# Contents

Welcome ........................................................................................................................................ ix
Introduction ................................................................................................................................... ix
TORQUE Administrator Guide overview ......................................................................................... xi

## Chapter 1: Overview ..................................................................................................................... 1
TORQUE installation overview ........................................................................................................ 1
  TORQUE architecture ................................................................................................................... 1
  Installing TORQUE .................................................................................................................... 2
  Compute nodes ............................................................................................................................ 5
  Enabling TORQUE as a service ................................................................................................... 7
Initializing/Configuring TORQUE on the server (pbs_server) ......................................................... 7
  Specifying compute nodes ......................................................................................................... 9
  Configuring TORQUE on compute nodes .................................................................................. 10
  Configuring Ports ....................................................................................................................... 11
  Configuring trqauthd for client commands .............................................................................. 13
  Finalizing configurations ........................................................................................................... 14

Advanced configuration .................................................................................................................. 14
  Customizing the install .............................................................................................................. 15
  Server configuration .................................................................................................................. 21

Manual setup of initial server configuration .................................................................................. 26
Server node file configuration ........................................................................................................ 27
  Basic node specification ............................................................................................................ 27
  Specifying virtual processor count for a node .......................................................................... 28
  Specifying GPU count for a node ............................................................................................. 28
  Specifying node features (node properties) .............................................................................. 29

Testing server configuration ......................................................................................................... 29
TORQUE on NUMA systems .......................................................................................................... 31
  TORQUE NUMA configuration ................................................................................................. 31
  Building TORQUE with NUMA support .................................................................................. 31
TORQUE Multi-MOM ...................................................................................................................... 35
  Multi-MOM configuration ......................................................................................................... 35
  Stopping pbs_mom in Multi-MOM mode ................................................................................... 37

## Chapter 2: Submitting and managing jobs .................................................................................. 39
Job submission ............................................................................................................................... 39
  Multiple job submission .......................................................................................................... 40
  Managing multi-node jobs ......................................................................................................... 42
  Requesting resources ................................................................................................................. 42
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Chapter 3: Managing nodes</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>Adding nodes</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>Node properties</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Changing node state</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>Host security</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>Linux cpuset support</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>Scheduling cores</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>Geometry request configuration</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>Geometry request usage</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>Geometry request considerations</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>Scheduling accelerator hardware</td>
<td>75</td>
</tr>
<tr>
<td>4</td>
<td>Chapter 4: Setting server policies</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Queue configuration</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Queue attributes</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>Example queue configuration</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>Setting a default queue</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>Mapping a queue to subset of resources</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>Creating a routing queue</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>Server high availability</td>
<td>91</td>
</tr>
<tr>
<td>TORQUE Quick Start Guide</td>
<td>301</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>BLCR acceptance tests</td>
<td>305</td>
<td></td>
</tr>
<tr>
<td>Test environment</td>
<td>305</td>
<td></td>
</tr>
<tr>
<td>Test 1 - Basic operation</td>
<td>305</td>
<td></td>
</tr>
<tr>
<td>Test 2 - Persistence of checkpoint images</td>
<td>308</td>
<td></td>
</tr>
<tr>
<td>Test 3 - Restart after checkpoint</td>
<td>308</td>
<td></td>
</tr>
<tr>
<td>Test 4 - Multiple checkpoint/restart</td>
<td>309</td>
<td></td>
</tr>
<tr>
<td>Test 5 - Periodic checkpoint</td>
<td>309</td>
<td></td>
</tr>
<tr>
<td>Test 6 - Restart from previous image</td>
<td>310</td>
<td></td>
</tr>
</tbody>
</table>
Welcome to the TORQUE Administrator Guide, version 4.2.6. This guide is intended as a reference for both users and system administrators.

Note: Advanced TORQUE Administration is a video tutorial of a session offered at Moab Con that offers further details on advanced TORQUE administration.

For more information about this guide, see these topics:

- **TORQUE Administrator Guide overview** on page xi
- **Introduction** on page ix

## Introduction

This section contains some basic introduction information to help you get started using TORQUE. It contains these topics:

- **What is a Resource Manager?** on page ix
- **What are Batch Systems?** on page ix
- **Basic Job Flow** on page x

### What is a Resource Manager?

While TORQUE has a built-in scheduler, pbs_sched, it is typically used solely as a resource manager with a scheduler making requests to it. Resources managers provide the low-level functionality to start, hold, cancel, and monitor jobs. Without these capabilities, a scheduler alone cannot control jobs.

### What are Batch Systems?

While TORQUE is flexible enough to handle scheduling a conference room, it is primarily used in batch systems. Batch systems are a collection of computers and other resources (networks, storage systems, license servers, and so forth) that operate under the notion that the whole is greater than the sum of the parts. Some batch systems consist of just a handful of machines running single-processor jobs, minimally managed by the users themselves. Other systems have thousands and thousands of machines executing users' jobs simultaneously while tracking software licenses and access to hardware equipment and storage systems.

Pooling resources in a batch system typically reduces technical administration of resources while offering a uniform view to users. Once configured properly, batch systems abstract away many of the details involved with running and managing jobs, allowing higher resource utilization. For example, users
typically only need to specify the minimal constraints of a job and do not need to know the individual machine names of each host on which they are running. With this uniform abstracted view, batch systems can execute thousands and thousands of jobs simultaneously.

Batch systems are comprised of four different components: (1) Master Node, (2) Submit/Interactive Nodes, (3) Compute Nodes, and (4) Resources.

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Master Node</td>
<td>A batch system will have a master node where <code>pbs_server</code> runs. Depending on the needs of the systems, a master node may be dedicated to this task, or it may fulfill the roles of other components as well.</td>
</tr>
<tr>
<td>Submit/Interactive Nodes</td>
<td>Submit or interactive nodes provide an entry point to the system for users to manage their workload. For these nodes, users are able to submit and track their jobs. Additionally, some sites have one or more nodes reserved for interactive use, such as testing and troubleshooting environment problems. These nodes have client commands (such as <code>qsub</code> and <code>qhold</code>).</td>
</tr>
<tr>
<td>Computer Nodes</td>
<td>Compute nodes are the workhorses of the system. Their role is to execute submitted jobs. On each compute node, <code>pbs_mom</code> runs to start, kill, and manage submitted jobs. It communicates with <code>pbs_server</code> on the master node. Depending on the needs of the systems, a compute node may double as the master node (or more).</td>
</tr>
<tr>
<td>Resources</td>
<td>Some systems are organized for the express purpose of managing a collection of resources beyond compute nodes. Resources can include high-speed networks, storage systems, license managers, and so forth. Availability of these resources is limited and needs to be managed intelligently to promote fairness and increased utilization.</td>
</tr>
</tbody>
</table>

Basic Job Flow

The life cycle of a job can be divided into four stages: (1) creation, (2) submission, (3) execution, and (4) finalization.
Stage | Description
--- | ---
Creation | Typically, a submit script is written to hold all of the parameters of a job. These parameters could include how long a job should run (walltime), what resources are necessary to run, and what to execute. The following is an example submit file:
```
#PBS -N localBlast
#PBS -S /bin/sh
#PBS -l nodes=1:ppn=2,walltime=240:00:00
#PBS -M user@my.organization.com
#PBS -m ea
source ~/.bashrc
cd $HOME/work/dir
sh myBlast.sh -i -v
```
This submit script specifies the name of the job (localBlast), what environment to use (/bin/sh), that it needs both processors on a single node (nodes=1:ppn=2), that it will run for at most 10 days, and that TORQUE should email "user@myorganization.com" when the job exits or aborts. Additionally, the user specifies where and what to execute.

Submission | A job is submitted with the qsub command. Once submitted, the policies set by the administration and technical staff of the site dictate the priority of the job and therefore, when it will start executing.

Execution | Jobs often spend most of their lifecycle executing. While a job is running, its status can be queried with qstat.

Finalization | When a job completes, by default, the stdout and stderr files are copied to the directory where the job was submitted.

Related topics
- [TORQUE Administrator Guide overview](#) on page xi

---

### TORQUE Administrator Guide overview

*Chapter 1: Overview on page 1* provides the details for installation and initialization, advanced configuration options, and (optional) qmgr option necessary to get the system up and running. System testing is also covered.

*Chapter 2: Submitting and managing jobs on page 39* covers different actions applicable to jobs. The first section details how to submit a job and request resources (nodes, software licenses, and so forth), and provides several examples. Other actions include monitoring, canceling, preemption, and keeping completed jobs.

*Chapter 3: Managing nodes on page 69* covers administrator tasks relating to nodes, which include the following: adding nodes, changing node properties, and identifying state. Also an explanation of how to configure restricted user access to nodes is covered in [Host security on page 71](#).
Chapter 4: Setting server policies on page 77 details server-side configurations of queue and high availability.

Chapter 5: Integrating schedulers for TORQUE on page 105 offers information about using the native scheduler versus an advanced scheduler.

Chapter 6: Configuring data management on page 107 deals with issues of data management. For non-network file systems, SCP setup on page 107 details setting up SSH keys and nodes to automate transferring data. NFS and other networked filesystems on page 110 covers configuration for these file systems. This chapter also addresses the use of file staging using the stagein and stageout directives of the qsub command.

Chapter 7: MPI (Message Passing Interface) support on page 113 offers details supporting MPI.

Chapter 8: Resources on page 115 covers configuration, utilization, and states of resources.

Chapter 9: Accounting records on page 119 explains how jobs are tracked by TORQUE for accounting purposes.

Chapter 10: Job logging on page 121 explains how to enable job logs that contain information for completed jobs.

Chapter 11: Troubleshooting on page 123 is a guide that offers help with general problems. It includes FAQ and instructions for how to set up and use compute node checks. It also explains how to debug TORQUE.

The appendices provide tables of commands, parameters, configuration options, error codes, the Quick Start Guide, and so forth.

- Commands overview on page 145
- Server parameters on page 227
- Node manager (MOM) configuration on page 245
- Diagnostics and error codes on page 263
- Considerations before upgrading on page 271
- Large cluster considerations on page 273
- Prologue and epilogue scripts on page 281
- Running multiple TORQUE servers and MOMs on the same node on page 289
- Security overview on page 291
- Job submission filter ("qsub wrapper") on page 293
- "torque.cfg" configuration file on page 295
- TORQUE Quick Start Guide on page 301
- BLCR acceptance tests on page 305

Related topics
- Introduction on page ix
Chapter 1: Overview

This section contains some basic information about TORQUE, including how to install and configure it on your system. For details, see these topics:

- TORQUE installation overview on page 1
- Initializing/Configuring TORQUE on the server (pbs_server) on page 7
- Advanced configuration on page 14
- Manual setup of initial server configuration on page 26
- Server node file configuration on page 27
- Testing server configuration on page 29
- TORQUE on NUMA systems on page 31
- TORQUE Multi-MOM on page 35

TORQUE installation overview

This section contains information about TORQUE architecture and explains how to install TORQUE. It also describes how to install tpackages on compute nodes and how to enable TORQUE as a service.

For details, see these topics:

- TORQUE architecture on page 1
- Installing TORQUE on page 2
- Compute nodes on page 5
- Enabling TORQUE as a service on page 7

Related topics

- Troubleshooting on page 123

TORQUE architecture

A TORQUE cluster consists of one head node and many compute nodes. The head node runs the pbs_server daemon and the compute nodes run the pbs_mom daemon. Client commands for submitting and managing jobs can be installed on any host (including hosts not running pbs_server or pbs_mom).
The head node also runs a scheduler daemon. The scheduler interacts with pbs_server to make local policy decisions for resource usage and allocate nodes to jobs. A simple FIFO scheduler, and code to construct more advanced schedulers, is provided in the TORQUE source distribution. Most TORQUE users choose to use a packaged, advanced scheduler such as Maui or Moab.

Users submit jobs to pbs_server using the qsub command. When pbs_server receives a new job, it informs the scheduler. When the scheduler finds nodes for the job, it sends instructions to run the job with the node list to pbs_server. Then, pbs_server sends the new job to the first node in the node list and instructs it to launch the job. This node is designated the execution host and is called Mother Superior. Other nodes in a job are called sister MOMs.

**Related topics**

- TORQUE installation overview on page 1
- Installing TORQUE on page 2

**Installing TORQUE**

Build the distribution on the machine that will act as the TORQUE server - the machine which monitors and controls all compute nodes by running the pbs_server daemon.

The built distribution package works only on compute nodes of a similar architecture. Nodes with different architecture must have the installation package built on them individually.

The following software is required to run TORQUE 4.2.6:

- libxml2-devel package (package name may vary)
- openssl-devel package (package name may vary)
- ANSI C compiler. The native C compiler is recommended if it is ANSI; otherwise, use gcc.
- A fully POSIX make. If you are unable to "make" PBS with your make, we suggest using gmake from GNU.
- Tcl/Tk version 8 or higher if you plan to build the GUI portion of TORQUE or use a Tcl based scheduler.

**Important:** If you intend to use TORQUE 4.2.6 with Moab, you must run Moab version 7.1 or later. TORQUE 4.2.6 will not work with versions earlier than Moab 7.1.

**To install TORQUE**

1. Install the gcc, libssl-devel, and libxml2-devel packages to build 4.2.6. The package names may vary.

   Use these commands to install the packages on the following operating systems:

   **RHEL, CentOS, and Scientific Linux:**
Chapter 1: Overview

SLES (You must have a licensed installation of SuSE and have installed the SuSE Linux Enterprise Software Development Kit and added the ISO to the repository):

```
[root]# yum update
[root]# yum install libxml2-devel openssl-devel gcc gcc-c++
```

2. Verify that the following ports are open for essential communication:
   - For client communication to pbs_server, all privileged ports must be open (ports under 1024).
   - For pbs_server communication to pbs_mom, the default port is 15003.
   - For pbs_mom to pbs_server, the default port is 15001.

For more information on configuring ports, see Configuring Ports on page 11.

3. TORQUE is now hosted at https://github.com under the adaptivecomputing organization. To download source, you will need to use the git utility. For example:

```
[root]# git clone https://github.com/adaptivecomputing/torque.git
```

To download a different version, replace each 4.2.6 with the desired version. After downloading a copy of the repository, you can list the current branches by typing git branch -a from within the directory of the branch you cloned.

If you're checking source out from git, read the README.building-40 file in the repository.

4. Extract the packaged file and navigate to the unpackaged directory.

```
[root]# tar -zxvf torque-4.2.6.tar.gz
[root]# cd torque-4.2.6/
```

5. Do the following to configure the package:
   a. Run ./configure. TORQUE has many options you can specify to configure your installation. For more information, see Customizing the Install on page 15.
   b. By default, the TORQUE directory is /var/spool/torque. If you want to change the location, use the --with-server-home configuration option.
   c. By default, make install installs all files in /usr/local/bin, /usr/local/lib, /usr/local/sbin, /usr/local/include, and /usr/local/man. You can specify an installation prefix other than /usr/local by using --prefix as an argument to ./configure. Note that TORQUE cannot be installed into a directory path that contains a space.

If you decide to use --prefix to specify a custom directory and that directory does not already exist, you must create it before running ./configure --prefix=....
Chapter 1: Overview

6. Run make and make install.

TORQUE must be installed by a root user. If running sudo fails, switch to root with su -.

OSX 10.4 users need to change #define _TDARWIN in src/include/pbs_config.h to #define _TDARWIN_8.

7. Configure the trqauthd daemon to start automatically at system boot (See Configuring trqauthd for client commands on page 13).

* If Debian distribution, do the following *
   [root]# cp contrib/init.d/debian.trqauthd /etc/init.d/trqauthd

* If SLES distribution, do the following *
   [root]# cp contrib/init.d/suse.trqauthd /etc/init.d/trqauthd

* If RHEL distribution, do the following *
   [root]# cp contrib/init.d/trqauthd /etc/init.d/

   [root]# chkconfig --add trqauthd
   [root]# echo /usr/local/lib > /etc/ld.so.conf.d/torque.conf
   [root]# ldconfig
   [root]# service trqauthd start

8. The make packages command can be used to create self-extracting packages that can be copied and executed on your nodes. For information on creating packages and deploying them, see Specifying compute nodes on page 9.

You will also want to scp the init.d scripts to the compute nodes and install them there.

The TORQUE initialization scripts are provided in the /init.d directory as a courtesy and may be modified at your discretion to work on your system.

9. Verify that the /var/spool/torque/server_name file exists and contains the correct name of the server.

   [root]# echo <pbs_server's_hostname> > /var/spool/torque/server_name

10. After installation, verify that you have the PATH environment variable configured to include /usr/local/bin/ and /usr/local/sbin/ for both the installation user and the root user.

By default, make install creates a directory at /var/spool/torque. This directory is referred to as TORQUE_HOME. TORQUE_HOME has several sub-directories, including server_priv/, server_logs/, mom_priv/, mom_logs/, and other directories used in the configuration and running of TORQUE.

11. Verify you have environment variables configured so your system can find the shared libraries and binary files for TORQUE. This step is not necessary if the shared libraries are in their default
locations.

To set the library path, add the directory where the TORQUE libraries are installed. For example, if your TORQUE libraries are installed in /usr/local/lib (if you changed the default library path at configure time, use that path instead.), execute the following:

```
[root]# echo '/usr/local/lib' > /etc/ld.so.conf.d/torque.conf
[root]# ldconfig
```

12. Initialize serverdb by executing the torque.setup script.
```
[root]# ./torque.setup root
```

If you are upgrading from TORQUE 2.5.9, run pbs_server -u before running torque.setup.
```
[root]# pbs_server -u
```

13. Add nodes to the /var/spool/torque/server_priv/nodes file. For information on syntax and options for specifying compute nodes, see Node manager (MOM) configuration on page 245.

14. Configure the MOMs if necessary. See Configuring TORQUE on compute nodes on page 10.

15. Configure pbs_server and pbs_mom to start automatically at system boot, then start their daemons. See Enabling TORQUE as a service on page 7.

While Adaptive Computing distributes the spec files to create RPMs, it doesn’t support RPM installation. Not every Linux distribution uses RPM. Adaptive Computing provides a single solution using make and make install that works across all Linux distributions and most UNIX systems. We recognize the RPM format provides many advantages for deployment but it is up to the individual site to repackage the TORQUE installation to match their individual needs.

If you have a multi-homed host, or if your host name resolves to the loopback address, you may encounter an “unauthorized request” error when running qmgr. The resolution is to as a manager manually add the host name of the specified interface to the serverdb file with the attributes tag.
```
<managers>root@napali.ib</managers>
```

Related topics

- TORQUE installation overview on page 1
- Compute nodes on page 5

Compute nodes

Use the Adaptive Computing tpackage system to create self-extracting tarballs which can be distributed and installed on compute nodes. The tpackages are customizable. See the INSTALL file for additional options and features.
To create tpackages

1. Configure and make as normal, and then run make packages.

```bash
> make packages
Building ./torque-package-clients-linux-i686.sh ...
Building ./torque-package-mom-linux-i686.sh ...
Building ./torque-package-server-linux-i686.sh ...
Building ./torque-package-gui-linux-i686.sh ...
Building ./torque-package-devel-linux-i686.sh ...
Done.
The package files are self-extracting packages that can be copied and executed on your production machines. Use --help for options.
```

2. Copy the desired packages to a shared location.

```bash
> cp torque-package-mom-linux-i686.sh /shared/storage/
> cp torque-package-clients-linux-i686.sh /shared/storage/
```

3. Install the tpackages on the compute nodes.

Adaptive Computing recommends that you use a remote shell, such as SSH, to install tpackages on remote systems. Set up shared SSH keys if you do not want to supply a password for each host.

```bash
for i in node01 node02 node03 node04; do scp torque-package-mom-linux-i686.sh ${i}:/tmp/; done
for i in node01 node02 node03 node04; do scp torque-package-clients-linux-i686.sh ${i}:/tmp/; done
for i in node01 node02 node03 node04; do ssh ${i} /tmp/torque-package-mom-linux-i686.sh --install; done
for i in node01 node02 node03 node04; do ssh ${i} /tmp/torque-package-clients-linux-i686.sh --install; done
```

Alternatively, you can use a tool like xCAT instead of dsh.

To use a tool like xCAT

1. Copy the tpackage to the nodes.

```bash
> prcp torque-package-linux-i686.sh noderange:/destinationdirectory/
```

2. Install the package.

```bash
> psh noderange /tmp/torque-package-linux-i686.sh --install
```

Although optional, it is possible to use the TORQUE server as a compute node and install a pbs_mom with the pbs_server daemon.
Enabling TORQUE as a service

Enabling TORQUE as a service is optional. In order to run TORQUE as a service, you must enable running client commands (for instructions, see Configuring trqauthd for client commands on page 13.

The method for enabling TORQUE as a service is dependent on the Linux variant you are using. Startup scripts are provided in the contrib/init.d/ directory of the source package. To enable TORQUE as a service, run the following on the host for the appropriate TORQUE daemon:

- **RedHat (as root)**

  ```
  > cp contrib/init.d/pbs_server /etc/init.d/pbs_server
  > chkconfig --add pbs_server
  ```

- **SuSE (as root)**

  ```
  > cp contrib/init.d/suse.pbs_server /etc/init.d/pbs_server
  > insserv -d pbs_server
  ```

- **Debian (as root)**

  ```
  > cp contrib/init.d/debian.pbs_server /etc/init.d/pbs_server
  > update-rc.d pbs_server defaults
  ```

You will need to customize these scripts to match your system.

These options can be added to the self-extracting packages. For more details, see the INSTALL file.

**Related topics**

- **TORQUE installation overview on page 1**

Initializing/Configuring TORQUE on the server (pbs_server)

The TORQUE server (pbs_server) contains all the information about a cluster. It knows about all of the MOM nodes in the cluster based on the information in the $TORQUE_HOME/server_priv/nodes file
Chapter 1: Overview

(See Configuring TORQUE on compute nodes on page 10). It also maintains the status of each MOM node through updates from the MOMs in the cluster (see pbsnodes on page 168). All jobs are submitted via qsub to the server, which maintains a master database of all jobs and their states.

Schedulers such as Moab Workload Manager receive job, queue, and node information from pbs_server and submit all jobs to be run to pbs_server.

The server configuration is maintained in a file named serverdb, located in $TORQUE_HOME/server_priv. The serverdb file contains all parameters pertaining to the operation of TORQUE plus all of the queues which are in the configuration. For pbs_server to run, serverdb must be initialized.

You can initialize serverdb in two different ways, but the recommended way is to use the ./torque.setup script:

- As root, execute ./torque.setup from the build directory (see ./torque.setup on page 8).
- Use pbs_server -t create (see Initializing/Configuring TORQUE on the server (pbs_server) on page 7).

Restart pbs_server after initializing serverdb.

```
> qterm
> pbs_server
```

./torque.setup

The torque.setup script uses pbs_server -t create to initialize serverdb and then adds a user as a manager and operator of TORQUE and other commonly used attributes. The syntax is as follows:

```
./torque.setup username

> ./torque.setup ken
> qmgr -c "p s"

# Create queues and set their attributes.
# #
# Create and define queue batch
# create queue batch
set queue batch queue_type = Execution
set queue batch resources_default.nodes = 1
set queue batch resources_default.walltime = 01:00:00
set queue batch enabled = True
set queue batch started = True
# # Set server attributes.
# set server scheduling = True
set server acl_hosts = kmn
set server managers = ken@kmn
set server operators = ken@kmn
set server default_queue = batch
set server log_events = 511
set server mail_from = adm
set server node_check_rate = 150
set server tcp_timeout = 6
set server mom_job_sync = True
set server keep_completed = 300
```
pbs_server -t create

The -t create option instructs pbs_server to create the serverdb file and initialize it with a minimum configuration to run pbs_server.

```
> pbs_server -t create
```

To see the configuration and verify that TORQUE is configured correctly, use qmgr:

```
> qmgr -c 'p s'
```

A single queue named batch and a few needed server attributes are created.

This section contains these topics:

- **Specifying compute nodes** on page 9
- **Configuring TORQUE on compute nodes** on page 10
- **Finalizing configurations** on page 14

Related topics

- **Node manager (MOM) configuration** on page 245
- **Advanced configuration** on page 14

### Specifying compute nodes

The environment variable **TORQUE_HOME** is where configuration files are stored. If you used the default locations during installation, you do not need to specify the **TORQUE_HOME** environment variable.

The pbs_server must recognize which systems on the network are its compute nodes. Specify each node on a line in the server's nodes file. This file is located at **TORQUE_HOME/server_priv/nodes**. In most cases, it is sufficient to specify just the names of the nodes on individual lines; however, various properties can be applied to each node.

**Only a root user can access the server_priv directory.**

Syntax of nodes file:

```
node-name[:ts] [np=] [gpus=] [properties]
```

- The **node-name** must match the hostname on the node itself, including whether it is fully qualified or shortened.
The [:ts] option marks the node as timeshared. Timeshared nodes are listed by the server in the node status report, but the server does not allocate jobs to them.

The [np=] option specifies the number of virtual processors for a given node. The value can be less than, equal to, or greater than the number of physical processors on any given node.

The [gpus=] option specifies the number of GPUs for a given node. The value can be less than, equal to, or greater than the number of physical GPUs on any given node.

The node processor count can be automatically detected by the TORQUE server if auto_node_np is set to TRUE. This can be set using this command:

```
qmgr -c set server auto_node_np = True
```

Setting auto_node_np to TRUE overwrites the value of np set in TORQUE_HOME/server_priv/nodes.

The [properties] option allows you to specify arbitrary strings to identify the node. Property strings are alphanumeric characters only and must begin with an alphabetic character.

Comment lines are allowed in the nodes file if the first non-white space character is the pound sign (#).

The following example shows a possible node file listing.

```
TORQUE_HOME/server_priv/nodes:

# Nodes 001 and 003-005 are cluster nodes
# node001 np=2 cluster01 rackNumber22
# node002 will be replaced soon
node002:ts waitingToBeReplaced
# node002 will be replaced soon
# node003 np=4 cluster01 rackNumber24
node004 cluster01 rackNumber25
node005 np=2 cluster01 rackNumber26 RAM16GB
node006
node007 np=2
node008:ts np=4
...
```

Related topics

- Initializing/Configuring TORQUE on the server (pbs_server) on page 7

Configuring TORQUE on compute nodes

If using TORQUE self-extracting packages with default compute node configuration, no additional steps are required and you can skip this section.

If installing manually, or advanced compute node configuration is needed, edit the TORQUE_HOME/mom_priv/config file on each node. The recommended settings follow.

```
TORQUE_HOME/mom_priv/config:
```
This file is identical for all compute nodes and can be created on the head node and distributed in parallel to all systems.

**Related topics**

- Initializing/Configuring TORQUE on the server (pbs_server) on page 7

## Configuring Ports

You can optionally configure the various ports that TORQUE uses for communication. Most ports can be configured multiple ways. The ports you can configure are:

- **pbs_server listening port**
- **pbs_mom listening port**
- **port pbs_server uses to communicate to the pbs_mom**
- **port pbs_mom uses to communicate to the pbs_server**
- **port client commands use to communicate to the pbs_server**
- **port trqauthd uses to communicate to the pbs_server**

**If you are running pbspro on the same system, be aware that it uses the same environment variables and /etc/services entries.**

### Configuring the pbs_server listening port

To configure the port the pbs_server listens on, follow any of these steps:

- Set an environment variable called PBS_BATCH_SERVICE_PORT to the port desired.
- Edit the /etc/services file and set pbs port_num/tcp.
- Start pbs_server with the -p option.

```bash
$ pbs_server -p port_num
```

- Edit the $PBS_HOME/server_name file and change server_name to server_name:<port_num>

- Start pbs_server with the -H option.

```bash
$ pbs_server -H server_name:port_num
```

### Configuring the pbs_mom listening port

To configure the port the pbs_mom listens on, follow any of these steps:
- Set an environment variable called `PBS_MOM_SERVICE_PORT` to the port desired.
- Edit the `/etc/services` file and set `pbs_mom port_num/tcp`.
- Start `pbs_mom` with the `-M` option.
- Edit the nodes file entry for that list: `add mom_service_port=port_num`.

**Configuring the port `pbs_server` uses to communicate with `pbs_mom`**

To configure the port the `pbs_server` uses to communicate with `pbs_mom`, follow any of these steps:

- Set an environment variable called `PBS_MOM_SERVICE_PORT` to the port desired.
- Edit the `/etc/services` file and set `pbs_mom port_num/tcp`.
- Start `pbs_mom` with the `-M` option.
- Edit the nodes file entry for that list: `add mom_service_port=port_num`.

**Configuring the port `pbs_mom` uses to communicate with `pbs_server`**

To configure the port the `pbs_mom` uses to communicate with `pbs_server`, follow any of these steps:

- Set an environment variable called `PBS_BATCH_SERVICE_PORT` to the port desired.
- Edit the `/etc/services` file and set `pbs_mom port_num/tcp`.
- Start `pbs_mom` with the `-S` option.
- Edit the nodes file entry for that list: `add mom_service_port=port_num`.

**Configuring the port client commands use to communicate with `pbs_server`**

To configure the port client commands use to communicate with `pbs_server`, follow any of these steps:

- Edit the `/etc/services` file and set `pbs_server port_num/tcp`.
- Edit the `$PBS_HOME/server_name` file and change `server_name` to `server_name:port_num`.

**Configuring the port trqauthd uses to communicate with `pbs_server`**

To configure the port `trqauthd` uses to communicate with `pbs_server`, follow any of these steps:

- Edit the `$PBS_HOME/server_name` file and change `server_name` to `server_name:port_num`.

**Related topics**

- [Initializing/Configuring TORQUE on the server (`pbs_server`) on page 7](#)
- `pbs_server`
Configuring trqauthd for client commands

trqauthd is a daemon used by TORQUE client utilities to authorize user connections to pbs_server. Once started, it remains resident. TORQUE client utilities then communicate with trqauthd on port 15005 on the loopback interface. It is multi-threaded and can handle large volumes of simultaneous requests.

Running trqauthd

trqauthd must be run as root. It must also be running on any host where TORQUE client commands will execute.

By default, trqauthd is installed to /usr/local/bin.

trqauthd can be invoked directly from the command line or by the use of init.d scripts which are located in the contrib/init.d directory of the TORQUE source.

There are three init.d scripts for trqauthd in the contrib/init.d directory of the TORQUE source tree:

<table>
<thead>
<tr>
<th>Script</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>debian.trqauthd</td>
<td>Used for apt-based systems (debian, ubuntu are the most common variations of this)</td>
</tr>
<tr>
<td>suse.trqauthd</td>
<td>Used for suse-based systems</td>
</tr>
<tr>
<td>trqauthd</td>
<td>An example for other package managers (Redhat, Scientific, CentOS, and Fedora are some common examples)</td>
</tr>
</tbody>
</table>

You should edit these scripts to be sure they will work for your site.

Inside each of the scripts are the variables PBS_DAEMON and PBS_HOME. These two variables should be updated to match your TORQUE installation. PBS_DAEMON needs to point to the location of trqauthd. PBS_HOME needs to match your TORQUE installation.

Choose the script that matches your dist system and copy it to /etc/init.d. If needed, rename it to trqauthd.

To start the daemon

```
/etc/init.d/trqauthd start
```

To stop the daemon

```
/etc/init.d/trqauthd stop
```
OR

```
[ ] service trqauthd start/stop
```

If you receive an error that says "Could not open socket in trq_simple_connect. error 97" and you use a CentOS, RedHat, or Scientific Linux 6+ operating system, check your `/etc/hosts` file for multiple entries of a single host name pointing to the same IP address. Delete the duplicate(s), save the file, and launch trqauthd again.

Related topics

- [Initializing/Configuring TORQUE on the server (pbs_server)](on page 7)

---

### Finalizing configurations

After configuring the `serverdb` and the `server_priv/nodes` files, and after ensuring minimal MOM configuration, restart the `pbs_server` on the server node and the `pbs_mom` on the compute nodes.

**Compute Nodes:**

```
> pbs_mom
```

**Server Node:**

```
> qterm -t quick
> pbs_server
```

After waiting several seconds, the `pbsnodes -a` command should list all nodes in state `free`.

Related topics

- [Initializing/Configuring TORQUE on the server (pbs_server)](on page 7)

---

### Advanced configuration

This section contains information about how you can customize the installation and configure the server to ensure that the server and nodes are communicating correctly. For details, see these topics:

- [Customizing the install](on page 15)
- [Server configuration](on page 21)

Related topics

- [Server parameters](on page 227)
Customizing the install

The TORQUE configure command has several options available. Listed below are some suggested options to use when running ./configure.

- By default, TORQUE does not install the admin manuals. To enable this, use --enable-docs.
- By default, only children MOM processes use syslog. To enable syslog for all of TORQUE, use --enable-syslog.

Table 1-1: Optional Features

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>--disable-clients</td>
<td>Directs TORQUE not to build and install the TORQUE client utilities such as qsub, qstat, qdel, etc.</td>
</tr>
<tr>
<td>--disable-FEATURE</td>
<td>Do not include FEATURE (same as --enable-FEATURE=no).</td>
</tr>
<tr>
<td>--disable-lib-tool-lock</td>
<td>Avoid locking (might break parallel builds).</td>
</tr>
<tr>
<td>--disable-mom</td>
<td>Do not include the MOM daemon.</td>
</tr>
<tr>
<td>--disable-mom-check-spool</td>
<td>Don’t check free space on spool directory and set an error.</td>
</tr>
<tr>
<td>--disable-posixmemlock</td>
<td>Disable the MOM’s use of mlockall. Some versions of OSs seem to have buggy POSIX MEMLOCK.</td>
</tr>
<tr>
<td>--disable-privports</td>
<td>Disable the use of privileged ports for authentication. Some versions of OSX have a buggy bind () and cannot bind to privileged ports.</td>
</tr>
<tr>
<td>--disable-qsub-keep-override</td>
<td>Do not allow the qsub -k flag to override -o -e.</td>
</tr>
<tr>
<td>--disable-server</td>
<td>Do not include server and scheduler.</td>
</tr>
<tr>
<td>--disable-shell-pipe</td>
<td>Give the job script file as standard input to the shell instead of passing its name via a pipe.</td>
</tr>
<tr>
<td>Option</td>
<td>Description</td>
</tr>
<tr>
<td>----------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>--disable-spool</td>
<td>If disabled, TORQUE will create output and error files directly in $HOME/pbs_spool if it exists or in $HOME otherwise. By default, TORQUE will spool files in TORQUE_HOME/spool and copy them to the users home directory when the job completes.</td>
</tr>
<tr>
<td>--disable-xopen-net-working</td>
<td>With HPUX and GCC, don’t force usage of XOPEN and libxnet.</td>
</tr>
<tr>
<td>--enable-acct-x</td>
<td>Enable adding x attributes to accounting log.</td>
</tr>
<tr>
<td>--enable-array</td>
<td>Setting this under IRIX enables the SGI Origin 2000 parallel support. Normally autodetected from the /etc/config/array file.</td>
</tr>
<tr>
<td>--enable-blcrc</td>
<td>Enable BLCR support.</td>
</tr>
<tr>
<td>--enable-cpa</td>
<td>Enable Cray’s CPA support.</td>
</tr>
<tr>
<td>--enable-cpuset</td>
<td>Enable Linux 2.6 kernel cpusets.</td>
</tr>
<tr>
<td></td>
<td>It is recommended that you turn on this feature to prevent a job from expanding across more CPU cores than it is assigned.</td>
</tr>
<tr>
<td>--enable-debug</td>
<td>Prints debug information to the console for pbs_server and pbs_mom while they are running. (This is different than --with-debug which will compile with debugging symbols)</td>
</tr>
<tr>
<td>--enable-dependency-tracking</td>
<td>Do not reject slow dependency extractors.</td>
</tr>
<tr>
<td>--enable-fast-install[=PKGS]</td>
<td>Optimize for fast installation [default=yes].</td>
</tr>
<tr>
<td>--enable-FEATURE [=ARG]</td>
<td>Include FEATURE [ARG=yes].</td>
</tr>
<tr>
<td>--enable-file-sync</td>
<td>Open files with sync on each write operation. This has a negative impact on TORQUE performance. This is disabled by default.</td>
</tr>
<tr>
<td>--enable-force-nodefile</td>
<td>Forces creation of nodefile regardless of job submission parameters. Not on by default.</td>
</tr>
<tr>
<td>Option</td>
<td>Description</td>
</tr>
<tr>
<td>----------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td><code>--enable-gcc-warnings</code></td>
<td>Enable gcc strictness and warnings. If using gcc, default is to error on any warning.</td>
</tr>
<tr>
<td><code>--enable-geometry-requests</code></td>
<td>TORQUE is compiled to use <code>procs_bitmap</code> during job submission.</td>
</tr>
<tr>
<td><code>--enable-gui</code></td>
<td>Include the GUI-clients.</td>
</tr>
<tr>
<td><code>--enable-main-tainer-mode</code></td>
<td>This is for the autoconf utility and tells autoconf to enable so called rebuild rules. See <a href="#">maintainer mode</a> for more information.</td>
</tr>
<tr>
<td><code>--enable-maxdefault</code></td>
<td>Turn on the RESOURCES_MAXDEFAULT flag.</td>
</tr>
<tr>
<td><code>--enable-nochildsignal</code></td>
<td>Turn on the NO_SIGCHLD flag.</td>
</tr>
<tr>
<td><code>--enable-pemask</code></td>
<td>Enable pemask-based scheduling on the Cray T3e.</td>
</tr>
<tr>
<td><code>--enable-plock-dae-mons[=ARG]</code></td>
<td>Enable daemons to lock themselves into memory: logical-or of 1 for pbs_server, 2 for pbs_scheduler, 4 for pbs_mom (no argument means 7 for all three).</td>
</tr>
<tr>
<td><code>--enable-quick-commit</code></td>
<td>Turn on the QUICKCOMMIT flag.</td>
</tr>
<tr>
<td><code>--enable-shared[=PKGS]</code></td>
<td>Build shared libraries [default=yes].</td>
</tr>
</tbody>
</table>

Version of TORQUE earlier than 2.4.5 attempted to apply queue and server defaults to a job that didn’t have defaults specified. If a setting still did not have a value after that, TORQUE applied the queue and server maximum values to a job (meaning, the maximum values for an applicable setting were applied to jobs that had no specified or default value). In TORQUE 2.4.5 and later, the queue and server maximum values are no longer used as a value for missing settings. To re-enable this behavior in TORQUE 2.4.5 and later, use `--enable-maxdefault`. 

