuset software enable you to partition a large, distributed memory host in a fashion that minimizes interactions between jobs running concurrently on the system. For more information about cpusets, see the following:

  HPE Performance Software - Message Passing Interface Cpuset Software Guide
- On a quiet, dedicated system, you can use the `dplace` command or the `MPI_DSM_CPULIST` environment variable to improve run-time performance repeatability. These approaches are not suited to shared, nondedicated systems.
- Use a workload manager such as Platform LSF from IBM or PBS Professional from Altair Engineering, Inc. These workload managers use cpusets to avoid oversubscribing the system and to avoid possible interference between applications.

Tuning MPI buffer resources

By default, MPI from HPE buffers messages that are longer than 64 bytes. The system buffers these longer messages in a series of 16 KB buffers. Messages that exceed 64 bytes are handled as follows:

- If the message is 128 K in length or shorter, the sender MPI process buffers the entire message.

  In this case, the sender MPI process delivers a message header, also called a **control message**, to a mailbox. When an MPI call is made, the MPI receiver polls the mail box. If the receiver finds a matching receive request for the sender control message, the following events occur:

  - The receiver copies the data out of the buffers into the application buffer indicated in the receive request.
  - The receiver sends a message header back to the sender process. This action indicates that the buffers are available for reuse.

- If the message is longer than 128 K, the software breaks the message into chunks that are 128 K in length.

  The smaller chunks allow the sender and receiver to overlap the copying of data in a pipelined fashion. Because there are a finite number of buffers, this activity can constrain overall application performance for certain communication patterns. You can use the `MPI_BUFS_PER_PROC` shell variable to adjust the number of buffers available for each process. To determine if the demand for buffering is high, use the MPI statistics counters.

  If you increase the number of buffers, you can often avoid excessive numbers of retries for buffers. However, when you increase the number of buffers, you consume more memory, and you might increase the probability for cache pollution. **Cache pollution** is the excessive filling of the cache with message buffers. Cache pollution can degrade performance during the compute phase of a message passing application.

For information about statistics counters, see the following:

MPI internal statistics on page 58

For information about buffering considerations when you run an MPI job over multiple hosts, see the following:

Tuning running applications that are spread across multiple hosts on page 44

For information about the programming implications of message buffering, see the following:

Buffering on page 29

Avoiding message buffering - enabling single copy

It is possible to avoid the need to buffer messages for the following:

- Message transfers between MPI processes within the same host
- Message transfers that use devices that support remote direct memory access (RDMA), such as InfiniBand.

The following topics provide more information about buffering:

- **Buffering and MPI_Send** on page 41
- **Using the XPMEM driver for single-copy optimization** on page 41

Buffering and `MPI_Send`

Many MPI applications are written to assume infinite buffering, so message buffering is not enabled by default for `MPI_Send`. For `MPI_Isend`, `MPI_Sendrecv`, and most collectives, this optimization is enabled by default for large message sizes. To disable this default, single-copy feature used for the collectives, use the `MPI_DEFAULT_SINGLE_COPY_OFF` environment variable.

Using the XPMEM driver for single-copy optimization

MPI uses the XPMEM driver to support single-copy message transfers between two processes within the same host.

Enabling single-copy transfers can increase performance because this technique improves MPI bandwidth. However, single-copy transfers can introduce additional synchronization points, which can reduce application performance.

The `MPI_BUFFER_MAX` environment variable specifies the threshold for message lengths. Set this environment variable to the message length, in bytes, beyond which you want MPI to use the single-copy method. In general, a value of 2000 or higher is beneficial for many applications.

During job startup, MPI uses the XPMEM driver, through the `xpmem` kernel module, to map memory from one MPI process to another. The mapped areas include the static (BSS) region, the private heap, the stack region, and (optionally) the symmetric heap region of each process.

Memory mapping allows each process to directly access memory from the address space of another process. This technique allows MPI to support single-copy transfers for contiguous data types from any of these mapped regions. For these transfers between processes residing on the same host, MPI uses the `bcopy` process to copy the data. The `bcopy` process also transfers data between two different executable files on the same host. For data residing outside of a mapped region (a `/dev/zero` region, for example), MPI uses a buffering technique to transfer the data.

Memory mapping is enabled by default. To disable it, set the `MPI_MEMMAP_OFF` environment variable. Memory mapping must be enabled to allow the following:

- Single-copy transfers
- MPI remote memory access (RMA) one-sided communication
- Support for the SHMEM model
- Certain collective optimizations

Memory placement and policies

The MPI library takes advantage of the placement functions that are available. Usually, the default placement is adequate. However, you can set one or more environment variables to modify the default behavior.

For a list of the environment variables that control memory placement, see the `MPI`(1) manpage.

The following topics contain information on environment variables and tools that enable you to tune memory placement:

- **MPI_DSM_CPULIST** on page 42
- **MPI_DSM_DISTRIBUTE** on page 42
- **MPI_DSM_VERBOSE** on page 43
- **Using dplace** on page 43

`MPI_DSM_CPULIST`

The `MPI_DSM_CPULIST` environment variable allows you to select the processors to use for an MPI application. At times, specifying a list of processors on which to run a job can be the best means to insure highly reproducible timings, particularly when running on a dedicated system.

The setting is an ordered list that uses commas (`,`) and hyphens (`-`) to specify a mapping of MPI processes to CPUs. If running across multiple hosts, separate the per-host components of the CPU list with a colon (`:`). When you use a hyphen-delineated list, you can specify CPU striding by specifying `/`*`stride_distance`* after the list.

The following table shows example settings.

Value	CPU assignment
`8,16,32`	Place three MPI processes on CPUs 8, 16, and 32.
`32,16,8`	Place the MPI process rank 0 on CPU 32, one on 16, and two on CPU 8.
`8-15/2`	Place the MPI processes 0 through 3 strided on CPUs 8, 10, 12, and 14.
`8-15,32-39`	Place the MPI processes 0 through 7 on CPUs 8 to 15. Place the MPI processes 8 through 15 on CPUs 32 to 39.
`39-32,8-15`	Place the MPI processes 0 through 7 on CPUs 39 to 32. Place the MPI processes 8 through 15 on CPUs 8 to 15.
`8-15:16-23`	Place the MPI processes 0 through 7 on the first host on CPUs 8 through 15. Place MPI processes 8 through 15 on CPUs 16 to 23 on the second host.

The process rank is the `MPI_COMM_WORLD` rank. The interpretation of the CPU values specified in the `MPI_DSM_CPULIST` depends on whether the MPI job is being run within a cpuset, as follows:

- If the job is run outside of a cpuset, the CPUs specify *cpunum* values. The *cpunum* values begin with 0 and range up to the number of CPUs in the system, minus one.
- If the job is run within a cpuset, the default behavior is to interpret the CPU values as relative processor numbers within the cpuset.

Make sure that the number of processors you specify equals the number of MPI processes used to run the application. The number of colon-delineated parts of the list must equal the number of hosts used for the MPI job. If an error occurs in processing the CPU list, the default placement policy is used.

`MPI_DSM_DISTRIBUTE`

The `MPI_DSM_DISTRIBUTE` environment variable ensures the following:

- Each MPI process gets a physical CPU on the node to which it was assigned
- Each MPI process gets memory on the node to which it was assigned.

`MPI_DSM_DISTRIBUTE` assigns MPI ranks, as follows:

- On systems that do not span hosts, `MPI_DSM_DISTRIBUTE` assigns MPI ranks starting at logical CPU 0 and incrementing until all ranks have been placed.
- Some systems include InfiniBand interconnect or Intel Omni-Path Architecture (OPA). On these systems, if the job spans hosts, `MPI_DSM_DISTRIBUTE` assigns MPI ranks starting with the CPU that is closest to the first InfiniBand host channel adapter (HCA) or OPA.

If you set both `MPI_DSM_DISTRIBUTE` and `MPI_DSM_CPULIST`, `MPI_DSM_CPULIST` overrides `MPI_DSM_DISTRIBUTE`.

`MPI_DSM_VERBOSE`

Setting the `MPI_DSM_VERBOSE` environment variable directs MPI to display a synopsis of the host placement options used at run time.

Using `dplace`

The `dplace` tool offers another way to specify the placement of MPI processes within a distributed memory host. Used together, the `dplace` tool and MPI improve the placement of certain shared memory data structures.

For information about `dplace` with MPI, see the following:

- **dplace command** on page 75
- The `dplace`(1) manpage
- The **Linux Application Tuning Guide**

Tuning MPI/OpenMP hybrid codes

A hybrid MPI/OpenMP application is one in which each MPI process itself is a parallel threaded program. These programs often exploit the OpenMP parallelism at the loop level while also implementing a higher-level parallel algorithm that uses MPI.

Process pinning can increase performance. Pin both the processes and the threads within them, for the duration of their execution, to specific processors. The effect is as follows:

- On HPE Superdome Flex Grid systems, pinning ensures that all local, nonshared memory is allocated on the same memory node as the processor referencing the memory.
- On all systems, pinning can ensure that some or all OpenMP threads stay on processors that share a bus or perhaps a processor cache. In this case, pinning can speed up thread synchronization.

The Message Passing Toolkit (MPT) provides the `omplace` command to help with the placement of OpenMP threads within an MPI program. The `omplace` command causes the threads in a hybrid MPI/OpenMP job to be placed on unique CPUs within the containing cpuset. For example, the threads in a two-process MPI program with two threads per process are placed as follows:

- Rank 0, thread 0 on CPU 0
- Rank 0, thread 1 on CPU 1
- Rank 1, thread 0 on CPU 2
- Rank 1, thread 1 on CPU 3

The CPU placement is performed by dynamically generating a `dplace` placement file and invoking `dplace`.

For more information, see the following:

- **Data placement tools** on page 75
- The `omplace`(1) manpage
- The `dplace`(1) manpage
- **Linux Application Tuning Guide for SGI X86-64 Based Systems**
- **HPE Performance Software - Message Passing Interface Cpuset Software Guide**

Example: Running a hybrid MPI/OpenMP application

The following command line runs a hybrid MPI/OpenMP application with eight MPI processes that are two-way threaded on two hosts:

```
mpirun host1,host2 -np 4 omplace -nt 2 ./a.out
```

- When using the PBS workload manager to schedule the hybrid MPI/OpenMP job, use the following resource allocation specification:

  ```
  #PBS -l select=8:ncpus=2
  ```

- In addition, use the following `mpiexec_mpt` command:

  ```
  mpiexec_mpt -n 8 omplace -nt 2 ./a.out
  ```

For more information about running MPT programs with PBS, see the following:

Using a workload manager to launch an MPI application on page 19

Tuning running applications that are spread across multiple hosts

When you run an MPI application across a cluster of hosts, you can use the environment variables in this topic to improve application performance across these hosts.

The following table shows the interconnect types and the run-time environment settings and configurations that you can use to improve performance.

Interconnect type	Default order of selection	Environment variable required
XPMEM	1	`MPI_USE_XPMEM`
Intel Omni-Path Architecture	2	`MPI_USE_OPA`
InfiniBand	3	`MPI_USE_IB`
InfiniBand Unreliable Datagram	4	`MPI_USE_UD`
TCP/IP	5	`MPI_USE_TCP`

The preceding table shows the different types of interconnects that systems can employ as the multihost interconnect. When launched as a distributed application, MPI probes for these interconnects at job startup.

When MPI detects a high-performance interconnect, MPI attempts to use this interconnect on every host the MPI job uses. Sometimes the first choice interconnect is not available. If the interconnect is not available for use on every host, the library attempts to use the next slower interconnect until this connectivity requirement is met. The second column in the preceding table specifies the order in which MPI probes for available interconnects.

The third column in the preceding table shows the environment variable that you can set to specify a nondefault interconnect. To insure the best application performance, allow MPI to pick the fastest available interconnect.

When using the TCP/IP interconnect, unless specified otherwise, MPI uses the default IP adapter for each host. To use a nondefault adapter, enter the adapter-specific host name on the `mpirun` command line.

The following table shows the environment variables that you can use to tune your application for multiple hosts.

Environment variable	Effect
MPI_IB_RAILS	When set to 1 and the MPI library uses the InfiniBand driver as the interhost interconnect, MPT sends InfiniBand traffic over the first fabric that it detects. Default on all HPE Superdome Flex Grid systems. When set to 1+, MPT sends all traffic across the first fabric, but if it encounters communication problems, it starts to use both fabrics. When set to 2, the library tries to use multiple, available, separate, InfiniBand fabrics and splits the traffic across them.
MPI_IB_SINGLE_COPY_BUFFER_MAX	MPI can send data in one of the following ways: • Directly between the buffers of a process • Through intermediate buffers inside the MPI library. If the following are both true, MPI sends the data directly between the process buffers: • MPI transfers data over InfiniBand AND • The size of the cumulative data is greater than the value of this environment variable. The default is 32767.
MPI_USE_IB	When set, the MPI library uses the InfiniBand driver as the interconnect when running across multiple hosts or running with multiple binaries. MPT requires the OFED software stack when the InfiniBand interconnect is used. If InfiniBand is used, the MPI_COREDUMP environment variable is forced to INHIBIT. This action complies with the InfiniBand driver restriction that no fork() actions occur after InfiniBand resources have been allocated. The default is false.

For more information about environment variables, see the ENVIRONMENT VARIABLES section of the mpi(1) manpage.

For information about how to launch a distributed application, see the following:

Using the mpirun Command to Launch an MPI Application on page 21

Tuning for running applications over the InfiniBand interconnect

You can set environment variables to improve application performance when you run an MPI application across a cluster of hosts using the InfiniBand interconnect. The following table shows these environment variables.

Environment variable	Effect
MPI_COLL_HCOLL	Enables or disables the Mellanox fabric collectives accelerator (FCA) offload. If FCA offload is configured on your cluster, set `MPI_COLL_HCOLL=true`. Make sure that the Mellanox HCOLL libraries are specified in your library path. Specify these libraries in one of the following ways: • Load the `hpcx` software module, if available. OR • Making sure that the location of `libhcoll.so` is in your `LD_LIBRARY_PATH` environment variable. The default is `MPI_COLL_HCOLL=false`.
MPI_CONNECTIONS_THRESHOLD	There is a time and resource cost to the act of creating a connection between pairs of ranks when a job starts. The larger the job, the larger the cost. When the number of ranks is set to at least the value of this environment variable, the MPI library creates InfiniBand connections on a demand basis. The default is `1025` ranks.
MPI_IB_FAILOVER	If a connection to another rank fails, MPT tries to restart the connection to the other rank a certain number of times. This restart occurs when the following conditions exist: • When an InfiniBand transmission error occurred AND • When the MPI library uses InfiniBand fabric AND • When the `MPI_IB_FAILOVER` variable is set The `MPI_IB_FAILOVER` variable specifies the number of times MPT tries to restart the connection. This variable defines the maximum number of reconnection attempts. The default is `32` times.
MPI_IB_PAYLOAD	When the MPI library uses InfiniBand fabric, it allocates memory for each message header that it uses for transfer. The memory allocation scheme used depends on the size of the data to be sent, as follows: • If the size is less than or equal to the value of `MPI_IB_PAYLOAD`, minus 64 bytes for the actual header, the data is inlined with the header. • If the size is greater than the value of `MPI_IB_PAYLOAD`, then the message is sent through remote direct memory access (RDMA) operations. The default is `16512` bytes.

Table Continued

Environment variable	Effect
`MPI_IB_RNR_TIMER`	A receiving host channel adapter (HCA) sends a negative acknowledgement (NAK) to a requestor when the following events occur: • A packet arrives at an InfiniBand host channel adapter (HCA) AND • There are no remaining receive buffers for the packet In this situation, the requesting HCA tries again after a certain time delay. The `MPI_IB_RNR_TIMER` variable controls the delay time. If you set a value higher than the default, performance can degrade. A higher value, however, is likely to improve fabric health significantly during high congestion. For precise translations of this value to delay times, see Table 45 of the official InfiniBand specification. The default is `14`.
`MPI_IB_TIMEOUT`	When an InfiniBand card sends a packet, it waits some amount of time for an `ACK` packet to be returned by the receiving InfiniBand card. If it does not receive one, it sends the packet again. This variable controls the wait period. The time spent is equal to 4.096 × 2^n, where *n* is specified by the `MPI_IB_TIMEOUT` variable. By default, the variable is set to 18, and the time spent is expressed in microseconds.
`MPI_IB_TM`	When set, this variable enables tag matching on Mellanox InfiniBand HCAs. In some cases, this variable reduces the number of internal buffer copies that MPT creates. HPE supports this variable on systems that include the following: • Mellanox OFED 4.1 and later • Mellanox ConnectX-5 host channel adapters (HCAs)
`MPI_NUM_MEMORY_REGIONS`	For zero-copy sends over the InfiniBand interconnect, MPT keeps a cache of application data buffers registered for these transfers. This environment variable controls the size of the cache. If the application rarely reuses data buffers, it may make sense to set this value to 0 to avoid cache trashing. By default, this variable is set to 1024 (1K). The possible range is from 0 to 8192 (8K).
`MPI_NUM_QUICKS`	Controls the number of other ranks that a rank can receive from over InfiniBand using a short message fast path. This variable is 8 by default and can be any value between 0 and 32.

Tuning for running applications over the Intel Omni-Path interconnect

You can set environment variables to improve application performance when you run an MPI application across a cluster of hosts using the Intel Omni-Path interconnect. The following table shows these environment variables.

Environment variable	Effect
`MPI_OPA_PAYLOAD`	When the MPI library uses the Intel Omni-Path Architecture, it allocates memory for each message header that it uses for transfer. The memory allocation scheme used depends on the size of the data to be sent, as follows: • If the size is less than or equal to the value of `MPI_OPA_PAYLOAD`, minus 64 bytes for the actual header, the data is inlined with the header. • If the size is greater than the value of `MPI_OPA_PAYLOAD`, the message is sent through remote direct memory access (RDMA) operations. The default is `16512` bytes.
`MPI_OPA_SINGLE_COPY_BUFFER_MAX`	MPI can send data in one of the following ways: • Directly between the buffers of a process • Through intermediate buffers inside the MPI library. If the following are both true, MPI sends the data directly between the process buffers: • MPI transfers data over Intel Omni-Path Architecture AND • The size of the cumulative data is greater than the value of this environment variable. The default is `32767`.
`MPI_WILDCARDS`	The MPT Intel Omni-Path Architecture optimizes applications when it knows that the application does not make certain MPI calls. If an MPI program uses variations of `MPI_Probe()` with long messages, the program can abort and generate an error that refers to `MPI_Probe()`. In this situation, for the cost of a small performance decrease, you can set `MPI_WILDCARDS=true` to enable the application to run. When this variable is set to `false`, MPT assumes that the application does not make these calls with rank wildcards. This assumption enables MPT to make some bandwidth optimizations in its Intel Omni-Path architecture code. If `MPI_WILDCARDS` is set to `false` and one of these situations occurs, the job might abort.

For more information, see the following:

About wildcards on page 32

MPI on HPE Superdome Flex Grid systems

The following topics contain more information about using MPI on HPE Superdome Flex Grid systems:

- **General considerations for MPI jobs on HPE Superdome Flex Grid systems** on page 50
- **Performance problems and corrective actions for MPI programs on HPE Superdome Flex Grid systems** on page 50
- **Other performance considerations for MPI programs on HPE Superdome Flex Grid systems** on page 51

General considerations for MPI jobs on HPE Superdome Flex Grid systems

To run an MPI job optimally, it is best to pin MPI processes to CPUs and isolate multiple MPI jobs onto different sets of sockets and hubs. You can configure a workload manager to create a cpuset for every MPI job. MPI pins its processes to the sequential list of logical processors within the containing cpuset by default. To control and alter the pinning pattern, use the following:

- `MPI_DSM_CPULIST`. For more information, see the following:

  MPI_DSM_CPULIST on page 42
- The `omplace` command.
- The `dplace` command.

Performance problems and corrective actions for MPI programs on HPE Superdome Flex Grid systems

The MPI library chooses buffer sizes and communication algorithms in an attempt to deliver the best performance to a wide variety of MPI applications automatically. The following list of performance problems can be remedied:

- Odd Hyper-Threads are idle.

  Most high performance computing MPI programs run best using only one Hyper-Thread per core. When an HPE Superdome Flex Grid system has multiple Hyper-Threads per core, logical CPUs are numbered such that odd Hyper-Threads are the high half of the logical CPU numbers. To schedule only on the even Hyper-Threads, schedule the MPI job as if only half the full number exist. The effect is to leaving the high logical CPUs idle. To determine if cores have multiple Hyper-Threads on your HPE Superdome Flex Grid system, use the `cpumap` command. The output shows the following:

  - The number of physical and logical processors
  - Whether Hyper-Threading is on or off
  - The way in which shared processors are paired. This information appears towards the bottom of `cpumap` command output.

- MPI large message bandwidth is inappropriate.

  Some programs use the `MPI_Send` function to transfer large messages. To use unbuffered, single-copy transport in these cases, set `MPI_BUFFER_MAX=0`. For more information, see `MPI`.
- MPI small or near messages are frequent.

  For small fabric hop counts, use shared memory message delivery. To deliver all messages within an HPE Superdome Flex Grid host through shared memory, set `MPI_SHARED_NEIGHBORHOOD=HOST`. For more information, see the `MPI`(1) manpage.

Other performance considerations for MPI programs on HPE Superdome Flex Grid systems

MPI application processes typically perform better when their local memory is allocated on the socket assigned to execute the process. This allocation cannot happen if memory on that socket is exhausted, either by the application itself or by other system consumption (for example, by file buffer cache).

You can use the `nodeinfo` command to view memory consumption on the nodes assigned to your job, and you can use the `bcfree` command to clear out excessive file buffer cache. PBS Professional workload manager installations can be configured to issue `bcfree` commands in the job prologue.

For more information, see the PBS Professional documentation and the `bcfree`(1) manpage.

Measuring parallelization and parallelizing your code

When tuning for performance, first assess the amount of code that is parallelized in your program. Use the following formula to calculate the amount of code that is parallelized:

```
p=N(T(1)-T(N)) / T(1)(N-1)
```

In this equation, T(1) is the time the code runs on a single CPU and T(N) is the time it runs on N CPUs. Speedup is defined as T(1)/T(N).

If *speedup*`/N` is less than 50% (that is, N>(2-*p*)/(1-*p*)), stop using more CPUs and tune for better scaling.

You can use one of the following to display CPU activity:

- The `top` command.
- The `vmstat` command.
- The open source Performance Co-Pilot tools. For example, `pmval` (`pmval kernel.percpu.cpu.user`) or the visualization command `pmchart`.

Next, focus on using one of the following parallelization methodologies:

- **Using MPI** on page 51
- **Using OpenMP** on page 52
- **Identifying OpenMP nested parallelism** on page 52
- **Using compiler options** on page 52
- **Identifying opportunities for loop parallelism in existing code** on page 53

Using MPI

The `-lmpi` compiler option compiles and links C and Fortran programs with the MPI library. For a list of environment variables that are supported, see the `mpi`(1) manpage.

The `MPIO_DIRECT_READ` and `MPIO_DIRECT_WRITE` environment variables are supported under Linux for local XFS file systems.

MPI provides the MPI-2 standard MPI I/O functions that provide file read and write capabilities. A number of environment variables are available to tune MPI I/O performance. The `mpi_io`(3) manpage describes these environment variables.

For information about performance tuning for MPI applications, see the following:

HPE Message Passing Interface (MPI) MPInside Reference Guide

Using OpenMP

OpenMP is a shared memory multiprocessing API, which standardizes existing practice. It is scales for fine-grained or coarse-grained parallelism with an emphasis on performance. It exploits the strengths of shared memory and is directive-based. The OpenMP implementation also contains library calls and environment variables. OpenMP is included with the C, C++, and Fortran compilers.

To use OpenMP directives, specify the `ifort -openmp` or `icc -openmp` compiler options. These options use the OpenMP front end that is built into the Intel compilers. The latest Intel compiler OpenMP run-time library name is `libiomp5.so`. The latest Intel compiler also supports the GNU OpenMP library as an either/or option, in other words, do not mix-and-match the GNU library with the Intel version.

For more information, see the OpenMP standard at the following website:

http://www.openmp.org

Identifying OpenMP nested parallelism

The following Open MP nested parallelism output shows two primary threads and four secondary threads, called master/nested:

```
% cat place_nested
firsttask cpu=0
thread name=a.out oncpu=0 cpu=4 noplace=1 exact onetime thread name=a.out
     oncpu=0
cpu=1-3 exact thread name=a.out oncpu=4 cpu=5-7 exact
```

```
% dplace -p place_nested a.out
Master thread 0 running on cpu 0
Master thread 1 running on cpu 4
Nested thread 0 of master 0 gets task 0 on cpu 0 Nested thread 1 of master 0
     gets task 1 on cpu 1
Nested thread 2 of master 0 gets task 2 on cpu 2 Nested thread 3 of master 0
     gets task 3 on cpu 3
Nested thread 0 of master 1 gets task 0 on cpu 4 Nested thread 1 of master 1
     gets task 1 on cpu 5
Nested thread 2 of master 1 gets task 2 on cpu 6 Nested thread 3 of master 1
     gets task 3 on cpu 7
```

NOTE: Lines in the preceding output examples are wrapped and indented for inclusion in this documentation.

For more information, see the `dplace`(1) manpage.

Using compiler options

You can use compiler options to invoke automatic parallelization. Use the `-parallel` or `-par_report` options to the `ifort` or `icc` compiler commands. These options show which loops were parallelized and the reasons why some loops were not parallelized. If a source file contains many loops, it might be necessary to add the `-override_limits` option to enable automatic parallelization. The code generated by the `–parallel` option is based on the OpenMP API. The standard OpenMP environment variables and Intel extensions apply.

The following are some limitations to automatic parallelization:

- For Fortran codes, only `DO` loops are analyzed.
- For C/C++ codes, only `for` loops using explicit array notation or those using pointer increment notation are analyzed. In addition, `for` loops using pointer arithmetic notation are not analyzed, nor does it analyze

`while` or `do while` loops. The compiler also does not check for blocks of code that can be run in parallel.

Identifying opportunities for loop parallelism in existing code

Another parallelization optimization technique is to identify loops that have a potential for parallelism, such as the following:

- Loops without data dependencies. A **data dependency conflict** occurs when results from one loop pass are needed in future passes of the same loop.
- Loops with data dependencies because of temporary variables, reductions, nested loops, or function calls or subroutines.

Loops that do not have a potential for parallelism are those loops with premature exits, too few iterations, or those loops where the programming effort to avoid data dependencies is too great.

Suspending MPI jobs

Internally, the MPI from HPE software uses the XPMEM kernel module to provide single-copy operations to local data. The XPMEM kernel module prevents any pages used by these operations from paging.

An administrator can suspend an MPI job temporarily to allow other applications to run. To suspend, unpin the pages so they can be swapped out and made available for other applications.

Each process of an MPI application that is using the XPMEM kernel module has a `/proc/xpmem/`*`pid`* file associated with it. File `/proc/xpmem/`*`pid`* includes the number of pages owned by this process that are prevented from paging by XPMEM. You can display the content of this file. For example:

```
# cat  /proc/xpmem/5562
pages pinned by XPMEM: 17
```

The following procedure explains how to unpin the pages for use by other processes.

Procedure

1. Log in as the system administrator.
2. Suspend all the processes in the application.
3. Use the `echo`(1) command to unpin the pages.

   You can echo any value into the `/proc/xpmem/`*`pid`* file.

   For *pid*, specify the process ID.

   The `echo` command does not return until the process pages are unpinned.

   For example:

   ```
   # echo 1 > /proc/xpmem/5562
   ```

When the MPI application is resumed, the `XPMEM` kernel module prevents the pages from paging as they are referenced by the application.

Performance profiling

Performance profiling occurs when you run your MPI program or SHMEM program with a tool that can aggregate run time statistics. Profiling tools gather statistics that show information such as the following:

- The amount of time that your program spends in MPI
- The number of messages sent
- The number of bytes sent

MPT includes profiling support in the `libmpi.so` library. When you use a profiling tool, the tool automatically replaces all `MPI_`*`XXX`* prototypes and function names with `PMPI_`*`XXX`* entry points.

This chapter describes the use of profiling tools to obtain performance information. Compared to the performance analysis of sequential applications, characterizing the performance of parallel applications can be challenging. HPE recommends that you focus first on improving MPI application performance at the single process level.

It may also be important to understand the message traffic generated by an application. The following are among the tools you can use to analyze this aspect of message passing application program performance:

- HPE MPInside
- Various third-party products

The following topics contain more information about profiling:

- **HPE Message Passing Interface (MPI) MPInside Reference Guide**. This manual explains how to use the MPInside profiling tool.
- **Using perfcatch** on page 54
- **Writing your own profiling interface** on page 57
- **Using third-party profilers** on page 58
- **MPI internal statistics** on page 58

Using `perfcatch`

You can use the `perfcatch` utility to profile the performance of an MPI program or SHMEM program. The `perfcatch` utility runs the MPI program with the wrapper library, `libmpi.so`, and writes MPI call profiling information to `MPI_PROFILING_STATS`.

The following topics contain more information about `perfcatch`:

- **The perfcatch command** on page 54
- **MPI_PROFILING_STATS results file example** on page 55
- **Environment variables used with perfcatch** on page 57

The `perfcatch` command

The following format shows how to use the `perfcatch` command:

```
mpiexec_mpt [ mpi_params ] perfcatch [ -i ] cmd [ args ]
```

By default, `perfcatch` assumes an MPT program. The `perfcatch` utility accepts the arguments in the following table.

Argument	Effect
mpi_params	Optional. Specifies the MPI parameters used to launch the program.
`-i`	Specifies to use Intel MPI.
cmd	Specifies the name of the executable program. For example, `a.out`.
args	Optional. Specifies additional command-line arguments.

To use `perfcatch` with an MPI from HPE program, insert the `perfcatch` command in front of the executable file name, as the following examples show:

- **`mpiexec_mpt -np 64 perfcatch a.out arg1`**
- **`mpiexec_mpt host1 32, host2 64 perfcatch a.out arg1`**

To use `perfcatch` with Intel MPI, add the `-i` option, as follows:

```
mpiexec -np 64 perfcatch -i a.out arg1
```

For more information, see the `perfcatch`(1) manpage.

`MPI_PROFILING_STATS` results file example

The `perfcatch` utility output file is called `MPI_PROFILING_STATS`. Upon program completion, the `MPI_PROFILING_STATS` file resides in the current working directory of the MPI process with rank 0.

This output file includes a summary statistics section followed by a rank-by-rank profiling information section. The summary statistics section reports some overall statistics. These statistics include the percent time each rank spent in MPI functions and the MPI process that spent the least and the most time in MPI functions. Similar reports are made about system time usage.

In the rank-by-rank profiling information, there is a list of every profiled MPI function called by a particular MPI process. The report includes the number of calls and the total time consumed by these calls. Some functions report additional information, such as average data counts and communication peer lists.