---

**Chapter 1: Overview**
### Table 1-2: Optional packages

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>--with-blc=DIR</td>
<td>BLCR installation prefix (Available in versions 2.5.6 and 3.0.2 and later).</td>
</tr>
<tr>
<td>--with-blc=include=DIR</td>
<td>Include path for libcr.h (Available in versions 2.5.6 and 3.0.2 and later).</td>
</tr>
<tr>
<td>--with-blc=lib=DIR</td>
<td>Lib path for libcr (Available in versions 2.5.6 and 3.0.2 and later).</td>
</tr>
<tr>
<td>--with-blc=bin=DIR</td>
<td>Bin path for BLCR utilities (Available in versions 2.5.6 and 3.0.2 and later).</td>
</tr>
<tr>
<td>--with-cpa=include=DIR</td>
<td>Include path for cpalib.h.</td>
</tr>
<tr>
<td>--with-cpa=lib=DIR</td>
<td>Lib path for libcpalib.</td>
</tr>
<tr>
<td>--with-debug=no</td>
<td>Do not compile with debugging symbols.</td>
</tr>
<tr>
<td>--with-default-server=r=HOSTNAME</td>
<td>Set the name of the computer that clients will access when no machine name is specified as part of the queue name. It defaults to the hostname of the machine on which PBS is being compiled.</td>
</tr>
<tr>
<td>Option</td>
<td>Description</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><code>--with-environ=PATH</code></td>
<td>Set the path containing the environment variables for the daemons. For SP2 and AIX systems, suggested setting is to <code>/etc/environment</code>. Defaults to the file &quot;pbs_environment&quot; in the server-home. Relative paths are interpreted within the context of the server-home.</td>
</tr>
<tr>
<td><code>--with-gnu-ld</code></td>
<td>Assume the C compiler uses GNU ld [default=no].</td>
</tr>
<tr>
<td><code>--with-mail-domain=MAILDOMAIN</code></td>
<td>Override the default domain for outgoing mail messages, i.e. &quot;user@maildomain&quot;. The default maildomain is the hostname where the job was submitted from.</td>
</tr>
<tr>
<td><code>--with-modulefiles=[DIR]</code></td>
<td>Use module files in specified directory [/etc/modulefiles].</td>
</tr>
<tr>
<td><code>--with-momlogdir</code></td>
<td>Use this directory for MOM logs.</td>
</tr>
<tr>
<td><code>--with-momlogsuffix</code></td>
<td>Use this suffix for MOM logs.</td>
</tr>
<tr>
<td><code>--without-PACKAGE</code></td>
<td>Do not use PACKAGE (same as <code>--with-PACKAGE=no</code>).</td>
</tr>
<tr>
<td><code>--without-readline</code></td>
<td>Do not include readline support (default: included if found).</td>
</tr>
<tr>
<td><code>--with-PACKAGE=[ARG]</code></td>
<td>Use PACKAGE [ARG=yes].</td>
</tr>
<tr>
<td><code>--with-pam=DIR</code></td>
<td>Directory that holds the system PAM modules. Defaults to <code>/lib(64)/security</code> on Linux.</td>
</tr>
<tr>
<td><code>--with-pic</code></td>
<td>Try to use only PIC/non-PIC objects [default=use both].</td>
</tr>
<tr>
<td><code>--with-qstatrc-file=FILE</code></td>
<td>Set the name of the file that qstat will use if there is no &quot;.qstatrc&quot; file in the directory where it is being invoked. Relative path names will be evaluated relative to the server home directory (see above). If this option is not specified, the default name for this file will be set to &quot;qstatrc&quot; (no dot) in the server home directory.</td>
</tr>
<tr>
<td><code>--with-rcp</code></td>
<td>One of &quot;scp&quot;, &quot;rcp&quot;, &quot;mom_rcp&quot;, or the full path of a remote file copy program. scp is the default if found, otherwise mom_rcp is used. Some rcp programs don’t always exit with valid error codes in case of failure. mom_rcp is a copy of BSD rcp included with this source that has correct error codes, but it is also old, unmaintained, and doesn’t have large file support.</td>
</tr>
<tr>
<td>Option</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>--with-sched=TYPE</td>
<td>Sets the scheduler type. If TYPE is &quot;c&quot;, the scheduler will be written in C. If TYPE is &quot;tcl&quot; the server will use a Tcl based scheduler. If TYPE is &quot;basl&quot;, TORQUE will use the rule based scheduler. If TYPE is &quot;no&quot;, then no scheduling is done. &quot;c&quot; is the default.</td>
</tr>
<tr>
<td>--with-sched-code=PATH</td>
<td>Sets the name of the scheduler to use. This only applies to BASL schedulers and those written in the C language. For C schedulers this should be a directory name and for BASL schedulers a filename ending in &quot;.basl&quot;. It will be interpreted relative to srctree/src/schedulers.SCHD_TYPE/samples. As an example, an appropriate BASL scheduler relative path would be &quot;nas.basl&quot;. The default scheduler code for &quot;C&quot; schedulers is &quot;fifo&quot;.</td>
</tr>
<tr>
<td>--with-scp</td>
<td>In TORQUE 2.1 and later, SCP is the default remote copy protocol. See --with-rcp if a different protocol is desired.</td>
</tr>
<tr>
<td>--with-sendmail[=FILE]</td>
<td>Sendmail executable to use.</td>
</tr>
<tr>
<td>--with-server-home=DIR</td>
<td>Set the server home/spool directory for PBS use. Defaults to /var/spool/torque.</td>
</tr>
<tr>
<td>--with-server-name-file=e=FILE</td>
<td>Set the file that will contain the name of the default server for clients to use. If this is not an absolute pathname, it will be evaluated relative to the server home directory that either defaults to /usr/spool/PBS or is set using the --with-server-home option to configure. If this option is not specified, the default name for this file will be set to &quot;server_name&quot;.</td>
</tr>
<tr>
<td>--with-tcl</td>
<td>Directory containing td configuration (tclConfig.sh).</td>
</tr>
<tr>
<td>--with-tclattrsep=CHAR</td>
<td>Set the Td attribute separator character this will default to &quot;.&quot; if unspecified.</td>
</tr>
<tr>
<td>--with-tclinclude</td>
<td>Directory containing the public Tcl header files.</td>
</tr>
<tr>
<td>--with-tclx</td>
<td>Directory containing tdx configuration (tdxConfig.sh).</td>
</tr>
<tr>
<td>--with-tk</td>
<td>Directory containing tk configuration (tkConfig.sh).</td>
</tr>
<tr>
<td>--with-tkinclude</td>
<td>Directory containing the public Tk header files.</td>
</tr>
<tr>
<td>--with-tlxx</td>
<td>Directory containing tkx configuration (tkxConfig.sh).</td>
</tr>
<tr>
<td>--with-tmpdir=DIR</td>
<td>Set the tmp directory that pbs_mom will use. Defaults to &quot;/tmp&quot;. This is a Cray-specific feature.</td>
</tr>
<tr>
<td>Option</td>
<td>Description</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>--with-xauth=PATH</td>
<td>Specify path to xauth program.</td>
</tr>
</tbody>
</table>

**HAVE_WORDEXP**

Wordxp() performs a shell-like expansion, including environment variables. By default, HAVE_WORDEXP is set to 1 in src/pbs_config.h. If set to 1, will limit the characters that can be used in a job name to those allowed for a file in the current environment, such as BASH. If set to 0, any valid character for the file system can be used.

If a user would like to disable this feature by setting HAVE_WORDEXP to 0 in src/include/pbs_config.h, it is important to note that the error and the output file names will not expand environment variables, including $PBS_JOBID. The other important consideration is that characters that BASH dislikes, such as (), will not be allowed in the output and error file names for jobs by default.

**Related topics**

- [Advanced configuration](#) on page 14
- [Server configuration](#) on page 21

**Server configuration**

See these topics for details:

- [Server configuration overview](#) on page 21
- [Name service configuration](#) on page 22
- [Configuring job submission hosts](#) on page 22
- [Configuring TORQUE on a multi-homed server](#) on page 23
- [Architecture specific notes](#) on page 23
- [Specifying non-root administrators](#) on page 23
- [Setting up email](#) on page 23
- [Using MUNGE authentication](#) on page 24
- [Setting up the MOM hierarchy](#) on page 25

**Server configuration overview**

There are several steps to ensure that the server and the nodes are completely aware of each other and able to communicate directly. Some of this configuration takes place within TORQUE directly using the qmgr command. Other configuration settings are managed using the pbs_server nodes file, DNS files such as /etc/hosts and the /etc/hosts.equiv file.
Name service configuration

Each node, as well as the server, must be able to resolve the name of every node with which it will interact. This can be accomplished using /etc/hosts, DNS, NIS, or other mechanisms. In the case of /etc/hosts, the file can be shared across systems in most cases.

A simple method of checking proper name service configuration is to verify that the server and the nodes can "ping" each other.

Configuring job submission hosts

Using RCmd authentication

When jobs can be submitted from several different hosts, these hosts should be trusted via the R* commands (such as rsh and rcp). This can be enabled by adding the hosts to the /etc/hosts.equiv file of the machine executing the pbs_server daemon or using other R* command authorization methods. The exact specification can vary from OS to OS (see the man page for ruserok to find out how your OS validates remote users). In most cases, configuring this file is as simple as adding a line to your /etc/hosts.equiv file, as in the following:

/etc/hosts.equiv:

```
# [+ | -] [hostname] [username]
mynode.myorganization.com
.....
```

Either of the hostname or username fields may be replaced with a wildcard symbol (+). The (+) may be used as a stand-alone wildcard but not connected to a username or hostname, e.g., +node01 or +user01. However, a (-) may be used in that manner to specifically exclude a user.

⚠️ Following the Linux man page instructions for hosts.equiv may result in a failure. You cannot precede the user or hostname with a (+). To clarify, node1 +user1 will not work and user1 will not be able to submit jobs.

For example, the following lines will not work or will not have the desired effect:

| +node02 user1
| node02 +user1

These lines will work:

| node03 +
| + jsmith
| node04 -tjones

The most restrictive rules must precede more permissive rules. For example, to restrict user tsmith but allow all others, follow this format:

| node01 -tsmith
| node01 +

Please note that when a hostname is specified, it must be the fully qualified domain name (FQDN) of the host. Job submission can be further secured using the server or queue `acl_hosts` and `acl_host_enabled` parameters (for details, see Queue attributes on page 78).
Using the "submit_hosts" service parameter

Trusted submit host access may be directly specified without using RCmd authentication by setting the server submit_hosts parameter via qmgr as in the following example:

```
> qmgr -c 'set server submit_hosts = host1'
> qmgr -c 'set server submit_hosts += host2'
> qmgr -c 'set server submit_hosts += host3'
```

| Use of submit_hosts is potentially subject to DNS spoofing and should not be used outside of controlled and trusted environments. |

Allowing job submission from compute hosts

If preferred, all compute nodes can be enabled as job submit hosts without setting .rhosts or hosts.equiv by setting the allow_node_submit parameter to true.

Configuring TORQUE on a multi-homed server

If the pbs_server daemon is to be run on a multi-homed host (a host possessing multiple network interfaces), the interface to be used can be explicitly set using the SERVERHOST parameter.

Architecture specific notes

With some versions of Mac OS/X, it is required to add the line $restricted *.<DOMAIN> to the pbs_mom configuration file. This is required to work around some socket bind bugs in the OS.

Specifying non-root administrators

By default, only root is allowed to start, configure and manage the pbs_server daemon. Additional trusted users can be authorized using the parameters managers and operators. To configure these parameters use the qmgr command, as in the following example:

```
> qmgr
Qmgr: set server managers += josh@*.fsc.com
Qmgr: set server operators += josh@*.fsc.com
```

All manager and operator specifications must include a user name and either a fully qualified domain name or a host expression.

To enable all users to be trusted as both operators and administrators, place the + (plus) character on its own line in the server_priv/acl_svr/operators and server_priv/acl_svr/managers files.

Setting up email

Moab relies on emails from TORQUE about job events. To set up email, do the following:
To set up email

1. Use the --with-sendmail configure option at configure time. TORQUE needs to know where the email application is. If this option is not used, TORQUE tries to find the sendmail executable. If it isn’t found, TORQUE cannot send emails.

   ```bash
   > ./configure --with-sendmail=<path_to_executable>
   ```

2. Set mail_domain in your server settings. If your domain is clusterresources.com, execute:

   ```bash
   > qmgr -c 'set server mail_domain=clusterresources.com'
   ```

3. (Optional) You can override the default mail_body_fmt and mail_subject_fmt values via qmgr:

   ```bash
   > qmgr -c 'set server mail_body_fmt=Job: %i
   \n   Name: %j \n   On host: %h \n   %m \n   %d'
   > qmgr -c 'set server mail_subject_fmt=Job %i - %r'
   ```

By default, users receive e-mails on job aborts. Each user can select which kind of e-mails to receive by using the qsub -m option when submitting the job. If you want to dictate when each user should receive e-mails, use a submit filter (for details, see Job submission filter (”qsub wrapper”) on page 293).

Using MUNGE authentication

MUNGE is an authentication service that creates and validates user credentials. It was developed by Lawrence Livermore National Laboratory (LLNL) to be highly scalable so it can be used in large environments such as HPC clusters. To learn more about MUNGE and how to install it, see http://code.google.com/p/munge/.

Configuring TORQUE to use MUNGE is a compile time operation. When you are building TORQUE, use --enable-munge-auth as a command line option with ./configure.

```bash
> ./configure --enable-munge-auth
```

You can use only one authorization method at a time. If --enable-munge-auth is configured, the privileged port ruserok method is disabled.

TORQUE does not link any part of the MUNGE library into its executables. It calls the MUNGE and UNMUNGE utilities which are part of the MUNGE daemon. The MUNGE daemon must be running on the server and all submission hosts. The TORQUE client utilities call MUNGE and then deliver the encrypted credential to pbs_server where the credential is then unmunged and the server verifies the user and host against the authorized users configured in serverdb.

Authorized users are added to serverdb using qmgr and the authorized_users parameter. The syntax for authorized_users is authorized_users=<user>@<host>. To add an authorized user to the server you can use the following qmgr command:

```bash
> qmgr -c 'set server authorized_users=user1@hosta
> qmgr -c 'set server authorized_users=user2@hostb
```

The previous example adds user1 and user2 from hosta to the list of authorized users on the server. Users can be removed from the list of authorized users by using the -= syntax as follows:

```bash
> qmgr -c 'set server authorized_users-=user1@hosta
```
Users must be added with the `<user>@<host>` syntax. The user and the host portion can use the '*' wildcard to allow multiple names to be accepted with a single entry. A range of user or host names can be specified using a [a-b] syntax where `a` is the beginning of the range and `b` is the end.

```
> qmgr -c 'set server authorized_users=user[1-10]@hosta
```

This allows user1 through user10 on hosta to run client commands on the server.

**Setting up the MOM hierarchy**

The MOM hierarchy allows you to override the compute nodes' default behavior of reporting status updates directly to the `pbs_server`. Instead, you configure compute nodes so that each node sends its status update information to another compute node. The compute nodes pass the information up a tree or hierarchy until eventually the information reaches a node that will pass the information directly to `pbs_server`. This can significantly reduce traffic and time required to keep the cluster status up to date.

The name of the file that contains the configuration information is named `mom_hierarchy`. By default, it is located in the `/var/spool/torque/server_priv` directory. The file uses syntax similar to XML:

```
<path>
  <level> comma-separated node list </level>
  <level> comma-separated node list </level>
  ...
</path>
...`

The `<path>` tag pair identifies a group of compute nodes. The `<level>` tag pair contains a comma-separated list of compute node names. Multiple paths can be defined with multiple levels within each path.

Within a `<path>` tag pair the levels define the hierarchy. All nodes in the top level communicate directly with the server. All nodes in lower levels communicate to the first available node in the level directly above it. If the first node in the upper level goes down, the nodes in the subordinate level will then communicate to the next node in the upper level. If no nodes are available in an upper level then the node will communicate directly to the server.

If an upper level node has fallen out and then becomes available, the lower level nodes will eventually find that the node is available and start sending their updates to that node.

If you want to specify MOMs on a different port than the default, you must list the node in the form: `hostname:mom_manager_port`.

For example:

```
<path>
  <level>hostname:mom_manager_port</level>
  ...
</path>
...`

Chapter 1: Overview

Putting the MOM hierarchy on the MOMs

You can put the MOM hierarchy file directly on the MOMs. This way, the pbs_server doesn't have to send the hierarchy to all the MOMs during each pbs_server startup. The hierarchy file still has to exist on the pbs_server and if the file versions conflict, the pbs_server version overwrites the local MOM file. Due to this, it is recommended that the hierarchy file either be symlinked to the MOMs or put on a global NFS share.

Once the hierarchy file exists on the MOMs, start pbs_server with the -n option which tells pbs_server to not send the hierarchy file on startup. Instead, pbs_server waits until a MOM requests it.

Related topics

- Advanced configuration on page 14

Manual setup of initial server configuration

On a new installation of TORQUE, the server database must be initialized using the command pbs_server -t create. This command creates a file in $TORQUEHOME/server_priv named serverdb which contains the server configuration information.

The following output from qmgr shows the base configuration created by the command pbs_server -t create:

```plaintext
# Set server attributes.
# set server acl_hosts = kmn
set server log_events = 511
set server mail_from = adm
set server node_check_rate = 150
set server tcp_timeout = 6
```

This is a bare minimum configuration and it is not very useful. By using qmgr, the server configuration can be modified to set up TORQUE to do useful work. The following qmgr commands will create a queue and enable the server to accept and run jobs. These commands must be executed by root.

```plaintext
pbs_server -t create
qmgr -c "set server scheduling=true"
qmgr -c "create queue batch queue_type=execution"
qmgr -c "set queue batch started=true"
qmgr -c "set queue batch enabled=true"
qmgr -c "set queue batch resources_default.nodes=1"
qmgr -c "set queue batch resources_default.walltime=3600"
qmgr -c "set server default_queue=batch"
```

When TORQUE reports a new queue to Moab a class of the same name is automatically applied to all nodes.

In this example, the configuration database is initialized and the scheduling interface is activated (using 'scheduling=true'). This option allows the scheduler to receive job and node events which allow it to be more responsive (See scheduling on page 243 for more information). The next command creates a
queue and specifies the queue type. Within PBS, the queue must be declared an ‘execution queue in order for it to run jobs. Additional configuration (i.e., setting the queue to started and enabled) allows the queue to accept job submissions, and launch queued jobs.

The next two lines are optional, setting default node and walltime attributes for a submitted job. These defaults will be picked up by a job if values are not explicitly set by the submitting user. The final line, default_queue=batch, is also a convenience line and indicates that a job should be placed in the batch queue unless explicitly assigned to another queue.

Additional information on configuration can be found in the admin manual and in the qmgr main page.

Related topics
- TORQUE installation overview on page 1

Server node file configuration

This section contains information about configuring server node files. It explains how to specify node virtual processor counts and GPU counts, as well as how to specify node features or properties. For details, see these topics:

- Basic node specification on page 27
- Specifying virtual processor count for a node on page 28
- Specifying GPU count for a node on page 28
- Specifying node features (node properties) on page 29

Related topics
- TORQUE installation overview on page 1
- Server parameters on page 227
- Moab node feature overview

Basic node specification

For the pbs_server to communicate with each of the MOMs, it needs to know which machines to contact. Each node that is to be a part of the batch system must be specified on a line in the server nodes file. This file is located at TORQUE_HOME/server_priv/nodes. In most cases, it is sufficient to specify just the node name on a line as in the following example:

server_priv/nodes:
```
node001
node002
node003
node004
```
Specifying virtual processor count for a node

By default each node has one virtual processor. Increase the number using the `np` attribute in the nodes file. The value of `np` can be equal to the number of physical cores on the node or it can be set to a value which represents available "execution slots" for the node. The value used is determined by the administrator based on hardware, system, and site criteria.

The following example shows how to set the `np` value in the nodes file. In this example, we are assuming that node001 and node002 have four physical cores. The administrator wants the value of `np` for node001 to reflect that it has four cores. However, node002 will be set up to handle multiple virtual processors without regard to the number of physical cores on the system.

```plaintext
server_priv/nodes:

node001 np=4
node002 np=12
...
```

Specifying GPU count for a node

Administrators can manually set the number of GPUs on a node or if they are using NVIDIA GPUs and drivers, they can have them detected automatically. For more information about how to set up TORQUE with GPUs, see the Moab Workload Manager accelerators documentation.

To manually set the number of GPUs on a node, use the `gpus` attribute in the nodes file. The value of GPUs is determined by the administrator based on hardware, system, and site criteria.

The following example shows how to set the `gpus` value in the nodes file. In the example, we assume node001 and node002 each have two physical GPUs. The administrator wants the value of node001 to reflect the physical GPUs available on that system and adds `gpus=2` to the nodes file entry for node001. However, node002 will be set up to handle multiple virtual GPUs without regard to the number of physical GPUs on the system.

```plaintext
server_priv/nodes:

node001 gpus=1
node002 gpus=4
...
```
Specifying node features (node properties)

Node features can be specified by placing one or more white space-delimited strings on the line for the associated host as in the following example:

```
server_priv/nodes:

| node001 np=2 fast ia64
| node002 np=4 bigmem fast ia64 smp
| ...
```

These features can be used by users to request specific nodes when submitting jobs. For example:

```
qu -l nodes=1:bigmem+1:fast job.sh
```

This job submission will look for a node with the bigmem feature (node002) and a node with the fast feature (either node001 or node002).

Related topics

- [Server node file configuration](#) on page 27

Testing server configuration

If you have initialized TORQUE using the torque.setup script or started TORQUE using pbs_server -t create and pbs_server is still running, terminate the server by calling qterm. Next, start pbs_server again without the -t create arguments. Follow the script below to verify your server configuration. The output for the examples below is based on the nodes file example in Specifying node features and Server configuration.
# verify all queues are properly configured
> qstat -q

server:kmn

<table>
<thead>
<tr>
<th>Queue</th>
<th>Memory</th>
<th>CPU Time</th>
<th>Walltime</th>
<th>Node</th>
<th>Run</th>
<th>Que</th>
<th>Lm</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>batch</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>0</td>
<td>0</td>
<td>--</td>
<td>ER</td>
</tr>
</tbody>
</table>

# view additional server configuration
> qmgr -c 'p s'

# Create queues and set their attributes

# Create and define queue batch
(create queue batch
set queue batch queue_type = Execution
set queue batch resources_default.nodes = 4
set queue batch resources_default.walltime = 01:00:00
set queue batch enabled = True
set queue batch started = True

# Set server attributes.
(set server scheduling = True
set server acl_hosts = kmn
set server managers = user1@knn
set server operators = user1@knn
set server log_events = 511
set server mail_from = adm
set server node_check_rate = 150
set server tcp_timeout = 300
set server job_stat_rate = 45
set server poll_jobs = True
set server mom_job_sync = True
set server keep_completed = 300
set server next_job_number = 0

# verify all nodes are correctly reporting
> pbsnodes -a

node001
  state=free	np=2
  properties=bigmem,fast,ia64,smp
  ntype=cluster
  status=rectime=1328810402,univar,netload=6814326158,gres,,loadave =0.21,ncpus=6,physmem=8193724kb,
availmem=13922548kb,totmem=16581304kb,loadtime=3,nusers=3,nsessions=18,sessions=1876
  1120 1912 1926 1937 1951 2019 2057 28399 2126 2140 2323 5419 17948 19356 27726 22254
  29569,uname=Linux kmn 2.6.38-11-generic #48-Ubuntu SMP Fri Jul 29 19:02:55 UTC 2011
  x86_64,opsys=linux
  mom_service_port = 15002
  mom_manager_port = 15003
  npus = 0

# submit a basic job - DO NOT RUN AS ROOT
> su - testuser
> echo "sleep 30" | qsub

# verify jobs display
> qstat

Job id | Name | User | Time Use | S | Queue
-------|------|------|----------|---|-------
At this point, the job should be in the Q state and will not run because a scheduler is not running yet. TORQUE can use its native scheduler by running `pbs_sched` or an advanced scheduler (such as Moab Workload Manager). See Integrating schedulers for details on setting up an advanced scheduler.

Related topics
- TORQUE installation overview on page 1

TORQUE on NUMA systems

Starting in TORQUE version 3.0, TORQUE can be configured to take full advantage of Non-Uniform Memory Architecture (NUMA) systems. The following instructions are a result of development on SGI Altix and UV hardware.

For details, see these topics:
- TORQUE NUMA configuration on page 31
- Building TORQUE with NUMA support on page 31

TORQUE NUMA configuration

There are three steps to configure TORQUE to take advantage of NUMA architectures:

1. Configure TORQUE with `--enable-numa-support`.
2. Create the `mom_priv/mom.layout` file.
3. Configure `server_priv/nodes`.

Related topics
- TORQUE on NUMA systems on page 31

Building TORQUE with NUMA support

To turn on NUMA support for TORQUE the `--enable-numa-support` option must be used during the configure portion of the installation. In addition to any other configuration options, add the `--enable-numa-support` option as indicated in the following example:

```
$ ./configure --enable-numa-support
```

Don't use MOM hierarchy with NUMA.

When TORQUE is enabled to run with NUMA support, there is only a single instance of `pbs_mom` (MOM) that is run on the system. However, TORQUE will report that there are multiple nodes running in the
cluster. While `pbs_mom` and `pbs_server` both know there is only one instance of `pbs_mom`, they manage the cluster as if there were multiple separate MOM nodes.

The `mom.layout` file is a virtual mapping between the system hardware configuration and how the administrator wants TORQUE to view the system. Each line in `mom.layout` equates to a node in the cluster and is referred to as a NUMA node.

**Automatically Creating mom.layout (Recommended)**

A perl script named `mom_gencfg` is provided in the `contrib/` directory that generates the `mom.layout` file for you. The script can be customized by setting a few variables in it. To automatically create the `mom.layout` file, follow these instructions (these instructions are also included in the script):

1. Verify hwloc version 1.1 or higher is installed - see contrib/hwloc_install.sh.
2. Install `Sys::Hwloc` from CPAN.
3. Verify `$PBS_HOME` is set to the proper value.
4. Update the variables in the 'Config Definitions' section of the script. Especially update `firstNodeId` and `nodesPerBoard` if desired. The `firstNodeId` variable should be set above 0 if you have a root cpuset that you wish to exclude and the `nodesPerBoard` variable is the number of NUMA nodes per board. Each node is defined in `/sys/devices/system/node`, in a subdirectory `node<node index>`
5. Back up your current file in case a variable is set incorrectly or neglected.
6. Run the script.
7. `./mom_gencfg`  

**Manually Creating mom.layout**

To properly set up the `mom.layout` file, it is important to know how the hardware is configured. Use the `topology` command line utility and inspect the contents of `/sys/devices/system/node`. The hwloc library can also be used to create a custom discovery tool.

Typing `topology` on the command line of a NUMA system produces something similar to the following:
Chapter 1: Overview

### Partition number: 0
6 Blades
72 CPUs
378.43 GB Memory Total

<table>
<thead>
<tr>
<th>Blade</th>
<th>ID</th>
<th>ASIC</th>
<th>NASID</th>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 r001i01b00</td>
<td>UVHub 1.0</td>
<td>0</td>
<td>67089152 kB</td>
<td></td>
</tr>
<tr>
<td>1 r001i01b01</td>
<td>UVHub 1.0</td>
<td>2</td>
<td>67092480 kB</td>
<td></td>
</tr>
<tr>
<td>2 r001i01b02</td>
<td>UVHub 1.0</td>
<td>4</td>
<td>67092480 kB</td>
<td></td>
</tr>
<tr>
<td>3 r001i01b03</td>
<td>UVHub 1.0</td>
<td>6</td>
<td>67092480 kB</td>
<td></td>
</tr>
<tr>
<td>4 r001i01b04</td>
<td>UVHub 1.0</td>
<td>8</td>
<td>67092480 kB</td>
<td></td>
</tr>
<tr>
<td>5 r001i01b05</td>
<td>UVHub 1.0</td>
<td>10</td>
<td>67092480 kB</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CPU</th>
<th>Blade</th>
<th>PhysID</th>
<th>CoreID</th>
<th>APIC-ID</th>
<th>Family</th>
<th>Model</th>
<th>Speed</th>
<th>L1(KiB)</th>
<th>L2(KiB)</th>
<th>L3(KiB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>r001i01b00</td>
<td>00</td>
<td>00</td>
<td>0</td>
<td>6</td>
<td>46</td>
<td>2666</td>
<td>32d/32i</td>
<td>256</td>
<td>18432</td>
</tr>
<tr>
<td>1</td>
<td>r001i01b00</td>
<td>00</td>
<td>02</td>
<td>4</td>
<td>6</td>
<td>46</td>
<td>2666</td>
<td>32d/32i</td>
<td>256</td>
<td>18432</td>
</tr>
<tr>
<td>2</td>
<td>r001i01b00</td>
<td>00</td>
<td>03</td>
<td>6</td>
<td>6</td>
<td>46</td>
<td>2666</td>
<td>32d/32i</td>
<td>256</td>
<td>18432</td>
</tr>
<tr>
<td>3</td>
<td>r001i01b00</td>
<td>00</td>
<td>08</td>
<td>16</td>
<td>6</td>
<td>46</td>
<td>2666</td>
<td>32d/32i</td>
<td>256</td>
<td>18432</td>
</tr>
<tr>
<td>4</td>
<td>r001i01b00</td>
<td>00</td>
<td>09</td>
<td>18</td>
<td>6</td>
<td>46</td>
<td>2666</td>
<td>32d/32i</td>
<td>256</td>
<td>18432</td>
</tr>
<tr>
<td>5</td>
<td>r001i01b00</td>
<td>00</td>
<td>11</td>
<td>22</td>
<td>6</td>
<td>46</td>
<td>2666</td>
<td>32d/32i</td>
<td>256</td>
<td>18432</td>
</tr>
<tr>
<td>6</td>
<td>r001i01b00</td>
<td>01</td>
<td>00</td>
<td>32</td>
<td>6</td>
<td>46</td>
<td>2666</td>
<td>32d/32i</td>
<td>256</td>
<td>18432</td>
</tr>
<tr>
<td>7</td>
<td>r001i01b00</td>
<td>01</td>
<td>02</td>
<td>36</td>
<td>6</td>
<td>46</td>
<td>2666</td>
<td>32d/32i</td>
<td>256</td>
<td>18432</td>
</tr>
<tr>
<td>8</td>
<td>r001i01b00</td>
<td>01</td>
<td>03</td>
<td>38</td>
<td>6</td>
<td>46</td>
<td>2666</td>
<td>32d/32i</td>
<td>256</td>
<td>18432</td>
</tr>
<tr>
<td>9</td>
<td>r001i01b00</td>
<td>01</td>
<td>08</td>
<td>48</td>
<td>6</td>
<td>46</td>
<td>2666</td>
<td>32d/32i</td>
<td>256</td>
<td>18432</td>
</tr>
<tr>
<td>10</td>
<td>r001i01b00</td>
<td>01</td>
<td>09</td>
<td>50</td>
<td>6</td>
<td>46</td>
<td>2666</td>
<td>32d/32i</td>
<td>256</td>
<td>18432</td>
</tr>
<tr>
<td>11</td>
<td>r001i01b00</td>
<td>01</td>
<td>11</td>
<td>54</td>
<td>6</td>
<td>46</td>
<td>2666</td>
<td>32d/32i</td>
<td>256</td>
<td>18432</td>
</tr>
<tr>
<td>12</td>
<td>r001i01b00</td>
<td>02</td>
<td>00</td>
<td>64</td>
<td>6</td>
<td>46</td>
<td>2666</td>
<td>32d/32i</td>
<td>256</td>
<td>18432</td>
</tr>
<tr>
<td>13</td>
<td>r001i01b01</td>
<td>02</td>
<td>02</td>
<td>68</td>
<td>6</td>
<td>46</td>
<td>2666</td>
<td>32d/32i</td>
<td>256</td>
<td>18432</td>
</tr>
<tr>
<td>14</td>
<td>r001i01b01</td>
<td>02</td>
<td>03</td>
<td>70</td>
<td>6</td>
<td>46</td>
<td>2666</td>
<td>32d/32i</td>
<td>256</td>
<td>18432</td>
</tr>
</tbody>
</table>

From this partial output, note that this system has 72 CPUs on 6 blades. Each blade has 12 CPUs grouped into clusters of 6 CPUs. If the entire content of this command were printed you would see each Blade ID and the CPU ID assigned to each blade.

The topology command shows how the CPUs are distributed, but you likely also need to know where memory is located relative to CPUs, so go to `/sys/devices/system/node`. If you list the node directory you will see something similar to the following:

```
# ls -al
total 0
drwxr-xr-x   14 root      0 Dec 3 12:14 .
-r--------   14 root      0 Dec 3 12:13 ..
-r--------   1 root       4096 Dec 3 14:58 has_cpu
-r--------   1 root       4096 Dec 3 14:58 has_normal_memory
drwxr-xr-x   2 root      0 Dec 3 12:14 node0
drwxr-xr-x   2 root      0 Dec 3 12:14 node1
drwxr-xr-x   2 root      0 Dec 3 12:14 node10
drwxr-xr-x   2 root      0 Dec 3 12:14 node11
```

The directory entries node0, node1,...node11 represent groups of memory and CPUs local to each other. These groups are a node board, a grouping of resources that are close together. In most cases, a node board is made up of memory and processor cores. Each bank of memory is called a memory node by the
operating system, and there are certain CPUs that can access that memory very rapidly. Note under the
directory for node board node0 that there is an entry called cpulist. This contains the CPU IDs of all
CPUs local to the memory in node board 0.

Now create the mom.layout file. The content of cpulist 0-5 are local to the memory of node board 0,
and the memory and cpus for that node are specified in the layout file by saying nodes=0. The cpulist for
node board 1 shows 6-11 and memory node index 1. To specify this, simply write nodes=1. Repeat this for
all twelve node boards and create the following mom.layout file for the 72 CPU system.

```
| nodes=0 |
| nodes=1 |
| nodes=2 |
| nodes=3 |
| nodes=4 |
| nodes=5 |
| nodes=6 |
| nodes=7 |
| nodes=8 |
| nodes=9 |
| nodes=10 |
| nodes=11 |
```

Each line in the mom.layout file is reported as a node to pbs_server by the pbs_mom daemon.

The mom.layout file does not need to match the hardware layout exactly. It is possible to combine node
boards and create larger NUMA nodes. The following example shows how to do this:

```
| nodes=0-1 |
```

The memory nodes can be combined the same as CPUs. The memory nodes combined must be contiguous.
You cannot combine mem 0 and 2.

Configuring server_priv/nodes

The pbs_server requires awareness of how the MOM is reporting nodes since there is only one MOM
daemon and multiple MOM nodes. So, configure the server_priv/nodes file with the num_node_boards
and numa_board_str attributes. The attribute num_node_boards tells pbs_server how many numa
nodes are reported by the MOM. Following is an example of how to configure the nodes file with num_node_boards:

```
| numa-10 np=72 num_node_boards=12 |
```

This line in the nodes file tells pbs_server there is a host named numa-10 and that it has 72 processors
and 12 nodes. The pbs_server divides the value of np (72) by the value for num_node_boards (12) and
determines there are 6 CPUs per NUMA node.

In this example, the NUMA system is uniform in its configuration of CPUs per node board, but a system
does not need to be configured with the same number of CPUs per node board. For systems with non-
uniform CPU distributions, use the attribute numa_board_str to let pbs_server know where CPUs are
located in the cluster.

The following is an example of how to configure the server_priv/nodes file for non-uniformly
distributed CPUs:

```
| Numa-11 numa_board_str=6,8,12 |
```
In this configuration, `pbs_server` knows it has three MOM nodes and the nodes have 6, 8, and 12 CPUs respectively. Note that the attribute np is not used. The np attribute is ignored because the number of CPUs per node is expressly given.

**Enforcement of memory resource limits**

TORQUE can better enforce memory limits with the use of the utility `memacctd`. The memacctd utility is provided by SGI on SuSe Linux Enterprise Edition (SLES). It is a daemon that caches memory footprints when it is queried. When configured to use the memory monitor, TORQUE queries memacctd. It is up to the user to make sure memacctd is installed. See the [SGI memacctd man page](https://www.sgi.com/tech/stuff/standards/man/memacctd.html) for more information.

**To configure TORQUE to use memacctd for memory enforcement**

1. Start `memacctd` as instructed by SGI.
2. Reconfigure TORQUE with `--enable-memacct`. This will link in the necessary library when TORQUE is recompiled.
3. Recompile and reinstall TORQUE.
4. Restart all MOM nodes.
5. (Optional) Alter the `qsub` filter to include a default memory limit for all jobs that are not submitted with memory limit.

**Related topics**

- [TORQUE NUMA configuration](#)
- [TORQUE on NUMA systems](#)

**TORQUE Multi-MOM**

Starting in TORQUE version 3.0 users can run multiple MOMs on a single node. The initial reason to develop a multiple MOM capability was for testing purposes. A small cluster can be made to look larger since each MOM instance is treated as a separate node.

When running multiple MOMs on a node each MOM must have its own service and manager ports assigned. The default ports used by the MOM are 15002 and 15003. With the multi-mom alternate ports can be used without the need to change the default ports for `pbs_server` even when running a single instance of the MOM.

For details, see these topics:

- [Multi-MOM configuration](#)
- [Stopping pbs_mom in Multi-MOM mode](#)

**Multi-MOM configuration**

There are three steps to setting up multi-MOM capability:
1. **Configure server_priv/nodes** on page 36
2. **/etc/hosts** on page 36
3. **Starting pbs_mom with multi-MOM options** on page 36

Configure server_priv/nodes

The attributes *mom_service_port* and *mom_manager_port* were added to the nodes file syntax to accommodate multiple MOMs on a single node. By default pbs_mom opens ports 15002 and 15003 for the service and management ports respectively. For multiple MOMs to run on the same IP address they need to have their own port values so they can be distinguished from each other. pbs_server learns about the port addresses of the different MOMs from entries in the *server_priv/nodes* file. The following is an example of a nodes file configured for multiple MOMs:

```
hosta  np=2
hosta-1 np=2 mom_service_port=30001 mom_manager_port=30002
hosta-2 np=2 mom_service_port=31001 mom_manager_port=31002
hosta-3 np=2 mom_service_port=32001 mom_manager_port=32002
```

Note that all entries have a unique host name and that all port values are also unique. The entry hosta does not have a mom_service_port or mom_manager_port given. If unspecified, then the MOM defaults to ports 15002 and 15003.

/etc/hosts file

Host names in the *server_priv/nodes* file must be resolvable. Creating an alias for each host enables the server to find the IP address for each MOM; the server uses the port values from the *server_priv/nodes* file to contact the correct MOM. An example /etc/hosts entry for the previous *server_priv/nodes* example might look like the following:

```
[192.65.73.10 hosta hosta-1 hosta-2 hosta-3]
```

Even though the host name and all the aliases resolve to the same IP address, each MOM instance can still be distinguished from the others because of the unique port value assigned in the *server_priv/nodes* file.

Starting pbs_mom with multi-MOM options

To start multiple instances of pbs_mom on the same node, use the following syntax (see pbs_mom on page 151 for details):

```
# pbs_mom -m -M <port value of MOM_service_port> -R <port value of MOM_manager_port> -A <name of MOM alias>
```

Continuing based on the earlier example, if you want to create four MOMs on hosta, type the following at the command line:

```
# pbs_mom -m -M 30001 -R 30002 -A hosta-1
# pbs_mom -m -M 31001 -R 31002 -A hosta-2
# pbs_mom -m -M 32001 -R 32002 -A hosta-3
# pbs_mom
```

Notice that the last call to pbs_mom uses no arguments. By default pbs_mom opens on ports 15002 and 15003. No arguments are necessary because there are no conflicts.
Related topics

- TORQUE Multi-MOM on page 35
- Stopping pbs_mom in Multi-MOM mode on page 37

Stopping pbs_mom in Multi-MOM mode

Terminate pbs_mom by using the momctl -s command (for details, see momctl). For any MOM using the default manager port 15003, the momctl -s command stops the MOM. However, to terminate MOMs with a manager port value not equal to 15003, you must use the following syntax:

```
  momctl -s -p <port value of MOM_manager_port>
```

The -p option sends the terminating signal to the MOM manager port and the MOM is terminated.

Related topics

- TORQUE Multi-MOM on page 35
- Multi-MOM configuration on page 35
Chapter 2: Submitting and managing jobs

This section contains information about how you can submit and manage jobs with TORQUE. For details, see the following topics:

- **Job submission** on page 39
- **Monitoring jobs** on page 51
- **Canceling jobs** on page 51
- **Job preemption** on page 52
- **Keeping completed jobs** on page 52
- **Job checkpoint and restart** on page 53
- **Job exit status** on page 63
- **Service jobs** on page 67

## Job submission

Job submission is accomplished using the **`qsub`** command, which takes a number of command line arguments and integrates such into the specified PBS command file. The PBS command file may be specified as a filename on the `qsub` command line or may be entered via STDIN.

- The PBS command file does not need to be executable.
- The PBS command file may be **piped** into `qsub` (i.e., `cat pbs.cmd | qsub`).
- In the case of parallel jobs, the PBS command file is staged to, and executed on, the first allocated compute node only. (Use `pbsdsh` to run actions on multiple nodes.)
- The command script is executed from the user’s home directory in all cases. (The script may determine the submission directory by using the `$PBS_O_WORKDIR` environment variable)
- The command script will be executed using the default set of user environment variables unless the `-V` or `-v` flags are specified to include aspects of the job submission environment.
- PBS directives should be declared first in the job script.

```bash
#PBS -S /bin/bash
#PBS -m abe
#PBS -M <yourEmail@company.com>
echo sleep 300
```

*This is an example of properly declared PBS directives.*
Multiple job submission

Sometimes users will want to submit large numbers of jobs based on the same job script. Rather than using a script to repeatedly call qsub, a feature known as job arrays now exists to allow the creation of multiple jobs with one qsub command. Additionally, this feature includes a new job naming convention
that allows users to reference the entire set of jobs as a unit, or to reference one particular job from the set.

Job arrays are submitted through the -t option to qsub, or by using #PBS -t in your batch script. This option takes a comma-separated list consisting of either a single job ID number, or a pair of numbers separated by a dash. Each of these jobs created will use the same script and will be running in a nearly identical environment.

```
> qsub -t 0-4 job_script
1098[].hostname
> qstat -t
1098[0].hostname ...
1098[1].hostname ...
1098[2].hostname ...
1098[3].hostname ...
1098[4].hostname ...
```

Versions of TORQUE earlier than 2.3 had different semantics for the -t argument. In these versions, -t took a single integer number—a count of the number of jobs to be created.

Each 1098[x] job has an environment variable called PBS_ARRAYID, which is set to the value of the array index of the job, so 1098[0].hostname would have PBS_ARRAYID set to 0. This allows you to create job arrays where each job in the array performs slightly different actions based on the value of this variable, such as performing the same tasks on different input files. One other difference in the environment between jobs in the same array is the value of the PBS_JOBNAME variable.

```
# These two examples are equivalent in TORQUE 2.2
> qsub -t 0-99
> qsub -t 100

# You can also pass comma delimited lists of ids and ranges:
> qsub -t 0,10,20,30,40
> qsub -t 0-50,60,70,80
```

Running qstat displays a job summary, which provides an overview of the array's state. To see each job in the array, run qstat -t.

The qalter, qdel, qhold, and qrls commands can operate on arrays—either the entire array or a range of that array. Additionally, any job in the array may be accessed normally by using that job's ID, just as you would with any other job. For example, running the following command would run only the specified job:

```
qrun 1098[0].hostname
```

Slot Limit

The slot limit is a way for administrators to limit the number of jobs from a job array that can be eligible for scheduling at the same time. When a slot limit is used, TORQUE puts a hold on all jobs in the array that exceed the slot limit. When an eligible job in the array completes, TORQUE removes the hold flag from the next job in the array. Slot limits can be declared globally with the max_slot_limit parameter, or on a per-job basis with qsub -t.
Managing multi-node jobs

By default, when a multi-node job runs, the Mother Superior manages the job across all the sister nodes by communicating with each of them and updating `pbs_server`. Each of the sister nodes sends its updates and stdout and stderr directly to the Mother Superior. When you run an extremely large job using hundreds or thousands of nodes, you may want to reduce the amount of network traffic sent from the sisters to the Mother Superior by specifying a job radix. Job radix sets a maximum number of nodes with which the Mother Superior and resulting intermediate MOMs communicate and is specified using the `-W on page 215` option for `qsub`.

For example, if you submit a smaller, 12-node job and specify `job_radix=3`, Mother Superior and each resulting intermediate MOM is only allowed to receive communication from 3 subordinate nodes.

Image 2-1: Job radix example

The Mother Superior picks three sister nodes with which to communicate the job information. Each of those nodes (intermediate MOMs) receives a list of all sister nodes that will be subordinate to it. They each contact up to three nodes and pass the job information on to those nodes. This pattern continues until the bottom level is reached. All communication is now passed across this new hierarchy. The stdout and stderr data is aggregated and sent up the tree until it reaches the Mother Superior, where it is saved and copied to the `.o` and `.e` files.

```
Job radix is meant for extremely large jobs only. It is a tunable parameter and should be adjusted according to local conditions in order to produce the best results.
```

Requesting resources

Various resources can be requested at the time of job submission. A job can request a particular node, a particular node attribute, or even a number of nodes with particular attributes. Either native TORQUE resources or external scheduler resource extensions may be specified. The native TORQUE resources are listed in the following table:
<table>
<thead>
<tr>
<th>Resource</th>
<th>Format</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>arch</td>
<td>string</td>
<td>Specifies the administrator defined system architecture required. This defaults to whatever the <strong>PBS_MACH</strong> string is set to in &quot;local.mk&quot;.</td>
</tr>
<tr>
<td>cput</td>
<td>seconds, or [HH:]MM:[SS]</td>
<td>Maximum amount of CPU time used by all processes in the job.</td>
</tr>
<tr>
<td>epilogue</td>
<td>string</td>
<td>Specifies a user owned epilogue script which will be run before the system epilogue and epilogueuser scripts at the completion of a job. The syntax is <code>epilogue=&lt;file&gt;</code>. The file can be designated with an absolute or relative path. For more information, see <a href="#">Prologue and epilogue scripts on page 281</a>.</td>
</tr>
<tr>
<td>feature</td>
<td>string</td>
<td>Specifies a property or feature for the job. Feature corresponds to TORQUE node properties and Moab features.</td>
</tr>
<tr>
<td>file</td>
<td>size</td>
<td>The amount of total disk requested for the job. (Ignored on Unicos.)</td>
</tr>
<tr>
<td>host</td>
<td>string</td>
<td>Name of the host on which the job should be run. This resource is provided for use by the site’s scheduling policy. The allowable values and effect on job placement is site dependent.</td>
</tr>
<tr>
<td>mem</td>
<td>size</td>
<td>Maximum amount of physical memory used by the job. (Ignored on Darwin, Digital Unix, Free BSD, HPUX 11, IRIX, NetBSD, and SunOS. Also ignored on Linux if number of nodes is not 1. Not implemented on AIX and HPUX 10.)</td>
</tr>
<tr>
<td>nice</td>
<td>integer</td>
<td>Number between -20 (highest priority) and 19 (lowest priority). Adjust the process execution priority.</td>
</tr>
</tbody>
</table>
## Resource Specification

<table>
<thead>
<tr>
<th>Resource</th>
<th>Format</th>
<th>Description</th>
</tr>
</thead>
</table>
| nodes    | \(<node\_count>\ [\<hostname>]\) [\<ppn>=<ppn>] [\<gpus>=<gpu>] [\<property>\ [\<property>\+] \+] | Number and/or type of nodes to be reserved for exclusive use by the job. The value is one or more node\_specs joined with the + (plus) character: node\_spec [\+node\_spec\_]. Each node\_spec is a number of nodes required of the type declared in the node\_spec and a name of one or more properties desired for the nodes. The number, the name, and each property in the node\_spec are separated by a : (colon). If no number is specified, one (1) is assumed. The name of a node is its hostname. The properties of nodes are:  
  - **ppn=#** - Specify the number of virtual processors per node requested for this job.
    The number of virtual processors available on a node by default is 1, but it can be configured in the $TORQUE\_HOME/server\_priv/nodes file using the np attribute (see Server node file configuration on page 27). The virtual processor can relate to a physical core on the node or it can be interpreted as an "execution slot" such as on sites that set the node np value greater than the number of physical cores (or hyper-thread contexts). The ppn value is a characteristic of the hardware, system, and site, and its value is to be determined by the administrator.
  - **gpus=#** - Specify the number of GPUs per node requested for this job.
    The number of GPUs available on a node can be configured in the $TORQUE\_HOME/server\_priv/nodes file using the gpu attribute (see Server node file configuration on page 27). The GPU value is a characteristic of the hardware, system, and site, and its value is to be determined by the administrator.
  - **property** - A string assigned by the system administrator specifying a node’s features. Check with your administrator as to the node names and properties available to you. |
| opsys    | string | Specifies the administrator defined operating system as defined in the MOM configuration file. |

TORQUE does not have a TPN (tasks per node) property. You can specify TPN in Moab Workload Manager with TORQUE as your resource manager, but TORQUE does not recognize the property when it is submitted directly to it via qsub.

See qsub -I nodes on page 46 for examples.

By default, the node resource is mapped to a virtual node (that is, directly to a processor, not a full physical compute node). This behavior can be changed within Maui or Moab by setting the JOBNODEMATCHPOLICY parameter. See "Appendix F: Parameters" of the Moab Workload Manager Administrator Guide for more information.
<table>
<thead>
<tr>
<th>Resource</th>
<th>Format</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>other</td>
<td>string</td>
<td>Allows a user to specify site specific information. This resource is provided for use by the site's scheduling policy. The allowable values and effect on job placement is site dependent.</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>This does not work for msub using Moab and Maui.</em></td>
</tr>
<tr>
<td>pcput</td>
<td>seconds, or [[HH:]MM:]SS</td>
<td>Maximum amount of CPU time used by any single process in the job.</td>
</tr>
<tr>
<td>pmem</td>
<td>size</td>
<td>Maximum amount of physical memory used by any single process of the job. (Ignored on Fujitsu. Not implemented on Digital Unix and HPUX.)</td>
</tr>
<tr>
<td>procs</td>
<td>procs=&lt;integer&gt;</td>
<td><em>(Applicable in version 2.5.0 and later.)</em> The number of processors to be allocated to a job. The processors can come from one or more qualified node(s). Only one procs declaration may be used per submitted <code>qsub</code> command. &gt; qsub -l nodes=3 -l procs=2</td>
</tr>
<tr>
<td>procs_bitmap</td>
<td>string</td>
<td>A string made up of 1's and 0's in reverse order of the processor cores requested. A <code>procs_bitmap=1110</code> means the job requests a node that has four available cores, but the job runs exclusively on cores two, three, and four. With this bitmap, core one is not used. For more information, see <em>Scheduling cores on page 74.</em></td>
</tr>
<tr>
<td>prologue</td>
<td>string</td>
<td>Specifies a user owned prologue script which will be run after the system prologue and prologueuser scripts at the beginning of a job. The syntax is <code>prologue=&lt;file&gt;</code>. The file can be designated with an absolute or relative path. For more information, see <em>Prologue and epilogue scripts on page 281.</em></td>
</tr>
<tr>
<td>pvmem</td>
<td>size</td>
<td>Maximum amount of virtual memory used by any single process in the job. (Ignored on Unicos.)</td>
</tr>
<tr>
<td>size</td>
<td>integer</td>
<td>For TORQUE, this resource has no meaning. It is passed on to the scheduler for interpretation. In the Moab scheduler, the size resource is intended for use in Cray installations only.</td>
</tr>
<tr>
<td>software</td>
<td>string</td>
<td>Allows a user to specify software required by the job. This is useful if certain software packages are only available on certain systems in the site. This resource is provided for use by the site's scheduling policy. The allowable values and effect on job placement is site dependent. (See &quot;Scheduler License Manager&quot; in the <em>Moab Workload Manager Administrator Guide</em> for more information.)</td>
</tr>
</tbody>
</table>
### Chapter 2: Submitting and managing jobs

<table>
<thead>
<tr>
<th>Resource</th>
<th>Format</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>vmem</td>
<td>size</td>
<td>Maximum amount of virtual memory used by all concurrent processes in the job. (Ignored on Unicos.)</td>
</tr>
<tr>
<td>walltime</td>
<td>seconds, or [[HH:]MM:]SS</td>
<td>Maximum amount of real time during which the job can be in the running state.</td>
</tr>
</tbody>
</table>

#### size

The *size* format specifies the maximum amount in terms of bytes or words. It is expressed in the form `integer[suffix]`. The suffix is a multiplier defined in the following table ("b" means bytes [the default] and "w" means words). The size of a word is calculated on the execution server as its word size.

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Multiplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>w</td>
</tr>
<tr>
<td>kb</td>
<td>kw</td>
</tr>
<tr>
<td>mb</td>
<td>mw</td>
</tr>
<tr>
<td>gb</td>
<td>gw</td>
</tr>
<tr>
<td>tb</td>
<td>tw</td>
</tr>
</tbody>
</table>

**Example 2-1: qsub -l nodes**

<table>
<thead>
<tr>
<th>Usage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; qsub -l nodes=12</td>
<td>Request 12 nodes of any type</td>
</tr>
<tr>
<td>&gt; qsub -l nodes=2:server+14</td>
<td>Request 2 &quot;server&quot; nodes and 14 other nodes (a total of 16) - this specifies two node specs, &quot;2:server&quot; and &quot;14&quot;</td>
</tr>
<tr>
<td>&gt; qsub -l nodes-s=server:hippi+10:noserver+3:bigmem:hippi</td>
<td>Request (a) 1 node that is a &quot;server&quot; and has a &quot;hippi&quot; interface, (b) 10 nodes that are not servers, and (c) 3 nodes that have a large amount of memory and have hippi</td>
</tr>
<tr>
<td>&gt; qsub -l nodes=b2005+b1803+b1813</td>
<td>Request 3 specific nodes by hostname</td>
</tr>
</tbody>
</table>
## Requesting generic resources

When *generic* resources have been assigned to nodes using the server's nodes file, these resources can be requested at the time of job submission using the `other` field. (See "Consumable Generic Resources" in the Moab Workload Manager Administrator Guide for details on configuration within Moab).

### Example 2-4: Generic

This job will run on any node that has the generic resource *matlab*.

```
> qsub -l other=matlab /home/user/script.sh
```

This can also be requested at the time of job submission using the `-W x=GRES:matlab` flag.

### Related topics

- [Requesting resources](#) on page 42
- [Job submission](#) on page 39
Requesting floating resources

When floating resources have been set up inside Moab, they can be requested in the same way as generic resources. Moab will automatically understand that these resources are floating and will schedule the job accordingly. (See "Floating Generic Resources" in the Moab Workload Manager Administrator Guide for details on configuration within Moab.)

Example 2-5: Floating

This job will run on any node when there are enough floating resources available.

```bash
> qsub -l other=matlab /home/user/script.sh
```

This can also be requested at the time of job submission using the `-W x=GRES:matlab` flag.

Related topics

- Requesting resources on page 42
- Job submission on page 39

Requesting other resources

Many other resources can be requested at the time of job submission using the Moab Workload Manager. See "Resource Manager Extensions" in the Moab Workload Manager Administrator Guide for a list of these supported requests and correct syntax.

Related topics

- Requesting resources on page 42
- Job submission on page 39

Exported batch environment variables

When a batch job is started, a number of variables are introduced into the job's environment that can be used by the batch script in making decisions, creating output files, and so forth. These variables are listed in the following table:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBS_JOBNAME</td>
<td>User specified jobname</td>
</tr>
<tr>
<td>PBS_ARRAYID</td>
<td>Zero-based value of job array index for this job (in version 2.2.0 and later)</td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>PBS_GPUFILE</td>
<td>Line-delimited list of GPUs allocated to the job located in $TORQUE_HOME/aux/jobidgpu. Each line follows the following format: &lt;host&gt;-gpu&lt;number&gt; For example, myhost-gpu1.</td>
</tr>
<tr>
<td>PBS_O_WORKDIR</td>
<td>Job’s submission directory</td>
</tr>
<tr>
<td>PBS_ENVIRONMENT</td>
<td>N/A</td>
</tr>
<tr>
<td>PBS_TASKNUM</td>
<td>Number of tasks requested</td>
</tr>
<tr>
<td>PBS_O_HOME</td>
<td>Home directory of submitting user</td>
</tr>
<tr>
<td>PBS_MOMPORT</td>
<td>Active port for MOM daemon</td>
</tr>
<tr>
<td>PBS_O_LOGNAME</td>
<td>Name of submitting user</td>
</tr>
<tr>
<td>PBS_O_LANG</td>
<td>Language variable for job</td>
</tr>
<tr>
<td>PBS_JOBCOOKIE</td>
<td>Job cookie</td>
</tr>
<tr>
<td>PBS_JOBID</td>
<td>Unique pbs job id</td>
</tr>
<tr>
<td>PBS_NODENUM</td>
<td>Node offset number</td>
</tr>
<tr>
<td>PBS_NUM_NODES</td>
<td>Number of nodes allocated to the job</td>
</tr>
<tr>
<td>PBS_NUM_PPN</td>
<td>Number of procs per node allocated to the job</td>
</tr>
<tr>
<td>PBS_O_SHELL</td>
<td>Script shell</td>
</tr>
<tr>
<td>PBS_O_HOST</td>
<td>Host on which job script is currently running</td>
</tr>
<tr>
<td>PBS_QUEUE</td>
<td>Job queue</td>
</tr>
</tbody>
</table>
### Variable Description

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBS_NODEFILE</td>
<td>File containing line delimited list on nodes allocated to the job</td>
</tr>
<tr>
<td>PBS_NP</td>
<td>Number of execution slots (cores) for the job</td>
</tr>
<tr>
<td>PBS_O_PATH</td>
<td>Path variable used to locate executables within job script</td>
</tr>
</tbody>
</table>

#### Related topics
- [Requesting resources on page 42](#)
- [Job submission on page 39](#)

### Enabling trusted submit hosts

By default, only the node running the `pbs_server` daemon is allowed to submit jobs. Additional nodes can be trusted as submit hosts by taking any of the following steps:

- Set the `allow_node_submit` server parameter (see [Allowing job submission from compute hosts on page 23](#)).
  
  Allows any host trusted as a compute host to also be trusted as a submit host.

- Set the `submit_hosts` server parameter (see [Using the "submit_hosts" service parameter on page 23](#)).
  
  Allows specified hosts to be trusted as a submit host.

- Use `.rhosts` to enable `ruserok()` based authentication (see [Using RCmd authentication on page 22](#)).

See [Configuring job submission hosts on page 22](#) for more information.

When you enable `allow_node_submit` on page 228, you must also enable the `allow_proxy_user` on page 228 parameter to allow user proxying when submitting and running jobs.

#### Related topics
- [Job submission on page 39](#)

### Example submit scripts

The following is an example job test script:
Chapter 2: Submitting and managing jobs

Monitoring jobs

TORQUE allows users and administrators to monitor submitted jobs with the `qstat` command. If the command is run by a non-administrative user, it will output just that user's jobs. For example:

```
> qstat
Job id Name User    Time Use S Queue
----- -------- ---- -------- ------ ------
 4807 scatter user01 12:56:34 R batch
...
```

Related topics

- **Job submission** on page 39

Canceling jobs

TORQUE allows users and administrators to cancel submitted jobs with the `qdel` command. The job will be sent TERM and KILL signals killing the running processes. When the top-level job script exits, the job will exit. The only parameter is the ID of the job to be canceled.

If a job is canceled by an operator or manager, an email notification will be sent to the user. Operators and managers may add a comment to this email with the `--m` option.
Related topics

- [Submitting and managing jobs on page 39](#)
Job execution queue or the **keep_completed** on page 233 parameter on the server. This should be set to the number of seconds that jobs should be held in the queue. If you set keep_completed on the job execution queue, completed jobs will be reported in the C state and the exit status is seen in the exit_status job attribute.

If the Mother Superior and TORQUE server are on the same server, expect the following behavior:

- When keep_completed is set, the job spool files will be deleted when the specified time arrives and TORQUE purges the job from memory.
- When keep_completed is not set, TORQUE deletes the job spool files upon job completion.
- If you manually purge a job (qdel -p) before the job completes or time runs out, TORQUE will never delete the spool files.

By maintaining status information about completed (or canceled, failed, etc.) jobs, administrators can better track failures and improve system performance. This allows TORQUE to better communicate with Moab Workload Manager and track the status of jobs. This gives Moab the ability to track specific failures and to schedule the workload around possible hazards. (See NODEFAILURERESERVETIME in "Appendix F: Parameters" of the Moab Workload Manager Administrator Guide for more information.)

**Related topics**

- [Submitting and managing jobs on page 39](#)

---

**Job checkpoint and restart**

While TORQUE has had a job checkpoint and restart capability for many years, this was tied to machine specific features. Now TORQUE supports BLCR—an architecture independent package that provides for process checkpoint and restart.

The support for BLCR is only for serial jobs, not for any MPI type jobs.

This section contains these topics:

- [Introduction to BLCR on page 54](#)
- [Configuration files and scripts on page 54](#)
- [Starting a checkpointable job on page 61](#)
- [Checkpointing a job on page 62](#)
- [Restarting a job on page 62](#)
- [Acceptance tests on page 63](#)

**Related topics**

- [Submitting and managing jobs on page 39](#)
Introduction to BLCR

BLCR is a kernel level package. It must be downloaded and installed from BLCR.

After building and making the package, it must be installed into the kernel with commands as follows. These can be installed into the file /etc/modules but all of the testing was done with explicit invocations of modprobe.

Installing BLCR into the kernel:

```bash
# /sbin/insmod /usr/local/lib/blcr/2.6.12-1.234/blcr_imports.ko
# /sbin/insmod /usr/local/lib/blcr/2.6.12-1.234/blcr_vmadump.ko
# /sbin/insmod /usr/local/lib/blcr/2.6.12-1.234/blcr.ko
```

The BLCR system provides four command line utilities:

- `cr_checkpoint`
- `cr_info`
- `cr_restart`
- `cr_run`

For more information about BLCR, see the BLCR Administrator's Guide.

Related topics

- Job checkpoint and restart on page 53

Configuration files and scripts

Configuring and Building TORQUE for BLCR:

```bash
> ./configure --enable-unixsockets=no --enable-blcrc
> make
> sudo make install
```

Depending on where BLCR is installed you may also need to use the following configure options to specify BLCR paths:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>--with-blcrc-inclu=DIR</td>
<td>include path for libcrh</td>
</tr>
<tr>
<td>--with-blcrc-lib=DIR</td>
<td>lib path for libcr</td>
</tr>
<tr>
<td>--with-blcrc-bin=DIR</td>
<td>bin path for BLCR utilities</td>
</tr>
</tbody>
</table>
The PBS configuration file located in `/var/spool/torque/mom_priv` must be modified to identify the script names associated with invoking the BLCR commands. The following variables should be used in the configuration file when using BLCR checkpointing.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$checkpoint_interval</td>
<td>How often periodic job checkpoints will be taken (minutes)</td>
</tr>
<tr>
<td>$checkpoint_script</td>
<td>The name of the script file to execute to perform a job checkpoint</td>
</tr>
<tr>
<td>$restart_script</td>
<td>The name of the script file to execute to perform a job restart</td>
</tr>
<tr>
<td>$checkpoint_run_exe</td>
<td>The name of an executable program to be run when starting a checkpointable job (for BLCR, cr_run)</td>
</tr>
</tbody>
</table>

The following example shows the contents of the configuration file used for testing the BLCR feature in TORQUE.

```
Example 2-6: Script file permissions

# chmod 754 blcr*
# ls -l
total 20
-rwxr-xr-- 1 root root 2112 2008-03-11 13:14 blcr_checkpoint_script
-rw-r--r-- 1 root 215 2008-03-11 13:13 config
drwxr-x-x 2 root root 4096 2008-03-11 13:21 jobs
-rw-r--r-- 1 root 7 2008-03-11 13:15 mom.lock
```

```
Example 2-7: mom_priv/config

$checkpoint_script /var/spool/torque/mom_priv/blcr_checkpoint_script
$restart_script /var/spool/torque/mom_priv/blcr_restart_script
$checkpoint_run_exe /usr/local/bin/cr_run
$pbsserver makua.cridomain
$loglevel 7
```
Example 2-8: \texttt{mom_prv/blcr\_checkpoint\_script}
# /usr/bin/perl
#
# Usage: checkpoint_script
# This script is invoked by pbs_mom to checkpoint a job.
#
use strict;
use Sys::Syslog;

# Log levels:
# 0 = none -- no logging
# 1 = fail -- log only failures
# 2 = info -- log invocations
# 3 = debug -- log all subcommands
my $logLevel = 3;
logPrint(2, "Invoked: $0 " . join(' ', @ARGV) . "\n");

my ($sessionId, $jobId, $userId, $signalNum, $checkpointDir, $checkpointName);
my $usage = "Usage: $0
"
my $cmd = "cr_checkpoint";
$cmd .= " --signal $signalNum" if $signalNum;
$cmd .= " --tree $sessionId";
$cmd .= " --file $checkpointName";
my $output = `$cmd 2>&1`;
my $rc = $? >> 8;
logDie(1, "Subcommand ($cmd) failed with rc=$rc\n$output") if $rc && $logLevel >= 1;
logPrint(3, "Subcommand ($cmd) yielded rc=$rc\n$output") if $logLevel >= 3;
exit 0;

# logPrint($message)
# Write a message (to syslog) and die
sub logPrint
{

Chapter 2: Submitting and managing jobs
my ($level, $message) = @_; 
my @severity = ('none', 'warning', 'info', 'debug'); 
return if $level > $logLevel; 
openlog('checkpoint_script', '', 'user'); 
syslog($severity[$level], $message); 
closelog(); 
}

 confessed
confessed

### logDie($message)
# Write a message (to syslog) and die
### logDie
{my ($level, $message) = @_; 
logPrint($level, $message);
die($message); 
}
Chapter 2: Submitting and managing jobs

Example 2-9: `mom_priv/blcr_restart_script`
Usage:
restart_script

This script is invoked by pbs_mom to restart a job.

use strict;
use Sys::Syslog;

# Log levels:
# 0 = none -- no logging
# 1 = fail -- log only failures
# 2 = info -- log invocations
# 3 = debug -- log all subcommands
my $logLevel = 3;

logPrint(2, "Invoked: $0 " . join( ' ', @ARGV) . "\n");

my ($sessionId, $jobId, $userId, $checkpointDir, $restartName);
my $usage = "Usage: $0 \n";
if (@ARGV == 5) {
    ($sessionId, $jobId, $userId, $checkpointDir, $restartName) = @ARGV;
} else {
    logDie(1, $usage);
}

# Change to the checkpoint directory where we want the checkpoint to be created
chdir $checkpointDir
    or logDie(1, "Unable to cd to checkpoint dir ($checkpointDir): $!");

my $cmd = "cr_restart";
$cmd .= " $restartName";
my $output = `"$cmd 2>&1";
my $rc = $? >> 8;

logDie(1, "Subcommand ($cmd) failed with rc=$rc:\n$output")
    if $rc && $logLevel >= 1;

logPrint(3, "Subcommand ($cmd) yielded rc=$rc:\n$output")
    if $logLevel >= 3;
exit 0;

# logPrint($message)
# Write a message (to syslog) and die
sub logPrint {
    my ($level, $message) = @_; 
    my @severity = ('none', 'warning', 'info', 'debug');

    return if $level > $logLevel;
    openlog('restart_script', '', 'user');
    syslog($severity{$level}, $message);
    closelog();
}

# logDie($message)
# Write a message (to syslog) and die
sub logDie {
    my ($level, $message) = @_; 
    logPrint($level, $message);
Starting a checkpointable job

Not every job is checkpointable. A job for which checkpointing is desirable must be started with the `-c` command line option. This option takes a comma-separated list of arguments that are used to control checkpointing behavior. The list of valid options available in the 2.4 version of TORQUE is shown below.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>none</code></td>
<td>No checkpointing (not highly useful, but included for completeness).</td>
</tr>
<tr>
<td><code>enabled</code></td>
<td>Specify that checkpointing is allowed, but must be explicitly invoked by either the <code>qhold</code> or <code>qchkpt</code> commands.</td>
</tr>
<tr>
<td><code>shutdown</code></td>
<td>Specify that checkpointing is to be done on a job at pbs_mom shutdown.</td>
</tr>
<tr>
<td><code>periodic</code></td>
<td>Specify that periodic checkpointing is enabled. The default interval is 10 minutes and can be changed by the <code>$checkpoint_interval</code> option in the MOM configuration file, or by specifying an interval when the job is submitted.</td>
</tr>
<tr>
<td><code>interval=minutes</code></td>
<td>Specify the checkpoint interval in minutes.</td>
</tr>
<tr>
<td><code>depth=number</code></td>
<td>Specify a number (depth) of checkpoint images to be kept in the checkpoint directory.</td>
</tr>
<tr>
<td><code>dir=path</code></td>
<td>Specify a checkpoint directory (default is <code>/var/spool/torque/checkpoint</code>).</td>
</tr>
</tbody>
</table>

Example 2-10: Sample test program

```c
#include "stdio.h"
int main( int argc, char *argv[] )
{
    int i;
    for (i=0; i<100; i++)
    {
        printf("i = %d\n", i);
        fflush(stdout);
        sleep(1);
    }
}
```

Example 2-11: Instructions for building test program

```
> gcc -o test test.c
```
Example 2-12: Sample test script

```bash
#!/bin/bash
./test
```

Example 2-13: Starting the test job

```bash
> qstat
> qsub -c enabled,periodic,shutdown,interval=1 test.sh
77.jakaa.cridomain
> qstat
<table>
<thead>
<tr>
<th>Job id</th>
<th>Name</th>
<th>User</th>
<th>Time Use S Queue</th>
</tr>
</thead>
<tbody>
<tr>
<td>77.jakaa</td>
<td>test.sh</td>
<td>jsmith</td>
<td>0 Q batch</td>
</tr>
</tbody>
</table>
```

If you have no scheduler running, you might need to start the job with `qrun`.

As this program runs, it writes its output to a file in `/var/spool/torque/spool`. This file can be observed with the command `tail -f`.

**Related topics**
- [Job checkpoint and restart](#) on page 53

## Checkpointing a job

Jobs are checkpointed by issuing a `qhold` command. This causes an image file representing the state of the process to be written to disk. The directory by default is `/var/spool/torque/checkpoint`.

This default can be altered at the queue level with the `qmgr` command. For example, the command `qmgr -c set queue batch checkpoint_dir=/tmp` would change the checkpoint directory to `/tmp` for the queue 'batch'.

The default directory can also be altered at job submission time with the `-c dir=/tmp` command line option.

The name of the checkpoint directory and the name of the checkpoint image file become attributes of the job and can be observed with the command `qstat -f`. Notice in the output the names `checkpoint_dir` and `checkpoint_name`. The variable `checkpoint_name` is set when the image file is created and will not exist if no checkpoint has been taken.

A job can also be checkpointed without stopping or holding the job with the command `qchkpt`.

**Related topics**
- [Job checkpoint and restart](#) on page 53

## Restarting a job

### Restarting a job in the Held state

The `qrls` command is used to restart the hibernated job. If you were using the `tail -f` command to watch the output file, you will see the test program start counting again.
It is possible to use the `qalter` command to change the name of the checkpoint file associated with a job. This could be useful if there were several job checkpoints and it restarting the job from an older image was specified.

**Restarting a job in the Completed state**

In this case, the job must be moved to the Queued state with the `qrerun` command. Then the job must go to the Run state either by action of the scheduler or if there is no scheduler, through using the `qrun` command.

**Related topics**

- [Job checkpoint and restart](#) on page 53

---

## Acceptance tests

A number of tests were made to verify the functioning of the BLCR implementation. See [BLCR](#) acceptance tests on page 305 for a description of the testing.

**Related topics**

- [Job checkpoint and restart](#) on page 53

---

## Job exit status

Once a job under TORQUE has completed, the `exit_status` attribute will contain the result code returned by the job script. This attribute can be seen by submitting a `qstat -f` command to show the entire set of information associated with a job. The `exit_status` field is found near the bottom of the set of output lines.
Example 2-14: `qstat -f` (job failure)

```
Job Id: 179.host
  Job Name = STDIN
  Job Owner = user@host
  job_state = C
  queue = batchq server = host
  Error Path = host:/opt/moab/STDIN.e179
  exec_host = node1/0
  Hold Types = n
  Join_Path = n
  Mail_Points = a
  Output Path = host:/opt/moab/STDIN.o179
  Priority = 0
  Rerunable = True
  Resource_List.ncpus = 2
  Resource_List.nodect = 1
  Resource_List.nodes = node1
  Variable List = PBS_O_HOME=/home/user,PBS_O_LOGNAME=user,
PBS_O_PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:
sbin:,PBS_O_SHELL=/bin/bash,PBS_O_HOST=host,
PBS_O_WORKDIR=/opt/moab,PBS_O_QUEUE=batchq
  sched_hmode = Post job file processing error; job 179.host on host node1/0Bad
  UID for job execution REJHOST=pala.cridomain MSG=cannot find user 'user' in
  password file
  exit_status = -1
```

The value of `Resource_List.*` is the amount of resources requested.

This code can be useful in diagnosing problems with jobs that may have unexpectedly terminated.

If TORQUE was unable to start the job, this field will contain a negative number produced by the `pbs_mom`. Otherwise, if the job script was successfully started, the value in this field will be the return value of the script.

Example 2-15: TORQUE supplied exit codes

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>JOB_EXEC_OK</td>
<td>0</td>
<td>Job execution successful</td>
</tr>
<tr>
<td>JOB_EXEC_FAIL1</td>
<td>-1</td>
<td>Job execution failed, before files, no retry</td>
</tr>
<tr>
<td>JOB_EXEC_FAIL2</td>
<td>-2</td>
<td>Job execution failed, after files, no retry</td>
</tr>
<tr>
<td>JOB_EXEC_RETRY</td>
<td>-3</td>
<td>Job execution failed, do retry</td>
</tr>
<tr>
<td>JOB_EXEC_INITABT</td>
<td>-4</td>
<td>Job aborted on MOM initialization</td>
</tr>
<tr>
<td>JOB_EXEC_INITRST</td>
<td>-5</td>
<td>Job aborted on MOM init, chkpt, no migrate</td>
</tr>
</tbody>
</table>
### Chapter 2: Submitting and managing jobs

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>JOB_EXEC_INITRMG</td>
<td>-6</td>
<td>Job aborted on MOM init, chkpt, ok migrate</td>
</tr>
<tr>
<td>JOB_EXEC_BADRESRT</td>
<td>-7</td>
<td>Job restart failed</td>
</tr>
<tr>
<td>JOB_EXEC_CMDFAIL</td>
<td>-8</td>
<td>Exec() of user command failed</td>
</tr>
<tr>
<td>JOB_EXEC_STDOUTFAIL</td>
<td>-9</td>
<td>Could not create/open stdout stderr files</td>
</tr>
<tr>
<td>JOB_EXEC_OVERLIMIT_MEM</td>
<td>-10</td>
<td>Job exceeded a memory limit</td>
</tr>
<tr>
<td>JOB_EXEC_OVERLIMIT_WT</td>
<td>-11</td>
<td>Job exceeded a walltime limit</td>
</tr>
<tr>
<td>JOB_EXEC_OVERLIMIT_CPUT</td>
<td>-12</td>
<td>Job exceeded a CPU time limit</td>
</tr>
</tbody>
</table>
Example 2-16: Exit code from C program

```c
#include
#include

int main(int argc, char *argv)
{
    exit(256+11);
}
```

```bash
$ gcc -o error error.c
$ echo ./error | qsub 180.xxx.yyy
```

```bash
$ qstat -f
Job Id: 180.xxx.yyy
Job Name = STDIN
Job.Owner = test.xxx.yyy
resources_used.cput = 00:00:00
resources_used.mem = 0kb
resources_used.vmem = 0kb
resources_used.walltime = 00:00:00
job_state = C
queue = batch
server = xxx.yyy
Checkpoint = u
cctime = Wed Apr 30 11:29:37 2008
Error_Path = xxx.yyy:/home/test/STDIN.e180
eexec_host = node01/0
Hold_Types = n
Join_Path = n
Keep_Files = n
Mail_Points = a
mtime = Wed Apr 30 11:29:37 2008
Output_Path = xxx.yyy:/home/test/STDIN.o180
Priority = 0
qctime = Wed Apr 30 11:29:37 2008
Rerunnable = True
Resource_List.neednodes = 1
Resource_List.nodect = 1
Resource_List.nodes = 1
Resource_List.walltime = 01:00:00
session_Id = 14107
substate = 59
Variable List = PBS_O_HOME=/home/test,PBS_O_LANG=en_US.UTF-8,
PBS_O_LOGNAME=test,
PBS_O_PATH=/usr/local/perltests/bin:/home/test/bin:/usr/local/sbin:
/usr/local/bin:/sbin:/usr/sbin:/usr/bin:/bin:/sbin:/bin:/bin:/bin:
PBS_O_SHELL=/bin/bash,PBS_SERVER=xxx.yyy,
PBS_O_HOST=xxx.yyy,PBS_O_WORKDIR=/home/test,
PBS_O_QUEUE=batch
euser = test
egroup = test
hashname = 180.xxx.yyy
queue_rank = 8
queue_type = E
comment = Job started on Wed Apr 30 at 11:29
etime = Wed Apr 30 11:29:37 2008
exit_status = 11
start_time = Wed Apr 30 11:29:37 2008
start_count = 1
```
Chapter 2: Submitting and managing jobs

Notice that the C routine `exit` passes only the low order byte of its argument. In this case, 256+11 is really 267 but the resulting exit code is only 11 as seen in the output.