The following is an example `MPI_PROFILING_STATS` results file:

```
============================================================
PERFCATCHER version 22
(C) Copyright Hewlett Packard Enterprise Development LP.
This library may only be used on HPE hardware platforms.
See LICENSE file for details.
============================================================
MPI program profiling information
Job profile recorded Wed Jan 17 13:05:24 2007
Program command line:   /home/estes01/michel/sastest/mpi_hello_linux
Total MPI processes                                  2

Total MPI job time, avg per rank                   0.0054768 sec
Profiled job time, avg per rank                    0.0054768 sec
Percent job time profiled, avg per rank            100%

Total user time, avg per rank                      0.001 sec
```

```
Percent user time, avg per rank                     18.2588%
Total system time, avg per rank                     0.0045 sec
Percent system time, avg per rank                   82.1648%

Time in all profiled MPI routines, avg per rank     5.75004e-07 sec
Percent time in profiled MPI routines, avg per rank 0.0104989%

Rank-by-Rank Summary Statistics
-------------------------------

Rank-by-Rank: Percent in Profiled MPI routines
        Rank:Percent
        0:0.0112245%    1:0.00968502%
  Least:  Rank 1      0.00968502%
  Most:   Rank 0      0.0112245%
  Load Imbalance:  0.000771%

Rank-by-Rank: User Time
        Rank:Percent
        0:17.2683%      1:19.3699%
  Least:  Rank 0      17.2683%
  Most:   Rank 1      19.3699%

Rank-by-Rank: System Time
        Rank:Percent
        0:86.3416%      1:77.4796%
  Least:  Rank 1      77.4796%
  Most:   Rank 0      86.3416%

Notes
-----

Wtime resolution is                   5e-08 sec

Rank-by-Rank MPI Profiling Results
----------------------------------

Activity on process rank 0

           Single-copy checking was not enabled.
comm_rank            calls:      1  time: 6.50005e-07 s  6.50005e-07 s/call

Activity on process rank 1

           Single-copy checking was not enabled.
comm_rank            calls:      1  time: 5.00004e-07 s  5.00004e-07 s/call

------------------------------------------------

recv profile

             cnt/sec for all remote ranks
local   ANY_SOURCE        0            1
 rank

------------------------------------------------
```

```
recv wait for data profile

             cnt/sec for all remote ranks
local        0            1
 rank
------------------------------------------------

recv wait for data profile

             cnt/sec for all remote ranks
local        0            1
 rank

------------------------------------------------

send profile

             cnt/sec for all destination ranks
  src        0            1
 rank

------------------------------------------------

ssend profile

             cnt/sec for all destination ranks
  src        0            1
 rank

------------------------------------------------

ibsend profile

             cnt/sec for all destination ranks
  src        0            1
 rank
```

Environment variables used with `perfcatch`

The following table shows the MPI performance-profiling environment variables.

Variable	Effect
`MPI_PROFILE_AT_INIT`	Activates MPI profiling immediately, that is, at the start of MPI program execution. To use this environment variable, set it to any value. For example, set `MPI_PROFILE_AT_INIT` to 1.
`MPI_PROFILING_STATS_FILE`	Specifies the `perfcatch` output file. MPI writes the profiling results to this file. By default, the profiler writes to `MPI_PROFILING_STATS`.

Writing your own profiling interface

You can write your own profiler by using the MPI standard `PMPI_*` calls. In addition, you can use the `MPI_Wtime` function call to time specific calls or sections of your code in the following ways:

- Within your own profiling library
- From within the application itself

The following example output is for a single rank. The program was run on 128 processors using a user-created profiling library that performs call counts and timings of common MPI calls. Notice that for this rank, most of the MPI time is spent in `MPI_Waitall` and `MPI_Allreduce`.

```
Total job time 2.203333e+02 sec
Total MPI processes 128
Wtime resolution is 8.000000e-07 sec

activity on process rank 0
comm_rank calls 1      time 8.800002e-06
get_count calls 0      time 0.000000e+00
ibsend calls    0      time 0.000000e+00
probe calls     0      time 0.000000e+00
recv calls      0   time 0.00000e+00 avg datacnt 0 waits 0 wait time 0.00000e+00
irecv calls     22039  time 9.76185e-01   datacnt 23474032 avg datacnt 1065
send calls      0      time 0.000000e+00
ssend calls     0      time 0.000000e+00
isend calls     22039  time 2.950286e+00
wait calls      0      time 0.00000e+00   avg datacnt 0
waitall calls   11045  time 7.73805e+01   # of Reqs 44078  avg data  cnt 137944
barrier calls   680    time 5.133110e+00
alltoall calls  0      time 0.0e+00       avg datacnt 0
alltoallv calls 0      time 0.000000e+00
reduce calls    0      time 0.000000e+00
allreduce calls 4658   time 2.072872e+01
bcast calls     680    time 6.915840e-02
gather calls    0      time 0.000000e+00
gatherv calls   0      time 0.000000e+00
scatter calls   0      time 0.000000e+00
scatterv calls  0      time 0.000000e+00

activity on process rank 1
...
```

Using third-party profilers

You can use third-party profiling tools with MPI from HPE. The following are examples of tools to consider:

- The TAU Performance System profiler from the University of Oregon. This software is a portable profiling and tracing toolkit. You can use it for performance analysis of parallel programs written in Fortran, C, C++, UPC, Java, and Python.
- The Allinea MAP profiler. The Allinea MAP profiler is part of the Allinea Forge toolkit

MPI internal statistics

MPI tracks certain resource utilization statistics. You can use these statistics to determine potential performance problems caused by a lack of MPI message buffers or other MPI internal resources.

To display MPI internal statistics, use the `MPI_STATS` environment variable or the `-stats` option on the `mpirun` command. MPI internal statistics are always being gathered, so displaying them does not cause significant additional overhead. In addition, one can sample the MPI statistics counters from within an application, allowing for finely grained measurements.

You can set the `MPI_STATS_FILE` environment variable to capture internal statistics upon program completion. When set, the system writes internal statistics to the file specified by this variable. These statistics can be useful in optimizing codes when you want to determine the following:

- Whether there are enough internal buffers and if processes are waiting (retries) to acquire them
- Whether single copy optimization is being used for point-to-point or collective calls

For additional information on how to use the MPI statistics counters to help tune the run-time environment for an MPI application, see the following:

Run-time tuning on page 39

Troubleshooting and frequently asked questions

Why is the `mpiexec_mpt` command failing?

If the `mpiexec_mpt` command fails, investigate the following:

- Look in `/var/log/messages` for any suspicious errors or warnings. For example, if your application tries to load a library that it cannot find, a message typically appears here. Only the administrator user can view this file.
- Be sure that you did not misspell the name of your application.
- To find dynamic link errors, try the following:
  - Run your program without `mpiexec_mpt`. When you do not use `mpiexec_mpt`, the output includes dynamic link errors that might not otherwise be displayed. In addition, the output includes the following message:

    ```
    mpiexec_mpt must be used to launch all MPI applications
    ```

  - Run your program with `mpiexec_mpt`. Set the `LD_DEBUG` environment variable to `all`, which generates output that includes a set of messages for each symbol that `rld` resolves. This variable produces numerous output messages, but it can help you find the cause of the link error.
- Verify that you set your remote directory properly. By default, `mpiexec_mpt` attempts to place your processes on all machines into the directory that has the same name as `$PWD`. This placement is the common case, but sometimes different functionality is required. For more information, see the `mpiexec_mpt`(1) manpage sections on `$MPI_DIR` and/or the `-dir` option.
- If you use a relative pathname for your application, verify that it appears in the `$PATH` environment variable. In particular, the `mpiexec_mpt` command does not look in the working directory (`"."`) for your application unless `"."` appears in `$PATH`.
- To verify that your array is configured correctly, run the following command:

  ```
  /usr/sbin/ascheck
  ```

- To verify the version of MPI that is running, run the `mpiexec_mpt -verbose` command.
- MPI creates the equivalent of a fresh login session for every job. Be careful when you set MPI environment variables from within your `.cshrc` or `.login` files. The `.cshrc` or `.login` files override any settings that you might later set from within your shell.

  The safe way to set up your environment is as follows:

  - Test for the existence of `$MPI_ENVIRONMENT` in your scripts
  - Set the other MPI environment variables only if it is undefined.
- If you are running in a Kerberos environment, you can experience unpredictable results because `mpiexec_mpt` cannot pass tokens. For example, in some cases, if you use `telnet` to connect to a host and then try to run `mpiexec_mpt` on that host, it fails. However, if you instead use `rsh` to connect to the host, `mpiexec_mpt` succeeds. This outcome might occur because `telnet` is secured with Kerberos, but `rsh` is not. If you are running under such conditions, talk to your local administrator about the proper way to launch MPI jobs.

- Look in the following directory on all the machines you are using:

  `/tmp/.arraysvcs`

  In some cases, you might find a helpful `errlog` file.
- You can increase the verbosity of the Array Services daemon, `arrayd`, when you use the `-v` option to generate more debugging information. For more information, see the `arrayd`(8) manpage.
- Check for error messages in the `/var/run/arraysvcs` directory.

Why does my code run correctly until it reaches `MPI_Finalize()` and then hang?

A code hang is almost always caused by a `send` request or a `recv` request that is either unmatched or not completed. These request types are as follows:

- An **unmatched request** is any blocking `send` request for which a corresponding `recv` request is never posted.
- An **incomplete request** is any nonblocking `send` or `recv` request that was never freed by a call to `MPI_Test()`, `MPI_Wait()`, or `MPI_Request_free()`.

Common examples are applications that call `MPI_Isend()` and then use internal means to determine when it is safe to reuse the send buffer. These applications never call `MPI_Wait()`. To fix such codes, update your code in one of the following ways:

- Insert a call to `MPI_Request_free()` immediately after all such `isend` operations.
- Add a call to `MPI_Wait()` at a later place in the code. Add this call prior to the point at which the send buffer must be reused.
- Set `MPI_REQUEST_DEBUG=true`, which causes MPT to check for this condition at `MPI_Finalize()` time.

Why does my hybrid code (using OpenMP) stall on the `mpirun` command?

If your application was compiled with the Open64 compiler, follow the linking instructions in the following topic:

Compiling and linking the MPI program on page 17

Why do I keep receiving warning messages about the `MPI_REQUEST_MAX` value being too small?

The MPI library generates the following warning message when the `MPI_REQUEST_MAX` value is not set appropriately:

```
MPT Warning:  MPT has run out of preallocated request entries.
This may slow performance or fragment memory.
Please increase MPI_REQUEST_MAX.
```

The following are the conditions under which the MPI library generates the preceding message:

- The number of simultaneous transfer requests in the program exceeds the number of requests for which MPT preallocates. To fix this situation, use the `MPI_REQUEST_MAX` shell variable to increase the number of preallocated requests.
- The application calls `MPI_Isend()` or `MPI_Irecv()` and does not complete or free the requested objects. To fix this situation, use `MPI_Request_free()`, as described in the following:

  Why does my code run correctly until it reaches MPI_Finalize() and then hang? on page 61

Why is it that I do I not see any `stdout` and/or `stderr` output from my MPI application?

All `stdout` output and `stderr` output is line-buffered, which means that the `mpirun` command does not print any partial lines of output. This situation sometimes causes problems for codes that prompt the user for input parameters but do not end their prompts with a newline character. As a remedy, append a newline character to each prompt.

You can set the `MPI_UNBUFFERED_STDIO` environment variable to disable line-buffering. For more information, see the `MPI`(1) and `mpirun`(1) manpages.

Where can I find more information about the OpenSHMEM programming model?

See the `intro_shmem`(3) manpage.

Why does the `ps` command say that my memory use (`SIZE`) is higher than expected?

At job start-up, the MPI and OpenSHMEM libraries cross-map all the user static and heap memory of the processes on the local host to provide optimization opportunities. The result is large virtual memory usage.

The `ps` command `SIZE` statistic conveys the amount of virtual address space used, not the amount of memory consumed.

Even if all the pages that you could reference were faulted in, most of the virtual address regions point to multiply mapped (shared) data regions. Even in these cases, actual per-process memory usage would be far lower than that indicated by `SIZE`.

What does the `MPI: could not run executable` message mean?

This message means that something happened while `mpiexec_mpt` was trying to launch your application. The `mpiexec_mpt` command failed before all the MPI processes were able to handshake with it.

The `mpiexec_mpt` command directs `arrayd` to launch a shepherd process on each host and listens on a socket for those shepherds to connect back to it. Because the shepherds are children of `arrayd`, `arrayd` traps `SIGCHLD` and passes that signal back to `mpiexec_mpt` whenever one of the shepherds terminates. If `mpiexec_mpt` receives a signal before it establishes connections with every host in the job, it knows that something has gone wrong.

How do I combine MPI with other tools?

Different MPI implementations use different methods to launch their worker processes. Some tools expect a method that is different from the MPT default. If you set the shell variable `MPI_SHEPHERD=true`, then MPT attempts to use a launch method that is similar to some other MPI implementations. Use of this option can disable some less-common features, such as spawning and checkpoint-restart support.

In general, the rule to follow is to run the `mpiexec_mpt` command on your tool and then run the tool on your application. Do not try to run the tool on `mpiexec_mpt`.

Also, because of the way that `mpiexec_mpt` sets up `stdio`, viewing the output from your tool might require a bit of effort. The most ideal case is when the tool directly supports an option to redirect its output to a file. However, many tools, for example, `dplace`, do not support such an option. However, you might be able to wrap a shell script around the tool and have the script do the redirection, as in the following example:

```
> cat myscript
#!/bin/sh
#################################################################
# NOTE: The example shown is for illustrative purposes only and #
# has not been evaluated for use in a production environment.   #
#################################################################
setenv MPI_DSM_OFF
dplace -verbose a.out 2> outfile
> mpirun -np 4 myscript
hello world from process 0
hello world from process 1
hello world from process 2
hello world from process 3
> cat outfile
there are now 1 threads
Setting up policies and initial thread.
Migration is off.
Data placement policy is PlacementDefault.
Creating data PM.
Data pagesize is 16k.
Setting data PM.
Creating stack PM.
Stack pagesize is 16k.
Stack placement policy is PlacementDefault.
Setting stack PM.
there are now 2 threads
there are now 3 threads
there are now 4 threads
there are now 5 threads
```

CAUTION: The preceding script example is for illustrative purposes only and has not been evaluated for use in a production environment.

Why do I see stack traceback information when my MPI job aborts?

For information, see the `MPI_COREDUMP` environment variable description and the `MPI_COREDUMP_DEBUGGER` environment variable description on the `MPI`(1) manpage.

Array Services

The HPE Array Services software enables parallel applications to run on multiple hosts in a cluster, or *array*. Array Services provides cluster job launch capabilities for Message Passing Toolkit (MPT) jobs.

The array can consist of the following:

- Multiple server nodes on a cluster computing system
- Multiple physical machines

An array system is bound together with a high-speed network and the Array Services software. Array users can access the system with familiar commands for job control, authentication, and remote execution. Array Services facilitates the following:

- Global session management
- Array configuration management
- Batch processing
- Message passing
- System administration
- Performance visualization

The Array Services software package includes the following:

- An array daemon that runs on each node. The daemon groups logically related processes together across multiple nodes. The process groups create a global process namespace across the array, facilitate accounting, and facilitate administration.

  The daemon maintains information about node configuration, process IDs, and process groups. Array daemons on the nodes cooperate with each other.
- Array configuration files. The files describe the array configuration and provide reference information for array daemons and user programs. Each node hosts a copy of each array configuration file.
- Commands, libraries, and utilities such as `ainfo`, `arshell`, and others.

The Message Passing Interface (MPI) of the MPI for HPE software uses Array Services to launch parallel applications.

The MPI from HPE software distribution includes the MUNGE software. This optional, open-source product provides secure Array Services functionality. MUNGE allows a process to authenticate the UID and GID of another local or remote process. The other local or remote process must reside on a group of hosts that have common users and groups. MUNGE authentication, which also includes the Array Services data exchanged in the array, is encrypted. For more information about MUNGE, see the MUNGE website at the following location:

http://dun.github.io/munge/

The Array Services package requires that the process sets service is installed and running. This package is provided in the `sgi-procset` RPM. To verify that the process sets service is installed and running, use one of the following command sequences:

- On RHEL 7.X, SLES 15, or SLES 12 SPX systems, enter the following commands:

```
# rpm -q sgi-procset
# systemctl status procset
```

- On RHEL 6.X or SLES 11 SPX systems, enter the following commands:

```
# rpm -q sgi-procset
# /etc/init.d/procset status
```

The following topics contain end-user information about Array Services:

- **Installing and configuring Array Services** on page 65
- **Array Services commands and arguments** on page 65
- **Array Services environment variables** on page 67
- **Obtaining information about the array** on page 67

NOTE: For Array Services information that pertains to system administration, see the following:

Array Services system administration information on page 94

Installing and configuring Array Services

As the system administrator user, install and configure the Array Services software before end users do the following:

- Use the Array Services software
- Run MPI from HPE programs

See one of the following for automated installation information:

- **HPE Performance Cluster Manager Installation Guide**
- **HPE SGI Management Suite Installation and Configuration Guide**
- **Installing the Array Services software on clusters that use the HPE Insight Cluster Management Utility (CMU)** on page 9

For information about the manual installation procedure, see the following:

Array Services system administration information on page 94

Array Services commands and arguments

When an application starts multiple processes on multiple nodes, the following are no longer adequate to manage the application:

- A Linux process identifier (PID)
- A process group identifier (PGID)

The Array Services commands enable you to view the entire array and to control processes on multiple nodes. You can enter Array Services commands from any workstation connected to an array system. You do not have to log in to an array node.