Related topics

- Job checkpoint and restart on page 53
- Submitting and managing jobs on page 39

Service jobs

TORQUE service jobs are a special kind of job that is treated differently by TORQUE than normal batch jobs. TORQUE service jobs are not related to Moab's dynamic service jobs. A TORQUE service job cannot dynamically grow and shrink in size over time.

Jobs are marked as service jobs at the time they are submitted to Moab or TORQUE. Just like a normal job, a script file is specified with the job. In a batch job, the contents of the script file are taken by TORQUE and executed on the compute nodes. For a service job, however, the script file is assumed to respond to certain command-line arguments. Instead of just executing the script, TORQUE will use these command-line arguments to start, stop, and check on the status of the job. Listed below are the three command-line arguments that must be supported by any script submitted as part of a TORQUE service job:

- **start**: The script should take this argument and launch its service/workload. The script should remain executing/running until the service stops.
- **stop**: The script should take this argument and stop the service/workload that was earlier started.
- **status**: The script should take this argument and return, via standard out, either "running" if the service/workload is running as expected or "stopped" if the service is not running.

This feature was created with long-running services in mind. The command-line arguments should be familiar to users who interact with Unix services, as each of the service scripts found in `/etc/init.d/` also accept and respond to the arguments as explained above.

For example, if a user wants to start the Apache 2 server on a compute node, they can use a TORQUE service job and specify a script which will start, stop, and check on the status of the "httpd" daemon--possibly by using the already present `/etc/init.d/httpd` script.

⚠️ If you wish to submit service jobs only through TORQUE, no special version of Moab is required. If you wish to submit service jobs using Moab's msub, then Moab 5.4 is required.

For details, see these topics:

- Submitting service jobs on page 68
- Submitting service jobs in MCM on page 68
- Managing service jobs on page 68
Submitting service jobs

There is a new option to qsub, "-s" which can take either a 'y' or 'n' (yes or no, respectively). When "-s y" is present, then the job is marked as a service job.

```
qsub -l walltime=100:00:00,nodes=1 -s y service_job.py
```

The example above submits a job to TORQUE with a walltime of 100 hours, one node, and it is marked as a service job. The script "service_job.py" will be used to start, stop, and check the status of the service/workload started on the compute nodes.

Moab, as of version 5.4, is able to accept the "-s" option when msub is used for submission. Moab will then pass this information to TORQUE when the job is migrated.

Related topics

- [Service jobs on page 67](#)

Submitting service jobs in MCM

Submitting a service job in MCM requires the latest Adaptive Computing Suite snapshot of MCM. It also requires MCM to be started with the "--future=2" option.

Once MCM is started, open the Create Workload window and verify Show Advanced Options is checked. Notice that there is a Service checkbox that can be selected in the Flags/Options area. Use this to specify the job is a service job.

Related topics

- [Service jobs on page 67](#)

Managing service jobs

Managing a service job is done much like any other job; only a few differences exist.

Examining the job with qstat -f will reveal that the job has the service = True attribute. Non-service jobs will not make any mention of the "service" attribute.

Canceling a service job is done with qdel, mjobctl -c, or through any of the GUI's as with any other job. TORQUE, however, cancels the job by calling the service script with the "stop" argument instead of killing it directly. This behavior also occurs if the job runs over its wallclock and TORQUE/Moab is configured to cancel the job.

If a service job completes when the script exits after calling it with "start," or if TORQUE invokes the script with "status" and does not get back "running," it will not be terminated by using the "stop" argument.

Related topics

- [Service jobs on page 67](#)
Chapter 3: Managing nodes

This section contains information about adding and configuring compute nodes. It explains how to work with host security for systems that require dedicated access to compute nodes. It also contains information about scheduling specific cores on a node at job submission.

For details, see these topics:

- Adding nodes on page 69
- Node properties on page 70
- Changing node state on page 71
- Host security on page 71
- Linux cpuset support on page 72
- Scheduling cores on page 74

Adding nodes

TORQUE can add and remove nodes either dynamically with `qmgr` or by manually editing the TORQUE_HOME/server_priv/nodes file (see Initializing/Configuring TORQUE on the server (pbs_server) on page 7).

Run-time node changes

TORQUE can dynamically add nodes with the `qmgr` command. For example, the following command will add node node003:

```
> qmgr -c "create node node003"
```

The above command appends the $TORQUE_HOME/server_priv/nodes file with:

```
node003
```

Nodes can also be removed with a similar command:

```
> qmgr -c "delete node node003"
```

Typically, an administrator will want to change the state of a node instead of remove it (for details, see Changing node state on page 71).
When you make changes to nodes – whether by using qmgr or directly editing the nodes file – you must restart pbs_server for those changes to take effect.

Related topics
- Managing nodes on page 69

Node properties
TORQUE can associate properties with nodes to aid in identifying groups of nodes. It's typical for a site to conglomerate a heterogeneous set of resources. To identify the different sets, properties can be given to each node in a set. For example, a group of nodes that has a higher speed network connection could have the property "ib". TORQUE can set, update, or remove properties either dynamically with qmgr or by manually editing the nodes file.

Run-time node changes
TORQUE can dynamically change the properties of a node with the qmgr command. For example, note the following to give node001 the properties of "bigmem" and "dualcore":

```
> qmgr -c "set node node001 properties = bigmem"
> qmgr -c "set node node001 properties += dualcore"
```

To relinquish a stated property, use the "-=" operator.

Manual node changes
The properties of each node are enumerated in TORQUE_HOME/server_priv/nodes. The feature(s) must be in a space delimited list after the node name. For example, to give node001 the properties of "bigmem" and "dualcore" and node002 the properties of "bigmem" and "matlab," edit the nodes file to contain the following:

```
server_priv/nodes:

node001 bigmem dualcore
node002 np=4 bigmem matlab
```

For changes to the nodes file to be activated, pbs_server must be restarted.

For a full description of this file, please see the PBS Administrator Guide.

Related topics
- Job submission on page 39
- Managing nodes on page 69
Changing node state

A common task is to prevent jobs from running on a particular node by marking it offline with `pbsnodes -o nodename`. Once a node has been marked offline, the scheduler will no longer consider it available for new jobs. Simply use `pbsnodes -c nodename` when the node is returned to service.

Also useful is `pbsnodes -l`, which lists all nodes with an interesting state, such as down, unknown, or offline. This provides a quick glance at nodes that might be having a problem. (See `pbsnodes` for details.)

Related topics

- Managing nodes on page 69

Host security

Enabling PAM with TORQUE

TORQUE is able to take advantage of the authentication services provided through Pluggable Authentication Modules (PAM) to help administrators manage access to compute nodes by users. The PAM module available in TORQUE is located in the PAM security directory. This module, when used in conjunction with other PAM modules, restricts access to the compute node unless the user has a job currently running on the node. The following configurations are examples only. For more information about PAM, see the PAM (Pluggable Authentication Modules) documentation from LinuxDocs.

To enable TORQUE PAM configure TORQUE using the `--with-pam` option. Using `--with-pam` is sufficient but if your PAM security modules are not in the default `/lib/security` or `/lib64/security` directory, you can specify the location using `--with-pam=<DIR>` where `<DIR>` is the directory where you want the modules to be installed. When TORQUE is installed the files `pam_pbssimpleauth.la` and `pam_pbssimpleauth.so` appear in `/lib/security`, `/lib64/security`, or the directory designated on the configuration line.

PAM is very flexible and policies vary greatly from one site to another. The following example restricts users trying to access a node using SSH. Administrators need to assess their own installations and decide how to apply the TORQUE PAM restrictions.

In this example, after installing TORQUE with PAM enabled, you would add the following two lines to `/etc/pam.d/sshd`:

```bash
account required pam_pbssimpleauth.so
account required pam_access.so
```

In `/etc/security/access.conf` make sure all users who access the compute node are added to the configuration. This is an example which allows the users root, george, allen, and michael access.

```bash
-:ALL EXCEPT root george allen michael torque:ALL
```

With this configuration, if user george has a job currently running on the compute node, george can use ssh to login to the node. If there are currently no jobs running, george is disconnected when attempting to login.
Chapter 3: Managing nodes

TORQUE PAM is good at keeping users out who do not have jobs running on a compute node. However, it does not have the ability to force a user to log out once they are in. To accomplish this use epilogue or prologue scripts to force users off the system.

Legacy TORQUE PAM configuration

There is an alternative PAM configuration for TORQUE that has been available since 2006. It can be found in the contrib/pam_authuser directory of the source tree. Adaptive Computing does not currently support this method but the instructions are given here for those who are currently using it and for those who wish to use it.

For systems requiring dedicated access to compute nodes (for example, users with sensitive data), TORQUE prologue and epilogue scripts provide a vehicle to leverage the authentication provided by linux-PAM modules. (See Prologue and epilogue scripts on page 281 for more information.)

To allow only users with running jobs (and root) to access compute nodes

1. Untar contrib/pam_authuser.tar.gz (found in the src tar ball).
2. Compile pam_authuser.c with make and make install on every compute node.
3. Edit /etc/system-auth as described in README.pam_authuser, again on every compute node.
4. Either make a tarball of the epilogue* and prologue* scripts (to preserve the symbolic link) and untar it in the mom_priv directory, or just copy epilogue* and prologue* to mom_priv/.

The prologue* scripts are Perl scripts that add the user of the job to /etc/authuser. The epilogue* scripts then remove the first occurrence of the user from /etc/authuser. File locking is employed in all scripts to eliminate the chance of race conditions. There is also some commented code in the epilogue* scripts, which, if uncommented, kills all processes owned by the user (using pkill), provided that the user doesn't have another valid job on the same node.

prologue and epilogue scripts were added to the pam_authuser tarball in version 2.1 of TORQUE.

Related topics

- Managing nodes on page 69

Linux cpuset support

- Cpuset overview on page 72
- Cpuset support on page 73
- Cpuset configuration on page 73
- Cpuset advantages / disadvantages on page 73

Cpuset overview

Linux kernel 2.6 Cpusets are logical, hierarchical groupings of CPUs and units of memory. Once created, individual processes can be placed within a cpuset. The processes will only be allowed to run/access the
specified CPUs and memory. Cpusets are managed in a virtual file system mounted at /dev/cpuset. New cpusets are created by simply making new directories. Cpusets gain CPUs and memory units by simply writing the unit number to files within the cpuset.

Cpuset support

All nodes using cpusets must have the hwloc library version 1.2 or higher installed. If you use CentOS 6, an automatic download -- yum install hwloc-devel, for instance -- will install the correct version; however, if you use CentOS 5 or SLES11 SP2, you will need to build it from the source using contrib/hwloc_install.sh or equivalent.

When started, pbs_mom will create an initial top-level cpuset at /dev/cpuset/torque. This cpuset contains all CPUs and memory of the host machine. If this "torqueset" already exists, it will be left unchanged to allow the administrator to override the default behavior. All subsequent cpusets are created within the torqueset.

When a job is started, the jobset is created at /dev/cpuset/torque/$jobid and populated with the CPUs listed in the exec_host job attribute. Also created are individual tasksets for each CPU within the jobset. This happens before prologue, which allows it to be easily modified, and it happens on all nodes.

The top-level batch script process is executed in the jobset. Tasks launched through the TM interface (pbsdsh and PW's mpiexec) will be executed within the appropriate taskset.

On job exit, all tasksets and the jobset are deleted.

Cpuset configuration

To configure cpuset

1. As root, mount the virtual filesystem for cpusets:

   ```bash
   mkdir /dev/cpuset
   mount -t cpuset none /dev/cpuset
   ```

   Do this for each MOM that is to use cpusets.

2. Because cpuset usage is a build-time option in TORQUE, you must add --enable-cpuset to your configure options:

   ```bash
   ./configure --enable-cpuset
   ```

3. Use this configuration for the MOMs across your system.

Cpuset advantages / disadvantages

Presently, any job can request a single CPU and proceed to use everything available in the machine. This is occasionally done to circumvent policy, but most often is simply an error on the part of the user. Cpuset support will easily constrain the processes to not interfere with other jobs.
Jobs on larger NUMA systems may see a performance boost if jobs can be intelligently assigned to specific CPUs. Jobs may perform better if striped across physical processors, or contained within the fewest number of memory controllers.

TM tasks are constrained to a single core, thus a multi-threaded process could seriously suffer.

Related topics
- Managing nodes on page 69
- Geometry request configuration on page 74

Scheduling cores

In TORQUE 2.4 and later, you can request specific cores on a node at job submission by using geometry requests. To use this feature, specify the procs_bitmap resource request of qsub-1 (see qsub) at job submission.

For details about scheduling cores, see these topics:
- Geometry request configuration on page 74
- Geometry request usage on page 75
- Geometry request considerations on page 75

Geometry request configuration

A Linux kernel of 2.6 or later is required to use geometry requests, because this feature uses Linux cpuset in its implementation. In order to use this feature, the cpuset directory has to be mounted. For more information on how to mount the cpuset directory, see Linux cpuset support on page 72. If the operating environment is suitable for geometry requests, configure TORQUE with the --enable-geometry-requests option.

```
> ./configure --prefix=/home/john/torque --enable-geometry-requests
```

TORQUE is configured to install to /home/john/torque and to enable the geometry requests feature.

The geometry request feature uses a subset of the cpuset feature. When you configure TORQUE using --enable-cpuset and --enable-geometry-requests at the same time, and use -l procs_bitmap=X, the job will get the requested cpuset. Otherwise, the job is treated as if only --enable-cpuset was configured.

Related topics
- Scheduling cores on page 74
Geometry request usage

Once enabled, users can submit jobs with a geometry request by using the `procs_bitmap=<string>` resource request. `procs_bitmap` requires a numerical string made up of 1's and 0's. A 0 in the bitmap means the job cannot run on the core that matches the 0's index in the bitmap. The index is in reverse order of the number of cores available. If a job is submitted with `procs_bitmap=1011`, then the job requests a node with four free cores, and uses only cores one, two, and four.

The geometry request feature requires a node that has all cores free. A job with a geometry request cannot run on a node that has cores that are busy, even if the node has more than enough cores available to run the job.

```
qsub -l procs_bitmap=0011 ossl.sh
```

The job `ossl.sh` is submitted with a geometry request of `0011`.

In the above example, the submitted job can run only on a node that has four cores. When a suitable node is found, the job runs exclusively on cores one and two.

Related topics
- [Scheduling cores](#) on page 74

Geometry request considerations

As previously stated, jobs with geometry requests require a node with all of its cores available. After the job starts running on the requested cores, the node cannot run other jobs, even if the node has enough free cores to meet the requirements of the other jobs. Once the geometry requesting job is done, the node is available to other jobs again.

Related topics
- [Scheduling cores](#) on page 74

Scheduling accelerator hardware

TORQUE works with accelerators (such as NVIDIA GPUs and Intel MICs) and can collect and report metrics from them or submit workload to them. This feature requires the use of the Moab scheduler. Refer to the [Moab Workload Manager accelerators documentation](#) for information on configuring accelerators in TORQUE.
Chapter 4: Setting server policies

This section explains how to set up and configure your queue. It lists the queue attributes and describes how to set up a routing queue. This section also explains how to set up TORQUE to run in high availability mode. For details, see these topics:

- Queue configuration on page 77
- Server high availability on page 91

Queue configuration

Under TORQUE, queue configuration is accomplished using the Server high availability command. With this tool, the first step is to create the queue. This is accomplished using the create subcommand of qmgr as in the following example:

```
> qmgr -c "create queue batch queue_type=execution"
```

Once created, the queue must be configured to be operational. At a minimum, this includes setting the options started and enabled. Further configuration is possible using any combination of the attributes listed in what follows.

For Boolean attributes, T, t, I, Y, and y are all synonymous with "TRUE," and F, f, 0, N, and n all mean "FALSE."

For queue_type, E and R are synonymous with "Execution" and "Routing" (respectively). See these topics for more details:

- Queue attributes on page 78
- Example queue configuration on page 88
- Setting a default queue on page 89
- Mapping a queue to subset of resources on page 89
- Creating a routing queue on page 89

Related topics

- Server parameters on page 227
- qalter on page 170 - command which can move jobs from one queue to another
Queue attributes

This section lists the following queue attributes:

- `acl_groups` on page 79
- `acl_group_enable` on page 79
- `acl_group_sloppy` on page 79
- `acl_hosts` on page 80
- `acl_host_enable` on page 80
- `acl_logic_or` on page 80
- `acl_users` on page 80
- `acl_user_enable` on page 81
- `disallowed_types` on page 81
- `enabled` on page 81
- `features_required` on page 82
- `keep_completed` on page 82
- `kill_delay` on page 82
- `max_queuable` on page 83
- `max_running` on page 83
- `max_user_queuable` on page 83
- `max_user_run` on page 84
- `priority` on page 84
- `queue_type` on page 84
- `required_login_property` on page 85
- `resources_available` on page 85
- `resources_default` on page 85
- `resources_max` on page 85
- `resources_min` on page 86
- `route_destinations` on page 86
- `started` on page 87

This section also lists some queue resource limits (see Assigning queue resource limits on page 87).

For Boolean attributes, T, t, 1, Y, and y are all synonymous with "TRUE," and F, f, 0, N, and n all mean "FALSE."
### acl_groups

<table>
<thead>
<tr>
<th>Format</th>
<th><code>&lt;GROUP&gt;[@&lt;HOST&gt;]+&lt;USER&gt;[@&lt;HOST&gt;]</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td><code>---</code></td>
</tr>
<tr>
<td>Description</td>
<td>Specifies the list of groups which may submit jobs to the queue. If acl_group_enable is set to true, only users with a primary group listed in acl_groups may utilize the queue.</td>
</tr>
</tbody>
</table>
| Example    | `> qmgr -c "set queue batch acl_groups=staff"`  
`> qmgr -c "set queue batch acl_groups+=ops@h2"`  
`> qmgr -c "set queue batch acl_groups+=staff@h3"` |

> If the PBSACLEUSEGROUPLIST variable is set in the pbs_server environment, acl_groups checks against all groups of which the job user is a member.

> Used in conjunction with `acl_group_enable`.

### acl_group_enable

<table>
<thead>
<tr>
<th>Format</th>
<th><code>&lt;BOOLEAN&gt;</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td><code>FALSE</code></td>
</tr>
<tr>
<td>Description</td>
<td>If <code>TRUE</code>, constrains TORQUE to only allow jobs submitted from groups specified by the <code>acl_groups</code> parameter.</td>
</tr>
<tr>
<td>Example</td>
<td><code>&gt; qmgr -c &quot;set queue batch acl_group_enable=true&quot;</code></td>
</tr>
</tbody>
</table>

### acl_group_sloppy

<table>
<thead>
<tr>
<th>Format</th>
<th><code>&lt;BOOLEAN&gt;</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td><code>FALSE</code></td>
</tr>
<tr>
<td>Description</td>
<td>If <code>TRUE</code>, <code>acl_groups</code> will be checked against all groups of which the job users is a member.</td>
</tr>
<tr>
<td>Example</td>
<td><code>---</code></td>
</tr>
</tbody>
</table>
### acl_hosts

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;HOST&gt;[+&lt;HOST&gt;]...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Specifies the list of hosts that may submit jobs to the queue.</td>
</tr>
<tr>
<td>Example</td>
<td><code>qmgr -c &quot;set queue batch acl_hosts=h1+h2+h3&quot;</code></td>
</tr>
</tbody>
</table>

*Used in conjunction with acl_host_enable.*

### acl_host_enable

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;BOOLEAN&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>FALSE</td>
</tr>
<tr>
<td>Description</td>
<td>If TRUE, constrains TORQUE to only allow jobs submitted from hosts specified by the acl_hosts parameter.</td>
</tr>
<tr>
<td>Example</td>
<td><code>qmgr -c &quot;set queue batch acl_host_enable=true&quot;</code></td>
</tr>
</tbody>
</table>

### acl_logic_or

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;BOOLEAN&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>FALSE</td>
</tr>
<tr>
<td>Description</td>
<td>If TRUE, user and group acls are logically OR'd together, meaning that either acl may be met to allow access. If FALSE or unset, then both acls are AND'd, meaning that both acls must be satisfied.</td>
</tr>
<tr>
<td>Example</td>
<td><code>qmgr -c &quot;set queue batch acl_logic_or=true&quot;</code></td>
</tr>
</tbody>
</table>

### acl_users

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;USER&gt;@&lt;HOST&gt;[+&lt;USER&gt;@&lt;HOST&gt;]...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td></td>
</tr>
</tbody>
</table>
### acl_users

| Description | Specifies the list of users who may submit jobs to the queue. If `acl_user_enable` is set to `TRUE`, only users listed in `acl_users` may use the queue. |
| Example | ```
> qmgr -c "set queue batch acl_users=John"
> qmgr -c "set queue batch acl_users+=Steve@h2"
> qmgr -c "set queue batch acl_users+=Stevek@h3"
```

> *Used in conjunction with `acl_user_enable`.*

### acl_user_enable

| Format | `<BOOLEAN>` |
| Default | FALSE |
| Description | If `TRUE`, constrains TORQUE to only allow jobs submitted from users specified by the `acl_users` parameter. |
| Example | ```
qmgr -c "set queue batch acl_user_enable=true"
```

### disallowed_types

| Format | `<type> [ + <type> ] .` |
| Default | `---` |
| Description | Specifies classes of jobs that are not allowed to be submitted to this queue. Valid types are interactive, batch, rerunnable, nonrerunnable, fault_tolerant (as of version 2.4.0 and later), fault_intolerant (as of version 2.4.0 and later), and job_array (as of version 2.4.1 and later). |
| Example | ```
qmgr -c "set queue batch disallowed_types = interactive"
qmgr -c "set queue batch disallowed_types += job_array"
```
## enabled

<table>
<thead>
<tr>
<th>Description</th>
<th>Specifies whether the queue accepts new job submissions.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td><code>qmgr -c &quot;set queue batch enabled=true&quot;</code></td>
</tr>
</tbody>
</table>

## features_required

<table>
<thead>
<tr>
<th>Format</th>
<th>feature1[feature2[feature3...]]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>---</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies that all jobs in this queue will require these features in addition to any they may have requested. A feature is a synonym for a property.</td>
</tr>
<tr>
<td>Example</td>
<td><code>qmgr -c 's q batch features_required=fast'</code></td>
</tr>
</tbody>
</table>

## keep_completed

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>0</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies the number of seconds jobs should be held in the Completed state after exiting. For more information, see Keeping completed jobs on page 52.</td>
</tr>
<tr>
<td>Example</td>
<td><code>qmgr -c &quot;set queue batch keep_completed=120&quot;</code></td>
</tr>
</tbody>
</table>

## kill_delay

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>2</td>
</tr>
</tbody>
</table>
### kill_delay

**Description:** Specifies the number of seconds between sending a SIGTERM and a SIGKILL to a job in a specific queue that you want to cancel. It is possible that the job script, and any child processes it spawns, can receive several SIGTERM signals before the SIGKILL signal is received.

> All MOMs must be configured with `$exec with exec true` in order for `kill_delay` to work even when relying on default `kill_delay` settings.

**Example:**
```
qmgr -c "set queue batch kill_delay=30"
```

### max_queuable

**Format:** `<INTEGER>`

**Default:** unlimited

**Description:** Specifies the maximum number of jobs allowed in the queue at any given time (includes idle, running, and blocked jobs).

**Example:**
```
qmgr -c "set queue batch max_queuable=20"
```

### max_running

**Format:** `<INTEGER>`

**Default:** unlimited

**Description:** Specifies the maximum number of jobs in the queue allowed to run at any given time.

**Example:**
```
qmgr -c "set queue batch max_running=20"
```

### max_user_queuable

**Format:** `<INTEGER>`

**Default:** unlimited
### max_user_queuable

<table>
<thead>
<tr>
<th>Description</th>
<th>Specifies the maximum number of jobs, per user, allowed in the queue at any given time (includes idle, running, and blocked jobs). Version 2.1.3 and greater.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>qmgr -c &quot;set queue batch max_user_queuable=20&quot;</td>
</tr>
</tbody>
</table>

### max_user_run

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>unlimited</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies the maximum number of jobs, per user, in the queue allowed to run at any given time.</td>
</tr>
<tr>
<td>Example</td>
<td>qmgr -c &quot;set queue batch max_user_run=10&quot;</td>
</tr>
</tbody>
</table>

### priority

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>0</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies the priority value associated with the queue.</td>
</tr>
<tr>
<td>Example</td>
<td>qmgr -c &quot;set queue batch priority=20&quot;</td>
</tr>
</tbody>
</table>

### queue_type

<table>
<thead>
<tr>
<th>Format</th>
<th>One of e, execution, r, or route (see Creating a routing queue on page 89)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>---</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies the queue type.</td>
</tr>
<tr>
<td>Example</td>
<td>qmgr -c &quot;set queue batch queue_type=execution&quot;</td>
</tr>
</tbody>
</table>
**required_login_property**

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>---</td>
</tr>
<tr>
<td>Description</td>
<td>Adds the specified login property as a requirement for all jobs in this queue.</td>
</tr>
<tr>
<td>Example</td>
<td>qmgr -c 's q &lt;queuename&gt; required_login_property=INDUSTRIAL'</td>
</tr>
</tbody>
</table>

**resources_available**

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>---</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies to cumulative resources available to all jobs running in the queue. See <a href="#">qsub will not allow the submission of jobs requesting many processors</a> on page 132 for more information.</td>
</tr>
<tr>
<td>Example</td>
<td>qmgr -c &quot;set queue batch resources_available.nodect=20&quot;</td>
</tr>
</tbody>
</table>

You must restart pbs_server for changes to take effect.

Also, resources_available is constrained by the smallest of queue.resources_available and server.resources_available.

**resources_default**

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>---</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies default resource requirements for jobs submitted to the queue.</td>
</tr>
<tr>
<td>Example</td>
<td>qmgr -c &quot;set queue batch resources_default.walltime=3600&quot;</td>
</tr>
</tbody>
</table>

**resources_max**

| Format   | <STRING> |
### resources_max

| Default | --- |
| Description | Specifies the maximum resource limits for jobs submitted to the queue. |
| Example | ```bash
qmgr -c "set queue batch resources_max.nodect=16"
``` |

### resources_min

| Format | <STRING> |
| Default | --- |
| Description | Specifies the minimum resource limits for jobs submitted to the queue. |
| Example | ```bash
qmgr -c "set queue batch resources_min.nodect=2"
``` |

### route_destinations

| Format | <queue>[@<host>] |
| Default | --- |
| Description | Specifies the potential destination queues for jobs submitted to the associated routing queue. |
| Example | ```bash
> qmgr -c "set queue route route_destinations=fast"
> qmgr -c "set queue route route_destinations=slow"
> qmgr -c "set queue route route_destinations=medium@hostname"
``` |

To set multiple queue specifications, use multiple commands:

```bash
> qmgr -c 's s route_destinations=batch'
> qmgr -c 's s route_destinations=long'
> qmgr -c 's s route_destinations=short'
```
Chapter 4: Setting server policies

**Queue configuration**

<table>
<thead>
<tr>
<th>started</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
</tr>
<tr>
<td>Default</td>
</tr>
<tr>
<td>Description</td>
</tr>
<tr>
<td>Example</td>
</tr>
</tbody>
</table>

Assigning queue resource limits

Administrators can use resource limits to help direct what kind of jobs go to different queues. There are four queue attributes where resource limits can be set: `resources_available`, `resources_default`, `resources_max`, and `resources_min`. The list of supported resources that can be limited with these attributes are `arch`, `mem`, `nodect`, `nodes`, `procc`, `pvmem`, `vmem`, and `walltime`.

<table>
<thead>
<tr>
<th>Resource</th>
<th>Format</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>arch</td>
<td>string</td>
<td>Specifies the administrator defined system architecture required.</td>
</tr>
<tr>
<td>mem</td>
<td>size</td>
<td>Amount of physical memory used by the job. (Ignored on Darwin, Digital Unix, FreeBSD, HP-UX 11, IRIX, NetBSD, and SunOS. Also ignored on Linux if number of nodes is not 1. Not implemented on AIX and HP-UX 10.)</td>
</tr>
<tr>
<td>nodect</td>
<td>integer</td>
<td>Sets the number of nodes available. By default, TORQUE will set the number of nodes available to the number of nodes listed in the <code>$TORQUE_HOME/server_priv/nodes</code> file. <code>nodect</code> can be set to be greater than or less than that number. Generally, it is used to set the node count higher than the number of physical nodes in the cluster.</td>
</tr>
<tr>
<td>nodes</td>
<td>integer</td>
<td>Specifies the number of nodes.</td>
</tr>
<tr>
<td>procc</td>
<td>integer</td>
<td>Sets limits on the total number of execution slots (procs) allocated to a job. The number of procs is calculated by summing the products of all node and ppn entries for a job. For example, <code>qsub -l nodes=2:ppn=2+3:ppn=4 job.sh</code> would yield a procc of 16. 2<em>2 (2:ppn=2) + 3</em>4 (3:ppn=4).</td>
</tr>
<tr>
<td>pvmem</td>
<td>size</td>
<td>Amount of virtual memory used by any single process in a job.</td>
</tr>
<tr>
<td>vmem</td>
<td>size</td>
<td>Amount of virtual memory used by all concurrent processes in the job.</td>
</tr>
</tbody>
</table>
### Resource | Format | Description
--- | --- | ---
walltime | seconds, or [[HH:]MM:]SS | Amount of real time during which a job can be in a running state.

### Size
The size format specifies the maximum amount in terms of bytes or words. It is expressed in the form `integer[suffix]`. The suffix is a multiplier defined in the following table ("b" means bytes [the default] and "w" means words). The size of a word is calculated on the execution server as its word size.

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Multiplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>w</td>
</tr>
<tr>
<td>kb</td>
<td>kw</td>
</tr>
<tr>
<td>mb</td>
<td>mw</td>
</tr>
<tr>
<td>gb</td>
<td>gw</td>
</tr>
<tr>
<td>tb</td>
<td>tw</td>
</tr>
</tbody>
</table>

### Related topics
- [Queue configuration on page 77](#)
- [Example queue configuration on page 88](#)

### Example queue configuration

The following series of `qmgr` commands will create and configure a queue named `batch`:

```bash
qmgr -c "create queue batch queue_type=execution"
qmgr -c "set queue batch started=true"
qmgr -c "set queue batch enabled=true"
qmgr -c "set queue batch resources_default.nodes=1"
qmgr -c "set queue batch resources_default.walltime=3600"
```

This queue will accept new jobs and, if not explicitly specified in the job, will assign a nodecount of 1 and a walltime of 1 hour to each job.

### Related topics
- [Queue configuration on page 77](#)
Setting a default queue

By default, a job must explicitly specify which queue it is to run in. To change this behavior, the server parameter `default_queue` may be specified as in the following example:

```
qmgr -c "set server default_queue=batch"
```

Related topics

- **Queue configuration** on page 77

Mapping a queue to subset of resources

TORQUE does not currently provide a simple mechanism for mapping queues to nodes. However, schedulers such as [Moab](https://moabproject.org) and [Maui](https://www.mcs.anl.gov/research/projects/moab) can provide this functionality.

The simplest method is using `default_resources.neednodes` on an execution queue, setting it to a particular node attribute. Maui/Moab will use this information to ensure that jobs in that queue will be assigned nodes with that attribute. For example, suppose we have some nodes bought with money from the chemistry department, and some nodes paid by the biology department.

```bash
$TORQUE_HOME/server_priv/nodes:
node01 np=2 chem
node02 np=2 chem
node03 np=2 bio
node04 np=2 bio
qmgr:
set queue chem resources_default.neednodes=chem
set queue bio resources_default.neednodes=bio
```

This example does not preclude other queues from accessing those nodes. One solution is to use some other generic attribute with all other nodes and queues.

More advanced configurations can be made with standing reservations and QoSs.

Related topics

- **Queue configuration** on page 77

Creating a routing queue

A routing queue will steer a job to a destination queue based on job attributes and queue constraints. It is set up by creating a queue of `queue_type "Route"` with a `route_destinations` attribute set, as in the following example.
In this example, the compute nodes are dual processors and default walltimes are set according to the number of processors/nodes of a job. Jobs with 32 nodes (63 processors) or more will be given a default walltime of 6 hours. Also, jobs with 16-31 nodes (31-62 processors) will be given a default walltime of 12 hours. All other jobs will have the server default walltime of 24 hours.

The ordering of the route_destinations is important. In a routing queue, a job is assigned to the first possible destination queue based on the resources_max, resources_min, acl_users, and acl_groups attributes. In the preceding example, the attributes of a single processor job would first be checked against the reg_64 queue, then the reg_32 queue, and finally the reg queue.

Adding the following settings to the earlier configuration elucidates the queue resource requirements:

```
qmgr

set queue reg resources_max.ncpus = 30
set queue reg resources_max.nodect = 15
set queue reg 16 resources_max.ncpus = 62
set queue reg 16 resources_max.nodect = 31
```
The time of enforcement of server and queue defaults is important in this example. TORQUE applies server and queue defaults differently in job centric and queue centric modes. For job centric mode, TORQUE waits to apply the server and queue defaults until the job is assigned to its final execution queue. For queue centric mode, it enforces server defaults before it is placed in the routing queue. In either mode, queue defaults override the server defaults. TORQUE defaults to job centric mode. To set queue centric mode, set `queue_centric_limits`, as in what follows:

```
qmgr
set_server_queue_centric_limits = true
```

An artifact of job centric mode is that if a job does not have an attribute set, the server and routing queue defaults are not applied when queue resource limits are checked. Consequently, a job that requests 32 nodes (not `ncpus=32`) will not be checked against a `min_resource.ncpus` limit. Also, for the preceding example, a job without any attributes set will be placed in the `reg.64` queue, since the server `ncpus` default will be applied after the job is assigned to an execution queue.

### Routine queue defaults are not applied to job attributes in versions 2.1.0 and before.

### If the error message “qsub: Job rejected by all possible destinations” is reported when submitting a job, it may be necessary to add queue location information, (i.e., in the routing queue’s `route_destinations` attribute, change "batch" to "batch@localhost").

**Related topics**

- [Queue configuration](#) on page 77
- [Queue attributes](#) on page 78

---

### Server high availability

You can now run TORQUE in a redundant or high availability mode. This means that there can be multiple instances of the server running and waiting to take over processing in the event that the currently running server fails.

**The high availability feature is available in the 2.3 and later versions of TORQUE. TORQUE 2.4 includes several enhancements to high availability (see [Server high availability](#) on page 91).**

For more details, see these sections:

- [Redundant server host machines](#) on page 92
- [Server high availability](#) on page 91
- [Enhanced high availability with Moab](#) on page 93
- [How commands select the correct server host](#) on page 93
- [Job names](#) on page 93
Chapter 4: Setting server policies

- Persistence of the pbs_server process on page 94
- High availability of the NFS server on page 94
- Installing TORQUE in high availability mode on page 94
- Installing TORQUE in high availability mode on headless nodes on page 98
- Example setup of high availability on page 102

Redundant server host machines

High availability enables TORQUE to continue running even if pbs_server is brought down. This is done by running multiple copies of pbs_server which have their torque/server_priv directory mounted on a shared file system. The torque/server_name must include the host names of all nodes that run pbs_server. All MOM nodes also must include the host names of all nodes running pbs_server in their torque/server_name file. The syntax of the torque/server_name is a comma delimited list of host names.

For example:

```bash
[ host1,host2,host3
```

When configuring high availability, do not use $pbsserver to specify the host names. You must use the $TORQUEHOMEDIR/server_name file.

All instances of pbs_server need to be started with the --ha command line option that allows the servers to run at the same time. Only the first server to start will complete the full startup. The second server to start will block very early in the startup when it tries to lock the file torque/server_priv/server.lock. When the second server cannot obtain the lock, it will spin in a loop and wait for the lock to clear. The sleep time between checks of the lock file is one second.

Notice that not only can the servers run on independent server hardware, there can also be multiple instances of the pbs_server running on the same machine. This was not possible before as the second one to start would always write an error and quit when it could not obtain the lock.

Enabling high availability

To use high availability, you must start each instance of pbs_server with the --ha option.

Prior to version 4.0, TORQUE with HA was configured with an --enable-high-availability option. That option is no longer required.

Three server options help manage high availability. The server parameters are lock_file, lock_file_update_time, and lock_file_check_time.

The lock_file option allows the administrator to change the location of the lock file. The default location is torque/server_priv. If the lock_file option is used, the new location must be on the shared partition so all servers have access.

The lock_file_update_time and lock_file_check_time parameters are used by the servers to determine if the primary server is active. The primary pbs_server will update the lock file based on the lock_file_update_time (default value of 3 seconds). All backup pbs_servers will check the lock file as indicated by the lock_file_check_time parameter (default value of 9 seconds). The lock_file_update_time must be less
than the lock_file_check_time. When a failure occurs, the backup pbs_server takes up to the lock_file_check_time value to take over.

```
> qmgr -c "set server lock_file_check_time=5"
```

In the above example, after the primary pbs_server goes down, the backup pbs_server takes up to 5 seconds to take over. It takes additional time for all MOMs to switch over to the new pbs_server.

The clock on the primary and redundant servers must be synchronized in order for high availability to work. Use a utility such as NTP to ensure your servers have a synchronized time.

Enhanced high availability with Moab

When TORQUE is run with an external scheduler such as Moab, and the pbs_server is not running on the same host as Moab, pbs_server needs to know where to find the scheduler. To do this, use the following syntax (the port is required and the default is 15004):

```
> pbs_server --ha -l <moabhost:port>
```

If Moab is running in HA mode, add a -l option for each redundant server.

```
> pbs_server --ha -l <moabhost1:port> -l <moabhost2:port>
```

The root user of each Moab host must be added to the operators and managers lists of the server. This enables Moab to execute root level operations in TORQUE.

How commands select the correct server host

The various commands that send messages to pbs_server usually have an option of specifying the server name on the command line, or if none is specified will use the default server name. The default server name comes either from the environment variable PBS_DEFAULT or from the file torque/server_name.

When a command is executed and no explicit server is mentioned, an attempt is made to connect to the first server name in the list of hosts from PBS_DEFAULT or torque/server_name. If this fails, the next server name is tried. If all servers in the list are unreachable, an error is returned and the command fails.

Note that there is a period of time after the failure of the current server during which the new server is starting up where it is unable to process commands. The new server must read the existing configuration and job information from the disk, so the length of time that commands cannot be received varies.

Commands issued during this period of time might fail due to timeouts expiring.

Job names

One aspect of this enhancement is in the construction of job names. Job names normally contain the name of the host machine where pbs_server is running. When job names are constructed, only the first name from the server specification list is used in building the job name.
Persistence of the pbs_server process

The system administrator must ensure that pbs_server continues to run on the server nodes. This could be as simple as a cron job that counts the number of pbs_server's in the process table and starts some more if needed.

High availability of the NFS server

One consideration of this implementation is that it depends on NFS file system also being redundant. NFS can be set up as a redundant service. See the following:

- Setting Up A Highly Available NFS Server
- Making NFS Work On Your Network
- Sourceforge Linux NFS FAQ
- NFS v4 main site

There are also other ways to set up a shared file system. See the following:

- Red Hat Global File System
- Data sharing with a GFS storage cluster

Installing TORQUE in high availability mode

The following procedure demonstrates a TORQUE installation in high availability (HA) mode.

Requirements

- gcc (GCC) 4.1.2
- BASH shell
- Servers (these cannot be two VMs on the same hypervisor) configured the following way:
  - 2 main servers with identical architecture:
    - server1 — Primary server running TORQUE with a shared file system (this example uses NFS)
    - server2 — Secondary server running with TORQUE with a shared file system (this example uses NFS)
  - fileServer — Shared file system (this example uses NFS)
  - Compute nodes

These systems can be CentOS 5.7 or higher, RHEL 5.7 or higher, or SLES 6.3 or higher.

To install TORQUE in HA mode

1. Stop all firewalls or update your firewall to allow traffic from TORQUE services.

   ```
   > service iptables stop
   > chkconfig iptables off
   ```
Chapter 4: Setting server policies

If you are unable to stop the firewall due to infrastructure restriction, open the following ports:

- 15001 [tcp, udp]
- 15002 [tcp, udp]
- 15003 [tcp, udp]

2. Disable SELinux

```
> vi /etc/sysconfig/selinux
SELINUX=disabled
```

3. Update your main ~/.bashrc profile to ensure you are always referencing the applications to be installed on all servers.

```
# TORQUE
export TORQUEHOME=/var/spool/torque
# Library Path
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:$TORQUEHOME/lib
# Update system paths
export PATH=$TORQUEHOME/bin:$TORQUEHOME/sbin:$ {PATH}
```

4. Verify server1 and server2 are resolvable via either DNS or looking for an entry in the /etc/hosts file.

5. Configure the NFS Mounts by following these steps:
   a. Create mount point folders on fileServer.

```
fileServer# mkdir -m 0755 /var/spool/torque
fileServer# mkdir -m 0750 /var/spool/torque/server_priv
```
   b. Update /etc/exports on fileServer. The IP addresses should be that of server2.

```
/var/spool/torque/server_priv 192.168.0.0/255.255.255.0 (rw,sync,no_root_squash)
```
   c. Update the list of NFS exported file systems.

```
fileServer# exportfs -r
```

6. If the NFS daemons are not already running on fileServer, start them.

```
> systemctl restart rpcbind.service
> systemctl start nfs-server.service
> systemctl start nfs-lock.service
> systemctl start nfs-idmap.service
```

7. Mount the exported file systems on server1 by following these steps:
   a. Create the directory reference and mount them.

```
server1# mkdir /var/spool/torque/server_priv
```

Repeat this process for server2.
b. Update /etc/fstab on server1 to ensure that NFS mount is performed on startup.

```
fileServer:/var/spool/torque/server_priv /var/spool/torque/server_priv nfs
    rsize=8192,wsize=8192,timeo=14,intr
```

Repeat this step for server2.

8. Install TORQUE by following these steps:

   a. Download and extract TORQUE 4.2.6 on server1.

```
server1# wget http://github.com/adaptivecomputing/torque/branches/4.2.6/torque-4.2.6.tar.gz
server1# tar -xvzf torque-4.2.6.tar.gz
```

   b. Navigate to the TORQUE directory and compile TORQUE with the HA flag on server1.

```
server1# configure
server1# make
server1# make install
server1# make packages
```

   c. If the installation directory is shared on both head nodes, then run `make install` on server1.

```
server1# make install
```

   If the installation directory is not shared, repeat step 8a-b (downloading and installing TORQUE) on server2.


```
sERVER1# /etc/init.d/trqauthd start
```

10. Configure TORQUE for HA

   a. List the host names of all nodes that run pbs_server in the torque/server_name file. You must also include the host names of all nodes running pbs_server in the torque/server_name file of each MOM node. The syntax of torque/server_name is a comma-delimited list of host names.

```
server1
server2
```

   b. Create a simple queue configuration for TORQUE job queues on server1.

```
server1# pbs_server -t create
server1# qmgr -c "set server scheduling=true"
server1# qmgr -c "create queue batch queue_type=execution"
server1# qmgr -c "set queue batch started=true"
server1# qmgr -c "set queue batch enabled=true"
server1# qmgr -c "set queue batch resources_default.nodes=1"
server1# qmgr -c "set queue batch resources_default.walltime=3600"
server1# qmgr -c "set server default_queue=Batch"
```
Because `server_priv/*` is a shared drive, you do not need to repeat this step on server2.

c. Add the root users of TORQUE to the TORQUE configuration as an operator and manager.

```
server1# qmgr -c "set server managers += root@server1"
server1# qmgr -c "set server managers += root@server2"
server1# qmgr -c "set server operators += root@server1"
server1# qmgr -c "set server operators += root@server2"
```

Because `server_priv/*` is a shared drive, you do not need to repeat this step on server2.

d. You must update the lock file mechanism for TORQUE in order to determine which server is the primary. To do so, use the `lock_file_update_time` and `lock_file_check_time` parameters. The primary `pbs_server` will update the lock file based on the specified `lock_file_update_time` (default value of 3 seconds). All backup `pbs_servers` will check the lock file as indicated by the `lock_file_check_time` parameter (default value of 9 seconds). The `lock_file_update_time` must be less than the `lock_file_check_time`. When a failure occurs, the backup `pbs_server` takes up to the `lock_file_check_time` value to take over.

```
server1# qmgr -c "set server lock_file_check_time=5"
server1# qmgr -c "set server lock_file_update_time=3"
```

Because `server_priv/*` is a shared drive, you do not need to repeat this step on server2.

e. List the servers running `pbs_server` in the TORQUE `acl_hosts` file.

```
server1# qmgr -c "set server acl_hosts += server2"
```

Because `server_priv/*` is a shared drive, you do not need to repeat this step on server2.

f. Restart the running `pbs_server` in HA mode.

```
server1# qterm
```

g. Start the `pbs_server` on the secondary server.

```
server1# pbs_server --ha -l server2:port
server2# pbs_server --ha -l server1:port
```

11. Check the status of TORQUE in HA mode.

```
server1# qmgr -c "p s"
server2# qmgr -c "p s"
```

The commands above returns all settings from the active TORQUE server from either node.
Drop one of the pbs_servers to verify that the secondary server picks up the request.

```
server1# qterm
server2# qmgr -c "p s"
```

Stop the pbs_server on server2 and restart pbs_server on server1 to verify that both nodes can handle a request from the other.

12. Install a pbs_mom on the compute nodes.
   a. Copy the install scripts to the compute nodes and install.
   b. Navigate to the shared source directory of TORQUE and run the following:

```
node1# torque-package-mom-linux-x86_64.sh --install
node2# torque-package-clients-linux-x86_64.sh --install
```

Repeat this for each compute node. Verify that the /var/pool/torque/server-name file shows all your compute nodes.

c. On server1 or server2, configure the nodes file to identify all available MOMs. To do so, edit the /var/spool/torque/server_priv/nodes file.

```
node1 np=2
node2 np=2
```

**Change the np flag to reflect number of available processors on that node.**

d. Recycle the pbs_servers to verify that they pick up the MOM configuration.

```
server1# qterm; pbs_server --ha -l server2:port
server2# qterm; pbs_server --ha -l server1:port
```

e. Start the pbs_mom on each execution node.

```
node5# pbs_mom
node6# pbs_mom
```

Installing TORQUE in high availability mode on headless nodes

The following procedure demonstrates a TORQUE installation in high availability (HA) mode on nodes with no local hard drive.

Chapter 4: Setting server policies
Requirements

- gcc (GCC) 4.1.2
- BASH shell
- Servers (these cannot be two VMs on the same hypervisor) configured the following way:
  - 2 main servers with identical architecture
    - server1 — Primary server running TORQUE with a file system share (this example uses NFS)
    - server2 — Secondary server running with TORQUE with a file system share (this example uses NFS)
  - Compute nodes
  - fileServer — A shared file system server (this example uses NFS)

These systems can be CentOS 5.7 or higher, RHEL 5.7 or higher, or SLES 6.3 or higher.

To install TORQUE in HA mode on a node with no local hard drive

1. Stop all firewalls or update your firewall to allow traffic from TORQUE services.

```
> service iptables stop
> chkconfig iptables off
```

If you are unable to stop the firewall due to infrastructure restriction, open the following ports:

- 15001 [tcp,udp]
- 15002 [tcp,udp]
- 15003 [tcp,udp]

2. Disable SELinux

```
> vi /etc/sysconfig/selinux
```

```
SELINUX=disabled
```

3. Update your main ~/.bashrc profile to ensure you are always referencing the applications to be installed on all servers.

```
# TORQUE
export TORQUEHOME=/var/spool/torque
# Library Path
export LD_LIBRARY_PATH=${LD_LIBRARY_PATH}:${TORQUEHOME}/lib
# Update system paths
export PATH=${TORQUEHOME}/bin:${TORQUEHOME}/sbin:${PATH}
```

4. Verify server1 and server2 are resolvable via either DNS or looking for an entry in the /etc/hosts file.

5. Configure the NFS Mounts by following these steps:
a. Create mount point folders on fileServer.

```
fileServer# mkdir -m 0755 /var/spool/torque
```

b. Update `/etc/exports` on fileServer. The IP addresses should be that of server2.

```
/var/spool/torque/ 192.168.0.0/255.255.255.0(rw,sync,no_root_squash)
```

c. Update the list of NFS exported file systems.

```
fileServer# exportfs -r
```

6. If the NFS daemons are not already running on fileServer, start them.

```
> systemctl restart rpcbind.service
> systemctl start nfs-server.service
> systemctl start nfs-lock.service
> systemctl start nfs-idmap.service
```

7. Mount the exported file systems on server1 by following these steps:

a. Create the directory reference and mount them.

```
server1# mkdir /var/spool/torque
```

Repeat this process for server2.

b. Update `/etc/fstab` on server1 to ensure that NFS mount is performed on startup.

```
fileServer:/var/spool/torque/server_priv /var/spool/torque/server_priv nfs rsize=8192,wsize=8192,timeo=14,intr
```

Repeat this step for server2.

8. Install TORQUE by following these steps:

a. Download and extract TORQUE 4.2.6 on server1.

```
server1# wget http://github.com/adaptivecomputing/torque/branches/4.2.6/torque-4.2.6.tar.gz
server1# tar -xvzf torque-4.2.6.tar.gz
```

b. Navigate to the TORQUE directory and compile TORQUE with the HA flag on server1.

```
server1# configure --prefix=/var/spool/torque
server1# make
server1# make install
server1# make packages
```

c. If the installation directory is shared on both head nodes, then run `make install` on server1.

```
server1# make install
```

If the installation directory is not shared, repeat step 8a-b (downloading and installing TORQUE) on server2.
   ```bash
   server1# /etc/init.d/trqauthd start
   ```

10. Configure TORQUE for HA.
   a. List the host names of all nodes that run pbs_server in the `torque/server_name` file. You must also include the host names of all nodes running pbs_server in the `torque/server_name` file of each MOM node. The syntax of `torque/server_name` is a comma-delimited list of host names.
   ```bash
   server1,server2
   ```
   
   b. Create a simple queue configuration for TORQUE job queues on server1.
   ```bash
   server1# pbs_server -t create
   server1# qmgr -c "set server scheduling=true"
   server1# qmgr -c "create queue batch queue type=execution"
   server1# qmgr -c "set queue batch started=true"
   server1# qmgr -c "set queue batch enabled=true"
   server1# qmgr -c "set queue batch resources_default.nodes=1"
   server1# qmgr -c "set queue batch resources_default.walltime=3600"
   ```
   
   Because TORQUE_HOME is a shared drive, you do not need to repeat this step on server2.

c. Add the root users of TORQUE to the TORQUE configuration as an operator and manager.
   ```bash
   server1# qmgr -c "set server managers += root@server1"
   server1# qmgr -c "set server managers += root@server2"
   server1# qmgr -c "set server operators += root@server1"
   server1# qmgr -c "set server operators += root@server2"
   ```
   
   Because TORQUE_HOME is a shared drive, you do not need to repeat this step on server2.

d. You must update the lock file mechanism for TORQUE in order to determine which server is the primary. To do so, use the `lock_file_update_time` and `lock_file_check_time` parameters. The primary pbs_server will update the lock file based on the specified `lock_file_update_time` (default value of 3 seconds). All backup pbs_servers will check the lock file as indicated by the `lock_file_check_time` parameter (default value of 9 seconds). The `lock_file_update_time` must be less than the `lock_file_check_time`. When a failure occurs, the backup pbs_server takes up to the `lock_file_check_time` value to take over.
   ```bash
   server1# qmgr -c "set server lock_file_check_time=5"
   server1# qmgr -c "set server lock_file_update_time=3"
   ```
   
   Because TORQUE_HOME is a shared drive, you do not need to repeat this step on server2.

e. List the servers running pbs_server in the TORQUE `acl_hosts` file.
   ```bash
   server1# qmgr -c "set server acl_hosts += server1"
   server1# qmgr -c "set server acl_hosts += server2"
   ```
Because `TORQUEHOME` is a shared drive, you do not need to repeat this step on `server2`.

f. Restart the running `pbs_server` in HA mode.

   ```
   server1# qterm
   server2# qterm
   ```

g. Start the `pbs_server` on the secondary server.

   ```
   server1# pbs_server --ha -l server2:port
   server2# pbs_server --ha -l server1:port
   ```

11. Check the status of `TORQUE` in HA mode.

   ```
   server1# qmgr -c "p s"
   server2# qmgr -c "p s"
   ```

   *The commands above returns all settings from the active `TORQUE` server from either node.*

Drop one of the `pbs_servers` to verify that the secondary server picks up the request.

   ```
   server1# qterm
   server2# qmgr -c "p s"
   ```

Stop the `pbs_server` on `server2` and restart `pbs_server` on `server1` to verify that both nodes can handle a request from the other.

12. Install a `pbs_mom` on the compute nodes.

   a. On `server1` or `server2`, configure the `nodes` file to identify all available MOMs. To do so, edit the `/var/spool/torque/server_priv/nodes` file.

      ```
      node1 np=2
      node2 np=2
      ```

      *Change the `np` flag to reflect number of available processors on that node.*

   b. Recycle the `pbs_servers` to verify that they pick up the MOM configuration.

      ```
      server1# qterm; pbs_server --ha -l server2:port
      server2# qterm; pbs_server --ha -l server1:port
      ```

   c. Start the `pbs_mom` on each execution node.

      ```
      server1# pbs_mom -d <mom-server1>
      server2# pbs_mom -d <mom-server2>
      ```

**Example setup of high availability**

1. The machines running `pbs_server` must have access to a shared `server_priv` directory (usually an NFS share on a MoM).

2. All MoMs must have the same content in their `server_name` file. This can be done manually or via an NFS share. The `server_name` file contains a comma-delimited list of the hosts that run `pbs_server`.  

---

**Server high availability**
3. The machines running pbs_server must be listed in acl_hosts.

```bash
> qmgr -c "set server acl_hosts += server1"
> qmgr -c "set server acl_hosts += server2"
```

4. Start pbs_server with the --ha option.

```bash
[root@server1] $ pbs_server --ha
[root@server2] $ pbs_server --ha
```

Related topics

- Setting server policies on page 77
- Queue configuration on page 77
Chapter 5: Integrating schedulers for TORQUE

Selecting the cluster scheduler is an important decision and significantly affects cluster utilization, responsiveness, availability, and intelligence. The default TORQUE scheduler, pbs_sched, is very basic and will provide poor utilization of your cluster’s resources. Other options, such as Maui Scheduler or Moab Workload Manager, are highly recommended. If you are using Maui or Moab, refer to the Moab-PBS Integration Guide. If using pbs_sched, simply start the pbs_sched daemon.

If you are installing Moab Cluster Suite, TORQUE and Moab were configured at installation for interoperability and no further action is required.
Chapter 6: Configuring data management

This section contains information about SCP-based data management with TORQUE. It describes how to use TORQUE with NFS and other networked file systems. It also outlines file staging requirements. For details, see these topics:

- SCP setup on page 107
- NFS and other networked filesystems on page 110
- File stage-in/stage-out on page 111

SCP setup

To use SCP-based data management, TORQUE must be authorized to migrate data to any of the compute nodes. If this is not already enabled within the cluster, this can be achieved with the process described below. This process enables uni-directional access for a particular user from a source host to a destination host.

These directions were written using OpenSSH version 3.6 and may not transfer correctly to older versions.

To set up TORQUE for SCP, follow the directions in each of these topics:

- Generating SSH key on source host on page 107
- Copying public SSH key to each destination host on page 108
- Configuring the SSH daemon on each destination host on page 108
- Validating correct SSH configuration on page 109
- Enabling bi-directional SCP access on page 109
- Compiling TORQUE to support SPC on page 109
- Troubleshooting on page 110

Related topics

- Configuring data management on page 107

Generating SSH key on source host

On the source host as the transfer user, execute the following:
This will prompt for a passphrase (optional) and create two files (id_rsa and id_rsa.pub) inside ~/.ssh/.

Related topics

- SCP setup on page 107
- Copying public SSH key to each destination host on page 108

Copying public SSH key to each destination host

Transfer public key to each destination host as the transfer user:

Easy key copy:

```
> ssh-copy-id [-i [identity_file]] [user@]machine
```

Manual steps to copy keys:

```
> scp ~/.ssh/id_rsa.pub destHost:~ (enter password)
```

Create an authorized_keys file on each destination host:

```
> ssh destHost (enter password)
> cat id_rsa.pub >> .ssh/authorized_keys
```

If the .ssh directory does not exist, create it with 700 privileges (mkdir .ssh; chmod 700 .ssh):

```
> chmod 700 .ssh/authorized_keys
```

Related topics

- Generating SSH key on source host on page 107
- SCP setup on page 107

Configuring the SSH daemon on each destination host

Some configuration of the SSH daemon may be required on the destination host. (Because this is not always the case, see Validating correct SSH configuration on page 109 and test the changes made to this point. If the tests fail, proceed with this step and then try testing again.) Typically, this is done by editing the /etc/ssh/sshd_config file (root access needed). To verify correct configuration, see that the following attributes are set (not commented):

```
RSAAuthentication yes
PubkeyAuthentication yes
```

If configuration changes were required, the SSH daemon will need to be restarted (root access needed):

```
> /etc/init.d/sshd restart
```
Validating correct SSH configuration

If all is properly configured, the following command issued on the *source* host should succeed and not prompt for a password:

```bash
> scp destHost:/etc/motd /tmp
```

> If this is your first time accessing *destination* from *source*, it may ask you if you want to add the fingerprint to a file of known hosts. If you specify yes, this message should no longer appear and should not interfere with scp copying via TORQUE. Also, it is important that the full hostname appear in the *known_hosts* file. To do this, use the full hostname for *destHost*, as in *machine.domain.org* instead of just *machine*.

Related topics

- SCP setup on page 107

Enabling bi-directional SCP access

The preceding steps allow *source* access to *destination* without prompting for a password. The reverse, however, is not true. Repeat the steps, but this time using the *destination* as the *source*, etc. to enable bi-directional SCP access (i.e. *source* can send to *destination* and *destination* can send to *source* without password prompts.)

Related topics

- SCP setup on page 107

Compiling TORQUE to support SPC

> In TORQUE 2.1 and later, SCP is the default remote copy protocol. These instructions are only necessary for earlier versions.

TORQUE must be re-configure (and then rebuilt) to use SCP by passing in the `--with-scp` flag to the configure script:

```bash
> ./configure --prefix=xxx --with-scp
> make
```

> If special SCP flags are required in your local setup, these can be specified using the `$rcpcmd` parameter.
Troubleshooting

If, after following all of the instructions in this section (see SCP setup on page 107), TORQUE is still having problems transferring data with SCP, set the PBSDEBUG environment variable and restart the pbs_mom for details about copying. Also check the MOM log files for more details.

Related topics

- SCP setup on page 107

NFS and other networked filesystems

When a batch job starts, its stdin file (if specified) is copied from the submission directory on the remote submission host. This file is placed in the $PBSMOMHOME directory on the mother superior node (i.e., /usr/spool/PBS/spool). As the job runs, stdout and stderr files are generated and placed in this directory using the naming convention $JOBID.OU and $JOBID.ER.

When the job completes, the MOM copies the files into the directory from which the job was submitted. By default, this file copying will be accomplished using a remote copy facility such as rcp or scp.

If a shared file system such as NFS, DFS, or AFS is available, a site can specify that the MOM should take advantage of this by specifying the $usecp directive inside the MOM configuration file (located in the $PBSMOMHOME/mom_priv directory) using the following format:

$usecp <HOST>:<SRCDIR> <DSTDIR>

$<HOST> can be specified with a leading wildcard (*) character. The following example demonstrates this directive:

```
mom_priv/config
# /home is NFS mounted on all hosts
$usecp *:/home /home
# submission hosts in domain fte.com should map '/data' directory on submit host to
# '/usr/local/data' on compute host
$usecp *:fte.com:/data /usr/local/data
```

If for any reason the MOM daemon is unable to copy the output or error files to the submission directory, these files are instead copied to the undelivered directory also located in $PBSMOMHOME.

Related topics

- Configuring data management on page 107
File stage-in/stage-out

File staging requirements are specified using the `stagein` and `stageout` directives of the `qsub` command. Stagein requests occur before the job starts execution, while stageout requests happen after a job completes.

On completion of the job, all staged-in and staged-out files are removed from the execution system. The `file_list` is in the form `local_file@hostname:remote_file[,...]` regardless of the direction of the copy. The name `local_file` is the name of the file on the system where the job executed. It may be an absolute path or relative to the home directory of the user. The name `remote_file` is the destination name on the host specified by hostname. The name may be absolute or relative to the user's home directory on the destination host. The use of wildcards in the file name is not recommended.

The file names map to a remote copy program (rcp/scp/cp, depending on configuration) called on the execution system in the following manner:

For `stagein`: `rcp/scp hostname:remote_file local_file`

For `stageout`: `rcp/scp local_file hostname:remote_file`

Examples

```bash
# stage /home/john/input_source.txt from node13.fsc to /home/john/input_
destination.txt on master compute node
> qsub -l nodes=1,walltime=100 -W stagein=input_
source.txt@node13.fsc:/home/john/input_destination.txt

# stage /home/bill/output_source.txt on master compute node to /tmp/output_
destination.txt on node15.fsc
> qsub -l nodes=1,walltime=100 -W stageout=/tmp/output_
source.txt@node15.fsc:/home/bill/output_destination.txt

$ fortune >xxx;echo cat xxx|qsub -W stagein=xxx@`hostname`:xxx
199.myhost.mydomain
$ cat STDIN*199
Anyone who has had a bull by the tail knows five or six more things
than someone who hasn't.
-- Mark Twain
```

Related topics

- [Configuring data management](#) on page 107
Chapter 7: MPI (Message Passing Interface) support

A message passing library is used by parallel jobs to augment communication between the tasks distributed across the cluster. TORQUE can run with any message passing library and provides limited integration with some MPI libraries.

For more information, see these topics:

- **MPICH** on page 113
- **Open MPI** on page 114

### MPICH

One of the most popular MPI libraries is MPICH available from Argonne National Lab. If using this release, you may want to consider also using the mpiexec tool for launching MPI applications. Support for mpiexec has been integrated into TORQUE.

#### MPIExec Overview

mpiexec is a replacement program for the script mpirun, which is part of the mpich package. It is used to initialize a parallel job from within a PBS batch or interactive environment. mpiexec uses the task manager library of PBS to spawn copies of the executable on the nodes in a PBS allocation.

Reasons to use mpiexec rather than a script (mpirun) or an external daemon (mpd):

- Starting tasks with the task manager (TM) interface is much faster than invoking a separate rsh * once for each process.
- Resources used by the spawned processes are accounted correctly with mpiexec, and reported in the PBS logs, because all the processes of a parallel job remain under the control of PBS, unlike when using mpirun-like scripts.
- Tasks that exceed their assigned limits of CPU time, wallclock time, memory usage, or disk space are killed cleanly by PBS. It is quite hard for processes to escape control of the resource manager when using mpiexec.
- You can use mpiexec to enforce a security policy. If all jobs are forced to spawn using mpiexec and the PBS execution environment, it is not necessary to enable rsh or ssh access to the compute nodes in the cluster.

For more information, see the mpiexec homepage.
MPIExec Troubleshooting

Although problems with mpiexec are rare, if issues do occur, the following steps may be useful:

- Determine current version using `mpiexec --version` and review the change log available on the [MPI homepage](https://www.mpi.gov) to determine if the reported issue has already been corrected.
- Send email to the mpiexec mailing list at [mpiexec@osc.edu](mailto:mpiexec@osc.edu).
- Browse the mpiexec user list archives for similar problems and resolutions.
- Read the FAQ contained in the README file and the mpiexec man pages contained within the mpiexec distribution.
- Increase the logging of mpiexec operation with `mpiexec --verbose` (reports messages to stderr).
- Increase logging of the master and slave resource manager execution daemons associated with the job (with TORQUE, use `$loglevel to 5 or higher in $TORQUEROOT/mom_priv/config` and look for 'tm' messages after associated join job messages).
- Use `tracejob` (included with TORQUE) or `qtracejob` (included with OSC's pbstools package) to isolate failures within the cluster.
- If the message 'exec: Error: get_hosts: pbs_connect: Access from host not allowed, or unknown host' appears, this indicates that mpiexec cannot communicate with the pbs_server daemon. In most cases, this indicates that the $TORQUEROOT/server_name file points to the wrong server or the node cannot resolve the server's name. The `qstat` command can be run on the node to test this.

General MPI Troubleshooting

When using MPICH, some sites have issues with orphaned MPI child processes remaining on the system after the master MPI process has been terminated. To address this, TORQUE epilogue scripts can be created that properly clean up the orphaned processes (see Prologue and epilogue scripts on page 281).

Related topics
- [MPI (Message Passing Interface) support](chapter7.html#mpich) on page 113

Open MPI

Open MPI is a new MPI implementation that combines technologies from multiple projects to create the best possible library. It supports the TM interface for integration with TORQUE. More information is available in the FAQ.

Related topics
- [MPI (Message Passing Interface) support](chapter7.html#mpich) on page 113
Chapter 8: Resources

A primary task of any resource manager is to monitor the state, health, configuration, and utilization of managed resources. TORQUE is specifically designed to monitor compute hosts for use in a batch environment. TORQUE is not designed to monitor non-compute host resources such as software licenses, networks, file systems, and so forth, although these resources can be integrated into the cluster using some scheduling systems.

With regard to monitoring compute nodes, TORQUE reports about a number of attributes broken into three major categories:

- **Configuration** on page 115
- **Utilization** on page 116
- **Node states** on page 116

### Configuration

Configuration includes both detected hardware configuration and specified batch attributes.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture (arch)</td>
<td>operating system of the node</td>
<td>The value reported is a derivative of the operating system installed.</td>
</tr>
<tr>
<td>Node Features (properties)</td>
<td>arbitrary string attributes associated with the node</td>
<td>No node features are specified by default. If required, they are set using the nodes file located in the TORQUE_HOME/server_priv directory. They may specify any string and are most commonly used to allow users to request certain subsets of nodes when submitting jobs.</td>
</tr>
<tr>
<td>Local Disk (size)</td>
<td>configured local disk</td>
<td>By default, local disk space is not monitored. If the MOM configuration size [fs=&lt;FS&gt;] parameter is set, TORQUE will report, in kilobytes, configured disk space within the specified directory.</td>
</tr>
<tr>
<td>Memory (physmem)</td>
<td>local memory/RAM</td>
<td>Local memory/RAM is monitored and reported in kilobytes.</td>
</tr>
</tbody>
</table>
### Processors (ncpus/np)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processors</td>
<td>real/virtual processors</td>
<td>The number of processors detected by TORQUE is reported via the <em>ncpus</em> attribute. However, for scheduling purposes, other factors are taken into account. In its default configuration, TORQUE operates in &quot;dedicated&quot; mode with each node possessing a single virtual processor. In dedicated mode, each job task will consume one virtual processor and TORQUE will accept workload on each node until all virtual processors on that node are in use. While the number of virtual processors per node defaults to 1, this may be configured using the nodes file located in the TORQUE_HOME/server_priv directory. An alternative to dedicated mode is &quot;timeshared&quot; mode. If TORQUE's timeshared mode is enabled, TORQUE will accept additional workload on each node until the node's maxload limit is reached.</td>
</tr>
<tr>
<td>Swap (tot-mem)</td>
<td>virtual memory/swap</td>
<td>Virtual memory/swap is monitored and reported in kilobytes.</td>
</tr>
</tbody>
</table>

### Utilization

Utilization includes information regarding the amount of node resources currently in use as well as information about who or what is consuming it.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk (size)</td>
<td>local disk availability</td>
<td>By default, local disk space is not monitored. If the MOM configuration <em>size [fs=&lt;FS&gt;]</em> parameter is set, TORQUE will report configured and currently available disk space within the specified directory in kilobytes.</td>
</tr>
<tr>
<td>Memory (availmem)</td>
<td>real memory/RAM</td>
<td>Available real memory/RAM is monitored and reported in kilobytes.</td>
</tr>
<tr>
<td>Network (netload)</td>
<td>local network adapter usage</td>
<td>Reports total number of bytes transferred in or out by the network adapter.</td>
</tr>
<tr>
<td>Processor Utilization (loadave)</td>
<td>node's cpu load average</td>
<td>Reports the node's 1 minute bsd load average.</td>
</tr>
</tbody>
</table>

### Node states

State information includes administrative status, general node health information, and general usage status.
<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idle Time (idletime)</td>
<td>time since local keyboard/mouse activity has been detected</td>
<td>Time in seconds since local keyboard/mouse activity has been detected.</td>
</tr>
<tr>
<td>State (state)</td>
<td>monitored/admin node state</td>
<td>A node can be in one or more of the following states:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- busy - node is full and will not accept additional work</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- down - node is failing to report, is detecting local failures with node</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- free - node is ready to accept additional work</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- job-exclusive - all available virtual processors are assigned to jobs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- job-sharing - node has been allocated to run multiple shared jobs and will remain in this state until jobs are complete</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- offline - node has been instructed by an admin to no longer accept work</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- reserve - node has been reserved by the server</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- time-shared - node always allows multiple jobs to run concurrently</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- unknown - node has not been detected</td>
</tr>
</tbody>
</table>
Chapter 9: Accounting records

TORQUE maintains accounting records for batch jobs in the following directory:

$TORQUEROOT/server_priv/accounting/<TIMESTAMP>

$TORQUEROOT defaults to /usr/spool/PBS and <TIMESTAMP> is in the format: YYYYMMDD.

These records include events, time stamps, and information on resources requested and used. Records for four different event types are produced and are described in the following table:

<table>
<thead>
<tr>
<th>Record marker</th>
<th>Record type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>abort</td>
<td>Job has been aborted by the server</td>
</tr>
<tr>
<td>C</td>
<td>checkpoint</td>
<td>Job has been checkpointed and held</td>
</tr>
<tr>
<td>D</td>
<td>delete</td>
<td>Job has been deleted</td>
</tr>
<tr>
<td>E</td>
<td>exit</td>
<td>Job has exited (either successfully or unsuccessfully)</td>
</tr>
<tr>
<td>Q</td>
<td>queue</td>
<td>Job has been submitted/queued</td>
</tr>
<tr>
<td>R</td>
<td>rerun</td>
<td>Attempt to rerun the job has been made</td>
</tr>
<tr>
<td>S</td>
<td>start</td>
<td>Attempt to start the job has been made (if the job fails to properly start, it may have multiple job start records)</td>
</tr>
<tr>
<td>T</td>
<td>restart</td>
<td>Attempt to restart the job (from checkpoint) has been made (if the job fails to properly start, it may have multiple job start records)</td>
</tr>
</tbody>
</table>

Accounting Variables

The following table offers accounting variable descriptions. Descriptions for accounting variables not indicated in the table, particularly those prefixed with Resources_List, are available at Job submission on page 39.
## Variable Description

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ctime</td>
<td>Time job was created</td>
</tr>
<tr>
<td>etime</td>
<td>Time job became eligible to run</td>
</tr>
<tr>
<td>qtime</td>
<td>Time job was queued</td>
</tr>
<tr>
<td>start</td>
<td>Time job started to run</td>
</tr>
</tbody>
</table>

A sample record in this file can look like the following:

```
06/06/2005 14:04:25;D;408.ign1.zeta2.org;requestor=guest@ign1.zeta2.org
06/06/2005 14:04:35;Q;409.ign1.zeta2.org;queue=batch
06/06/2005 14:04:44;Q;410.ign1.zeta2.org;queue=batch
06/06/2005 14:06:06;S;407.ign1.zeta2.org;user=guest
                             group=guest
                             jobname=STDIN
                             queue=batch
                             ctime=1118087915
                             qtime=1118087915
                             etime=1118087915
                             start=1118088366
                             exec_host=ign1.zeta2.org/0
                             Resource_List.nodect=1
                             Resource_List.nodes=1
                             Resource_List.walltime=00:16:40
06/06/2005 14:07:17;D;407.ign1.zeta2.org;requestor=guest@ign1.zeta2.org
06/06/2005 14:07:17;E;407.ign1.zeta2.org;user=guest
                             group=guest
                             jobname=STDIN
                             queue=batch
                             ctime=1118087915
                             qtime=1118087915
                             etime=1118087915
                             start=1118088366
                             exec_host=ign1.zeta2.org/0
                             Resource_List.nodect=1
                             Resource_List.nodes=1
                             Resource_List.walltime=00:16:40
                             session=6365
                             end=1118088437
                             Exit_status=271
                             resources_used.cput=00:00:00
                             resources_used.mem=3068kb
                             resources_used.vmem=16080kb
                             resources_used.walltime=00:01:11
```

> The value of Resource_List.* is the amount of resources requested, and the value of resources_used.* is the amount of resources actually used.
Chapter 10: Job logging

New in TORQUE 2.5.3 is the ability to log job information for completed jobs. The information stored in the log file is the same information produced with the command `qstat -f`. The log file data is stored using an XML format. Data can be extracted from the log using the utility `showjobs` found in the `contrib/` directory of the TORQUE source tree. Custom scripts that can parse the XML data can also be used.

For details about job logging, see these topics:

- Job log location and name on page 121
- Enabling job logs on page 121

Job log location and name

When job logging is enabled (See Enabling job logs on page 121), the job log is kept at `$TORQUE_HOME/job_logs`. The naming convention for the job log is the same as for the server log or MOM log. The log name is created from the current year/month/day.

For example, if today's date is 26 October, 2010 the log file is named 20101026.

A new log file is created each new day that data is written to the log.

Related topics

- Enabling job logs on page 121
- Job logging on page 121

Enabling job logs

There are five new server parameters used to enable job logging. These parameters control what information is stored in the log and manage the log files.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>record_job_info</td>
<td>This must be set to true in order for job logging to be enabled. If not set to true, the remaining server parameters are ignored.</td>
</tr>
</tbody>
</table>
### Parameter Description

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>record_job_script</td>
<td>If set to true, this adds the contents of the script executed by a job to the log.</td>
</tr>
<tr>
<td>job_log_file_max_size</td>
<td>This specifies a soft limit (in kilobytes) for the job log’s maximum size. The file size is checked every five minutes and if the current day file size is greater than or equal to this value, it is rolled from <code>&lt;filename&gt;</code> to <code>&lt;filename.1&gt;</code> and a new empty log is opened. If the current day file size exceeds the maximum size a second time, the <code>&lt;filename.1&gt;</code> log file is rolled to <code>&lt;filename.2&gt;</code>, the current log is rolled to <code>&lt;filename.1&gt;</code>, and a new empty log is opened. Each new log causes all other logs to roll to an extension that is one greater than its current number. Any value less than 0 is ignored by pbs_server (meaning the log will not be rolled).</td>
</tr>
<tr>
<td>job_log_file_roll_depth</td>
<td>This sets the maximum number of new log files that are kept in a day if the job_log_file_max_size parameter is set. For example, if the roll depth is set to 3, no file can roll higher than <code>&lt;filename.3&gt;</code>. If a file is already at the specified depth, such as <code>&lt;filename.3&gt;</code>, the file is deleted so it can be replaced by the incoming file roll, <code>&lt;filename.2&gt;</code>.</td>
</tr>
<tr>
<td>job_log_keep_days</td>
<td>This maintains logs for the number of days designated. If set to 4, any log file older than 4 days old is deleted.</td>
</tr>
</tbody>
</table>

### Related topics

- [Job log location and name](#) on page 121
- [Job logging](#) on page 121
Chapter 11: Troubleshooting

There are a few general strategies that can be followed to determine the cause of unexpected behavior. These are a few of the tools available to help determine where problems occur. See these topics for details:

- **Host resolution** on page 123
- **Firewall configuration** on page 124
- **TORQUE log files** on page 124
- **Using "tracejob" to locate job failures** on page 125
- **Using GDB to locate job failures** on page 127
- **Other diagnostic options** on page 128
- **Stuck jobs** on page 128
- **Frequently asked questions (FAQ)** on page 129
- **Compute node health check** on page 134
- **Debugging** on page 136

**Host resolution**

The TORQUE server host must be able to perform both forward and reverse name lookup on itself and on all compute nodes. Likewise, each compute node must be able to perform forward and reverse name lookup on itself, the TORQUE server host, and all other compute nodes. In many cases, name resolution is handled by configuring the node's `/etc/hosts` file although DNS and NIS services may also be used. Commands such as `nslookup` or `dig` can be used to verify proper host resolution.