To retrieve overview information about Array Services online, see the following manpage:

`array_services`(5)

The following topics contain more information about Array Services commands:

- **Array Services commands** on page 66
- **Additional information for the ainfo command and the array command** on page 66

Array Services commands

The following are the Array Services commands:

- `ainfo`

  Retrieves information about the different arrays at your site and about the nodes included in each array. At most sites, there is only one array, but you can have multiple arrays at your site. The command output includes the hostnames for each node in each array at your site.

- `array`

  Runs a system command on one or more nodes and returns output to `stdout`. As arguments, `array` accepts several options and the one system command that you want to run on the array. There is a default set of system commands, but your system administrator determines the list of commands available to you at your site.

- `arshell`

  Runs a system command remotely on a different node. As arguments, `arshell` accepts several options and the one system command that you want to run on the remote node.

  The `arshell` command is like `rsh` in that it runs a command on another machine under the user ID of the invoking user. Use of authentication codes makes Array Services more secure than `rsh`.

The `ainfo`(1), `array`(1), and `arshell`(1) online manpages explain the arguments that each command accepts.

Additional information for the `ainfo` command and the `array` command

The `ainfo` command and `array` command support several common command-line arguments. For comprehensive information about these commands, see the `ainfo`(1) and `array`(1) online manpages.

The following information supplements the information on the manpages:

- Your array administrator might have established an authentication code for all or some array nodes. The authentication code is a 64-bit number.

  The `ainfo` and `array` commands accept the following options:

  - `-Kl` *`number`*
  - `-Kr` *`number`*

  For *number*, specify the 64-bit authentication key number for the local node or the remote node that you want to target with the command. The code applies to any command entered at that node or addressed to that node.

  If it is necessary to specify an authentication code, your system administrator can tell you.

NOTE: For the `-Kl` *`number`* option, the option letter is a lowercase letter “L”, for “local”.

- The `-l` and `-s` options work together. The `-l` option restricts the scope of a command to the node upon which the command is run. This option is a lowercase letter “L”, for “local”. By default, that is the node where the command is entered. When `-l` is not used, the command queries all nodes of the array. The `-s` option runs the command on a specified node of the array. These options work together as follows:
    - To query all nodes as seen by the local node, use neither option.
    - To query only the local node, use only `-l`.
    - To query all nodes as seen by a specified node, use only `-s`.
    - To query only a particular node, use both `-s` and `-l`.

Array Services environment variables

The Array Services commands depend on environment variables to define default values for the less-common command options. The following table summarizes these variables.

Table 2: Array Services environment variables

Variable name	Use	Default when undefined
`ARRAYD_FORWARD`	When defined with a string starting with the letter *y*, all commands default to forwarding through the array daemon (option `-F`).	Commands default to direct communication (option `-D`).
`ARRAYD_PORT`	The port (socket) number monitored by the array daemon on the destination node.	The standard number of 5434, or the number given with option `-p`.
`ARRAYD_LOCALKEY`	Authentication key for the local node (option `-Kl`).	No authentication unless the `-Kl` option is used.
`ARRAYD_REMOTEKEY`	Authentication key for the destination node (option `-Kr`).	No authentication unless `-Kr` option is used.
`ARRAYD`	The destination node, when not specified by the `-s` option.	The local node, or the node given with `-s`.

Obtaining information about the array

You can use Array Services commands and system commands to retrieve information about the array. For example, you can use the `ainfo` command and the `array` command to check the hardware components and the software workload on the array. In addition, you can use system commands, such as `who`, `top`, and `uptime`, to retrieve information about users and workload on a node. To obtain information about the entire array, use these commands with the `array` command.

The following topics include examples that show how to retrieve information about the array:

Retrieving array names

The following command shows how to retrieve the names of all arrays configured at your site:

```
homegrown% ainfo arrays
Arrays known to array services daemon
ARRAY DevArray
    IDENT 0x3381
ARRAY BigDevArray
    IDENT 0x7456
ARRAY test
    IDENT 0x655e
```

Your system administrator configures the arrays at your site. Different arrays might know different sets of other array names.

Retrieving node names

The following command uses the `-b` option of `ainfo` command to retrieve a brief version of the information about all the nodes in the current array:

```
homegrown 175% ainfo -b machines
machine homegrown homegrown 5434 192.48.165.36 0
machine disarray disarray 5434 192.48.165.62 0
machine datarray datarray 5434 192.48.165.64 0
machine tokyo tokyo 5434 150.166.39.39 0
```

Retrieving user names

Example 1. The following `array who` command retrieves the names of all users logged in to the array:

```
mynode% array who
frederik corfu        rummage.eng.sgi  -tcsh
stefaan  sf           yoga.eng.sgi  -tcsh
timo     tokyo        frost.ued.sgi   /bin/tcsh
wim      boston       sig.eng.sgi   vi +153 fs/procfs/prd
ruben    paris        mountain.eng.sgi   -tcsh
...
```

Example 2. The following command retrieves the names of users logged in to the node named `tokyo`. This command uses the `-l` and `-s` options.

```
homegrown 180% array -s tokyo -l who
joecd    tokyo        frummage.eng.sgi -tcsh
joecd    tokyo        frummage.eng.sgi -tcsh
benf     tokyo        einstein.ued.sgi. /bin/tcsh
yohn     tokyo        rayleigh.eng.sg vi +153 fs/procfs/prd
...
```

The preceding examples have been edited for brevity and security.

Retrieving workload information

Example 1. The following command shows how to use the `uptime` command to retrieve information for the entire array:

```
homegrown 181% array uptime
   homegrown:  up 1 day,  7:40,  26 users,  load average: 7.21, 6.35, 4.72
    disarray:  up  2:53,  0 user,  load average: 0.00, 0.00, 0.00
    datarray:  up  5:34,  1 user,  load average: 0.00, 0.00, 0.00
       tokyo:  up 7 days,  9:11,  17 users,  load average: 0.15, 0.31, 0.29
```

Example 2. The following command shows how to use the `uptime` command to retrieve information about a single node:

```
homegrown 182% array -l -s tokyo uptime
       tokyo:  up 7 days,  9:11,  17 users,  load average: 0.12, 0.30, 0.28
```

Example 3. The following command shows how to use the `top` command to list the processes that are currently using the most CPU time:

```
homegrown 183% array top
        ASH        Host           PID User           %CPU Command
----------------------------------------------------------------
0x1111ffff00000000 homegrown        5 root           1.20 vfs_sync
0x1111ffff000001e9 homegrown     1327 arraysvcs      1.19 atop
0x1111ffff000001e9 tokyo        19816 arraysvcs      0.73 atop
0x1111ffff000001e9 disarray      1106 arraysvcs      0.47 atop
0x1111ffff000001e9 datarray      1423 arraysvcs      0.42 atop
0x1111ffff00000000 homegrown       20 root           0.41 ShareII
0x1111ffff000000c0 homegrown    29683 kchang         0.37 ld
0x1111ffff0000001e homegrown     1324 root           0.17 arrayd
0x1111ffff00000000 homegrown      229 root           0.14 routed
0x1111ffff00000000 homegrown       19 root           0.09 pdflush
0x1111ffff000001e9 disarray      1105 arraysvcs      0.02 atopm
```

The output identifies each process by its internal array session handle (ASH) value. As an alternative, you could use the `-l` and `-s` options to select data about a single node.

Using the Message Passing Toolkit (MPT) plugin for Nagios

Nagios is a web-based system monitoring tool that HPE automatically installs on cluster computing systems. Nagios enables you to monitor the cluster infrastructure. When you install the optional MPT plugin for Nagios, the MPT system log messages that typically appear in `/var/log/messages` also appear in the Nagios GUI. The plugin scans the system log for messages that MPT has logged. In the Nagios GUI, the plugin displays the number of error messages and warning messages that the plugin encountered in the scan.

Preparing to install the MPT Nagios plugin

Procedure

1. Locate the HPE Performance Software installation DVD, and insert the DVD into the DVD reader on the admin node.
2. Log into the admin node as the administrator user.
3. Change to the RPM repository directory.
4. Proceed to the following:

   Installing the MPT Nagios plugin on page 70

Installing the MPT Nagios plugin

Prerequisites

Make sure that you completed the following procedure:

Preparing to install the MPT Nagios plugin on page 70

Procedure

1. Enter one of the following commands to install the plugin:

   - On RHEL systems, enter the following command:

     ```
     # yum install checkmpt-plugin
     ```
   - On SLES systems, enter the following command:

     ```
     # zypper in checkmpt-plugin
     ```

   The preceding commands install the following files:

   ```
   /opt/hpe/hpc/mpt/checkmpt-plugin/README
   /opt/sgi/nagios/libexec/check_mpt
   ```
2. Use a text editor to open file `/opt/hpe/hpc/mpt/checkmpt-plugin/README`, and leave the file open in a window on your desktop.

   This file contains a shorthand version of these installation instructions. Some steps in this installation procedure require you to insert specific lines into specific files. It is easiest to copy the lines out of the README file and modify them as this procedure explains.

3. Enter the following command to edit file `sudoers`:

   ```
   # visudo
   ```

4. Copy the following lines from the `README` file to the end of the `sudoers` file, and replace `<nagiosuser>` and `<PLUGINSDIR>` with values that are valid at your site:

   ```
   # check_mpt plugin for Nagios (needs access to syslogs)
   <nagiosuser> ALL=NOPASSWD: <PLUGINSDIR>/check_mpt
   # end check_mpt
   ```

   Replace the variables in the preceding lines as follows:

   - Replace `<nagiosuser>` with the Nagios username assigned when Nagios was installed. By default, this username is `nagios`.
   - Replace `<PLUGINSDIR>` with the directory in which the Nagios plugin resides. By default, this directory is `/opt/sgi/nagios/libexec`.

5. Save and close the `sudoers` file.

6. Use a text editor to open file `commands.cfg`.

   By default, this file resides in the following directory:

   ```
   /opt/sgi/nagios/etc/objects
   ```

7. Copy the following lines from the `README` file to the end of the `commands.cfg` file:

   ```
   # check_mpt command definition
       define command {
               command_name check_mpt
               command_line sudo $USER1$/check_mpt -W $ARG1$ -E $ARG2$
       }
       # end check_mpt
   ```

   You do not need to assign values to `$ARG1$` or `$ARG2$`. A later step in this procedure populates these arguments with values.

8. Save and close the `commands.cfg` file.

9. Use a text editor to open file `localhost.cfg`.

   By default, this file resides in the following directory:

   ```
   /opt/sgi/nagios/etc/objects
   ```

10. Copy the following lines from the `README` file to the end of the `localhost.cfg` file:

    ```
    # check_mpt service definition
    define service {
            use                     local-service
            host_name               localhost
            service_description     check_mpt
            check_command           check_mpt!10!5
            max_check_attempts      2
            normal_check_interval   2
            retry_check_interval    1
    }
    # end of check_mpt
    ```

    The key lines in the preceding module have the following effects:

Line	Comment
`use local-service`	Use the generic Nagios template.
`host_name localhost`	Run on localhost or similar.
`service_description check_mpt`	Declare the service name.
`check_command check_mpt!10!5`	Is CRITICAL if >10 warnings / >5 errors.
`max_check_attempts 2`	If OK, try check again.
`normal_check_interval 2`	Run check every 2 minutes.
`retry_check_interval 1`	Retry every 1 minute.

11. Save and close file `localhost.cfg`.
12. Enter the following command to verify the changes that you made and to make sure that there are no conflicts:

    ```
    nagios_dir/bin/nagios -v nagios_dir/etc/nagios.cfg
    ```

    For *nagios_dir*, specify the Nagios home directory. By default, this directory is `/opt/sgi/nagios`.
13. Restart Nagios on the node.

    On RHEL 7, SLES 15, and SLES 12 platforms, enter the following command:

    ```
    # systemctl restart Nagios
    ```

    On RHEL 6 platforms and SLES 11 platforms, enter the following command:

    ```
    # service nagios restart
    ```

    If you change any of the Nagios `.cfg` files, restart Nagios.
14. On the admin node, use a shell command to set the following environment variable:

    ```
    MPI_SYSLOG_COPY=1
    ```

    For example:

    ```
    # set MPI_SYSLOG_COPY=1
    ```

    Make sure to set this value in your shell before you run any MPI for HPE or SHMEM applications.
15. (Optional) Leave the DVD in the admin node DVD reader, and install the Nagios plugin on one or more rack leader controllers (RLCs).

    After you install the plugin on an RLC, you can start Nagios on that RLC and monitor the following:

    - The messages on that RLC
    - The messages related to that compute nodes associated with that RLC

    Complete the following steps to install the Nagios plugin on an RLC:

a. From the admin node, use the `ssh` command to log into one of the RLCs as the administrator user.

b. Complete this procedure again starting with **installing the plugin software**.

Viewing MPT messages from within Nagios and clearing the messages

The following procedure explains how to retrieve MPT messages and clear MPT messages.

Procedure

1. Log into one of the cluster nodes.

   If you log into the admin node and start Nagios from the admin node, Nagios displays information for the whole cluster.

   If you log into one of the RLCs and start Nagios from one of the RLCs, Nagios displays information for that RLC. Nagios also displays information for the nodes that are subordinate to that RLC.

2. Start Nagios.

   Enter one of the following URLs into your browser:

   - To start Nagios on the admin node, enter the following:

     `http://`*`admin_name`*`/nagios/`*`rlc_name`*

     For *admin_name*, enter the hostname or IP address of the admin node.

   - To start Nagios on one of the RLCs, enter the following:

     `http://`*`admin_name`*`/nagios/`*`rlc_name`*

     For *admin_name*, type the hostname or IP address of the admin node.

     For *rlc_name*, type the hostname or IP address of the RLC. For example, `r1lead`.

3. Enter the username and password for the Nagios user.

   By default, the username is `nagiosadmin`. By default, the password is `sgisgi`.

4. Look for MPT information in the Nagios interface.

   By default, the plugin scans the messages in the `/var/log/messages` and reports messages to Nagios, as follows:

   - If you installed the plugin on the admin node, the plugin sends messages to Nagios for the admin node.
   - If you installed the plugin on one or more RLCs, the plugin sends messages to Nagios for the RLC and the RLC compute nodes. To observe the messages related to an RLC, start Nagios on that RLC.

   If you click an MPT message from within the Nagios interface, you retrieve more information about the message.

5. Use administrator commands to remedy the error conditions, if needed.

6. On the admin node, run the `check_mpt` command to clear the messages that Nagios reported.

   If you installed the plugin on the RLCs, run the `check_mpt` on RLCs, too.

The MPT plugin works by scanning `/var/log/messages`, from beginning to end. To stop the plugin from repeatedly scanning the log file, a file offset is preserved. After you run the `check_mpt` command, the changes appear in Nagios after the next scan.

The following examples show how to use options to the `check_mpt` command to direct the plugin to scan the system log according to your site preferences.

Example 1. To direct the plugin to scan for only newly logged messages, use the `-C` option. The `-C` option clears all current message counts and requests that Nagios continue its scan for new messages. Also, the `-C` parameter changes the Nagios `CRITICAL` and `WARNING` status back to `OK` after you correct the reported error condition. To use this option, enter the following command:

```
# check_mpt -C
```

Example 2. The `-X` parameter directs the plugin to start a new scan of `/var/log/messages`, clears the MPT message counts, and resets the offsets to 0. You can run `check_mpt` with the `-X` parameter after each log rotation. This command is as follows:

```
# check_mpt -X
```

The `check_mpt` command accepts additional parameters. For more information on these parameters, enter the following command to retrieve a usage statement:

```
# check_mpt -h
```

(Optional) Modifying the notification email

Nagios sends notifications to the Nagios GUI. Nagios also sends email notifications of alert conditions. If you modify the Nagios email configuration file, the Nagios email can include hostname information in the notifications. After you modify the configuration in this manner, you can identify the node upon which the error condition occurred more easily.

The `commands.cfg file` contains the following:

```
# 'notify-service-by-email-long' command definition
define command {
        command_name    notify-service-by-email-long
        command_line    /usr/bin/printf "%b" "***** Nagios *****\n\nNotification
Type: $NOTIFICATIONTYPE$\n\nService: $SERVICEDESC$\nHost: $HOSTALIAS$ \nAddress:
$HOSTADDRESS$\nState: $SERVICESTATE$\n\nDate/Time: $LONGDATETIME$\n\nAdditional
Info:\n\n$SERVICEOUTPUT$\n\n$LONGSERVICEOUTPUT$" | /usr/bin/mail -s "**
$NOTIFICATIONTYPE$ Service Alert: $HOSTALIAS$/$SERVICEDESC$ is $SERVICESTATE$ **"
$CONTACTEMAIL$
}
```

If you change `$HOSTALIAS$` to hostname, the Nagios emails include the hostname of the node upon which the error condition occurred. For example, the following file shows this enhancement:

```
# 'notify-service-by-email-long' command definition
define command {
        command_name    notify-service-by-email-long
        command_line    /usr/bin/printf "%b" "***** Nagios *****\n\nNotification
Type: $NOTIFICATIONTYPE$\n\nService: $SERVICEDESC$\nHost: `hostname` \nAddress:
$HOSTADDRESS$\nState: $SERVICESTATE$\n\nDate/Time: $LONGDATETIME$\n\nAdditional
Info:\n\n$SERVICEOUTPUT$\n\n$LONGSERVICEOUTPUT$" | /usr/bin/mail -s "**
$NOTIFICATIONTYPE$ Service Alert: $HOSTALIAS$/$SERVICEDESC$ is $SERVICESTATE$ **"
$CONTACTEMAIL$
}
```

For more information about Nagios and the Nagios email reporting feature, see your Nagios documentation.

High-performance computing tools (HPC) tools

The HPC tools include the data placement tools and the flexible file I/O (FFIO) tools.

Data placement tools

The data placement tools place data in specific memory locations, which minimizes communication overhead within an application.

The `dplace` tool and the cpuset tools are built upon the cpusets API. You can use these tools avoid poor data locality in your application. Poor data locality can be caused by process or thread drift from CPU to CPU. The `omplace` tool works like the `dplace` tool and is designed for use with OpenMP applications. The differences among these tools are as follows:

- The `taskset` command restricts execution to the listed set of CPUs when you specify one of the following options:
    - `-c`
    - `--cpu-list`

  The process is free to move among the CPUs that you specify.
- The `dplace` tool differs from `taskset` in that `dplace` binds processes to specified CPUs in round-robin fashion. After a process is pinned, it does not migrate. To increase the performance and reproducibility of parallel codes, use `dplace`.
- Cpusets are named subsets of system cpus/memories and are used extensively in batch environments. Use any of the following to manage cpusets:
    - The cpuset controller in Linux cgroups
    - The cpuset interfaces that HPE provides

For information about the data placement tools, see the following:

- **dplace command** on page 75
- **omplace command** on page 80
- The cpuset information in the following:

  HPE Performance Software - Message Passing Interface Cpuset Software Guide
- `taskset` command information in the following manual:

  Linux Application Tuning Guide for SGI X86-64 Based Systems

`dplace` command

By default, memory is allocated to a process on the node on which the process is running. When a process moves from node to node while it is running, memory references are made to remote nodes. Because remote accesses typically have higher access times, performance can degrade. Also, MPI has to reload CPU instruction pipelines.

The `dplace` command specifies scheduling and memory placement policies for the process. You can use the `dplace` command to bind a related set of processes to specific CPUs or nodes to prevent process migrations. In some cases, this binding improves performance because more memory accesses are made to local nodes.

Processes always execute within a cpuset. The cpuset specifies the CPUs on which a process can run. By default, processes usually execute in a cpuset that contains all the CPUs in the system.

The `dplace` command creates a placement container that includes all the CPUs, or a subset of CPUs, of a cpuset. The `dplace` process is placed in this container. By default, the `dplace` process is bound to the first CPU of the cpuset associated with the container. Then `dplace` invokes `exec` to run the command.

The command runs within this placement container and remains bound to the first CPU of the container. As the command forks child processes, the child processes inherit the container and are bound to the next available CPU of the container.

If you do not specify a placement file, `dplace` binds the processes sequentially in a round-robin fashion to CPUs of the placement container. For example, assume that the current cpuset consists of physical CPUs 2, 3, 8, and 9. The first process launched by `dplace` is bound to CPU 2. The first child process forked by this process is bound to CPU 3. The next process, regardless of whether it is forked by a parent or a child, is bound to CPU 8, and so on. If more processes are forked than there are CPUs in the cpuset, binding starts over with the first CPU in the cpuset.

For more information about `dplace`, see the `dplace`(1) manpage, which also includes examples of how to use the command.

For information about cpusets, see the following:

HPE Performance Software - Message Passing Interface Cpuset Software Guide

Example: Using the `dplace` command with MPI programs

The following command improves the placement of MPI programs and verifies placement of certain data structures of a long-running MPI program:

```
% mpirun -np 64 /usr/bin/dplace -s1 -c 0-63 ./a.out
```

The `-s1` parameter causes `dplace` to start placing processes with the second process, `p1`. The first process, `p0`, is not placed because it is associated with the job launch, not with the job itself. The `-c 0-63` parameter causes `dplace` to use processors 0-63.

In another window, you can use the `dlook` command to verify placement of the data structures on one of the slave thread PIDs. For more information about the `dlook` command, see the `dlook`(1) manpage.

Example: Using the `dplace` command with OpenMP programs

The following command runs an OpenMP program on logical CPUs 4 through 7 within the current cpuset:

```
% efc -o prog -openmp -O3 program.f
% setenv OMP_NUM_THREADS 4
% dplace -c4-7 ./prog
```

Example: Using the `dplace` command with OpenMP programs

The `dplace` command has a static load balancing feature, so you do not have to supply a CPU list. To place `prog1` on logical CPUs 0 through 3 and `prog2` on logical CPUs 4 through 7, enter the following:

```
% setenv OMP_NUM_THREADS 4
% dplace ./prog1 &
% dplace ./prog2 &
```

You can use the `dplace -q` command to display the static load information.

Example: Using the `dplace` command with Linux commands

The examples in the following table assume that you run the `dplace` commands from a shell that runs in a cpuset. The cpuset consists of physical CPUs 8 through 15.

Command	Run location
`dplace -c2 date`	Runs the `date` command on physical CPU 10.
`dplace make linux`	Runs `gcc` and related processes on physical CPUs 8 through 15.
`dplace -c0-4,6 make linux`	Runs `gcc` and related processes on physical CPUs 8 through 12 or 14.
`taskset 4,5,6,7 dplace app`	The `taskset` command restricts execution to physical CPUs 12 through 15. The `dplace` command sequentially binds processes to CPUs 12 through 15.

Example: Using the `dplace` command and a debugger for verification

To use the `dplace` command accurately, be aware of how your placed tasks are created in terms of the `fork`, `exec`, and `pthread_create` calls. Determine whether each worker call is an MPI rank task or a group of pthreads created by rank tasks. Here is an example of two MPI ranks, each creating three threads:

```
cat <<EOF > placefile
firsttask cpu=0
exec name=mpiapp cpu=1
fork   name=mpiapp cpu=4-8:4 exact
thread name=mpiapp oncpu=4 cpu=5-7 exact
thread name=mpiapp oncpu=8
cpu=9-11 exact EOF

#  mpirun is placed on cpu 0 in this example
#  the MPI shepherd is placed on cpu 1 in this example

# or, if your version of dplace supports the "cpurel=" option:
# firsttask cpu=0
# fork   name=mpiapp cpu=4-8:4 exact
# thread name=mpiapp oncpu=4 cpurel=1-3 exact


# create 2 rank tasks, each will pthread_create 3 more
# ranks will be on 4 and 8
#  thread children on 5,6,7   9,10,11
dplace -p placefile mpirun -np 2 ~cpw/bin/mpiapp -P 3 -l


exit
```

You can use the debugger to determine whether it is working. The following output shows two MPI processes, each with three pthreads:

```
>> pthreads | grep mpiapp
px *(task_struct *)e00002343c528000   17769   17769   17763  0       mpiapp
      member task: e000013817540000   17795   17769   17763  0     5 mpiapp
      member task: e000013473aa8000   17796   17769   17763  0     6 mpiapp
      member task: e000013817c68000   17798   17769   17763  0       mpiapp
px *(task_struct *)e0000234704f0000   17770   17770   17763  0       mpiapp
      member task: e000023466ed8000   17794   17770   17763  0     9 mpiapp
      member task: e00002384cce0000   17797   17770   17763  0       mpiapp
      member task: e00002342c448000   17799   17770   17763  0       mpiapp
```

You can also use the debugger to see the MPT shepherd, the parent of the two MPI processes, as follows:

```
>> ps | grep mpiapp
0xe00000340b300000   1139  17763  17729      1  0xc800000   -  mpiapp
0xe00002343c528000   1139  17769  17763      0  0xc800040   -  mpiapp
0xe0000234704f0000   1139  17770  17763      0  0xc800040   8  mpiapp
```

These processes are placed as specified:

```
>> oncpus e00002343c528000 e000013817540000 e000013473aa8000
>> e000013817c68000 e0
000234704f0000 e000023466ed8000 e00002384cce0000 e00002342c448000
task: 0xe00002343c528000  mpiapp cpus_allowed: 4
task: 0xe000013817540000  mpiapp cpus_allowed: 5
task: 0xe000013473aa8000  mpiapp cpus_allowed: 6
task: 0xe000013817c68000  mpiapp cpus_allowed: 7
task: 0xe0000234704f0000  mpiapp cpus_allowed: 8
task: 0xe000023466ed8000  mpiapp cpus_allowed: 9
task: 0xe00002384cce0000  mpiapp cpus_allowed: 10
task: 0xe00002342c448000  mpiapp cpus_allowed: 11
```

Example: Using the `dplace` command for compute thread placement troubleshooting

Sometimes compute threads do not end up on unique processors when using commands such a `dplace` or `profile.pl`.

In this example, assume that the `dplace -s1 -c0-15` command bound 16 processes to run on 0-15 CPUs. However, output from the `top` command shows the following:

- Only 13 CPUs running
- CPUs 13, 14, and 15 still idle
- CPUs 0, 1 and 2 are shared with six processes

```
263 processes: 225 sleeping, 18 running, 3 zombie, 17 stopped
CPU states:  cpu    user    nice  system    irq  softirq  iowait    idle
             total  1265.6%    0.0%   28.8%   0.0%    11.2%    0.0%  291.2%

             cpu00  100.0%    0.0%    0.0%   0.0%     0.0%    0.0%    0.0%

             cpu01   90.1%    0.0%    0.0%   0.0%     9.7%    0.0%    0.0%

             cpu02   99.9%    0.0%    0.0%   0.0%     0.0%    0.0%    0.0%

             cpu03   99.9%    0.0%    0.0%   0.0%     0.0%    0.0%    0.0%

             cpu04  100.0%    0.0%    0.0%   0.0%     0.0%    0.0%    0.0%

             cpu05  100.0%    0.0%    0.0%   0.0%     0.0%    0.0%    0.0%

             cpu06  100.0%    0.0%    0.0%   0.0%     0.0%    0.0%    0.0%

             cpu07   88.4%    0.0%   10.6%   0.0%     0.8%    0.0%    0.0%

             cpu08  100.0%    0.0%    0.0%   0.0%     0.0%    0.0%    0.0%
```

```
            cpu09   99.9%    0.0%    0.0%   0.0%     0.0%    0.0%    0.0%

            cpu10   99.9%    0.0%    0.0%   0.0%     0.0%    0.0%    0.0%

            cpu11   88.1%    0.0%   11.2%   0.0%     0.6%    0.0%    0.0%

            cpu12   99.7%    0.0%    0.2%   0.0%     0.0%    0.0%    0.0%

            cpu13    0.0%    0.0%    2.5%   0.0%     0.0%    0.0%   97.4%

            cpu14    0.8%    0.0%    1.6%   0.0%     0.0%    0.0%   97.5%

            cpu15    0.0%    0.0%    2.4%   0.0%     0.0%    0.0%   97.5%
 Mem:  60134432k av, 15746912k used, 44387520k free,       0k shrd,
672k buff
        351024k active,            13594288k inactive

  Swap: 2559968k av,       0k used, 2559968k free
  2652128k cached

    PID USER     PRI  NI  SIZE  RSS SHARE STAT %CPU %MEM   TIME CPU COMMAND

   7653 ccao      25   0  115G 586M  114G R    99.9  0.9   0:08   3 mocassin

  7656 ccao      25   0  115G 586M  114G R    99.9  0.9   0:08   6 mocassin

   7654 ccao      25   0  115G 586M  114G R    99.8  0.9   0:08   4 mocassin

   7655 ccao      25   0  115G 586M  114G R    99.8  0.9   0:08   5 mocassin

   7658 ccao      25   0  115G 586M  114G R    99.8  0.9   0:08   8 mocassin

   7659 ccao      25   0  115G 586M  114G R    99.8  0.9   0:08   9 mocassin

   7660 ccao      25   0  115G 586M  114G R    99.8  0.9   0:08  10 mocassin

   7662 ccao      25   0  115G 586M  114G R    99.7  0.9   0:08  12 mocassin

   7657 ccao      25   0  115G 586M  114G R    88.5  0.9   0:07   7 mocassin

   7661 ccao      25   0  115G 586M  114G R    88.3  0.9   0:07  11 mocassin

   7649 ccao      25   0  115G 586M  114G R    55.2  0.9   0:04   2 mocassin

   7651 ccao      25   0  115G 586M  114G R    54.1  0.9   0:03   1 mocassin

   7650 ccao      25   0  115G 586M  114G R    50.0  0.9   0:04   0 mocassin

   7647 ccao      25   0  115G 586M  114G R    49.8  0.9   0:03   0 mocassin

   7652 ccao      25   0  115G 586M  114G R    44.7  0.9   0:04   2 mocassin

   7648 ccao      25   0  115G 586M  114G R    35.9  0.9   0:03   1 mocassin
```

Even if an application starts some threads executing for a short time, the threads still have taken a token in the CPU list. Then, when the compute threads are finally started, the list is exhausted and restarts from the beginning. Consequently, some threads share a CPU. To bypass this condition, work to eliminate the ghost thread creation, as follows:

- Check for a call to the system function. A call to the system function is often responsible for placement failure due to unexpected thread creation. If all the compute processes have the same name, issue a command such as the following:

  ```
  % dplace -c0-15 -n compute-process-name ...
  ```

- You can run `dplace -e -c0-32` on 16 CPUs to understand the pattern of the thread creation. If this pattern is the same from one run to the other (unfortunately race between thread creation often occurs), you can find the right option to `dplace`. For example, assume the following:

  - You want to run on CPUs 0-3, with `dplace -e -C0-16`.
  - You see that threads are always placed on CPU 0, 1, 5, and 6.

  In this situation, use one of the following commands to place your threads correctly:

  ```
  dplace -e -c0,1,x,x,x,2,3
  ```

  or

  ```
  dplace -x24 -c0-3  # x24 =11000, place the 2 first and skip 3 before placing
  ```

`omplace` command

The `omplace` command controls the placement of MPI processes and OpenMP threads. This command is a wrapper script for `dplace`. Use `omplace`, rather than `dplace`, if your application uses MPI, OpenMP, pthreads, or hybrid MPI/OpenMP and MPI/pthreads codes. The `omplace` command generates the proper `dplace` placement file syntax automatically. It also supports some unique options, such as block-strided CPU lists.

The `omplace` command causes the successive threads in a hybrid MPI/OpenMP job to be placed on unique CPUs. The CPUs are assigned in order from the effective CPU list within the containing cpuset. The CPU placement is performed by dynamically generating a placement file and invoking `dplace` with the MPI job launch.

For example, to run two MPI processes with four threads per process, and to display the generated placement file, enter a command similar to the following:

```
# mpirun -np 2 omplace -nt 4 -vv ./a.out
```

The preceding command places the threads as follows:

```
rank 0 thread 0 on CPU 0
rank 0 thread 1 on CPU 1
rank 0 thread 2 on CPU 2
rank 0 thread 3 on CPU 3
rank 1 thread 0 on CPU 4
rank 1 thread 1 on CPU 5
rank 1 thread 2 on CPU 6
rank 1 thread 3 on CPU 7
```

For more information, see the `omplace`(1) manpage.