Invalid host resolution may exhibit itself with compute nodes reporting as down within the output of `pbsnodes -a` and with failure of the `momctl -d3` command.

**Related topics**

- **Troubleshooting** on page 123
Firewall configuration

Be sure that, if you have firewalls running on the server or node machines, you allow connections on the appropriate ports for each machine. TORQUE pbs_mom daemons use UDP ports 1023 and below if privileged ports are configured (privileged ports is the default). The pbs_server and pbs_mom daemons use TCP and UDP ports 15001-15004 by default.

Firewall based issues are often associated with server to MOM communication failures and messages such as ‘premature end of message’ in the log files.

Also, the tcpdump program can be used to verify the correct network packets are being sent.

Related topics
  - Troubleshooting on page 123

TORQUE log files

pbs_server and pbs_mom log files

The pbs_server keeps a daily log of all activity in the TORQUE_HOME/server_logs directory. The pbs_mom also keeps a daily log of all activity in the TORQUE_HOME/mom_logs/ directory. These logs contain information on communication between server and MOM as well as information on jobs as they enter the queue and as they are dispatched, run, and terminated. These logs can be very helpful in determining general job failures. For MOM logs, the verbosity of the logging can be adjusted by setting the $loglevel parameter in the mom_priv/config file. For server logs, the verbosity of the logging can be adjusted by setting the server log_level attribute in qmgr.

For both pbs_mom and pbs_server daemons, the log verbosity level can also be adjusted by setting the environment variable PBSLOGLEVEL to a value between 0 and 7. Further, to dynamically change the log level of a running daemon, use the SIGUSR1 and SIGUSR2 signals to increase and decrease the active log level by one. Signals are sent to a process using the kill command.

For example, kill -USR1 `pgrep pbs_mom` would raise the log level up by one.

The current loglevel for pbs_mom can be displayed with the command momctl -d3.

trqauthd log files

As of TORQUE 4.1.3, trqauthd logs its events in the $TORQUE_HOME/client_logs directory. It names the log files in the format <YYYYMMDD>, creating a new log daily as events occur.
Using "tracejob" to locate job failures

Overview

The tracejob utility extracts job status and job events from accounting records, MOM log files, server log files, and scheduler log files. Using it can help identify where, how, a why a job failed. This tool takes a job id as a parameter as well as arguments to specify which logs to search, how far into the past to search, and other conditions.

Syntax

```
tracejob [-a|s|l|m|q|v|z] [-c count] [-w size] [-p path] [-n <DAYS>] [-f filter_type] <JOBID>
```

- `-p` : path to PBS_SERVER_HOME
- `-w` : number of columns of your terminal
- `-n` : number of days in the past to look for job(s) [default 1]
- `-f` : filter out types of log entries, multiple -f's can be specified error, system, admin, job, job_usage, security, sched, debug,
debug2, or absolute numeric hex equivalent
-z : toggle filtering excessive messages
-c : what message count is considered excessive
-a : don't use accounting log files
-s : don't use server log files
-l : don't use scheduler log files
-m : don't use MOM log files
-q : quiet mode - hide all error messages
-v : verbose mode - show more error messages

Example

> tracejob -n 10 1131

Job: 1131.icluster.org

Job: 1131.icluster.org
03/02/2005 17:58:28 S enqueuing into batch, state 1 hop 1
03/02/2005 17:58:28 S Job Queued at request of dev@icluster.org, owner =
            dev@icluster.org, job name = STDIN, queue = batch
03/02/2005 17:58:28 A queue=batch
03/02/2005 17:58:41 S Job Run at request of dev@icluster.org
03/02/2005 17:58:41 M evaluating limits for job
03/02/2005 17:58:41 M phase 2 of job launch successfully completed
03/02/2005 17:58:41 M saving task (TMomFinalizeJob3)
03/02/2005 17:58:41 M job successfully started
03/02/2005 17:58:41 M job 1131.koa.icluster.org reported successful start on 1 node
(a)
03/02/2005 17:58:41 A user=dev group=dev jobname=STDIN
qtime=1109811508 etime=1109811508 start=1109811521
03/02/2005 17:58:41
03/02/2005 18:02:11 M walltime 210 exceeded limit 100
03/02/2005 18:02:11 M kill_task: killing pid 14060 task 1 with sig 15
03/02/2005 18:02:11 M kill_task: killing pid 14061 task 1 with sig 15
03/02/2005 18:02:11 M kill_task: killing pid 14063 task 1 with sig 15
03/02/2005 18:02:11 M kill_job done
03/02/2005 18:04:11 M kill_job
03/02/2005 18:04:11 M kill_job found a task to kill
03/02/2005 18:04:11 M sending signal 15 to task
03/02/2005 18:04:11 M kill_job
03/02/2005 18:04:11 M sending signal 15 to task
03/02/2005 18:04:11 M kill_job
03/02/2005 18:04:11 M performing job clean-up
03/02/2005 18:06:27 A user=dev group=dev jobname=STDIN queue=batch ctime=1109811508
qtime=1109811508 etime=1109811508 start=1109811521
exec_host=icluster.org/0 Resource_List.nodes=1 Resource_List.nodect=1
Resource_List.nodes=1 Resource_List.nodes=1 Resource_List.nodes=1
Resource_List.nodes=1
session=14060
end=1109811987 Exit_status=265 resources_used.cput=00:00:00
resources_used.mem=3544kb resources_used.vmem=10632kb
resources_used.walltime=00:07:46

...
Using GDB to locate job failures

If either the pbs_mom or pbs_server fail unexpectedly (and the log files contain no information on the failure) gdb can be used to determine whether or not the program is crashing. To start pbs_mom or pbs_server under GDB export the environment variable PBSDEBUG=yes and start the program (i.e., gdb pbs_mom and then issue the run subcommand at the gdb prompt).

GDB may run for some time until a failure occurs and at which point, a message will be printed to the screen and a gdb prompt again made available. If this occurs, use the gdb where subcommand to determine the exact location in the code. The information provided may be adequate to allow local diagnosis and correction. If not, this output may be sent to the mailing list or to help for further assistance.

See the PBSCOREDUMP parameter for enabling creation of core files (see Debugging on page 136).

Related topics
- Troubleshooting on page 123
Other diagnostic options

When PBSDEBUG is set, some client commands will print additional diagnostic information.

```
$ export PBSDEBUG=yes
$ cmd
```

To debug different kinds of problems, it can be useful to see where in the code time is being spent. This is called profiling and there is a Linux utility "gprof" that will output a listing of routines and the amount of time spent in these routines. This does require that the code be compiled with special options to instrument the code and to produce a file, gmon.out, that will be written at the end of program execution.

The following listing shows how to build TORQUE with profiling enabled. Notice that the output file for pbs_mom will end up in the mom_priv directory because its startup code changes the default directory to this location.

```
# ./configure "CFLAGS=-pg -lgcov -fPIC"
# make -j5
# make install
# pbs_mom ... do some stuff for a while ...
# momctl -s
# cd /var/spool/torque/mom_priv
# gprof -b `which pbs_mom` gmon.out |less
#
```

Another way to see areas where a program is spending most of its time is with the valgrind program. The advantage of using valgrind is that the programs do not have to be specially compiled.

```
# valgrind --tool=callgrind pbs_mom
```

Related topics

- Troubleshooting on page 123

Stuck jobs

If a job gets stuck in TORQUE, try these suggestions to resolve the issue:

- Use the qdel command to cancel the job.

```
> qdel <JOBID>
```

- Force the MOM to send an obituary of the job ID to the server.

```
> qsig -s 0 <JOBID>
```

- You can try clearing the stale jobs by using the momctl command on the compute nodes where the jobs are still listed.

```
> momctl -c 58925 -h compute-5-20
```

- Setting the qmgr server setting mom_job_sync to True might help prevent jobs from hanging.

```
> qmgr -c "set server mom_job_sync = True"
```
To check and see if this is already set, use:

```
$ qmgr -c "p s"
```

- If the suggestions above cannot remove the stuck job, you can try `qdel -p`. However, since the `-p` option purges all information generated by the job, this is not a recommended option unless the above suggestions fail to remove the stuck job.

```
$ qdel -p <JOBID>
```

- The last suggestion for removing stuck jobs from compute nodes is to restart the `pbs_mom`.

For additional troubleshooting, run a `tracejob` on one of the stuck jobs. You can then create an [online support ticket](#) with the full server log for the time period displayed in the trace job.

### Related topics

- [Troubleshooting](#) on page 123

### Frequently asked questions (FAQ)

- [Cannot connect to server: error=15034](#) on page 130
- [Deleting 'stuck' jobs](#) on page 130
- [Which user must run TORQUE?](#) on page 130
- [Scheduler cannot run jobs - rc: 15003](#) on page 130
- [PBS Server: pbsd_init, Unable to read server database](#) on page 131
- [qsub will not allow the submission of jobs requesting many processors](#) on page 132
- [qsub reports 'Bad UID for job execution'](#) on page 132
- [Why does my job keep bouncing from running to queued?](#) on page 132
- [How do I use PVM with TORQUE?](#) on page 133
- [My build fails attempting to use the TCL library](#) on page 133
- [My job will not start, failing with the message 'cannot send job to mom, state=PRERUN'](#) on page 133
- [How do I determine what version of TORQUE I am using?](#) on page 133
- [How do I resolve autogen.sh errors that contain "error: possibly undefined macro: AC_MSG_ERROR"](#) on page 133
- [How do I resolve compile errors with libssl or libcrypto for TORQUE 4.0 on Ubuntu 10.04?](#) on page 133
- [Why are there so many error messages in the client logs (trqauthd logs) when I don’t notice client commands failing?](#) on page 134
Cannot connect to server: error=15034

This error occurs in TORQUE clients (or their APIs) because TORQUE cannot find the server_name file and/or the PBS_DEFAULT environment variable is not set. The server_name file or PBS_DEFAULT variable indicate the pbs_server's hostname that the client tools should communicate with. The server_name file is usually located in TORQUE's local state directory. Make sure the file exists, has proper permissions, and that the version of TORQUE you are running was built with the proper directory settings. Alternatively you can set the PBS_DEFAULT environment variable. Restart TORQUE daemons if you make changes to these settings.

Deleting 'stuck' jobs

To manually delete a "stale" job which has no process, and for which the mother superior is still alive, sending a sig 0 with qsig will often cause MOM to realize the job is stale and issue the proper JobObit notice. Failing that, use momctl -c to forcefully cause MOM to purge the job. The following process should never be necessary:

- Shut down the MOM on the mother superior node.
- Delete all files and directories related to the job from TORQUE_HOME/mom_priv/jobs.
- Restart the MOM on the mother superior node.

If the mother superior MOM has been lost and cannot be recovered (i.e. hardware or disk failure), a job running on that node can be purged from the output of qstat using the qdel on page 180 –p command or can be removed manually using the following steps:

To remove job X

1. Shut down pbs_server.
   ```bash
   ] > qterm
   ```

2. Remove job spool files.
   ```bash
   ] > rm TORQUE_HOME/server_priv/jobs/X.SC TORQUE_HOME/server_priv/jobs/X.JB
   ```

3. Restart pbs_server.
   ```bash
   ] > pbs_server
   ```

Which user must run TORQUE?

TORQUE (pbs_server & pbs_mom) must be started by a user with root privileges.

Scheduler cannot run jobs - rc: 15003

For a scheduler, such as Moab or Maui, to control jobs with TORQUE, the scheduler needs to be run be a user in the server operators / managers list (see qmgr). The default for the server operators / managers list is root@localhost. For TORQUE to be used in a grid setting with Silver, the scheduler needs to be run as root.
PBS Server: pbsd_init, Unable to read server database

If this message is displayed upon starting pbs_server it means that the local database cannot be read. This can be for several reasons. The most likely is a version mismatch. Most versions of TORQUE can read each other's databases. However, there are a few incompatibilities between OpenPBS and TORQUE. Because of enhancements to TORQUE, it cannot read the job database of an OpenPBS server (job structure sizes have been altered to increase functionality). Also, a compiled in 32-bit mode cannot read a database generated by a 64-bit pbs_server and vice versa.

To reconstruct a database (excluding the job database)

1. First, print out the old data with this command:

```bash
%> qmgr -c "p s"
#
# Create queues and set their attributes.
#
# Create and define queue_type = Execution
set queue batch acl_host_enable = False
set queue batch resources_max.nodect = 6
set queue batch resources_default.nodes = 1
set queue batch resources_default.walltime = 01:00:00
set queue batch resources_available.nodect = 18
set queue batch enabled = True
set queue batch started = True
#
# Set server attributes.
#
set server scheduling = True
set server managers = griduser@oahu.icluster.org
set server managers += scott*@*.icluster.org
set server managers += wightman*@*.icluster.org
set server operators = griduser@oahu.icluster.org
set server operators += scott*@*.icluster.org
set server operators += wightman*@*.icluster.org
set server default_queue = batch
set server log_events = 511
set server mail_from = adm
set server resources_available.nodect = 80
set server node_ping_rate = 300
set server node_check_rate = 600
set server tcp_timeout = 6
```

2. Copy this information somewhere.

3. Restart pbs_server with the following command:

```bash
> pbs_server -t create
```

4. When you are prompted to overwrite the previous database, enter y, then enter the data exported by the qmgr command as in this example:

```bash
> cat data | qmgr
```

5. Restart pbs_server without the flags:

```bash
> qterm
> pbs_server
```
Chapter 11: Troubleshooting

This will reinitialize the database to the current version.

```plaintext
> qmgr
Qmgr: set server resources_available.nodect=2048
Qmgr: set queue batch resources_available.nodect=2048
```

The `pbs_server` daemon will need to be restarted before these changes will take effect.

qsub will not allow the submission of jobs requesting many processors

TORQUE's definition of a node is context sensitive and can appear inconsistent. The `qsub -l nodes=<X>` expression can at times indicate a request for X processors and other time be interpreted as a request for X nodes. While `qsub` allows multiple interpretations of the keyword nodes, aspects of the TORQUE server's logic are not so flexible. Consequently, if a job is using `-l nodes` to specify processor count and the requested number of processors exceeds the available number of physical nodes, the server daemon will reject the job.

To get around this issue, the server can be told it has an inflated number of nodes using the `resources_available` attribute. To take effect, this attribute should be set on both the server and the associated queue as in the example below. (See `resources_available` for more information.)

```plaintext
> qmgr
Qmgr: set queue batch resources_available.nodect=2048
Qmgr: set server resources_available.nodect=2048
```

qsub reports 'Bad UID for job execution'

```plaintext
[guest@login2]$ qsub test.job
qsub: Bad UID for job execution
```

Job submission hosts must be explicitly specified within TORQUE or enabled via RCmd security mechanisms in order to be trusted. In the example above, the host 'login2' is not configured to be trusted. This process is documented in Configuring job submission hosts on page 22.

Why does my job keep bouncing from running to queued?

There are several reasons why a job will fail to start. Do you see any errors in the MOM logs? Be sure to increase the loglevel on MOM if you don't see anything. Also be sure TORQUE is configured with `--enable-syslog` and look in `/var/log/messages` (or wherever your syslog writes).

Also verify the following on all machines:

- DNS resolution works correctly with matching forward and reverse
- Time is synchronized across the head and compute nodes
- User accounts exist on all compute nodes
- User home directories can be mounted on all compute nodes
- Prologue scripts (if specified) exit with 0

If using a scheduler such as Moab or Maui, use a scheduler tool such as `checkjob` to identify job start issues.
How do I use PVM with TORQUE?

- Start the master pvmd on a compute node and then add the slaves
- mpiexec can be used to launch slaves using rsh or ssh (use export PVM_RSH=/usr/bin/ssh to use ssh)

Access can be managed by rsh/ssh without passwords between the batch nodes, but denying it from anywhere else, including the interactive nodes. This can be done with xinetd and ssd configuration (root is allowed to ssh everywhere). This way, the pvm daemons can be started and killed from the job script.

The problem is that this setup allows the users to bypass the batch system by writing a job script that uses rsh/ssh to launch processes on the batch nodes. If there are relatively few users and they can more or less be trusted, this setup can work.

My build fails attempting to use the TCL library

TORQUE builds can fail on TCL dependencies even if a version of TCL is available on the system. TCL is only utilized to support the xpbsmon client. If your site does not use this tool (most sites do not use xpbsmon), you can work around this failure by rerunning configure with the `--disable-gui` argument.

My job will not start, failing with the message 'cannot send job to mom, state=PRERUN'

If a node crashes or other major system failures occur, it is possible that a job may be stuck in a corrupt state on a compute node. TORQUE 2.2.0 and higher automatically handle this when the `mom_job_sync` parameter is set via `qmgr` (the default). For earlier versions of TORQUE, set this parameter and restart the pbs_mom daemon.

This error can also occur if not enough free space is available on the partition that holds TORQUE.

How do I determine what version of TORQUE I am using?

There are times when you want to find out what version of TORQUE you are using. An easy way to do this is to run the following command:

```
qmgr
> qmgr -c "p s" | grep pbs_ver
```

How do I resolve autogen.sh errors that contain "error: possibly undefined macro: AC_MSG_ERROR"?

Verify the `pkg-config` package is installed.

How do I resolve compile errors with libssl or libcrypto for TORQUE 4.0 on Ubuntu 10.04?

When compiling TORQUE 4.0 on Ubuntu 10.04 the following errors might occur:
Chapter 11: Troubleshooting

To resolve the compile issue, use these commands:

```
> cd /usr/lib
> ln -s /lib/libcrypto.so.0.9. libcrypto.so
> ln -s /lib/libssl.so.0.9.8 libssl.so
```

Why are there so many error messages in the client logs (trqauthd logs) when I don't notice client commands failing?

If a client makes a connection to the server and the trqauthd connection for that client command is authorized before the client’s connection, the trqauthd connection is rejected. The connection is retried, but if all retry attempts are rejected, trqauthd logs a message indicating a failure. Some client commands then open a new connection to the server and try again. The client command fails only if all its retries fail.

Related topics
- Troubleshooting on page 123

**Compute node health check**

TORQUE provides the ability to perform health checks on each compute node. If these checks fail, a failure message can be associated with the node and routed to the scheduler. Schedulers (such as Moab) can forward this information to administrators by way of scheduler triggers, make it available through scheduler diagnostic commands, and automatically mark the node down until the issue is resolved. (See the RMMSIGNORE parameter in the "Parameters" Appendix of the Moab Workload Manager Administrator's Guide for more information.)

Additionally, Michael Jennings at LBNL has authored an open-source bash node health check script project. It offers an easy way to perform some of the most common node health checking tasks, such as verifying network and filesystem functionality. More information is available on the project's page.

For more information about node health checks, see these topics:
- Configuring MOMs to launch a health check on page 135
- Creating the health check script on page 135
Chapter 11: Troubleshooting

- Adjusting node state based on the health check output on page 136
- Example health check script on page 136

Related topics
- Troubleshooting on page 123

Configuring MOMs to launch a health check

The health check feature is configured via the `mom_priv/config` file using the parameters described below:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Format</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>$node_check_script</code></td>
<td>STRING</td>
<td>N/A</td>
<td>(Required) Specifies the fully qualified pathname of the health check script to run</td>
</tr>
<tr>
<td><code>$node_check_interval</code></td>
<td>INTEGER</td>
<td>1</td>
<td>(Optional) Specifies the number of MOM intervals between health checks (by default, each MOM interval is 45 seconds long - this is controlled via the <code>DEFAULT_SERVER_STAT_UPDATES</code> #define located in <code>TORQUE_HOME/src/resmom/mom_main.c</code>). The integer may be followed by a list of event names (currently supported are <code>jobstart</code> and <code>jobend</code>). (For more information, see <code>pbs_mom</code>).</td>
</tr>
</tbody>
</table>

Related topics
- Compute node health check on page 134

Creating the health check script

The health check script is executed directly by the `pbs_mom` daemon under the root user id. It must be accessible from the compute node and may be a script or compile executable program. It may make any needed system calls and execute any combination of system utilities but should not execute resource manager client commands. Also, as of TORQUE 1.0.1, the `pbs_mom` daemon blocks until the health check is completed and does not possess a built-in timeout. Consequently, it is advisable to keep the launch script execution time short and verify that the script will not block even under failure conditions.

If the script detects a failure, it should return the keyword `ERROR` to stdout followed by an error message. When a failure is detected, the `ERROR` keyword should be printed to stdout before any other data. The message (up to 1024 characters) immediately following the `ERROR` keyword must all be contained on the same line. The message is assigned to the node attribute `message` of the associated node.
Adjusting node state based on the health check output

If the health check reports an error, the node attribute "message" is set to the error string returned. Cluster schedulers can be configured to adjust a given node's state based on this information. For example, by default, Moab sets a node's state to down if a node error message is detected and restores the state as soon as the failure disappears.

Example health check script

As mentioned, the health check can be a shell script, PERL, Python, C-executable, or anything which can be executed from the command line capable of setting STDOUT. The example below demonstrates a very simple health check:

```bash
#!/bin/sh
/bin/mount | grep global
if [ $? != "0"
    then
    echo "ERROR cannot locate filesystem global"
fi
```

Debugging

TORQUE supports a number of diagnostic and debug options including the following:

**PBSDEBUG** environment variable - If set to 'yes', this variable will prevent pbs_server, pbs_mom, and/or pbs_sched from backgrounding themselves allowing direct launch under a debugger. Also, some client commands will provide additional diagnostic information when this value is set.

**PBSLOGLEVEL** environment variable - Can be set to any value between 0 and 7 and specifies the logging verbosity level (default = 0)

**PBSOREDUMP** environment variable - If set, it will cause the offending resource manager daemon to create a core file if a SIGSEGV, SIGILL, SIGFPE, SIGSYS, or SIGTRAP signal is received. The core dump will be placed in the daemon's home directory ($PBSHOME/mom_priv for pbs_mom and $PBSHOME/server_priv for pbs_server).
To enable core dumping in a Red Hat system, you must add the following line to the /etc/init.d/pbs_mom and /etc/init.d/pbs_server scripts:
```
export DAEMON_COREFILE_LIMIT=unlimited
```

_NDEBUG #define - if set at build time, will cause additional low-level logging information to be output to stdout for pbs_server and pbs_mom daemons.

_tracejob_ reporting tool - can be used to collect and report logging and accounting information for specific jobs (for more information, see _Using "tracejob" to locate job failures_ on page 125)

_PBSLOGLEVEL_ and _PBSCOREDUMP_ must be added to the $PBSHOME/pbs_environment file, not just the current environment. To set these variables, add a line to the pbs_environment file as either "variable=value" or just "variable". In the case of "variable=value", the environment variable is set up as the value specified. In the case of "variable", the environment variable is set based upon its value in the current environment.

TORQUE error codes

<table>
<thead>
<tr>
<th>Error code name</th>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBSE_FLOOR</td>
<td>15000</td>
<td>No error</td>
</tr>
<tr>
<td>PBSE_UNKJOBID</td>
<td>15001</td>
<td>Unknown job identifier</td>
</tr>
<tr>
<td>PBSE_NOATTR</td>
<td>15002</td>
<td>Undefined attribute</td>
</tr>
<tr>
<td>PBSE_ATTRRO</td>
<td>15003</td>
<td>Attempt to set READ ONLY attribute</td>
</tr>
<tr>
<td>PBSEIVALREQ</td>
<td>15004</td>
<td>Invalid request</td>
</tr>
<tr>
<td>PBSE_UNKREQ</td>
<td>15005</td>
<td>Unknown batch request</td>
</tr>
<tr>
<td>PBSE_TOOMANY</td>
<td>15006</td>
<td>Too many submit retries</td>
</tr>
<tr>
<td>PBSE_PERM</td>
<td>15007</td>
<td>No permission</td>
</tr>
<tr>
<td>PBSE_IFF_NOT_FOUND</td>
<td>15008</td>
<td>&quot;pbs_iff&quot; not found; unable to authenticate</td>
</tr>
<tr>
<td>PBSE_MUNGE_NOT_FOUND</td>
<td>15009</td>
<td>&quot;munge&quot; executable not found; unable to authenticate</td>
</tr>
<tr>
<td>PBSE_BADHOST</td>
<td>15010</td>
<td>Access from host not allowed</td>
</tr>
<tr>
<td>Error code name</td>
<td>Number</td>
<td>Description</td>
</tr>
<tr>
<td>-------------------</td>
<td>--------</td>
<td>-------------------------------------------------------</td>
</tr>
<tr>
<td>PBSE_JOBEXIST</td>
<td>15011</td>
<td>Job already exists</td>
</tr>
<tr>
<td>PBSE_SYSTEM</td>
<td>15012</td>
<td>System error occurred</td>
</tr>
<tr>
<td>PBSE_INTERNAL</td>
<td>15013</td>
<td>Internal server error occurred</td>
</tr>
<tr>
<td>PBSE_REGROUTE</td>
<td>15014</td>
<td>Parent job of dependent in rte queue</td>
</tr>
<tr>
<td>PBSE_UNKSIG</td>
<td>15015</td>
<td>Unknown signal name</td>
</tr>
<tr>
<td>PBSE_BADATVAL</td>
<td>15016</td>
<td>Bad attribute value</td>
</tr>
<tr>
<td>PBSE_MODATRRUN</td>
<td>15017</td>
<td>Cannot modify attribute in run state</td>
</tr>
<tr>
<td>PBSE_BADSTATE</td>
<td>15018</td>
<td>Request invalid for job state</td>
</tr>
<tr>
<td>PBSE_UNQQUE</td>
<td>15020</td>
<td>Unknown queue name</td>
</tr>
<tr>
<td>PBSE_BADCRED</td>
<td>15021</td>
<td>Invalid credential in request</td>
</tr>
<tr>
<td>PBSE_EXPIRED</td>
<td>15022</td>
<td>Expired credential in request</td>
</tr>
<tr>
<td>PBSE_QUNOENB</td>
<td>15023</td>
<td>Queue not enabled</td>
</tr>
<tr>
<td>PBSE_QACCESS</td>
<td>15024</td>
<td>No access permission for queue</td>
</tr>
<tr>
<td>PBSE_BADUSER</td>
<td>15025</td>
<td>Bad user - no password entry</td>
</tr>
<tr>
<td>PBSE_HOPCOUNT</td>
<td>15026</td>
<td>Max hop count exceeded</td>
</tr>
<tr>
<td>PBSE_QUEEXIST</td>
<td>15027</td>
<td>Queue already exists</td>
</tr>
<tr>
<td>PBSE_ATTRTYPE</td>
<td>15028</td>
<td>Incompatible queue attribute type</td>
</tr>
<tr>
<td>PBSE_QUEBUSY</td>
<td>15029</td>
<td>Queue busy (not empty)</td>
</tr>
<tr>
<td>PBSE_QUENBIG</td>
<td>15030</td>
<td>Queue name too long</td>
</tr>
<tr>
<td>PBSE_NOSUP</td>
<td>15031</td>
<td>Feature/function not supported</td>
</tr>
<tr>
<td>Error code name</td>
<td>Number</td>
<td>Description</td>
</tr>
<tr>
<td>-------------------</td>
<td>--------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>PBSE_QUENOEN</td>
<td>15032</td>
<td>Cannot enable queue, needs add def</td>
</tr>
<tr>
<td>PBSE_PROTOCOL</td>
<td>15033</td>
<td>Protocol (ASN.1) error</td>
</tr>
<tr>
<td>PBSE_BADATLST</td>
<td>15034</td>
<td>Bad attribute list structure</td>
</tr>
<tr>
<td>PBSE_NOCONNECTS</td>
<td>15035</td>
<td>No free connections</td>
</tr>
<tr>
<td>PBSE_NOSERVER</td>
<td>15036</td>
<td>No server to connect to</td>
</tr>
<tr>
<td>PBSE_UNKRES</td>
<td>15037</td>
<td>Unknown resource</td>
</tr>
<tr>
<td>PBSE_EXCQRESC</td>
<td>15038</td>
<td>Job exceeds queue resource limits</td>
</tr>
<tr>
<td>PBSE_QUENODFLT</td>
<td>15039</td>
<td>No default queue defined</td>
</tr>
<tr>
<td>PBSE_NORERUN</td>
<td>15040</td>
<td>Job not rerunnable</td>
</tr>
<tr>
<td>PBSE_ROUTEREJ</td>
<td>15041</td>
<td>Route rejected by all destinations</td>
</tr>
<tr>
<td>PBSE_ROUTEEXPD</td>
<td>15042</td>
<td>Time in route queue expired</td>
</tr>
<tr>
<td>PBSE_MOMREJECT</td>
<td>15043</td>
<td>Request to MOM failed</td>
</tr>
<tr>
<td>PBSE_BADSCRIPT</td>
<td>15044</td>
<td>(qsub) Cannot access script file</td>
</tr>
<tr>
<td>PBSE_STAGEIN</td>
<td>15045</td>
<td>Stage-In of files failed</td>
</tr>
<tr>
<td>PBSE_RESCUNAV</td>
<td>15046</td>
<td>Resources temporarily unavailable</td>
</tr>
<tr>
<td>PBSE_BADGRP</td>
<td>15047</td>
<td>Bad group specified</td>
</tr>
<tr>
<td>PBSE_MAXQUED</td>
<td>15048</td>
<td>Max number of jobs in queue</td>
</tr>
<tr>
<td>PBSE_CKPBSY</td>
<td>15049</td>
<td>Checkpoint busy, may be retries</td>
</tr>
<tr>
<td>PBSE_EXLIMIT</td>
<td>15050</td>
<td>Limit exceeds allowable</td>
</tr>
<tr>
<td>PBSE_BADACCT</td>
<td>15051</td>
<td>Bad account attribute value</td>
</tr>
<tr>
<td>Error code name</td>
<td>Number</td>
<td>Description</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------</td>
<td>-------------------------------------------------------</td>
</tr>
<tr>
<td>PBSE_ALRDYEXIT</td>
<td>15052</td>
<td>Job already in exit state</td>
</tr>
<tr>
<td>PBSE_NOCOPYFILE</td>
<td>15053</td>
<td>Job files not copied</td>
</tr>
<tr>
<td>PBSE_CLEANEDOUT</td>
<td>15054</td>
<td>Unknown job id after clean init</td>
</tr>
<tr>
<td>PBSE_NOSYNCMSTR</td>
<td>15055</td>
<td>No master in sync set</td>
</tr>
<tr>
<td>PBSE_BADDEPEND</td>
<td>15056</td>
<td>Invalid dependency</td>
</tr>
<tr>
<td>PBSE_DUPLIST</td>
<td>15057</td>
<td>Duplicate entry in list</td>
</tr>
<tr>
<td>PBSE_DISPROTO</td>
<td>15058</td>
<td>Bad DIS based request protocol</td>
</tr>
<tr>
<td>PBSE_EXECTXHERE</td>
<td>15059</td>
<td>Cannot execute there</td>
</tr>
<tr>
<td>PBSE_SISREJECT</td>
<td>15060</td>
<td>Sister rejected</td>
</tr>
<tr>
<td>PBSE_SISCOMM</td>
<td>15061</td>
<td>Sister could not communicate</td>
</tr>
<tr>
<td>PBSE_SVRDOW</td>
<td>15062</td>
<td>Requirement rejected -server shutting down</td>
</tr>
<tr>
<td>PBSE_CKPSHORT</td>
<td>15063</td>
<td>Not all tasks could checkpoint</td>
</tr>
<tr>
<td>PBSE_UNKNODE</td>
<td>15064</td>
<td>Named node is not in the list</td>
</tr>
<tr>
<td>PBSE_UNKNODEATR</td>
<td>15065</td>
<td>Node-attribute not recognized</td>
</tr>
<tr>
<td>PBSE_NONODES</td>
<td>15066</td>
<td>Server has no node list</td>
</tr>
<tr>
<td>PBSE_NODENBIG</td>
<td>15067</td>
<td>Node name is too big</td>
</tr>
<tr>
<td>PBSE_NODEEXIST</td>
<td>15068</td>
<td>Node name already exists</td>
</tr>
<tr>
<td>PBSE_BADNDATVAL</td>
<td>15069</td>
<td>Bad node-attribute value</td>
</tr>
<tr>
<td>PBSE_MUTUALEX</td>
<td>15070</td>
<td>State values are mutually exclusive</td>
</tr>
<tr>
<td>PBSE_GMODERR</td>
<td>15071</td>
<td>Error(s) during global modification of nodes</td>
</tr>
</tbody>
</table>
## Error code name

<table>
<thead>
<tr>
<th>Error code name</th>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBSE_NORELYMOM</td>
<td>15072</td>
<td>Could not contact MOM</td>
</tr>
<tr>
<td>PBSE_NOTSNODE</td>
<td>15073</td>
<td>No time-shared nodes</td>
</tr>
<tr>
<td>PBSE_JOBTYPE</td>
<td>15074</td>
<td>Wrong job type</td>
</tr>
<tr>
<td>PBSE_BADAACLHOST</td>
<td>15075</td>
<td>Bad ACL entry in host list</td>
</tr>
<tr>
<td>PBSE_MAXUSERQUED</td>
<td>15076</td>
<td>Maximum number of jobs already in queue for user</td>
</tr>
<tr>
<td>PBSE_BADDISALLOWTYPE</td>
<td>15077</td>
<td>Bad type in &quot;disallowed_types&quot; list</td>
</tr>
<tr>
<td>PBSE_NOINTERACTIVE</td>
<td>15078</td>
<td>Interactive jobs not allowed in queue</td>
</tr>
<tr>
<td>PBSE_NOBATCH</td>
<td>15079</td>
<td>Batch jobs not allowed in queue</td>
</tr>
<tr>
<td>PBSE_NORERUNABLE</td>
<td>15080</td>
<td>Rerunnable jobs not allowed in queue</td>
</tr>
<tr>
<td>PBSE_NONONRERUNABLE</td>
<td>15081</td>
<td>Non-rerunnable jobs not allowed in queue</td>
</tr>
<tr>
<td>PBSEUNKARRAYID</td>
<td>15082</td>
<td>Unknown array ID</td>
</tr>
<tr>
<td>PBSE_BAD_ARRAY_REQ</td>
<td>15083</td>
<td>Bad job array request</td>
</tr>
<tr>
<td>PBSE_TIMEOUT</td>
<td>15084</td>
<td>Time out</td>
</tr>
<tr>
<td>PBSE_JOBNOTFOUND</td>
<td>15085</td>
<td>Job not found</td>
</tr>
<tr>
<td>PBSE_NOFAULTTOLERANT</td>
<td>15086</td>
<td>Fault tolerant jobs not allowed in queue</td>
</tr>
<tr>
<td>PBSE_NOFAULTINTOLERANT</td>
<td>15087</td>
<td>Only fault tolerant jobs allowed in queue</td>
</tr>
<tr>
<td>PBSE_NOJOBARRAYS</td>
<td>15088</td>
<td>Job arrays not allowed in queue</td>
</tr>
<tr>
<td>PBSE_RELAYED_TO_MOM</td>
<td>15089</td>
<td>Request was relayed to a MOM</td>
</tr>
<tr>
<td>PBSE_MEM_MALLOC</td>
<td>15090</td>
<td>Failed to allocate memory for memmgr</td>
</tr>
<tr>
<td>PBSE_MUTEX</td>
<td>15091</td>
<td>Failed to allocate controlling mutex (lock/unlock)</td>
</tr>
</tbody>
</table>
## Related topics

- [Troubleshooting on page 123](#)
Appendices

The appendices provide tables of commands, parameters, configuration options, error codes, the Quick Start Guide, and so forth.