Flexible File I/O (FFIO)

Flexible File I/O (FFIO) can improve the file I/O performance of existing applications without having to resort to source code changes. The current executable remains unchanged. Knowledge of source code is not required. However, knowledge of how the source and the application software work can help you better interpret and optimize FFIO results. To take advantage of FFIO, simply set some environment variables before running your application.

The FFIO subsystem allows you to define one or more additional I/O buffer caches for specific files to augment the Linux kernel I/O buffer cache. The FFIO subsystem manages this buffer cache for you. To manage buffer cache, FFIO intercepts standard I/O calls such as open, read, and write, and replaces them with FFIO equivalent routines. These routines route I/O requests through the FFIO subsystem, which uses the user-defined FFIO buffer cache.

FFIO can bypass the Linux kernel I/O buffer cache by communicating with the disk subsystem through direct I/O. This bypass gives you precise control over cache I/O characteristics and allows for more efficient I/O requests. For example, doing direct I/O in large chunks (for example, 16 megabytes) allows the FFIO cache to amortize disk access. When FFIO is used with direct I/O enabled, all file buffering occurs in user space, which differs from the Linux buffer cache mechanism. To buffer data in kernel memory, the Linux buffer cache mechanism requires a context switch. Avoiding this kind of overhead helps FFIO to scale efficiently.

Another important distinction is that FFIO allows you to create an I/O buffer cache dedicated to a specific application. In contrast, the Linux kernel manages all the jobs on the entire system with a single I/O buffer cache. As a result, FFIO typically outperforms the Linux kernel buffer cache when it comes to I/O intensive throughput.

The following topics explain how to use FFIO:

- **Environment variables** on page 81
- **FFIO examples** on page 82
- **Multithreading considerations** on page 84
- **Application examples** on page 85
- **Event tracing** on page 85
- **System information** on page 85

Environment variables

To use FFIO, set one of the following environment variables: `LD_PRELOAD` or `FF_IO_OPTS`.

To enable FFIO to trap standard I/O calls, set the `LD_PRELOAD` environment variable, as follows:

```
# export LD_PRELOAD="/usr/lib64/libFFIO.so"
```

The `LD_PRELOAD` software is a Linux feature that instructs the linker to preload the indicated shared libraries. In this case, `libFFIO.so` is preloaded and provides the routines that replace the standard I/O calls. An application that is not dynamically linked with the `glibc` library cannot work with FFIO because the standard I/O calls cannot be intercepted. To disable FFIO, enter the following:

```
# unset LD_PRELOAD
```

The FFIO buffer cache is managed by the `FF_IO_OPTS` environment variable. The syntax for setting this variable can be complex. A simple format for defining this variable is as follows:

```
export FF_IO_OPTS  'string(eie.direct.mbytes:size:num:lead:share:stride:0)'
```

The following table shows the arguments that you can use with the `FF_IO_OPTS` environment variable.

Argument	Effect
string	Matches the names of files that can use the buffer cache.
size	Number of 4k blocks in each page of the I/O buffer cache.
num	Number of pages in the I/O buffer cache.
lead	The maximum number of read-ahead pages.
share	A value of 1 means a shared cache, 0 means private.
stride	The `stride` parameter is always 0.

Example 1. Assume that you want a shared buffer cache of 128 pages. Each page is to be 16 megabytes (that is, 4096*4k). The cache has a lead of six pages and uses a stride of one. The command is as follows:

```
% setenv FF_IO_OPTS 'test*(eie.direct.mbytes:4096:128:6:1:1:0)'
```

Each time the application opens a file, the FFIO code checks the file name to see if it matches the string supplied by `FF_IO_OPTS`. The full pathname to the file is not considered when checking for a match against the string. For example, file names of `/tmp/test16` and `/var/tmp/testit` both match.

Example 2. This more complicated usage of `FF_IO_OPTS` builds upon the previous example. Multiple file name types can share the cache, as the following example shows:

```
% setenv FF_IO_OPTS 'output* test*(eie.direct.mbytes:4096:128:6:1:1:0)'
```

Example 3. You can specify multiple caches with `FF_IO_OPTS`. In the example that follows, files of the form `output*` and `test*` share a 128 page cache of 16 megabyte pages. The file `special42` has a 256-page private cache of 32 megabyte pages. The command, which uses the backslash (`\`) continuation character, is as follows:

```
% setenv FF_IO_OPTS 'output* test*(eie.direct.mbytes:4096:128:6:1:1:0) \
special42(eie.direct.mbytes:8192:256:6:0:1:0)'
```

You can add parameters to `FF_IO_OPTS` to create feedback that is sent to standard output. For examples of this diagnostic output, see the following:

FFIO examples on page 82

FFIO examples

This topic includes some simple FFIO examples. Assume that `LD_PRELOAD` is set for the correct library, and `FF_IO_OPTS` is defined as follows:

```
% setenv FF_IO_OPTS 'test*(eie.direct.mbytes:4096:128:6:1:1:0)'
```

It can be difficult to tell what FFIO might or might not be doing even with a simple program. The examples in this topic use a small C program called `fio` that reads 4-megabyte chunks from a file for 100 iterations. When the program runs, it produces the following output:

```
% ./fio -n 100 /build/testit
Reading 4194304 bytes 100 times to /build/testit
Total time  = 7.383761
Throughput  = 56.804439 MB/sec
```

Example 1. You can direct a simple FFIO operations summary to standard output by making the following simple addition to `FF_IO_OPTS`:

```
% setenv FF_IO_OPTS 'test*(eie.direct.mbytes:4096:128:6:1:1:0, \
event.summary.mbytes.notrace )'
```

This new setting for `FF_IO_OPTS` generates the following summary on standard output when the program runs:

```
% ./fio -n 100 /build/testit
Reading 4194304 bytes 100 times to /build/testit
Total time  = 7.383761
Throughput  = 56.804439 MB/sec

event_close(testit)  eie <-->syscall (496 mbytes)/( 8.72 s)=  56.85 mbytes/
s
oflags=0x0000000000004042=RDWR+CREAT+DIRECT
sector size =4096(bytes)
cblks =0  cbits =0x0000000000000000
current file size =512 mbytes   high water file size =512 mbytes

function  times    wall   all     mbytes     mbytes      min      max     avg
          called   time  hidden  requested  delivered  request  request  request
 open          1   0.00
 read          2   0.61              32          32        16       16       16
 reada        29   0.01     0       464         464        16       16       16
 fcntl
    recall
    reada     29   8.11
    other      5   0.00
 flush         1   0.00
 close         1   0.00
```

Two synchronous reads of 16 megabytes each were issued, for a total of 32 megabytes. In addition, there were 29 asynchronous reads (`reada`) issued, for a total of 464 megabytes.

Example 2. You can generate additional diagnostic information by specifying the `.diag` modifier. The following is an example of the diagnostic output generated when the `.diag` modifier is used:

```
% setenv FF_IO_OPTS 'test*(eie.direct.diag.mbytes:4096:128:6:1:1:0 )'
% ./fio -n 100 /build/testit
Reading 4194304 bytes 100 times to /build/testit
Total time  = 7.383761
Throughput  = 56.804439 MB/sec

eie_close EIE final stats for file /build/testit
eie_close Used shared eie cache 1
eie_close 128 mem pages of 4096 blocks (4096 sectors), max_lead = 6 pages
eie_close advance reads used/started :   23/29    79.31%   (1.78 seconds wasted)
eie_close write hits/total           :    0/0      0.00%
eie_close read  hits/total           :   98/100   98.00%
eie_close mbytes transferred  parent --> eie --> child      sync        async
eie_close                            0           0        0            0
eie_close                          400         496        2           29 (0,0)
eie_close                     parent <-- eie <-- child

eie_close EIE stats for Shared cache 1
eie_close 128 mem pages of 4096 blocks
eie_close advance reads used/started :   23/29   79.31%   (0.00 seconds wasted)
eie_close write hits/total           :    0/0     0.00%
eie_close read  hits/total           :   98/100  98.00%
eie_close mbytes transferred  parent --> eie --> child    sync      async
eie_close                               0                 0           0
eie_close                             400         496     2          29 (0,0)
```

The preceding output lists information for both the file and the cache. In the `mbytes transferred` information, the lines in **bold** are for write and read operations, respectively. It is only for simple I/O patterns that the difference between `parent --> eie` and `eie --> child` read statistics can be explained by the number of read-ahead operations. For random reads of a large file over a long time period, this explanation is not accurate. All write operations count as `async`.

You can generate additional diagnostic information by specifying the `.diag` modifier and the `.event.summary` modifier. The two modifiers operate independently from one another. The following specification uses both modifiers:

```
% setenv FF_IO_OPTS 'test*(eie.diag.direct.mbytes:4096:128:6:1:1:0, \
event.summary.mbytes.notrace )'
```

Multithreading considerations

FFIO works with applications that use MPI for parallel processing. An MPI job assigns each thread a number or rank. The master thread has rank 0. The remaining slave threads have ranks from 1 to *N*-l where *N* is the total number of threads in the MPI job. Remember that the threads comprising an MPI job do not necessarily have access to the address spaces of other threads. As a result, there is no way for the different MPI threads to share an FFIO cache. By default, each thread defines a separate FFIO cache based on the parameters defined by `FF_IO_OPTS`.

Having each MPI thread define a separate FFIO cache, based on a single environment variable (`FF_IO_OPTS`), can waste much memory. Fortunately, FFIO provides a mechanism that allows you to specify a different FFIO cache for each MPI thread through the following environment variables:

```
setenv FF_IO_OPTS_RANK0 'result*(eie.direct.mbytes:4096:512:6:1:1:0)'
setenv FF_IO_OPTS_RANK1 'output*(eie.direct.mbytes:1024:128:6:1:1:0)'
setenv FF_IO_OPTS_RANK2 'input*(eie.direct.mbytes:2048:64:6:1:1:0)'
             .
             .
             .
setenv FF_IO_OPTS_RANKN-1 ...   (N = number of threads).
```

Each rank environment variable is set using the exact same syntax as `FF_IO_OPTS`, and each defines a distinct cache for the corresponding MPI rank. If the cache is designated as shared, all files within the same ranking thread can use the same cache. FFIO works with MPI from HPE. To work with MPI applications, FFIO determines the rank of callers by invoking the `mpi_comm_rank_()` MPI library routine. Therefore, set the following environment variable to enable FFIO to determine the location of the MPI library used by the application:

```
setenv SGI_MPI /opt/hpe/hpc/mpt/mpt 2.19/lib
```

To use the rank functionality, set both the `SGI_MPI` and `FF_IO_OPTS_RANK0` environment variables. If either variable is not set, then the MPI threads all use `FF_IO_OPTS`. If both the MPI and the `FF_IO_OPTS_RANK0` variables are defined but, for example, `FF_IO_OPTS_RANK2` is undefined, all rank 2 files generate a `no match` with FFIO. This analysis reveals that none of the rank 2 files are cached by FFIO. In this case, the software does not default to `FF_IO_OPTS`.

Fortran and C/C++ applications that use the `pthreads` interface create threads that share an address space. These threads can all use the single FFIO cache defined by `FF_IO_OPTS`.

Application examples

FFIO has been deployed successfully with several high-performance computing applications, such as Nastran and Abaqus. In one benchmark, I/O performance improved when the FFIO cache size was set to roughly 10-15% of the disk space that Abaqus required. For this benchmark, the `FF_IO_OPTS` environment variable was defined as follows:

```
% setenv FF_IO_OPTS '*.fct *.opr* *.ord *.fil *.mdl* *.stt* *.res *.sst \
*.hdx *.odb* *.023 *.nck* *.sct *.lop *.ngr *.elm *.ptn* *.stp* *.eig \
*.lnz* *.mass *.inp* *.scn* *.ddm *.dat* \
fort*(eie.direct.nodiag.mbytes:4096:512:6:1:1:0,event.summary.mbytes.notrace)'
```

For the MPI version of Abaqus, different caches were specified for each MPI rank, as follows:

```
% setenv FF_IO_OPTS_RANK0 '*.fct *.opr* *.ord *.fil *.mdl* *.stt* *.res \
*.sst *.hdx *.odb* *.023 *.nck* *.sct *.lop *.ngr *.ptn* *.stp* *.elm \
*.eig *.lnz* *.mass *.inp *.scn* *.ddm *.dat* \
fort*(eie.direct.nodiag.mbytes:4096:512:6:1:1:0,event.summary.mbytes.notrace)'

% setenv FF_IO_OPTS_RANK1 '*.fct *.opr* *.ord *.fil *.mdl* *.stt* *.res \
*.sst *.hdx *.odb* *.023 *.nck* *.sct *.lop *.ngr *.ptn* *.stp* *.elm \
*.eig *.lnz* *.mass *.inp *.scn* *.ddm *.dat* \
fort*(eie.direct.nodiag.mbytes:4096:16:6:1:1:0,event.summary.mbytes.notrace)'

% setenv FF_IO_OPTS_RANK2 '*.fct *.opr* *.ord *.fil *.mdl* *.stt* *.res \
*.sst *.hdx *.odb* *.023 *.nck* *.sct *.lop *.ngr *.ptn* *.stp* *.elm \
*.eig *.lnz* *.mass *.inp *.scn* *.ddm *.dat* \
fort*(eie.direct.nodiag.mbytes:4096:16:6:1:1:0,event.summary.mbytes.notrace)'

% setenv FF_IO_OPTS_RANK3 '*.fct *.opr* *.ord *.fil *.mdl* *.stt* *.res \
*.sst *.hdx *.odb* *.023 *.nck* *.sct *.lop *.ngr *.ptn* *.stp* *.elm \
*.eig *.lnz* *.mass *.inp *.scn* *.ddm *.dat* \
fort*(eie.direct.nodiag.mbytes:4096:16:6:1:1:0,event.summary.mbytes.notrace)'
```

Event tracing

If you specify the `.trace` option as part of the `event` parameter, you can enable the event tracing feature in FFIO.

For example:

```
% setenv FF_IO_OPTS 'test*(eie.direct.mbytes:4096:128:6:1:1:0, \
event.summary.mbytes.trace )'
```

This option generates files of the form `ffio.events.`*`pid`* for each process that is part of the application. By default, event files are placed in `/tmp`. To change this destination, set the `FFIO_TMPDIR` environment variable. These files contain timestamped events for files that use the FFIO cache. You can use the information in these files to trace I/O activity such as I/O sizes and offsets.

System information

The FFIO subsystem supports applications written in C, C++, and Fortran. C and C++ applications can be built with either the Intel or gcc compiler. Only Fortran codes built with the Intel compiler work with FFIO.

The following restrictions on FFIO must also be observed:

- The FFIO implementation of `pread`/`pwrite` is not correct. The file offset advances.
- Do not use FFIO for I/O on a socket.

- Do not link your application with the `librt` asynchronous I/O library.
- FFIO does not intercept calls that operate on files in `/proc`, `/etc`, and `/dev`.
- FFIO does not intercept calls that operate on `stdin`, `stdout`, and `stderr`.
- FFIO is not intended for generic I/O applications such as the following:
  - `vi`
  - `cp`
  - `mv`

Websites

General websites

Hewlett Packard Enterprise Information Library

www.hpe.com/info/EIL

Single Point of Connectivity Knowledge (SPOCK) Storage compatibility matrix

www.hpe.com/storage/spock

Storage white papers and analyst reports

www.hpe.com/storage/whitepapers

For additional websites, see **Support and other resources**.

MPI websites

Message Passing Interface Forum

http://www.mpi-forum.org

Support and other resources

Accessing Hewlett Packard Enterprise Support

- For live assistance, go to the Contact Hewlett Packard Enterprise Worldwide website:

  http://www.hpe.com/assistance

- To access documentation and support services, go to the Hewlett Packard Enterprise Support Center website:

  http://www.hpe.com/support/hpesc

Information to collect

- Technical support registration number (if applicable)
- Product name, model or version, and serial number
- Operating system name and version
- Firmware version
- Error messages
- Product-specific reports and logs
- Add-on products or components
- Third-party products or components

Accessing updates

- Some software products provide a mechanism for accessing software updates through the product interface. Review your product documentation to identify the recommended software update method.
- To download product updates:

  Hewlett Packard Enterprise Support Center

  www.hpe.com/support/hpesc

  Hewlett Packard Enterprise Support Center: Software downloads

  www.hpe.com/support/downloads

  Software Depot

  www.hpe.com/support/softwaredepot

- To subscribe to eNewsletters and alerts:

  www.hpe.com/support/e-updates

- To view and update your entitlements, and to link your contracts and warranties with your profile, go to the Hewlett Packard Enterprise Support Center **More Information on Access to Support Materials** page:

  www.hpe.com/support/AccessToSupportMaterials

IMPORTANT: Access to some updates might require product entitlement when accessed through the Hewlett Packard Enterprise Support Center. You must have an HPE Passport set up with relevant entitlements.

Customer self repair

Hewlett Packard Enterprise customer self repair (CSR) programs allow you to repair your product. If a CSR part needs to be replaced, it will be shipped directly to you so that you can install it at your convenience. Some parts do not qualify for CSR. Your Hewlett Packard Enterprise authorized service provider will determine whether a repair can be accomplished by CSR.

For more information about CSR, contact your local service provider or go to the CSR website:

http://www.hpe.com/support/selfrepair

Remote support

Remote support is available with supported devices as part of your warranty or contractual support agreement. It provides intelligent event diagnosis, and automatic, secure submission of hardware event notifications to Hewlett Packard Enterprise, which will initiate a fast and accurate resolution based on your product's service level. Hewlett Packard Enterprise strongly recommends that you register your device for remote support.

If your product includes additional remote support details, use search to locate that information.

Remote support and Proactive Care information

HPE Get Connected

www.hpe.com/services/getconnected

HPE Proactive Care services

www.hpe.com/services/proactivecare

HPE Proactive Care service: Supported products list

www.hpe.com/services/proactivecaresupportedproducts

HPE Proactive Care advanced service: Supported products list

www.hpe.com/services/proactivecareadvancedsupportedproducts

Proactive Care customer information

Proactive Care central

www.hpe.com/services/proactivecarecentral

Proactive Care service activation

www.hpe.com/services/proactivecarecentralgetstarted

Warranty information

To view the warranty for your product or to view the *Safety and Compliance Information for Server, Storage, Power, Networking, and Rack Products* reference document, go to the Enterprise Safety and Compliance website:

www.hpe.com/support/Safety-Compliance-EnterpriseProducts

Additional warranty information

HPE ProLiant and x86 Servers and Options

www.hpe.com/support/ProLiantServers-Warranties

HPE Enterprise Servers

www.hpe.com/support/EnterpriseServers-Warranties

HPE Storage Products

www.hpe.com/support/Storage-Warranties

HPE Networking Products

www.hpe.com/support/Networking-Warranties

Regulatory information

To view the regulatory information for your product, view the *Safety and Compliance Information for Server, Storage, Power, Networking, and Rack Products*, available at the Hewlett Packard Enterprise Support Center:

www.hpe.com/support/Safety-Compliance-EnterpriseProducts

Additional regulatory information

Hewlett Packard Enterprise is committed to providing our customers with information about the chemical substances in our products as needed to comply with legal requirements such as REACH (Regulation EC No 1907/2006 of the European Parliament and the Council). A chemical information report for this product can be found at:

www.hpe.com/info/reach

For Hewlett Packard Enterprise product environmental and safety information and compliance data, including RoHS and REACH, see:

www.hpe.com/info/ecodata

For Hewlett Packard Enterprise environmental information, including company programs, product recycling, and energy efficiency, see:

www.hpe.com/info/environment

Documentation feedback

Hewlett Packard Enterprise is committed to providing documentation that meets your needs. To help us improve the documentation, send any errors, suggestions, or comments to Documentation Feedback (**docsfeedback@hpe.com**). When submitting your feedback, include the document title, part number, edition, and publication date located on the front cover of the document. For online help content, include the product name, product version, help edition, and publication date located on the legal notices page.

Using the Message Passing Toolkit (MPT) on a virtual machine

You can configure a virtual machine (VM) on the following systems:

- On an SGI UV system
- On an HPE MC990 X system

The VM creates a general-purpose computer, and MPT can run on that computer. When you use MPT from within a VM, however, you expect the following:

- Differences in the computing environment
- Differences with regard to application behavior

NOTE: Support for MPT on VMs on HPE Superdome Flex Server platforms is deferred.

For administrator information about how to configure the VM environment on your hardware, see the following:

- **Installing software within the virtual machine (VM)** on page 91
- **Adjusting virtual machine system settings** on page 92

For user information about how a program might behave differently when running from within a VM, see the following:

Running MPI from HPE programs from within a virtual machine on page 93

Installing software within the virtual machine (VM)

The following procedure explains the software to install in the VM in order for MPI programs to run on the VM.

Procedure

1. Install and configure the operating system (RHEL or SLES) and the HPE System Foundation Software on the SGI UV system or on the HPE Integrity MC990 X Server.

   For installation information, see the software installation guide that pertains to your platform.
2. Install and configure the VM according to your operating system vendor instructions.

   NOTE: RHEL and SLES do not support InfiniBand technology from within a VM. Other OFED providers support InfiniBand technology from within a VM through single-root I/O virtualization (SR-IOV), but HPE does not support SR-IOV or other alternatives to the distribution-supplied OFED.
3. (Optional) Install the HPE System Foundation Software into the VM.

   For installation information, see the installation guide for your platform.
4. Install the HPE Performance Software into the VM.

   For installation information, see the HPE Performance Software release notes.
5. Install MPT into the VM.

For installation information, see **Getting started with MPI from HPE** on page 17.

Adjusting virtual machine system settings

For best performance, HPE recommends that you change certain operating system settings after the software installation is complete.

The following procedure explains how to adjust the number of files that can be open at a given time.

Procedure

1. Log into the system as the administrator user.
2. Enter the `cpumap` command to retrieve the number of cores.

   For example:

   ```
   # cpumap
   This is an SGI UV
   model name           : Genuine Intel(R) CPU @ 2.60GHz
   Architecture         : x86_64
   cpu MHz              : 2600.072
   cache size           : 20480 KB (Last Level)

   Total Number of Sockets                : 16
   Total Number of Cores                  : 128   (8 per socket)
   Hyperthreading                         : ON
   Total Number of Physical Processors    : 128
   Total Number of Logical Processors     : 256   (2 per Phys Processor)

   UV Information
    HUB Version:                           UVHub  3.0
    Number of Hubs:                        16
    Number of connected Hubs:              16
    Number of connected NUMAlink ports:    128
   =============================================================================
   . . .
   ```

   The `Total Number of Cores` line reveals that there are 128 cores, 8 per socket.

3. Display the contents of the `/etc/sysctl-conf` file.

   For example, enter the following command:

   ```
   # less /etc/sysctl.conf
   ...
   fs.file-max = 8204481
   ...
   ```

4. (Conditional) Use a text editor to open file `sysctl.conf` and increase the value of the `fs.file-max` parameter in the `/etc/sysctl.conf` file.

   Complete this step if the following are both true:

   - The number of cores on your system is greater than 512.

     and

   - The `fs.file-max` parameter is set to less than 10,000,000.

For optimum performance within a VM, set the `fs.file-max` parameter to be at least `10000000` on systems with 512 cores or more.

Running MPI from HPE programs from within a virtual machine

MPI and SHMEM programs behave differently when run from within a virtual machine (VM). The following is a list of differences:

- Hardware-dependent features might not exist on a VM.

  When you run an MPI program on a VM, the environment detects the virtual nature of the platform and ignores any hardware-specific features. The following hardware features are not available to an application that runs in a VM: NUMAlink, Superpages, the SGI UV timer, the HUB ASIC, and hardware performance counters. In addition, processor-specific performance diagnostics are limited.

  Some hardware technologies are not specific to particular hardware platforms. You can expect the VM to honor nonspecific technologies.

- Topology characteristics might be different.

  Some applications rely on specific hardware platform topologies. Run such applications on a VM that was configured for and mimics that specific topology. MPI programs do not automatically use special topology characteristics effectively. If the application requires special heuristics for locality and placement, configure that into the VM.

- XPMEM libraries are beneficial in large VMs.

  HPE has tested XPMEM on VMs. XPMEM loads, and your application can call XPMEM routines successfully. However, XPMEM is useful only on systems with large amounts of memory.

- No InfiniBand support.

  The RHEL and SLES operating systems do not support InfiniBand technology in VMs. To find out whether single-root I/O virtualization (SR-IOV) is configured on the VM, consult your system administrator.

Array Services system administration information

The following topics explain how to install Array Services manually and how to configure Array Services to suit your site needs:

Manually configuring Array Services on multiple hosts

You can configure Array Services in an automated way or manually. The following list shows where you can find the standard, automated procedures:

- For cluster systems, use the information in the following:
  - **HPE Performance Cluster Manager Installation Guide**
  - **HPE SGI Management Suite Installation and Configuration Guide**
  - **Installing the Array Services software on clusters that use the HPE Insight Cluster Management Utility (CMU)** on page 9
- For HPE Superdome Flex Grid systems, use the information in the following:

  Configuring the Message Passing Toolkit (MPT) on page 9

When you configure Array Services in a manual way, you can customize the software at installation time.

The following procedure explains how to configure Array Services to run on multiple hosts.

Procedure

1. Log in as the administrator user to one of the hosts you want to include in the array.

   You must be logged in as an administrator to perform this procedure.

   For example, on an HPW SGI 8600 system, log into one of the service nodes. You can include service nodes and compute nodes in the array.
2. (Optional) Install the MUNGE package from the MPI for HPE software distribution.

   The optional MUNGE software package enables additional security for Array Services operations.

   During MUNGE installation, make sure of the following:

   - The MUNGE key that is used is the same across all the nodes in the array.

     The MUNGE key resides in `/etc/munge/munge.key`.
   - You configure a good time clock source, such as an NTP server. MUNGE depends on time synchronization across all nodes in the array.

To install MUNGE, use one of the following commands:

- On Red Hat Enterprise Linux platforms: `yum install munge`
- On SUSE Linux Enterprise Server platforms: `zypper install munge`

For more information about how to install MUNGE, see the MPI for HPE release notes.

3. Open file `/etc/array/arrayd.conf` with a text editor.
4. Edit the `/etc/array/arrayd.conf` file to list the machines in the array.

   This file enables you to configure many characteristics of an Array Services environment. The required specifications are as follows:

   - The array name.
   - The hostnames of the array participants.
   - A default destination array.

   For more information about the additional characteristics that you can specify in the `arrayd.conf` file, see the `arrayd.conf`(4) manpage.

   For an example `arrayd.conf` file, see file `/usr/lib/array/arrayd.conf.template`.

   Example 1. The following lines specify an array name (`sgicluster`) and two hostnames. Specify each hostname on its own line. `array` and `machine` are keywords in the file.

   ```
   array sgicluster
                machine host1
                machine host2
   ```

   Example 2. The following line sets a default array name.

   ```
   destination array sgicluster
   ```

5. Save and close file `/etc/array/arrayd.conf`.
6. Use a text editor to open file `/etc/array/arrayd.auth`.
7. (Optional) Change the authentication method from the default of `NOREMOTE` to a method of your choosing.

   By default, the Array Services software does not allow remote access to the array. You can change this authentication method. For information about the various authentication methods, see the following:

   Changing the security access level in the AUTHENTICATION parameter on page 96

   To change the access method, complete the following steps:

   - Search for the string `AUTHENTICATION NOREMOTE`, and insert a `#` character in column 1 to comment out the line.
   - Enable the security level under which you want Array Services to operate.

     This step specifies the authentication mechanism to use when Array Services messages pass between the Array Services daemons. Possible security levels are `NONE`, `SIMPLE`, or `MUNGE`, as follows:

    - If no authentication is required, remove the `#` character from column 1 of the `AUTHENTICATION NONE` line.
    - To enable simple authentication, ensure that there is no `#` in column 1 of the `AUTHENTICATION SIMPLE` line. Default.
    - To enable authentication through MUNGE, remove the `#` character from column 1 of the `AUTHENTICATION MUNGE` line.

      Make sure that MUNGE has been installed, as prescribed earlier in this procedure.

  - Save and close file `/etc/array/arrayd.auth`.

8. (Optional) Reset the default user account or the default array port.

   The Array Services installation and configuration process sets the following defaults in the `/etc/array/arrayd.conf` configuration file:

   - A default user account of `arraysvcs`.

     Certain array services commands require the existence of a user account on all hosts in the array. If you create a different account, update the `arrayd.conf` file, and set the user account permissions correctly on all hosts.
   - A default port number of 5434.

     The `/etc/services` file contains a line that defines the `arrayd` service and port number as follows:

     ```
     sgi-arrayd   5434/tcp    # Array Services daemon
     ```

     You can set any value for the port number, but all systems mentioned in the `arrayd.conf` file must use the same value.

9. Enter one of the following commands to restart Array Services:

   - On RHEL 7.*X*, SLES 15, or SLES 12 SP*X* systems, enter the following command:

     ```
     systemctl restart array
     ```

   - On RHEL 6.*X* or SLES 11 SP*X* systems, enter the following command:

     ```
     /etc/init.d/array restart
     ```

10. Repeat the preceding steps on the other hosts or copy the `/etc/array/arrayd.conf` and `/etc/array/arrayd.auth` files to the other hosts.

    The Array Services feature requires that the configuration files on each participant host include the list of host participants and the authentication method. The files can contain additional, host-specific information.

Changing the security access level in the `AUTHENTICATION` parameter

The `AUTHENTICATION` parameter in the `/etc/array/arrayd.auth` file specifies access to the array. The `AUTHENTICATION` parameter can have one of the following settings:

- `NOREMOTE` (default).

When set to NOREMOTE, the arrayd daemon allows only local access to the array. That is, the arrayd daemon does not allow remote requests to access the array.

Use NOREMOTE if the array is attached to a public network or if individual machines cannot be trusted.

- NONE.

  When set to NONE, the arrayd daemon assumes that remote users identify themselves accurately and honestly when making requests. In other words, if a request claims to be coming from user abc, the arrayd daemon assumes that it is in fact from user abc and not somebody spoofing abc.

  All requests from remote systems are authenticated. The authentication mechanism involves private keys that are known only to the superusers on the local and remote systems. Requests originating on systems that do not have these private keys are rejected. For more information, see the section on authentication information in the arrayd.conf(4) manpage.

  This setting can be adequate for systems that are behind a network firewall or otherwise protected from hostile attack. In this situation, all the users inside the firewall are presumed to be nonhostile.

  Do not set AUTHENTICATION to NONE when either of the following conditions exist:

  - The array is attached to a public network
  - Individual machines cannot be trusted

- SIMPLE. Generates hostname/key pairs by using the OpenSSL rand command, 64-bit values (if available), or by using $RANDOM Bash facilities. For more information, see the arrayd.auth(5) manpage.

- MUNGE.

  When set to MUNGE, uses the MUNGE credential encoder. For more information, see the munge(1) manpage.

The Array Services daemon, arrayd, runs as the administrator user and does not support mapping of user, group, or project names between two different namespaces. All members of an array are assumed to share the namespace for users, groups, and projects. Thus, if systems A and B are members of the same array, username abc on system A is assumed to be the same user as username abc on system B. These naming conventions are most significant for the administrator username. Use authentication to prevent access to an array by machines using a different namespace.

Configuring nodes into arrays

The following topics contain examples that show how to specify the nodes in an array in the arrayd.conf file:

- **Specifying an array name and machine names** on page 97
- **Specifying IP addresses and ports** on page 98
- **Specifying additional attributes** on page 98

Specifying an array name and machine names

Often, the hostname of each node is the same as the node name to the site domain name services (DNS). The following example defines an array for which the names are the same:

```
array simple
        machine congo
        machine niger
        machine nile
```

To access this array, specify the array name, `simple`, as the argument to the `-a` option on the following commands:

- The `array` command
- The `ainfo` command

Specify one array name in a `DESTINATION ARRAY` local option as the default array. `ainfo dflt` reports the name you specify.

For information about local options, see the following:

Configuring local options on page 104

Specifying IP addresses and ports

At your site, it is possible that a machine IP address cannot be obtained from the given hostname. In this situation, edit the `arrayd.conf` file to provide a `hostname` subentry to specify either a fully qualified domain name (FQDN) or an IP address. For example:

```
array simple
        machine congo
            hostname congo.engr.hitech.com
            port 8820
        machine niger
            hostname niger.engr.hitech.com
        machine nile
            hostname "198.206.32.85"
```

The preceding example uses the `port` subentry to specify that `arrayd` in a particular machine use a different socket number than the default of 5434.