- Commands overview on page 145
- Server parameters on page 227
- Node manager (MOM) configuration on page 245
- Diagnostics and error codes on page 263
- Considerations before upgrading on page 271
- Large cluster considerations on page 273
- Prologue and epilogue scripts on page 281
- Running multiple TORQUE servers and MOMs on the same node on page 289
- Security overview on page 291
- Job submission filter ("qsub wrapper") on page 293
- "torque.cfg" configuration file on page 295
- TORQUE Quick Start Guide on page 301
- BLCR acceptance tests on page 305
## Commands overview

### Client commands

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>momctl</strong></td>
<td>Manage/diagnose MOM (node execution) daemon</td>
</tr>
<tr>
<td><strong>pbsdsh</strong></td>
<td>Launch tasks within a parallel job</td>
</tr>
<tr>
<td><strong>pbsnodes</strong></td>
<td>View/modify batch status of compute nodes</td>
</tr>
<tr>
<td><strong>qalter</strong></td>
<td>Modify queued batch jobs</td>
</tr>
<tr>
<td><strong>qchkpt</strong></td>
<td>Checkpoint batch jobs</td>
</tr>
<tr>
<td><strong>qdel</strong></td>
<td>Delete/cancel batch jobs</td>
</tr>
<tr>
<td><strong>qgpumode</strong></td>
<td>Specifies new mode for GPU</td>
</tr>
<tr>
<td><strong>qgpureset</strong></td>
<td>Reset the GPU</td>
</tr>
<tr>
<td><strong>qhold</strong></td>
<td>Hold batch jobs</td>
</tr>
<tr>
<td><strong>qmgr</strong></td>
<td>Manage policies and other batch configuration</td>
</tr>
<tr>
<td><strong>qmove on page 189</strong></td>
<td>Move batch jobs</td>
</tr>
<tr>
<td><strong>qorder on page 190</strong></td>
<td>Exchange order of two batch jobs in any queue</td>
</tr>
<tr>
<td><strong>qrerun</strong></td>
<td>Rerun a batch job</td>
</tr>
<tr>
<td><strong>qrls</strong></td>
<td>Release batch job holds</td>
</tr>
</tbody>
</table>
### Command Description

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>qrun</td>
<td>Start a batch job</td>
</tr>
<tr>
<td>qsig</td>
<td>Send a signal to a batch job</td>
</tr>
<tr>
<td>qstat</td>
<td>View queues and jobs</td>
</tr>
<tr>
<td>qsub</td>
<td>Submit jobs</td>
</tr>
<tr>
<td>qterm</td>
<td>Shutdown pbs server daemon</td>
</tr>
<tr>
<td>tracejob</td>
<td>Trace job actions and states recorded in TORQUE logs (see Using &quot;tracejob&quot; to locate job failures on page 125)</td>
</tr>
</tbody>
</table>

### Binary executables

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>pbs_sif</td>
<td>Interprocess authentication service</td>
</tr>
<tr>
<td>pbs_mom</td>
<td>Start MOM (node execution) daemon</td>
</tr>
<tr>
<td>pbs_server</td>
<td>Start server daemon</td>
</tr>
<tr>
<td>pbs_track</td>
<td>Tell pbs_mom to track a new process</td>
</tr>
</tbody>
</table>

### Related topics

- Node manager (MOM) configuration on page 245
- Server parameters on page 227

### momctl

*(PBS MOM Control)*

**Synopsis**

momctl -c { <JOBID> | all }
momctl -C
momctl -d { <INTEGER> | <JOBID> }
momctl -f <FILE>
momctl -h <HOST>[,<HOST>]...
momctl -p <PORT_NUMBER>
momctl -q <ATTRIBUTE>
momctl -r { <FILE> | LOCAL:<FILE> }
momctl -s

Overview

The **momctl** command allows remote shutdown, reconfiguration, diagnostics, and querying of the **pbs_mom** daemon.

### Format

<table>
<thead>
<tr>
<th>-c — Clear</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td><strong>Default</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
<tr>
<td><strong>Example</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>-C — Cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td><strong>Default</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
<tr>
<td><strong>Example</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>-d — Diagnose</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td><strong>Default</strong></td>
</tr>
</tbody>
</table>
### -d — Diagnose

| Description | Diagnose MOM(s)  
(For more details, see Diagnose detail on page 150 below.) |
|-------------|--------------------------------------------------|
| Example     | `momctl -n node1 -d 2`  
Print level 2 and lower diagnose information for the MOM on node1. |

### -f — Host File

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;FILE&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>---</td>
</tr>
<tr>
<td>Description</td>
<td>A file containing only comma or whitespace (space, tab, or new line) delimited hostnames</td>
</tr>
</tbody>
</table>
| Example | `momctl -f hosts.txt -d`  
Print diagnose information for the MOMs running on the hosts specified in hosts.txt. |

### -h — Host List

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;HOST&gt;[,&lt;HOST&gt;].</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>localhost</td>
</tr>
<tr>
<td>Description</td>
<td>A comma separated list of hosts</td>
</tr>
</tbody>
</table>
| Example | `momctl -h node1,node2,node3 -d`  
Print diagnose information for the MOMs running on node1, node2, and node3. |

### -p — Port

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;PORT_NUMBER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>TORQUE’s default port number</td>
</tr>
<tr>
<td>Description</td>
<td>The port number for the specified MOM(s)</td>
</tr>
</tbody>
</table>
### -p — Port

**Example**

```bash
momctl -p 5455 -h node1 -d
```

Request diagnose information over port 5455 on node1.

### -q — Query

**Format**

```bash
<ATTRIBUTE>
```

**Default**

```bash
---
```

**Description**

Query `<ATTRIBUTE>` on specified MOM, where `<ATTRIBUTE>` is a property listed by `pbsnodes -a` (see Query attributes on page 150 for a list of attributes)

**Example**

```bash
momctl -q physmem
```

Print the amount of physmem on localhost.

### -r — Reconfigure

**Format**

```bash
{ <FILE> | LOCAL:<FILE> }
```

**Default**

```bash
---
```

**Description**

Reconfigure MOM(s) with remote or local config file, `<FILE>`. This does not work if `$remote_reconfig` is not set to true when the MOM is started.

**Example**

```bash
momctl -r /home/user1/new.config
```

Reconfigure MOM on node1 with `/home/user1/new.config` on node1.

### -s — Shutdown

**Format**

```bash
---
```

**Default**

```bash
---
```

**Description**

Shutdown pbs_mom

**Example**

```bash
momctl -s
```

Terminates pbs_mom process on localhost.
## Query attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>arch</td>
<td>node hardware architecture</td>
</tr>
<tr>
<td>availmem</td>
<td>available RAM</td>
</tr>
<tr>
<td>loadave</td>
<td>1 minute load average</td>
</tr>
<tr>
<td>ncpus</td>
<td>number of CPUs available on the system</td>
</tr>
<tr>
<td>netload</td>
<td>total number of bytes transferred over all network interfaces</td>
</tr>
<tr>
<td>nsessions</td>
<td>number of sessions active</td>
</tr>
<tr>
<td>nusers</td>
<td>number of users active</td>
</tr>
<tr>
<td>physmem</td>
<td>configured RAM</td>
</tr>
<tr>
<td>sessions</td>
<td>list of active sessions</td>
</tr>
<tr>
<td>totmem</td>
<td>configured RAM plus configured swap</td>
</tr>
</tbody>
</table>

## Diagnose detail

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Display the following information:</td>
</tr>
<tr>
<td></td>
<td>• Local hostname</td>
</tr>
<tr>
<td></td>
<td>• Expected server hostname</td>
</tr>
<tr>
<td></td>
<td>• Execution version</td>
</tr>
<tr>
<td></td>
<td>• MOM home directory</td>
</tr>
<tr>
<td></td>
<td>• MOM config file version (if specified)</td>
</tr>
<tr>
<td></td>
<td>• Duration MOM has been executing</td>
</tr>
<tr>
<td></td>
<td>• Duration since last request from pbs_server daemon</td>
</tr>
<tr>
<td></td>
<td>• Duration since last request to pbs_server daemon</td>
</tr>
<tr>
<td></td>
<td>• RM failure messages (if any)</td>
</tr>
<tr>
<td></td>
<td>• Log verbosity level</td>
</tr>
<tr>
<td></td>
<td>• Local job list</td>
</tr>
<tr>
<td>Level</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
</tr>
</tbody>
</table>
| 1     | All information for level 0 plus the following:  
  - Interval between updates sent to server  
  - Number of initialization messages sent to pbs_server daemon  
  - Number of initialization messages received from pbs_server daemon  
  - Prolog/epilog alarm time  
  - List of trusted clients |
| 2     | All information from level 1 plus the following:  
  - PID  
  - Event alarm status |
| 3     | All information from level 2 plus the following:  
  - syslog enabled |

**Example A-1: MOM diagnostics**

```
  momctl -d 1
  Host: nsrf/nsrf.fllcl.com   Server: 10.10.10.113   Version: torque_1.1.0p4
  HomeDirectory: /usr/spool/PBS/mom_priv
  ConfigVersion: 147
  MOM active: 7390 seconds
  Last Msg From Server: 7389 seconds (CLUSTER ADDRS)
  Server Update Interval: 20 seconds
  Server Update Interval: 20 seconds
  Init Msgs Received: 0 hellos/1 cluster-addrs
  Init Msgs Sent: 1 hellos
  LOGLEVEL: 0 (use SIGUSR1/SIGUSR2 to adjust)
  Prolog Alarm Time: 300 seconds
  Trusted Client List: 12.14.213.113,127.0.0.1
  JobList: NONE

diagnostics complete
```

**Example A-2: System shutdown**

```
> moment -s -f /opt/clusterhostfile
  shutdown request successful on node001
  shutdown request successful on node002
  shutdown request successful on node003
  shutdown request successful on node004
  shutdown request successful on node005
  shutdown request successful on node006
```

**pbs_mom**

Start a pbs batch execution mini-server.
Synopsis

```
```

Description

The `pbs_mom` command is located within the `TORQUE_HOME` directory and starts the operation of a batch Machine Oriented Mini-server (MOM) on the execution host. To insure that the `pbs_mom` command is not runnable by the general user community, the server will only execute if its real and effective uid is zero.

The first function of `pbs_mom` is to place jobs into execution as directed by the server, establish resource usage limits, monitor the job's usage, and notify the server when the job completes. If they exist, `pbs_mom` will execute a prologue script before executing a job and an epilogue script after executing the job.

The second function of `pbs_mom` is to respond to resource monitor requests. This was done by a separate process in previous versions of PBS but has now been combined into one process. It provides information about the status of running jobs, memory available etc.

The last function of `pbs_mom` is to respond to task manager requests. This involves communicating with running tasks over a TCP socket as well as communicating with other MOMs within a job (a.k.a. a "sisterhood").

`pbs_mom` will record a diagnostic message in a log file for any error occurrence. The log files are maintained in the `mom_logs` directory below the home directory of the server. If the log file cannot be opened, the diagnostic message is written to the system console.

Options

<table>
<thead>
<tr>
<th>Flag</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-a</td>
<td>alarm</td>
<td>Used to specify the alarm timeout in seconds for computing a resource. Every time a resource request is processed, an alarm is set for the given amount of time. If the request has not completed before the given time, an alarm signal is generated. The default is 5 seconds.</td>
</tr>
<tr>
<td>-A</td>
<td>alias</td>
<td>Used to specify this multimom's alias name. The alias name needs to be the same name used in the mom hierarchy file. It is only needed when running multiple MOMs on the same machine. For more information, see <code>TORQUE Multi-MOM</code> on page 35.</td>
</tr>
<tr>
<td>-C</td>
<td>chkdirectory</td>
<td>Specifies the path of the directory used to hold checkpoint files. (Currently this is only valid on Cray systems.) The default directory is <code>TORQUE_HOME/spool/checkpoint</code> (see the -d option). The directory specified with the -C option must be owned by root and accessible (rwx) only by root to protect the security of the checkpoint files.</td>
</tr>
<tr>
<td>Flag</td>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>-------------</td>
</tr>
<tr>
<td>-c</td>
<td>config</td>
<td>Specifies an alternative configuration file, see description below. If this is a relative file name it will be relative to TORQUE_HOME/mom_priv (see the <code>-d</code> option). If the specified file cannot be opened, pbs_mom will abort. If the <code>-c</code> option is not supplied, pbs_mom will attempt to open the default configuration file &quot;config&quot; in TORQUE_HOME/mom_priv. If this file is not present, pbs_mom will log the fact and continue.</td>
</tr>
<tr>
<td>-d</td>
<td>directory</td>
<td>Specifies the path of the directory which is the home of the server's working files, TORQUE_HOME. This option is typically used along with <code>-M</code> when debugging MOM. The default directory is given by <code>$PBS_SERVER_HOME</code> which is typically <code>/usr/spool/PBS</code>.</td>
</tr>
<tr>
<td>-h</td>
<td>hostname</td>
<td>Set MOM's hostname. This can be useful on multi-homed networks.</td>
</tr>
<tr>
<td>-L</td>
<td>logfile</td>
<td>Specify an absolute path name for use as the log file. If not specified, MOM will open a file named for the current date in the TORQUE_HOME/mom_logs directory (see the <code>-d</code> option).</td>
</tr>
<tr>
<td>-M</td>
<td>port</td>
<td>Specifies the port number on which the mini-server (MOM) will listen for batch requests.</td>
</tr>
<tr>
<td>-p</td>
<td>n/a</td>
<td>Specifies the impact on jobs which were in execution when the mini-server shut down. On any restart of MOM, the new mini-server will not be the parent of any running jobs, MOM has lost control of her offspring (not a new situation for a mother). With the <code>-p</code> option, MOM will allow the jobs to continue to run and monitor them indirectly via polling. This flag is redundant in that this is the default behavior when starting the server. The <code>-p</code> option is mutually exclusive with the <code>-R</code> and <code>-q</code> options.</td>
</tr>
<tr>
<td>-P</td>
<td>purge</td>
<td>Specifies the impact on jobs which were in execution when the mini-server shut down. With the <code>-P</code> option, it is assumed that either the entire system has been restarted or the MOM has been down so long that it can no longer guarantee that the pid of any running process is the same as the recorded job process pid of a recovering job. Unlike the <code>-p</code> option, no attempt is made to try and preserve or recover running jobs. All jobs are terminated and removed from the queue.</td>
</tr>
<tr>
<td>-q</td>
<td>n/a</td>
<td>Specifies the impact on jobs which were in execution when the mini-server shut down. With the <code>-q</code> option, MOM will allow the processes belonging to jobs to continue to run, but will not attempt to monitor them. The <code>-q</code> option is mutually exclusive with the <code>-p</code> and <code>-R</code> options.</td>
</tr>
<tr>
<td>-R</td>
<td>port</td>
<td>Specifies the port number on which the mini-server (MOM) will listen for resource monitor requests, task manager requests and inter-MOM messages. Both a UDP and a TCP port of this number will be used.</td>
</tr>
<tr>
<td>Flag</td>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>-r</td>
<td>n/a</td>
<td>Specifies the impact on jobs which were in execution when the mini-server shut down. With the -r option, MOM will kill any processes belonging to jobs, mark the jobs as terminated, and notify the batch server which owns the job. The -r option is mutually exclusive with the -p and -q options. Normally the mini-server is started from the system boot file without the -p or the -r option. The mini-server will make no attempt to signal the former session of any job which may have been running when the mini-server terminated. It is assumed that on reboot, all processes have been killed. If the -r option is used following a reboot, process IDs (pids) may be reused and MOM may kill a process that is not a batch session.</td>
</tr>
<tr>
<td>-w</td>
<td>wait_for_server</td>
<td>When started with -w, pbs_moms wait until they get their MOM hierarchy file from pbs_server to send their first update, or until 10 minutes pass. This reduces network traffic on startup and can bring up clusters faster.</td>
</tr>
<tr>
<td>-x</td>
<td>n/a</td>
<td>Disables the check for privileged port resource monitor connections. This is used mainly for testing since the privileged port is the only mechanism used to prevent any ordinary user from connecting.</td>
</tr>
</tbody>
</table>

**Configuration file**

The configuration file may be specified on the command line at program start with the `-C` flag. The use of this file is to provide several types of run time information to pbs_mom: static resource names and values, external resources provided by a program to be run on request via a shell escape, and values to pass to internal set up functions at initialization (and re-initialization).

Each item type is on a single line with the component parts separated by white space. If the line starts with a hash mark (pound sign, #), the line is considered to be a comment and is skipped.

*Static Resources*

For static resource names and values, the configuration file contains a list of resource names/values pairs, one pair per line and separated by white space. An example of static resource names and values could be the number of tape drives of different types and could be specified by:

- tape3480 4
- tape3420 2
- tapedat 1
- tape8mm 1

*Shell Commands*

If the first character of the value is an exclamation mark (!), the entire rest of the line is saved to be executed through the services of the system(3) standard library routine.

The shell escape provides a means for the resource monitor to yield arbitrary information to the scheduler. Parameter substitution is done such that the value of any qualifier sent with the query, as
explained below, replaces a token with a percent sign (%) followed by the name of the qualifier. For example, here is a configuration file line which gives a resource name of "escape":

```
[ escape !echo %xxx %yyy]
```

If a query for "escape" is sent with no qualifiers, the command executed would be `echo %xxx %yyy`.

If one qualifier is sent, `escape[xxx=hi there]`, the command executed would be `echo hi there %yyy`.

If two qualifiers are sent, `escape[xxx=hi] [yyy=there]`, the command executed would be `echo hi there %yyy`.

If a qualifier is sent with no matching token in the command line, `escape[zzz=snafu]`, an error is reported.

```
size[fs=<FS>]
```

Specifies that the available and configured disk space in the `<FS>` filesystem is to be reported to the pbs_server and scheduler. To request disk space on a per job basis, specify the file resource, as in `qsub -l nodes=1, file=1000kb`. For example, the available and configured disk space in the `/localscratch` filesystem will be reported:

```
[size[fs=/localscratch]]
```

**Initialization Value**

An initialization value directive has a name which starts with a dollar sign ($) and must be known to the MOM via an internal table. The entries in this table now are:

<table>
<thead>
<tr>
<th>Entry</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>pbsclient</td>
<td>Causes a host name to be added to the list of hosts which will be allowed to connect to the MOM as long as they are using a privileged port for the purposes of resource monitor requests. For example, here are two configuration file lines which will allow the hosts &quot;fred&quot; and &quot;wilma&quot; to connect:</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><code>$pbsclient fred</code></td>
</tr>
<tr>
<td></td>
<td><code>$pbsclient wilma</code></td>
</tr>
<tr>
<td></td>
<td>Two host names are always allowed to connect to pbs_mom &quot;localhost&quot; and the name returned to pbs_mom by the system call gethostname(). These names need not be specified in the configuration file. The hosts listed as &quot;clients&quot; can issue Resource Manager (RM) requests. Other MOM nodes and servers do not need to be listed as clients.</td>
</tr>
<tr>
<td>Entry</td>
<td>Description</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>restricted</td>
<td>Causes a host name to be added to the list of hosts which will be allowed to connect to the MOM without needing to use a privileged port. These names allow for wildcard matching. For example, here is a configuration file line which will allow queries from any host from the domain &quot;ibm.com&quot;.</td>
</tr>
<tr>
<td></td>
<td>[ \text{&quot;restricted *.ibm.com&quot;} ]</td>
</tr>
<tr>
<td></td>
<td>The restriction which applies to these connections is that only internal queries may be made. No resources from a config file will be found. This is to prevent any shell commands from being run by a non-root process. This parameter is generally not required except for some versions of OSX.</td>
</tr>
<tr>
<td>logevent</td>
<td>Sets the mask that determines which event types are logged by pbs_mom. For example:</td>
</tr>
<tr>
<td></td>
<td>[ \text{&quot;$logevent 0x1fff $logevent 255&quot;} ]</td>
</tr>
<tr>
<td></td>
<td>The first example would set the log event mask to 0x1ff (511) which enables logging of all events including debug events. The second example would set the mask to 0x0ff (255) which enables all events except debug events.</td>
</tr>
<tr>
<td>cputmult</td>
<td>Sets a factor used to adjust cpu time used by a job. This is provided to allow adjustment of time charged and limits enforced where the job might run on systems with different cpu performance. If the MOM's system is faster than the reference system, set cputmult to a decimal value greater than 1.0. If the MOM's system is slower, set cputmult to a value between 1.0 and 0.0. For example:</td>
</tr>
<tr>
<td></td>
<td>[ \text{&quot;$cputmult 1.5 $cputmult 0.75&quot;} ]</td>
</tr>
<tr>
<td>usecp</td>
<td>Specifies which directories should be staged with cp instead of rcp/scp. If a shared filesystem is available on all hosts in a cluster, this directive is used to make these filesystems known to the MOM. For example, if /home is NFS mounted on all nodes in a cluster:</td>
</tr>
<tr>
<td></td>
<td>[ \text{&quot;$usecp *:/home /home&quot;} ]</td>
</tr>
<tr>
<td>wallmult</td>
<td>Sets a factor to adjust wall time usage by to job to a common reference system. The factor is used for walltime calculations and limits in the same way that cputmult is used for cpu time.</td>
</tr>
<tr>
<td>configversion</td>
<td>Specifies the version of the config file data, a string.</td>
</tr>
<tr>
<td>check_poll_time</td>
<td>Specifies the MOM interval in seconds that TORQUE polls the sisters for job information. The MOM checks each job for updated resource usages, exited processes, over-limit conditions, etc, once per interval. This value should be equal or lower to pbs_server's \text{job} \text{stat_rate}. High values result in stale information reported to pbs_server. Low values result in increased system usage by the MOM. Default is 45 seconds.</td>
</tr>
<tr>
<td>Entry</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>down_on_error</td>
<td>Causes the MOM to report itself as state &quot;down&quot; to pbs_server in the event of a failed health check. This feature is experimental (For more information, see Health check on page 160.)</td>
</tr>
<tr>
<td>ideal_load</td>
<td>Ideal processor load. Represents a low water mark for the load average. A node that is currently busy will consider itself free after falling below ideal_load.</td>
</tr>
<tr>
<td>loglevel</td>
<td>Specifies the verbosity of logging with higher numbers specifying more verbose logging. Values may range between 0 and 7.</td>
</tr>
<tr>
<td>log_file_max_size</td>
<td>If this is set to a value &gt; 0, then pbs_mom will roll the current log file to log-file-name.1 when its size is greater than or equal to the value of log_file_max_size. This value is interpreted as kilobytes.</td>
</tr>
<tr>
<td>log_file_roll_depth</td>
<td>If this is set to a value &gt;=1 and log_file_max_size is set, then pbs_mom will allow logs to be rolled up to the specified number of logs. At every roll, the oldest log will be the one to be deleted to make room for rolling. pbs_mom will continue rolling the log files to log-file-name.log_file_roll_depth.</td>
</tr>
<tr>
<td>max_load</td>
<td>Maximum processor load. Nodes over this load average are considered busy (see ideal_load above).</td>
</tr>
<tr>
<td>enablemomrestart</td>
<td>Enables automatic restarts of the MOM. If enabled, the MOM will check if its binary has been updated and restart itself at a safe point when no jobs are running; thus making upgrades easier. The check is made by comparing the mtime of the pbs_mom executable. Command-line args, the process name, and the PATH env variable are preserved across restarts. It is recommended that this not be enabled in the config file, but enabled when desired with momctl (see Resources on page 159 for more information.)</td>
</tr>
<tr>
<td>node_check_script</td>
<td>Specifies the fully qualified pathname of the health check script to run (see Health check on page 160 for more information).</td>
</tr>
<tr>
<td>node_check_interval</td>
<td>Specifies when to run the MOM health check. The check can be either periodic, event-driven, or both. The value starts with an integer specifying the number of MOM intervals between subsequent executions of the specified health check. After the integer is an optional comma-separated list of event names. Currently supported are &quot;jobstart&quot; and &quot;jobend&quot;. This value defaults to 1 with no events indicating the check is run every MOM interval. (see Health check on page 160 for more information.)</td>
</tr>
</tbody>
</table>

<p>| node_check_interval   | $node_check_interval 0,Disabled |
|                       | $node_check_interval 0,jobstartOnly |
|                       | $node_check_interval 10,jobstart,jobend |</p>
<table>
<thead>
<tr>
<th>Entry</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>prologalarm</td>
<td>Specifies maximum duration (in seconds) which the MOM will wait for the job prolog or job epilog to complete. This parameter defaults to 300 seconds (5 minutes).</td>
</tr>
</tbody>
</table>
| rcpcmd              | Specify the full path and argument to be used for remote file copies. This overrides the compile-time default found in configure. This must contain 2 words: the full path to the command and the options. The copy command must be able to recursively copy files to the remote host and accept arguments of the form "user@host:files." For example:  
  ```bash
  $rcpcmd /usr/bin/rcp -rp
  $rcpcmd /usr/bin/scp -rpB
  ``` |
| remote_checkpoint_dirs | Specifies which server checkpoint directories are remotely mounted. It tells the MOM which directories are shared with the server. Using remote checkpoint directories eliminates the need to copy the checkpoint files back and forth between the MOM and the server. All entries must be on the same line, separated by a space.  
  ```bash
  $remote_checkpoint_dirs /checkpointFiles /bigStorage /fast
  ```  
  *This informs the MOM that the /checkpointFiles, /bigStorage, and /fast directories are remotely mounted checkpoint directories.* |
| remote_reconfig     | Enables the ability to remotely reconfigure pbs_mom with a new config file. Default is disabled. This parameter accepts various forms of true, yes, and 1. |
| timeout             | Specifies the number of seconds before TCP messages will time out. TCP messages include obituaries, and TM requests if RPP is disabled. Default is 60 seconds. |
| tmpdir              | Sets the directory base name for a per-job temporary directory. Before job launch, the MOM appends the jobid of running jobs to the tmpdir base name and creates the directory. After the job exits, the MOM recursively deletes the directory with the jobid in its name. TORQUE creates and removes the directory as the job owner and group, so the owner must have write permissions to create the directory. The environment variable TMPDIR will be set for all prologue and epilogue scripts, the job script, and TM tasks for each job.  
  It is recommended that you create the tmpdir directory before running any jobs on the MOM and that you make it readable and writable to all users who will run jobs. If you do not specify a base tmpdir directory, the first job run on the MOM will create the directory and set the directory owner to its own. After that, no other users will be able to run jobs on that MOM. |
| status_update_time  | Specifies (in seconds) how often the MOM updates its status information to pbs_server. This value should correlate with the server’s scheduling interval and its “node_check_rate” attribute. High values for “status_update_time” cause pbs_server to report stale information, while low values increase the load of pbs_server and the network. Default is 45 seconds. |
### varattr
This is similar to a shell escape above, but includes a TTL. The command will only be run every TTL seconds. A TTL of -1 will cause the command to be executed only once. A TTL of 0 will cause the command to be run every time varattr is requested. This parameter may be used multiple times, but all output will be grouped into a single "varattr" attribute in the request and status output. If the command has no output, the name will be skipped in the output.

```bash
$varattrseta
$varattrsetb
```

### xauthpath
Specifies the path to the xauth binary to enable X11 forwarding.

### ignvmem
If set to true, then pbs_mom will ignore vmem/pvmem limit enforcement.

### ignwalltime
If set to true, then pbs_mom will ignore walltime limit enforcement.

### mom_host
Sets the local hostname as used by pbs_mom.

## Resources
Resource Manager queries can be made with `momctl -q` options to retrieve and set `pbs_mom` options. Any configured static resource may be retrieved with a request of the same name. These are resource requests not otherwise documented in the PBS ERS.

<table>
<thead>
<tr>
<th>Request</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>cycle</td>
<td>Forces an immediate MOM cycle.</td>
</tr>
<tr>
<td>status_update_time</td>
<td>Retrieve or set the <code>$status_update_time</code> parameter.</td>
</tr>
<tr>
<td>check_poll_time</td>
<td>Retrieve or set the <code>$check_poll_time</code> parameter.</td>
</tr>
<tr>
<td>configversion</td>
<td>Retrieve the config version.</td>
</tr>
<tr>
<td>jobstartblocktime</td>
<td>Retrieve or set the <code>$jobstartblocktime</code> parameter.</td>
</tr>
<tr>
<td>enablemomrestart</td>
<td>Retrieve or set the <code>$enablemomrestart</code> parameter.</td>
</tr>
<tr>
<td>loglevel</td>
<td>Retrieve or set the <code>$loglevel</code> parameter.</td>
</tr>
<tr>
<td>down_on_error</td>
<td>Retrieve or set the EXPERIMENTAL <code>$down_on_error</code> parameter.</td>
</tr>
<tr>
<td>Request</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>diag0 - diag4</td>
<td>Retrieves varied diagnostic information.</td>
</tr>
<tr>
<td>rcpcmd</td>
<td>Retrieve or set the $rcpcmd parameter.</td>
</tr>
<tr>
<td>version</td>
<td>Retrieves the pbs_mom version.</td>
</tr>
</tbody>
</table>

**Health check**

The health check script is executed directly by the pbs_mom daemon under the root user id. It must be accessible from the compute node and may be a script or compiled executable program. It may make any needed system calls and execute any combination of system utilities but should not execute resource manager client commands. Also, the pbs_mom daemon blocks until the health check is completed and does not possess a built-in timeout. Consequently, it is advisable to keep the launch script execution time short and verify that the script will not block even under failure conditions.

If the script detects a failure, it should return the keyword "Error" to stdout followed by an error message. The message (up to 256 characters) immediately following the Error string will be assigned to the node attribute message of the associated node.

If the script detects a failure when run from "jobstart", then the job will be rejected. You can use this behavior with an advanced scheduler, such as Moab Workload Manager, to cause the job to be routed to another node. TORQUE currently ignores Error messages by default, but you can configure an advanced scheduler to react appropriately.

If the experimental $down_on_error MOM setting is enabled, the MOM will set itself to state down and report to pbs_server. Additionally, the experimental $down_on_error server attribute can be enabled which has the same effect but moves the decision to pbs_server. It is redundant to have MOM's $down_on_error and pbs_servers down_on_error features enabled. See "down_on_error" in pbs_server_attributes(7B).

**Files**

<table>
<thead>
<tr>
<th>File</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$PBS_SERVER_HOME/server_name</td>
<td>Contains the hostname running pbs_server</td>
</tr>
<tr>
<td>$PBS_SERVER_HOME/mom_priv</td>
<td>The default directory for configuration files, typically (/usr/spool/pbs)/mom_priv</td>
</tr>
<tr>
<td>$PBS_SERVER_HOME/mom_logs</td>
<td>Directory for log files recorded by the server</td>
</tr>
<tr>
<td>$PBS_SERVER_HOME/mom_priv/-prologue</td>
<td>The administrative script to be run before job execution</td>
</tr>
</tbody>
</table>
### File

<table>
<thead>
<tr>
<th>File</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$PBS_SERVER_HOME/mom_priv/epilogue</td>
<td>The administrative script to be run after job execution</td>
</tr>
</tbody>
</table>

### Signal handling

`pbs_mom` handles the following signals:

<table>
<thead>
<tr>
<th>Signal</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIGHUP</td>
<td>Causes <code>pbs_mom</code> to re-read its configuration file, close and reopen the log file, and reinitialize resource structures.</td>
</tr>
<tr>
<td>SIGALRM</td>
<td>Results in a log file entry. The signal is used to limit the time taken by certain children processes, such as the prologue and epilogue.</td>
</tr>
<tr>
<td>SIGINT and SIGTERM</td>
<td>Results in <code>pbs_mom</code> exiting without terminating any running jobs. This is the action for the following signals as well: SIGXCPU, SIGFSZ, SIGCPUJLM, and SIGSHUTDN.</td>
</tr>
<tr>
<td>SIGUSR1, SIGUSR2</td>
<td>Causes the MOM to increase and decrease logging levels, respectively.</td>
</tr>
<tr>
<td>SIGPIPE, SIGINFO</td>
<td>Are ignored.</td>
</tr>
<tr>
<td>SIGBUS, SIGFPE, SIGILL, SIGTRAP, and SIGSYS</td>
<td>Cause a core dump if the PBSCOREDUMP environmental variable is defined.</td>
</tr>
</tbody>
</table>

All other signals have their default behavior installed.

### Exit status

If the `pbs_mom` command fails to begin operation, the server exits with a value greater than zero.

### Related topics

- `pbs_server`(8B)

### Non-Adaptive Computing topics

- `pbs_scheduler_basl`(8B)
- `pbs_scheduler_tcl`(8B)
- PBS External Reference Specification
- PBS Administrators Guide
**pbs_server**

(*PBS Server*) pbs batch system manager

**Synopsis**

```bash
pbs_server [ -a active] [ -c] [ -d config_path] [ -f force overwrite] [ -p port] [ -A acctfile] [ -L logfile] [ -S scheduler_port] [ -h hostname] [ -t type] [ --ha] [ -n don't send hierarchy] [ --about] [ -v] [ --version]
```

**Description**

The pbs_server command starts the operation of a batch server on the local host. Typically, this command will be in a local boot file such as /etc/rc.local. If the batch server is already in execution, pbs_server will exit with an error. To ensure that the pbs_server command is not runnable by the general user community, the server will only execute if its real and effective uid is zero.

The server will record a diagnostic message in a log file for any error occurrence. The log files are maintained in the `server_logs` directory below the home directory of the server. If the log file cannot be opened, the diagnostic message is written to the system console.

As of TORQUE 4.0, the pbs_server is multi-threaded which leads to quicker response to client commands, is more robust, and allows for higher job throughput.

**Options**

<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-A</td>
<td>acctfile</td>
<td>Specifies an absolute path name of the file to use as the accounting file. If not specified, the file name will be the current date in the PBS_HOME/server_priv/accounting directory.</td>
</tr>
<tr>
<td>-a</td>
<td>active</td>
<td>Specifies if scheduling is active or not. This sets the server attribute scheduling. If the option argument is &quot;true&quot; (&quot;True&quot;, &quot;t&quot;, &quot;T&quot;, or &quot;1&quot;), the server is active and the PBS job scheduler will be called. If the argument is &quot;false&quot; (&quot;False&quot;, &quot;f&quot;, &quot;F&quot;, or &quot;0&quot;), the server is idle, and the scheduler will not be called and no jobs will be run. If this option is not specified, the server will retain the prior value of the scheduling attribute.</td>
</tr>
<tr>
<td>-c</td>
<td>wait_for_moms</td>
<td>This directs pbs_server to send the MOM hierarchy only to MOMs that request it for the first 10 minutes. After 10 minutes, it attempts to send the MOM hierarchy to MOMs that haven't requested it already. This greatly reduces traffic on start up.</td>
</tr>
<tr>
<td>Option</td>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>-d</td>
<td>config_directory</td>
<td>Specifies the path of the directory which is home to the server's configuration files, PBS_HOME. A host may have multiple servers. Each server must have a different configuration directory. The default configuration directory is given by the symbol $PBS_SERVER_HOME which is typically var/spool/torque.</td>
</tr>
<tr>
<td>-f</td>
<td>force overwrite</td>
<td>Forces an overwrite of the server database. This can be useful to bypass the yes/no prompt when running something like pbs_server -t create and can ease installation and configuration of TORQUE via scripts.</td>
</tr>
<tr>
<td>-h</td>
<td>hostname</td>
<td>Causes the server to start under a different hostname as obtained from gethostname (2). Useful for servers with multiple network interfaces to support connections from clients over an interface that has a hostname assigned that differs from the one that is returned by gethost name(2).</td>
</tr>
<tr>
<td>-ha</td>
<td>high_availability</td>
<td>Starts server in high availability mode (for details, see Server high availability on page 91).</td>
</tr>
<tr>
<td>-L</td>
<td>logfile</td>
<td>Specifies an absolute path name of the file to use as the log file. If not specified, the file will be the current date in the PBS_HOME/server_logs directory (see the -d option).</td>
</tr>
<tr>
<td>-n</td>
<td>no send</td>
<td>This directs pbs_server to not send the hierarchy to all the MOMs on startup. Instead, the hierarchy is only sent if a MOM requests it. This flag works only in conjunction with the local MOM hierarchy feature.</td>
</tr>
<tr>
<td>-p</td>
<td>port</td>
<td>Specifies the port number on which the server will listen for batch requests. If multiple servers are running on a single host, each must have its own unique port number. This option is for use in testing with multiple batch systems on a single host.</td>
</tr>
<tr>
<td>-S</td>
<td>scheduler_port</td>
<td>Specifies the port number to which the server should connect when contacting the scheduler. The argument scheduler_conn is of the same syntax as under the -M option.</td>
</tr>
</tbody>
</table>
Option | Name | Description
---|---|---
-t | type | Specifies the impact on jobs which were in execution, running, when the server shut down. If the running job is not rerunnable or restartable from a checkpoint image, the job is aborted. If the job is rerunnable or restartable, then the actions described below are taken. When the type argument is:
- **hot** – All jobs are requeued except non-rerunnable jobs that were executing. Any rerunnable job which was executing when the server went down will be run immediately. This returns the server to the same state as when it went down. After those jobs are restarted, then normal scheduling takes place for all remaining queued jobs.
- **warm** – All rerunnable jobs which were running when the server went down are requeued. All other jobs are maintained. New selections are made for which jobs are placed into execution. Warm is the default if -t is not specified.
- **cold** – All jobs are deleted. Positive confirmation is required before this direction is accepted.
- **create** – The server will discard any existing configuration files, queues and jobs, and initialize configuration files to the default values. The server is idled.

Files

<table>
<thead>
<tr>
<th>File</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TORQUE_HOME/server_priv</td>
<td>Default directory for configuration files, typically /usr/spool/pbs/server_priv</td>
</tr>
<tr>
<td>TORQUE_HOME/server_logs</td>
<td>Directory for log files recorded by the server</td>
</tr>
</tbody>
</table>

Signal handling

On receipt of the following signals, the server performs the defined action:

<table>
<thead>
<tr>
<th>Action</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIGHUP</td>
<td>The current server log and accounting log are closed and reopened. This allows for the prior log to be renamed and a new log started from the time of the signal.</td>
</tr>
<tr>
<td>SIGINT</td>
<td>Causes an orderly shutdown of pbs_server.</td>
</tr>
</tbody>
</table>
### Action | Description
--- | ---
SIGUSR1, SIGUSR2 | Causes server to increase and decrease logging levels, respectively.
SIGTERM | Causes an orderly shutdown of pbs_server.
SIGSHUTDN | On systems (Unicos) where SIGSHUTDN is defined, it also causes an orderly shutdown of the server.
SIGPIPE | This signal is ignored.

All other signals have their default behavior installed.

**Exit status**

If the server command fails to begin batch operation, the server exits with a value greater than zero.

**Related topics**

- pbs_mom(8B)
- pbsnodes(8B)
- qmgr(1B)
- qrun(8B)
- qsub(1B)
- qterm(8B)

**Non-Adaptive Computing topics**

- pbs_connect(3B)
- pbs_sched_basl(8B)
- pbs_sched_tcl(8B)
- qdisable(8B)
- qenable(8B)
- qstart(8B)
- qstop(8B)
- PBS External Reference Specification

**pbs_track**

Starts a new process and informs pbs_mom to start tracking it.

**Synopsis**

```
bps_track -j <JOBID> [-b] <executable> [args]
```
Description

The `pbs_track` command tells a `pbs_mom` daemon to monitor the lifecycle and resource usage of the process that it launches using `exec()`. The `pbs_mom` is told about this new process via the Task Manager API, using `tm_adopt()`. The process must also be associated with a job that already exists on the `pbs_mom`.

By default, `pbs_track` will send its PID to TORQUE via `tm_adopt()`. It will then perform an `exec()`, causing `<executable>` to run with the supplied arguments. `pbs_track` will not return until the launched process has completed because it becomes the launched process.

This command can be considered related to the `pbsdsh` command which uses the `tm_spawn()` API call. The `pbsdsh` command asks a `pbs_mom` to launch and track a new process on behalf of a job. When it is not desirable or possible for the `pbs_mom` to spawn processes for a job, `pbs_track` can be used to allow an external entity to launch a process and include it as part of a job.