Specifying additional attributes

If you want the `ainfo` command to display certain strings, you can insert these values as subentries to the `array` entry. The following are some examples:

```
array simple
       array_attribute config_date="04/03/96"
       machine a_node
       machine_attribute aka="congo"
       hostname congo.engr.hitech.com
```

TIP: You can write code that fetches any array name, machine name, or attribute string from any node in the array.

About the Array Services configuration files

The Array Services configuration files are as follows:

- `/etc/array/arrayd.conf`
- `/etc/array/arrayd.auth`
- `/etc/sysconfig/array`

The configuration files contain array information, node information, authentication key information, and valid commands. The Array Services daemon reads each configuration file when it starts. Typically, the daemon starts on each node at boot time and then runs as a background process. The Array Services commands call the daemon process on each node to obtain information. You can also run the daemon from a command line. For example, you might want to run the daemon from a command line to check the syntax of a configuration file.

The following topics contain more configuration file information:

- **About configuration file formats and contents** on page 99
- **About loading configuration data** on page 99

About configuration file formats and contents

A configuration file is a readable text file that contains the following types of entries:

- Array definition information. This information describes this array and other known arrays and includes array names, node names, and node types.
- Command definitions. These definitions specify the usage and operation of commands that can be invoked through the `array` command.
- Authentication information, which specifies the authentication key numbers used to access the array. Not all arrays use authentication keys.
- Local options, which are options that modify the operation of the other entries or `arrayd`.

Within the configuration files, you can use blank lines, white space, and comment lines that begin with a pound character (`#`) for readability. Entries can be in any order in any of the Array Services configuration files.

Besides punctuation, entries have keyword-based syntax. Keyword recognition is not case-sensitive, but keywords appear in uppercase in this documentation and in the manpage. As the `arrayd.conf`(4) manpage describes, the entries are formed from keywords, numbers, and quoted strings.

About loading configuration data

When run as a command, the Array Services daemon, `arrayd`, accepts one or more file names as arguments. It reads them all and treats them like logical continuations. In effect, it concatenates them. If you do not specify any file names, it reads the following configuration files:

- `/etc/array/arrayd.conf`
- `/etc/array/arrayd.auth`
- `/etc/sysconfig/array`

  This file can contain a list of files and `arrayd` command-line options. The startup script that launches `arrayd` at boot time reads this file.

Because configuration data can reside in two or more files, you can combine different strategies. For example:

- One file can have different access permissions than another. Typically, `/etc/array/arrayd.conf` is world-readable and contains the available `array` commands. In contrast, the `/etc/array/arrayd.auth` file is readable only by the administrator user and contains authentication codes.
- One node can have different configuration data than another. For example, certain commands might be defined only on certain nodes, or only the nodes used for interactive logins might know the names of all other nodes.
- You can use NFS-mounted configuration files. You could put a small configuration file on each machine to define the array and authentication keys. Alternatively, you could have a larger file that is NFS-mounted from one node and defines `array` commands.

After you modify the configuration files, you can make `arrayd` reload them by killing and restarting the daemon on each machine, as follows:

- To terminate the daemon, use one of the following commands:
  - On RHEL 7, SLES 15, or SLES 12 systems, enter the following:
    ```
    systemctl stop array
    ```
  - On RHEL 6 or SLES 11 systems, enter the following:
    ```
    /etc/init.d/array stop
    ```
- To terminate and restart the daemon in one operation, use one of the following commands:
  - On RHEL 7, SLES 15, or SLES 12 systems, enter the following:
    ```
    systemctl restart array
    ```
  - On RHEL 6 or SLES 11 systems, enter the following:
    ```
    /etc/init.d/array restart
    ```

The Array Services daemon on any node can access only the information in the configuration files on that node. One advantage to this design is that you can limit the use of particular nodes. At the same time, though, you need insure that common information is synchronized. For information about how to ensure synchronization, see the following:

Designing new array commands on page 104

Designing Array Services commands

By default, most Array Services commands run using the user, group, and project ID of either the user that issued the original command or `arraysvcs`. When you add new array commands to `arrayd.conf`, or when you modify existing array commands, always use the most restrictive IDs possible. This practice minimizes trouble if a hostile or careless user were to run that command. Avoid adding commands that run with more powerful IDs, such as the administrator user or group `sys`, than the user. If such commands are necessary, analyze them carefully to ensure that an arbitrary user would not be granted any more privileges than expected. In other words, analyze these commands the same as you analyze a `setuid` program.

The user can invoke arbitrary system commands on single nodes using the `arshell` command. The user can also launch MPI programs that automatically distribute over multiple nodes. However, the only way to launch coordinated system programs on all nodes at once is to use the `array` command. This command does not accept any system command; it only permits execution of commands that the administrator has configured into the Array Services configuration file.

As the administrator, you can define any set of commands that your users need. You have complete control over how any single array node runs a command. For example, the definition can be different on different nodes. A command can simply invoke a standard system command. If you define a command as invoking a script, you can make a command arbitrarily complex.

About substitution syntax

The `arrayd.conf`(4) manpage explains the syntax rules for entries in the configuration files. An important feature of this syntax is the use of several kinds of text substitution by which variable text is substituted into entries when run.

Most of the supported substitutions are used in command entries. These substitutions are performed dynamically each time the `array` command invokes a subcommand. At that time, substitutions insert values that are unique to the invocation of that subcommand. For example, the value `%USER` inserts the user ID of the user who is invoking the `array` command. Such a substitution has meaning only when the command runs.

Substitutions in other configuration entries are performed only once, at the time `arrayd` reads the configuration file. Only environment variable substitution makes sense in these entries. The environment variable values that are substituted are the values inherited by `arrayd` from the script that invokes it, which is as follows:

- On RHEL 7, SLES 15, or SLES 12 systems, the script is as follows:

  ```
  /usr/lib/systemd/system/array.service
  ```

- On RHEL 6 or SLES 11 systems, the script is as follows:

  ```
  /etc/init.d/array
  ```

About array command operations

When a user runs an `array` command, it has the following format:

```
array [options] subcommand
```

The specified *subcommand* operates on nodes as follows:

- If the user does not specify any options, the *subcommand* runs on the whole array.
- If the user specifies the `-l` option, the *subcommand* runs on the local node.
- If the user specifies the `-s` *`node`* option, the command runs on all nodes that *node* knows about.

  Remember that the destination node can be configured with only a subset of nodes. At each node, `arrayd` searches the configuration file for a `COMMAND` entry with the same name as the subcommand.
- If the user specifies both `-l` and `-s` *`node`*, the subcommand runs on the specified *node*.

For example, in the following command, `arrayd` processes the `uptime` subcommand on node `tokyo`:

```
array -s tokyo uptime
```

When `arrayd` finds the subcommand to be valid, it distributes the subcommand to every node that is configured in the default array at node `tokyo`.

In the `/etc/array/arrayd.conf` on `tokyo`, the `COMMAND` entry for `uptime` is as follows:

```
command uptime
        invoke /usr/lib/array/auptime %LOCAL
```

The `INVOKE` subentry tells `arrayd` how to run this command. In this case, it runs a shell script, `/usr/lib/array/auptime`, and passes one argument, the name of the local node. This command runs on every node, with `%LOCAL` replaced by the node name.

Command definition syntax summary

The basic Array Services commands reside in `/etc/array/arrayd.conf`. Each `COMMAND` entry is defined using subentries. The following table shows the subentries. The `arrayd.conf`(4) manpage also describes the subentries.

Table 3: Subentries of a `COMMAND` definition

Keyword	Meaning of following values
`COMMAND`	The name of the command as the user gives it to `array`.
`INVOKE`	A system command to be run on every node. Specify the full path to the system command. The argument values can be literals, user-supplied arguments, or other substitution values.
`MERGE`	A system command to be run only on the distributing node. Specify the full path to the system command. Its purpose is to gather the streams of output from all nodes and combine them into a single stream.
`USER`	The user ID under which the `INVOKE` and `MERGE` commands run. Typically specified as `USER %USER`, so it runs as the user who invoked `array`.
`GROUP`	The group name under which the `INVOKE` and `MERGE` commands run. Typically specified as `GROUP %GROUP`, so it runs in the group of the user who invoked `array`. For more information, see the `groups`(1) manpage.
`PROJECT`	The project under which the `INVOKE` and `MERGE` commands run. Typically specified as `PROJECT %PROJECT`, so it runs in the project of the user who invoked `array`. For more information, see the `projects`(5) manpage.
`OPTIONS`	A variety of options to modify this command. For more information, see the following: **Table 5: Options of the COMMAND definition** on page 103

As with a shell script, system commands are often composed from a few literal values and many substitution strings. The following table shows the supported substitutions. The `arrayd.conf`(4) manpage describes the substitutions in more detail.

Table 4: Substitutions used in a `COMMAND` definition

Substitution	Replacement value
`%1..%9`; `%ARG(`*n*`)`; `%ALLARGS`; `%OPTARG(`*n*`)`	Argument tokens from the user subcommand that the user specified. If the specified argument is omitted, `%OPTARG` does not produce an error message.
`%USER`, `%GROUP`, `%PROJECT`	The effective user ID, effective group ID, and project of the user who invoked `array`.

Table Continued

Substitution	Replacement value
`%REALUSER`, `%REALGROUP`	The real user ID and real group ID of the user who invoked `array`.
`%ASH`	The internal array session handle (ASH) number under which the `INVOKE` or `MERGE` command is to run. The term **array session** includes all the processes for one application, regardless of where the processes run. Typically, an array session includes the login shell of the user and the programs that the user started from the login shell. A batch job is an array session.
`%PID(`*`ash`*`)`	List of process identifier (PID) values for a specified ASH. `%PID(%ASH)` is a common use.
`%ARRAY`	The array name, either default or as given in the `-a` option.
`%LOCAL`	The hostname of the executing node.
`%ORIGIN`	The full domain name of the node where the `array` command ran and where the output is to be viewed.
`%OUTFILE`	List of names of temporary files, each containing the output from the `INVOKE` command of one node. Valid only in the `MERGE` subentry.

The `OPTIONS` subentry permits a number of important modifications of the command execution. The following table summarizes these modifications.

Table 5: Options of the `COMMAND` definition

Keyword	Effect on command
`LOCAL`	Does not distribute to other nodes. Effectively forces the `-l` option.
`NEWSESSION`	Runs the `INVOKE` command under a newly created ASH. The `%ASH` in the `INVOKE` line is the new ASH. The `MERGE` command runs under the original ASH, and `%ASH` substitutes as the old ASH in that line.
`SETRUID`	Sets both the real and the effective user ID from the `USER` subentry. Typically, `USER` sets only the effective UID.
`SETRGID`	Sets both the real and effective group ID from the `GROUP` subentry. Typically, `GROUP` sets only the effective GID.

Table Continued

Keyword	Effect on command
QUIET	Discards the output of INVOKE, unless a MERGE subentry is present. If a MERGE subentry is present, passes INVOKE output to MERGE as usual, and discards the MERGE output.
NOWAIT	Discards the output and returns as soon as the processes are invoked. Does not wait for completion. A MERGE subentry is ineffective.

Configuring local options

The LOCAL entry specifies options to arrayd itself. The following table summarizes the most important options.

Table 6: Subentries of the LOCAL entry

Subentry	Purpose
DIR	Pathname for the arrayd working directory, which is the initial, current working directory of INVOKE and MERGE commands. The default is /usr/lib/array.
DESTINATION ARRAY	Name of the default array, used when the user omits the -a option. When only one ARRAY entry is specified, it is the default destination.
USER, GROUP, PROJECT	Default values for COMMAND execution when USER, GROUP, or PROJECT are omitted from the COMMAND definition.
HOSTNAME	Value returned in this node by %LOCAL. Default is the hostname.
PORT	Socket to be used by arrayd.

If you do not supply LOCAL USER, GROUP, and PROJECT values, the default values for USER and GROUP are arraysvcs.

The HOSTNAME entry is needed whenever the hostname command does not return a node name as specified in the ARRAY MACHINE entry. To supply a LOCAL HOSTNAME entry unique to each node, each node needs an individualized copy of at least one configuration file.

Designing new array commands

The /usr/lib/array/arrayd.conf.template file contains a basic set of commands. Examine this file carefully before defining commands of your own. Any new commands that you design become available to the users of the array system. You can develop new administrative commands, too.

Typically, a new command is defined with an INVOKE subentry that names a script written in sh, csh, or Perl syntax. You can use the substitution values to set up arguments to the script. You use the USER, GROUP, PROJECT, and OPTIONS subentries to establish the execution conditions of the script.

Within the invoked script, you can write any amount of logic to verify and validate the arguments and to run any command sequence. For an example of a script in Perl, see /usr/lib/array/aps, which is invoked by the array ps command.

NOTE: Perl is an interesting choice for `array` commands because Perl has native support for socket I/O. In principle, you can build a distributed application in Perl in which multiple instances are launched by `array` and coordinate and exchange data using sockets. Performance would not rival the highly tuned MPI libraries, but development would be simpler.

The following example shows an administrator command called `reinit`, which reinitializes the Array Services configuration file on all nodes at once:

- The shell script in file `/usr/lib/array/arrayd-reinit` reinitializes each Array Services configuration file, on each node, simultaneously. The script is designed for RHEL 7, SLES 15, and SLES 12 systems and is as follows:

  ```
  #!/bin/sh
  # Script to reinitialize arrayd with a new configuration file
  # Usage:  arrayd-reinit <hostname:new-config-file>
  sleep 10                 # Let old arrayd finish distributing
  scp $1 /etc/array/
  systemctl restart array
  exit 0
  ```

  The script uses `rcp` to copy a specified file, presumably a configuration file such as `arrayd.conf`, into `/etc/array`. The script fails if `%USER` is not privileged. Then the script restarts `arrayd` to reread configuration files.

- The following is the command definition:

  ```
  command reinit
     invoke /usr/lib/array/arrayd-reinit %ORIGIN:%1
     user   %USER
     group  %GROUP
     options nowait   # Exit before restart occurs!
  ```

  The `INVOKE` subentry calls the script shown previously. The `NOWAIT` option prevents the daemon from waiting for the script to finish. The script stops the daemon.

Testing configuration changes after creating Array Services commands

The configuration files contain many sections and options. You can use the `ascheck` command to perform a basic check of all configuration files in the array.

After making a change, you can run the `arrayd` command with the `-c` and `-f` options to test an individual configuration file for correct syntax. For example, assume that you added command definition to `/etc/array/arrayd.local`. You can enter the following command to check its syntax:

```
arrayd -c -f /usr/lib/array/arrayd.local
```

When testing new commands for correct operation, monitor the warning and error messages produced by the `arrayd` command and by the processes that the `arrayd` command can spawn. Typically, the `stderr` messages from a daemon are not visible.

Procedure

1. Notify the array users that you are able to start an array configuration update.

   Users might experience a lack of response to `ainfo` and `array` commands.

2. Log into one of the nodes as the administrator user, and enter one of the following commands:

On RHEL 7, SLES 15, and SLES 12 systems, enter the following command:

```
# systemctl stop array
```

On RHEL 6 systems and SLES 11 systems, enter the following command:

```
# /etc/init.d/array stop
```

3. In one shell window on that node, enter the following `arrayd` command:

   ```
   # /usr/sbin/arrayd -n -v
   ```

   The preceding command prevents the `arrayd` command from moving into the background. The command remains attached to the shell terminal.

   The `arrayd` command becomes functional in this mode, but it does not refer to the `/etc/sysconfig/array` file. Specify all command-line options, such as the names of nonstandard configuration files, explicitly.

4. Enter `ainfo` and `array` commands to test the new configuration data.

   Enter these commands as follows:

   - From another shell window on the same node
   - From another shell window on another node

   Observe that diagnostic output appears in the `arrayd` shell window.

5. From the shell window in which you entered the `/usr/sbin/arrayd -n -v` command, enter CTRL-c to terminate the `arrayd` daemon.

6. Enter one of the following commands to start the `arrayd` daemon:
   On RHEL 7, SLES 15, and SLES 12 systems, enter the following command:

   ```
   # systemctl start array
   ```

   On RHEL 6 and SLES 11 systems, enter the following command:

   ```
   # /etc/init.d/array start
   ```