This command improves integration with TORQUE and SGI's MPT MPI implementation.

Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>-j &lt;JOBID&gt;</code></td>
<td>Job ID the new process should be associated with.</td>
</tr>
<tr>
<td><code>-b</code></td>
<td>Instead of having <code>pbs_track</code> send its PID to TORQUE, it will fork() first, send the child PID to TORQUE, and then execute from the forked child. This essentially &quot;backgrounds&quot; <code>pbs_track</code> so that it will return after the new process is launched.</td>
</tr>
</tbody>
</table>

Operands

The `pbs_track` command accepts a path to a program/executable (`<executable>`) and, optionally, one or more arguments to pass to that program.

Exit status

Because the `pbs_track` command becomes a new process (if used without `-b`), its exit status will match that of the new process. If the `-b` option is used, the exit status will be zero if no errors occurred before launching the new process.

If `pbs_track` fails, whether due to a bad argument or other error, the exit status will be set to a non-zero value.

Related topics

- `pbsdsh`(1B)

Non-Adaptive Computing topics

- `tm_spawn`(3B)
pbsdsh

Distribute tasks to nodes under pbs.

Synopsis

pbsdsh [-h nodename] [-o] [-v] program [args]

Description

Executes (spawns) a normal Unix program on one or more nodes under control of the Portable Batch System, PBS. Pbsdsh uses the Task Manager API (see tm_spawn(3)) to distribute the program on the allocated nodes.

When run without the -c or the -n option, pbsdsh will spawn the program on all nodes allocated to the PBS job. The spawns take place concurrently - all execute at (about) the same time.

Users will find the PBS_TASKNUM, PBS_NODENUM, and the PBS_VNODENUM environmental variables useful. They contain the TM task id, the node identifier, and the cpu (virtual node) identifier.

Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-c</td>
<td>copies</td>
<td>The program is spawned on the first Copies nodes allocated. This option is mutually exclusive with -n.</td>
</tr>
<tr>
<td>-h</td>
<td>hostname</td>
<td>The program is spawned on the node specified.</td>
</tr>
<tr>
<td>-n</td>
<td>node</td>
<td>The program is spawned on one node which is the n-th node allocated. This option is mutually exclusive with -c.</td>
</tr>
<tr>
<td>-o</td>
<td>---</td>
<td>Capture stdout of the spawned program. Normally stdout goes to the job’s output.</td>
</tr>
<tr>
<td>-s</td>
<td>---</td>
<td>If this option is given, the program is run in turn on each node, one after the other.</td>
</tr>
<tr>
<td>-u</td>
<td>---</td>
<td>The program is run once on each node (unique). This ignores the number of allocated processors on a given node.</td>
</tr>
<tr>
<td>-v</td>
<td>---</td>
<td>Verbose output about error conditions and task exit status is produced.</td>
</tr>
</tbody>
</table>

Operands

The first operand, program, is the program to execute.

Additional operands are passed as arguments to the program.
Standard error

The **pbsdsh** command will write a diagnostic message to standard error for each error occurrence.

Exit status

Upon successful processing of all the operands presented to the command, the exit status will be a value of zero.

If the **pbsdsh** command fails to process any operand, or fails to contact the MOM daemon on the localhost the command exits with a value greater than zero.

Related topics

- [**qsub**(1B)]

Non-Adaptive Computing topics

- [**tm_spawn**(3B)]

### pbsnodes

**PBS node manipulation.**

**Synopsis**

```plaintext
pbsnodes [-{a|x}] [-q] [-s server] [node[:property]]
pbsnodes -l [-q] [-s server] [state] [nodename[:property] ...]
pbsnodes [-{c|d|o|r}] [-q] [-s server] [n -l] [N "note"] [node[:property]]
```

**Description**

The **pbsnodes** command is used to mark nodes down, free or offline. It can also be used to list nodes and their state. Node information is obtained by sending a request to the PBS job server. Sets of nodes can be operated on at once by specifying a node property prefixed by a colon. (For more information, see Node states.)

Nodes do not exist in a single state, but actually have a set of states. For example, a node can be simultaneously "busy" and "offline". The "free" state is the absence of all other states and so is never combined with other states.

In order to execute **pbsnodes** with other than the **-a** or **-l** options, the user must have PBS Manager or Operator privilege.

**Options**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-a</td>
<td>All attributes of a node or all nodes are listed. This is the default if no flag is given.</td>
</tr>
<tr>
<td>Option</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td><code>-x</code></td>
<td>Same as <code>-a</code>, but the output has an XML-like format.</td>
</tr>
<tr>
<td><code>-c</code></td>
<td>Clear OFFLINE from listed nodes.</td>
</tr>
<tr>
<td><code>-d</code></td>
<td>Print MOM diagnosis on the listed nodes. Not yet implemented. Use <code>momctl</code> instead.</td>
</tr>
<tr>
<td><code>-o</code></td>
<td>Add the OFFLINE state. This is different from being marked DOWN. OFFLINE prevents new jobs from running on the specified nodes. This gives the administrator a tool to hold a node out of service without changing anything else. The OFFLINE state will never be set or cleared automatically by <code>pbs_server</code>; it is purely for the manager or operator.</td>
</tr>
<tr>
<td><code>-p</code></td>
<td>Purge the node record from <code>pbs_server</code>. Not yet implemented.</td>
</tr>
<tr>
<td><code>-r</code></td>
<td>Reset the listed nodes by clearing OFFLINE and adding DOWN state. <code>pbs_server</code> will ping the node and, if they communicate correctly, free the node.</td>
</tr>
<tr>
<td><code>-l</code></td>
<td>List node names and their state. If no state is specified, only nodes in the DOWN, OFFLINE, or UNKNOWN states are listed. Specifying a state string acts as an output filter. Valid state strings are &quot;active&quot;, &quot;all&quot;, &quot;busy&quot;, &quot;down&quot;, &quot;free&quot;, &quot;job-exclusive&quot;, &quot;job-sharing&quot;, &quot;offline&quot;, &quot;reserve&quot;, &quot;state-unknown&quot;, &quot;time-shared&quot;, and &quot;up&quot;.</td>
</tr>
<tr>
<td></td>
<td>• Using <em>all</em> displays all nodes and their attributes.</td>
</tr>
<tr>
<td></td>
<td>• Using <em>active</em> displays all nodes which are job-exclusive, job-sharing, or busy.</td>
</tr>
<tr>
<td></td>
<td>• Using <em>up</em> displays all nodes in an &quot;up state&quot;. Up states include job-exclusive, job-sharing, reserve, free, busy and time-shared.</td>
</tr>
<tr>
<td></td>
<td>• All other strings display the nodes which are currently in the state indicated by the string.</td>
</tr>
<tr>
<td><code>-N</code></td>
<td>Specify a &quot;note&quot; attribute. This allows an administrator to add an arbitrary annotation to the listed nodes. To clear a note, use <code>-N &quot;&quot;</code> or <code>-N n</code>.</td>
</tr>
<tr>
<td><code>-n</code></td>
<td>Show the &quot;note&quot; attribute for nodes that are DOWN, OFFLINE, or UNKNOWN. This option requires <code>-l</code>.</td>
</tr>
<tr>
<td><code>-q</code></td>
<td>Suppress all error messages.</td>
</tr>
<tr>
<td><code>-s</code></td>
<td>Specify the PBS server’s hostname or IP address.</td>
</tr>
</tbody>
</table>

**Related topics**
- [`pbs_server`(BB)]

**Non-Adaptive Computing topics**
- PBS External Reference Specification
qalter

Alter batch job.

Synopsis

calter [-a date_time][-A account_string][-c interval][-e path_name]
[-h hold_list][-j join_list][-k keep_list][-l resource_list]
[-m mail_options][-M mail_list][-n][-N name][-o path_name]
[-p priority][-r y|n][-s path_name_list][-u user_list]
[-v variable_list][-w additional_attributes]
[-t array_range]
job_identifier ...

Description

The qalter command modifies the attributes of the job or jobs specified by job_identifier on the command line. Only those attributes listed as options on the command will be modified. If any of the specified attributes cannot be modified for a job for any reason, none of that job's attributes will be modified.

The qalter command accomplishes the modifications by sending a Modify Job batch request to the batch server which owns each job.

Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
</table>
| -a     | date_time  | Replaces the time at which the job becomes eligible for execution. The date_time argument syntax is:


  If the month, MM, is not specified, it will default to the current month if the specified day DD, is in the future. Otherwise, the month will be set to next month. Likewise, if the day, DD, is not specified, it will default to today if the time hhmm is in the future. Otherwise, the day will be set to tomorrow.

  This attribute can be altered once the job has begun execution, but it will not take effect unless the job is rerun. |
<p>| -A     | account_string | Replaces the account string associated with the job. This attribute cannot be altered once the job has begun execution. |</p>
<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
</table>
| -c     | checkpoint_interval | Replaces the interval at which the job will be checkpointed. If the job executes upon a host which does not support checkpointing, this option will be ignored.  
The interval argument is specified as:  
- n – No checkpointing is to be performed.  
- s – Checkpointing is to be performed only when the server executing the job is shutdown.  
- c – Checkpointing is to be performed at the default minimum CPU time for the queue from which the job is executing.  
- c=minutes – Checkpointing is performed at intervals of the specified amount of time in minutes. Minutes are the number of minutes of CPU time used, not necessarily clock time.  

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>This value must be greater than zero. If the number is less than the default checkpoint time, the default time will be used.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>This attribute can be altered once the job has begun execution, but the new value does not take effect unless the job is rerun.</td>
</tr>
</tbody>
</table>
| -e     | path_name     | Replaces the path to be used for the standard error stream of the batch job. The path argument is of the form:  
[hostname:]path_name  
where hostname is the name of a host to which the file will be returned and path_name is the path name on that host in the syntax recognized by POSIX 1003.1. The argument will be interpreted as follows:  
- path_name – Where path_name is not an absolute path name, then the qalter command will expand the path name relative to the current working directory of the command. The command will supply the name of the host upon which it is executing for the hostname component.  
- hostname:path_name – Where path_name is not an absolute path name, then the qalter command will not expand the path name. The execution server will expand it relative to the home directory of the user on the system specified by hostname.  
This attribute can be altered once the job has begun execution, but it will not take effect unless the job is rerun. |
### Option | Name | Description
--- | --- | ---
-h | hold_list | Updates the types of holds on the job. The `hold_list` argument is a string of one or more of the following characters:
- `u` – Add the USER type hold.
- `s` – Add the SYSTEM type hold if the user has the appropriate level of privilege. (Typically reserved to the batch administrator.)
- `o` – Add the OTHER (or OPERATOR ) type hold if the user has the appropriate level of privilege. (Typically reserved to the batch administrator and batch operator.)
- `n` – Set to none and clear the hold types which could be applied with the user's level of privilege. Repetition of characters is permitted, but "n" may not appear in the same option argument with the other three characters.

This attribute can be altered once the job has begun execution, but the hold will not take effect unless the job is rerun.

-j | join | Declares which standard streams of the job will be merged together. The `join` argument value may be the characters "oe" and "eo", or the single character "n".
An argument value of oe directs that the standard output and standard error streams of the job will be merged, intermixed, and returned as the standard output. An argument value of eo directs that the standard output and standard error streams of the job will be merged, intermixed, and returned as the standard error.
A value of n directs that the two streams will be two separate files. This attribute can be altered once the job has begun execution, but it will not take effect unless the job is rerun.

- If using either the `-e` or the `-o` option and the `-j eo|oe` option, the `-j` option takes precedence and all standard error and output messages go to the chosen output file.
<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-k</td>
<td>keep</td>
<td>Defines which if either of standard output or standard error of the job will be retained on the execution host. If set for a stream, this option overrides the path name for that stream. The argument is either the single letter &quot;e&quot;, &quot;o&quot;, or &quot;n&quot;, or one or more of the letters &quot;e&quot; and &quot;o&quot; combined in either order.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• <em>n</em> – No streams are to be retained.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• <em>e</em> – The standard error stream is to be retained on the execution host. The stream will be placed in the home directory of the user under whose user id the job executed. The file name will be the default file name given by:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>job_name.esequence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>where job_name is the name specified for the job, and sequence is the sequence number component of the job identifier.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• <em>o</em> – The standard output stream is to be retained on the execution host. The stream will be placed in the home directory of the user under whose user id the job executed. The file name will be the default file name given by:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>job_name.osequence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>where job_name is the name specified for the job, and sequence is the sequence number component of the job identifier.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• <em>eo</em> – Both the standard output and standard error streams will be retained.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• <em>oe</em> – Both the standard output and standard error streams will be retained. This attribute cannot be altered once the job has begun execution.</td>
</tr>
<tr>
<td>-l</td>
<td>resource_list</td>
<td>Modifies the list of resources that are required by the job. The resource_list argument is in the following syntax: resource_name=[{value}][,resource_name=[{value}],...]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If a requested modification to a resource would exceed the resource limits for jobs in the current queue, the server will reject the request.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If the job is running, only certain resources can be altered. Which resources can be altered in the run state is system dependent. A user may only lower the limit for those resources.</td>
</tr>
<tr>
<td>-m</td>
<td>mail_options</td>
<td>Replaces the set of conditions under which the execution server will send a mail message about the job. The mail_options argument is a string which consists of the single character &quot;n&quot;, or one or more of the characters &quot;a&quot;, &quot;b&quot;, and &quot;e&quot;.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If the character &quot;n&quot; is specified, no mail will be sent. For the letters &quot;a&quot;, &quot;b&quot;, and &quot;e&quot;:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• <em>a</em> – Mail is sent when the job is aborted by the batch system.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• <em>b</em> – Mail is sent when the job begins execution.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• <em>e</em> – Mail is sent when the job ends.</td>
</tr>
<tr>
<td>Option</td>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>-------------</td>
</tr>
<tr>
<td>-M</td>
<td>user_list</td>
<td>Replaces the list of users to whom mail is sent by the execution server when it sends mail about the job. The user_list argument is of the form: <code>user[@host],[user[@host],...]</code></td>
</tr>
<tr>
<td>-n</td>
<td>node-exclusive</td>
<td>Sets or unsets exclusive node allocation on a job. Use the y and n options to enable or disable the feature. This affects only cpusets and compatible schedulers.</td>
</tr>
<tr>
<td>-N</td>
<td>name</td>
<td>Renames the job. The name specified may be up to and including 15 characters in length. It must consist of printable, non-white space characters with the first character alphabetic.</td>
</tr>
<tr>
<td>-o</td>
<td>path</td>
<td>Replaces the path to be used for the standard output stream of the batch job. The path argument is of the form: <code>[hostname:]path_name</code> where <code>hostname</code> is the name of a host to which the file will be returned and <code>path_name</code> is the path name on that host in the syntax recognized by POSIX. The argument will be interpreted as follows:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• <code>path_name</code> – Where <code>path_name</code> is not an absolute path name, then the <code>qalter</code> command will expand the path name relative to the current working directory of the command. The command will supply the name of the host upon which it is executing for the hostname component.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• <code>hostname:path_name</code> – Where <code>path_name</code> is not an absolute path name, then the <code>qalter</code> command will not expand the path name. The execution server will expand it relative to the home directory of the user on the system specified by hostname.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>This attribute can be altered once the job has begun execution, but it will not take effect unless the job is rerun.</td>
</tr>
<tr>
<td>-p</td>
<td>priority</td>
<td>Replaces the priority of the job. The priority argument must be an integer between -1024 and +1023 inclusive. This attribute can be altered once the job has begun execution, but it will not take effect unless the job is rerun.</td>
</tr>
<tr>
<td>-r</td>
<td>[y/n]</td>
<td>Declares whether the job is rerunnable (see the <code>qrerun</code> command). The option argument c is a single character. PBS recognizes the following characters: y and n. If the argument is &quot;y&quot;, the job is marked rerunnable. If the argument is &quot;n&quot;, the job is marked as not rerunnable.</td>
</tr>
<tr>
<td>Option</td>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>---------------</td>
<td>-------------</td>
</tr>
<tr>
<td>-S</td>
<td>path</td>
<td>Declares the shell that interprets the job script. The option argument path_list is in the form: path[@host][,path[@host],...] Only one path may be specified for any host named. Only one path may be specified without the corresponding host name. The path selected will be the one with the host name that matched the name of the execution host. If no matching host is found, then the path specified (without a host) will be selected. If the -S option is not specified, the option argument is the null string or no entry from the path_list is selected, the execution will use the login shell of the user on the execution host. This attribute can be altered once the job has begun execution, but it will not take effect unless the job is rerun.</td>
</tr>
</tbody>
</table>
| -t     | array_range   | The array_range argument is an integer id or a range of integers. Multiple ids or id ranges can be combined in a comma delimited list. Examples: -t 1-100 or -t 1,10,50-100 If an array range isn’t specified, the command tries to operate on the entire array. The command acts on the array (or specified range of the array) just as it would on an individual job. An optional “slot limit” can be specified to limit the amount of jobs that can run concurrently in the job array. The default value is unlimited. The slot limit must be the last thing specified in the array_request and is delimited from the array by a percent sign (%). \[
\text{qalter weatherSimulationArray[]} -t \%20
\] Here, the array weatherSimulationArray[] is configured to allow a maximum of 20 concurrently running jobs. Slot limits can be applied at job submit time with qsub, or can be set in a global server parameter policy with max_slot_limit. |
| -u     | user_list     | Replaces the user name under which the job is to run on the execution system. The user_list argument is of the form: user[@host][,user[@host],...] Only one user name may be given for per specified host. Only one of the user specifications may be supplied without the corresponding host specification. That user name will be used for execution on any host not named in the argument list. This attribute cannot be altered once the job has begun execution. |
| -W     | additional_attributes | The -W option allows for the modification of additional job attributes. Note if white space occurs anywhere within the option argument string or the equal sign, "=" occurs within an attribute_value string, then the string must be enclosed with either single or double quote marks. To see the attributes PBS currently supports within the -W option, see Table A-1: -W additional_attributes on page 176. |
Table A-1: W additional_attributes
### Attribute

<table>
<thead>
<tr>
<th>depend=dependency_list</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
</table>

Redefines the dependencies between this and other jobs. The `dependency_list` is in the form:

```
type[:argument[:argument...][,type:argument...]]
```

The argument is either a numeric count or a PBS job id according to type. If argument is a count, it must be greater than 0. If it is a job id and is not fully specified in the form:

```
seq_number.server.name
```

it will be expanded according to the default server rules.

If argument is null (the preceding colon need not be specified), the dependency of the corresponding type is cleared (unset).

- **syncaccount:count** – This job is the first in a set of jobs to be executed at the same time. Count is the number of additional jobs in the set.
- **syncwith:jobid** – This job is an additional member of a set of jobs to be executed at the same time. In the above and following dependency types, `jobid` is the job identifier of the first job in the set.
- **after:jobid [jobid...]** – This job may be scheduled for execution at any point after jobs `jobid` have started execution.
- **afterok:jobid [jobid...]** – This job may be scheduled for execution only after jobs `jobid` have terminated with no errors. See the csh warning under “Extended Description”.
- **afternotok:jobid [jobid...]** – This job may be scheduled for execution only after jobs `jobid` have terminated with errors. See the csh warning under “Extended Description”.
- **afterany:jobid [jobid...]** – This job may be scheduled for execution after jobs `jobid` have terminated, with or without errors.
- **oncount** – This job may be scheduled for execution after count dependencies on other jobs have been satisfied. This dependency is used in conjunction with any of the 'before' dependencies shown below. If job A has on:2, it will wait for two jobs with 'before' dependencies on job A to be fulfilled before running.
- **before:jobid [jobid...]** – When this job has begun execution, then jobs `jobid...` may begin.
- **beforeok:jobid [jobid...]** – If this job terminates execution without errors, then jobs `jobid...` may begin. See the csh warning under "Extended Description".
- **beforenotok:jobid [jobid...]** – If this job terminates execution with errors, then jobs `jobid...` may begin. See the csh warning under "Extended Description".
- **beforeany:jobid [jobid...]** – When this job terminates execution, jobs `jobid...` may begin.

If any of the before forms are used, the job referenced by `jobid` must have been submitted with a dependency type of on.

If any of the before forms are used, the jobs referenced by `jobid` must have the same owner as the job being altered. Otherwise, the dependency will not take effect.

Error processing of the existence, state, or condition of the job specified to `qalter` is a deferred service, i.e. the check is performed after the job is queued. If an error is detected, the job will be deleted by the server. Mail will be sent to the job submitter stating the error.
### Attribute | Description
--- | ---
**group_list=g_list** | Alters the group name under which the job is to run on the execution system. The `g_list` argument is of the form:
```
group[@host][,group[@host],...]
```
Only one group name may be given per specified host. Only one of the group specifications may be supplied without the corresponding host specification. That group name will used for execution on any host not named in the argument list.

**stagein=file_list**  
**stageout=file_list** | Alters which files are staged (copied) in before job start or staged out after the job completes execution. The `file_list` is in the form:
```
local_file@hostname:remote_file[,]
```
The name `local_file` is the name on the system where the job executes. It may be an absolute path or a path relative to the home directory of the user. The name `remote_file` is the destination name on the host specified by `hostname`. The name may be absolute or relative to the user’s home directory on the destination host.

---

**Operands**

The `qalter` command accepts one or more job_identifier operands of the form:

```
sequence_number[.server_name][@server]
```

**Standard error**

Any error condition, either in processing the options or the operands, or any error received in reply to the batch requests will result in an error message being written to standard error.

**Exit status**

Upon successful processing of all the operands presented to the `qalter` command, the exit status will be a value of zero.

If the `qalter` command fails to process any operand, the command exits with a value greater than zero.

**Copyright**

Portions of this text are reprinted and reproduced in electronic form from IEEE Std 1003.1, 2003 Edition, Standard for Information Technology -- Portable Operating System Interface (POSIX), The Open Group Base Specifications Issue 6, Copyright © 2001-2003 by the Institute of Electrical and Electronics Engineers, Inc and The Open Group. In the event of any discrepancy between this version and the original IEEE and The Open Group Standard, the original IEEE and The Open Group Standard is the referee document. The original Standard can be obtained online at [http://www.opengroup.org/unix/online.html](http://www.opengroup.org/unix/online.html).

**Related topics**

- `qdel`
- `qhold`
Non-Adaptive Computing topics

- Batch Environment Services
- qmove
- touch

qchkpt

Checkpoint pbs batch jobs.

Synopsis

qchkpt <JOBID>[ <JOBID>] ...

Description

The `qchkpt` command requests that the PBS MOM generate a checkpoint file for a running job.

This is an extension to POSIX.2d.

The `qchkpt` command sends a Chkpt Job batch request to the server as described in the general section.

Options

None.

Operands

The `qchkpt` command accepts one or more job_identifier operands of the form:

`sequence_number[.server_name][@server]`

Examples

```
> qchkpt 3233 request a checkpoint for job 3233
```

Standard error

The `qchkpt` command will write a diagnostic message to standard error for each error occurrence.

Exit status

Upon successful processing of all the operands presented to the `qchkpt` command, the exit status will be a value of zero.

If the `qchkpt` command fails to process any operand, the command exits with a value greater than zero.
Related topics

- qhold(1B)
- qrls(1B)
- qalter(1B)
- qsub(1B)

Non-Adaptive Computing topics

- pbs_alterjob(3B)
- pbs_holdjob(3B),
- pbs_rlsjob(3B)
- pbs_job_attributes(7B)
- pbs_resources_unicos8(7B)

qdel

(delete job)

Synopsis

qdel [{-a <asynchronous delete>|-m <message>|-p|-W <delay>|-t <array_range>}] <JOBID>[ <JOBID>]... | 'all' | 'ALL'

Description

The qdel command deletes jobs in the order in which their job identifiers are presented to the command. A job is deleted by sending a Delete Job batch request to the batch server that owns the job. A job that has been deleted is no longer subject to management by batch services.

A batch job may be deleted by its owner, the batch operator, or the batch administrator.

A batch job being deleted by a server will be sent a SIGTERM signal following by a SIGKILL signal. The time delay between the two signals is an attribute of the execution queue from which the job was run (set table by the administrator). This delay may be overridden by the -W option.

See the PBS ERS section 3.1.3.3, "Delete Job Request", for more information.

Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-a</td>
<td>asynchronous</td>
<td>Performs an asynchronous delete. The server responds to the user before</td>
</tr>
<tr>
<td></td>
<td>delete</td>
<td>contacting the MOM. The option qdel -a all performs qdel all due to</td>
</tr>
<tr>
<td></td>
<td></td>
<td>restrictions from being single-threaded.</td>
</tr>
<tr>
<td>Option</td>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>----------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>-W</td>
<td>delay</td>
<td>Specifies the wait delay between the sending of the SIGTERM and SIGKILL signals. The argument is the length of time in seconds of the delay.</td>
</tr>
<tr>
<td>-p</td>
<td>purge</td>
<td>Forcibly purges the job from the server. This should only be used if a running job will not exit because its allocated nodes are unreachable. The admin should make every attempt at resolving the problem on the nodes. If a job's mother superior recovers after purging the job, any epilogue scripts may still run. This option is only available to a batch operator or the batch administrator.</td>
</tr>
<tr>
<td>-m</td>
<td>message</td>
<td>Specify a comment to be included in the email. The argument message specifies the comment to send. This option is only available to a batch operator or the batch administrator.</td>
</tr>
<tr>
<td>-t</td>
<td>array_range</td>
<td>The array_range argument is an integer id or a range of integers. Multiple ids or id ranges can be combined in a comma delimited list (examples: -t 1-100 or -t 1,10,50-100). If an array range isn’t specified, the command tries to operate on the entire array. The command acts on the array (or specified range of the array) just as it would on an individual job.</td>
</tr>
</tbody>
</table>

**Operands**

The `qdel` command accepts one or more job_identifier operands of the form:

```
sequence_number[.server_name][@server]
```

or

```
all
```

**Examples**

```
> qdel 1324
> qdel 1324-3 To delete one job of a job array
> qdel all To delete all jobs (Version 2.3.0 and later)
```

**Standard error**

The `qdel` command will write a diagnostic messages to standard error for each error occurrence.

**Exit status**

Upon successful processing of all the operands presented to the `qdel` command, the exit status will be a value of zero.

If the `qdel` command fails to process any operand, the command exits with a value greater than zero.
Related topics

- `qsub`(1B)
- `qeig`(1B)

Non-Adaptive Computing topics

- `pbs_deljob`(3B)

## qgpumode

*(GPU mode)*

### Synopsis

```
qgpumode -H host -g gpuid -m mode
```

### Description

The `qgpumode` command specifies the mode for the GPU. This command triggers an immediate update of the `pbs_server`.

### Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-H</td>
<td>Specifies the host where the GPU is located.</td>
</tr>
<tr>
<td>-g</td>
<td>Specifies the ID of the GPU. This varies depending on the version of the Nvidia driver used. For driver 260.x, it is 0, 1, and so on. For driver 270.x, it is the PCI bus address, i.e., 0:5:0.</td>
</tr>
</tbody>
</table>
### Option

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
</table>
| -m     | Specifies the new mode for the GPU:  
  - 0 (Default/Shared): Default/shared compute mode. Multiple threads can use cudaSetDevice() with this device.  
  - 1 (Exclusive Thread): Compute-exclusive-thread mode. Only one thread in one process is able to use cudaSetDevice() with this device.  
  - 2 (Prohibited): Compute-prohibited mode. No threads can use cudaSetDevice() with this device.  
  - 3 (Exclusive Process): Compute-exclusive-process mode. Many threads in one process are able to use cudaSetDevice() with this device. |

```
qgpumode -H node01 -g 0 -m 1
```

This puts the first GPU on node01 into mode 1 (exclusive)

```
qgpumode -H node01 -g 0 -m 0
```

This puts the first GPU on node01 into mode 0 (shared)

### Related topics
- **qgpureset** on page 183

---

### qgpureset

*(reset GPU)*

**Synopsis**

```
qgpureset -H host -g gpuid -p -v
```

**Description**

The `qgpureset` command resets the GPU.

**Options**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-H</td>
<td>Specifies the host where the GPU is located.</td>
</tr>
<tr>
<td>-g</td>
<td>Specifies the ID of the GPU. This varies depending on the version of the Nvidia driver used. For driver 260.x, it is 0, 1, and so on. For driver 270.x, it is the PCI bus address, i.e., 0:5:0.</td>
</tr>
<tr>
<td>Option</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>-p</td>
<td>Specifies to reset the GPU's permanent ECC error count.</td>
</tr>
<tr>
<td>-v</td>
<td>Specifies to reset the GPU's volatile ECC error count.</td>
</tr>
</tbody>
</table>

Related topics
- qgpumode on page 182

qhold

*(hold job)*

Synopsis

```
qhold [{-h <HOLD LIST>|-t <array_range>}] <JOBID>[ <JOBID>] ... 
```

Description

The `qhold` command requests that the server place one or more holds on a job. A job that has a hold is not eligible for execution. There are three supported holds: USER, OTHER (also known as operator), and SYSTEM.

A user may place a USER hold upon any job the user owns. An "operator", who is a user with "operator privilege," may place either an USER or an OTHER hold on any job. The batch administrator may place any hold on any job.

If no `-h` option is given, the USER hold will be applied to the jobs described by the job_identifier operand list.

If the job identified by job_identifier is in the queued, held, or waiting states, then the hold type is added to the job. The job is then placed into held state if it resides in an execution queue.

If the job is in running state, then the following additional action is taken to interrupt the execution of the job. If checkpoint/restart is supported by the host system, requesting a hold on a running job will (1) cause the job to be checkpointed, (2) the resources assigned to the job will be released, and (3) the job is placed in the held state in the execution queue.

If checkpoint/restart is not supported, `qhold` will only set the requested hold attribute. This will have no effect unless the job is rerun with the `qrerun` command.
Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-h</td>
<td>hold_list</td>
<td>The hold_list argument is a string consisting of one or more of the letters &quot;u&quot;, &quot;o&quot;, or &quot;s&quot; in any combination. The hold type associated with each letter is:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• u – USER</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• o – OTHER</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• s – SYSTEM</td>
</tr>
<tr>
<td>-t</td>
<td>array_range</td>
<td>The array_range argument is an integer id or a range of integers. Multiple ids or id ranges can be combined in a comma delimited list (examples: -t 1-100 or -t 1,10,50-100).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If an array range isn’t specified, the command tries to operate on the entire array. The command acts on the array (or specified range of the array) just as it would on an individual job.</td>
</tr>
</tbody>
</table>

Operands

The qhold command accepts one or more job_identifier operands of the form:

sequence_number[.server_name][@server]

Example

```
> qhold -h u 3233 place user hold on job 3233
```

Standard error

The qhold command will write a diagnostic message to standard error for each error occurrence.

Exit status

Upon successful processing of all the operands presented to the qhold command, the exit status will be a value of zero.

If the qhold command fails to process any operand, the command exits with a value greater than zero.

Related topics

- qrls(1B)
- qalter(1B)
- qsub(1B)

Non-Adaptive Computing topics

- pbs_alterjob(3B)
- pbs_holdjob(3B)
- pbs_rlsjob(3B)
pbs_job_attributes(7B)
pbs_resources_unicos8(7B)

qmgr

(PBS Queue Manager) PBS batch system manager.

Synopsis

qmgr [-a] [-c command] [-e] [-n] [-z] [server...]

Description

The qmgr command provides an administrator interface to query and configure batch system parameters (see Server parameters on page 227).

The command reads directives from standard input. The syntax of each directive is checked and the appropriate request is sent to the batch server or servers.

The list or print subcommands of qmgr can be executed by general users. Creating or deleting a queue requires PBS Manager privilege. Setting or unsetting server or queue attributes requires PBS Operator or Manager privilege.

By default, the user root is the only PBS Operator and Manager. To allow other users to be privileged, the server attributes operators and managers will need to be set (i.e., as root, issue 'qmgr -c 'set server managers +=' <USER1>@<HOST>'). See "TORQUE/PBS Integration Guide - RM Access Control" in the Moab Workload Manager Administrator's Guide.

If qmgr is invoked without the -c option and standard output is connected to a terminal, qmgr will write a prompt to standard output and read a directive from standard input.

Commands can be abbreviated to their minimum unambiguous form. A command is terminated by a new line character or a semicolon, ";", character. Multiple commands may be entered on a single line. A command may extend across lines by escaping the new line character with a back-slash "\".

Comments begin with the "#" character and continue to end of the line. Comments and blank lines are ignored by qmgr.

Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-a</td>
<td>---</td>
<td>Abort qmgr on any syntax errors or any requests rejected by a server.</td>
</tr>
<tr>
<td>-c</td>
<td>command</td>
<td>Execute a single command and exit qmgr.</td>
</tr>
<tr>
<td>-e</td>
<td>---</td>
<td>Echo all commands to standard output.</td>
</tr>
<tr>
<td>Option</td>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>-------------</td>
</tr>
<tr>
<td>-n</td>
<td>---</td>
<td>No commands are executed, syntax checking only is performed.</td>
</tr>
<tr>
<td>-z</td>
<td>---</td>
<td>No errors are written to standard error.</td>
</tr>
</tbody>
</table>

**Operands**

The *server* operands identify the name of the batch server to which the administrator requests are sent. Each *server* conforms to the following syntax:

```
host_name[:port]
```

where *host_name* is the network name of the host on which the server is running and *port* is the port number to which to connect. If *port* is not specified, the default port number is used.

If *server* is not specified, the administrator requests are sent to the local server.

**Standard input**

The *qmgr* command reads standard input for directives until end of file is reached, or the exit or quit directive is read.

**Standard output**

If Standard Output is connected to a terminal, a command prompt will be written to standard output when *qmgr* is ready to read a directive.

If the `-e` option is specified, *qmgr* will echo the directives read from standard input to standard output.

**Standard error**

If the `-z` option is not specified, the *qmgr* command will write a diagnostic message to standard error for each error occurrence.

**Directive syntax**

A *qmgr* directive is one of the following forms:

```
command server [names] [attr OP value[,attr OP value,...]]
command queue [names] [attr OP value[,attr OP value,...]]
command node [names] [attr OP value[,attr OP value,...]]
```

where *command* is the command to perform on an object.

Commands are:
<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>active</td>
<td>Sets the active objects. If the active objects are specified, and the name is not given in a <code>qmgr</code> cmd the active object names will be used.</td>
</tr>
<tr>
<td>create</td>
<td>Is to create a new object, applies to queues and nodes.</td>
</tr>
<tr>
<td>delete</td>
<td>Is to destroy an existing object, applies to queues and nodes.</td>
</tr>
<tr>
<td>set</td>
<td>Is to define or alter attribute values of the object.</td>
</tr>
<tr>
<td>unset</td>
<td>Is to clear the value of attributes of the object.</td>
</tr>
</tbody>
</table>

ℹ️ This form does not accept an OP and value, only the attribute name.

| list     | Is to list the current attributes and associated values of the object. |
| print    | Is to print all the queue and server attributes in a format that will be usable as input to the `qmgr` command. |
| names    | Is a list of one or more names of specific objects. The name list is in the form: `name[@server][,queue_name[@server]...]` with no intervening white space. The name of an object is declared when the object is first created. If the name is `@server`, then all the objects of specified type at the server will be affected. |
| attr     | Specifies the name of an attribute of the object which is to be set or modified. If the attribute is one which consist of a set of resources, then the attribute is specified in the form: `attribute_name.resource_name` |
| OP       | Operation to be performed with the attribute and its value: |
|         | - "=" – set the value of the attribute. If the attribute has an existing value, the current value is replaced with the new value. |
|         | - "+=" – increase the current value of the attribute by the amount in the new value. |
|         | - "-=" – decrease the current value of the attribute by the amount in the new value. |
| value   | The value to assign to an attribute. If the value includes white space, commas or other special characters, such as the "#" character, the value string must be enclosed in quote marks ("). |

The following are examples of `qmgr` directives:
create queue fast priority=10,queue_type=e,enabled = true,max_running=0
set queue fast max_running +=2
create queue little
set queue little resources_max.mem=8mw,resources_max.cput=10
unset queue fast max_running
set node state = "down,offline"
active server s1,s2,s3
list queue @server1
set queue max_running = 10 - uses active queues

Exit status

Upon successful processing of all the operands presented to the qmgr command, the exit status will be a value of zero.

If the qmgr command fails to process any operand, the command exits with a value greater than zero.

Related topics

- pbs_server(8B)

Non-Adaptive Computing topics

- pbs_queue_attributes (7B)
- pbs_server_attributes (7B)
- qstart (8B), qstop (8B)
- qenable (8B), qdisable (8)
- PBS External Reference Specification

qmove

Move PBS batch jobs.

Synopsis

qmove destination jobId [jobId ...]

Description

To move a job is to remove the job from the queue in which it resides and instantiate the job in another queue. The qmove command issues a Move Job batch request to the batch server that currently owns each job specified by jobId.

A job in the Running, Transiting, or Exiting state cannot be moved.

Operands

The first operand, the new destination, is one of the following:

queue
@server
queue@server

If the destination operand describes only a queue, then qmove will move jobs into the queue of the specified name at the job's current server. If the destination operand describes only a batch server, then qmove will move jobs into the default queue at that batch server. If the destination operand describes both a queue and a batch server, then qmove will move the jobs into the specified queue at the specified server.

All following operands are job ids which specify the jobs to be moved to the new destination. The qmove command accepts one or more job ids of the form: sequenceNumber [.serverName] [@server]

Standard error

The qmove command will write a diagnostic message to standard error for each error occurrence.

Exit status

Upon successful processing of all the operands presented to the qmove command, the exit status will be a value of zero.

If the qmove command fails to process any operand, the command exits with a value greater than zero.

Related topics

- qsub on page 204

Related topics (non-Adaptive Computing topics)

- pbs_movejob(3B)

qorder

Exchange order of two PBS batch jobs in any queue.

Synopsis

qorder job1_identifier job2_identifier

Description

To order two jobs is to exchange the jobs' positions in the queue(s) in which the jobs reside. The two jobs must be located on the same server. No attribute of the job, such as priority, is changed. The impact of changing the order in the queue(s) is dependent on local job schedule policy. For information about your local job schedule policy, contact your systems administrator.

ℹ️ A job in the running state cannot be reordered.
Operands

Both operands are job_identifiers that specify the jobs to be exchanged. The qorder command accepts two job_identifier operands of the following form:

sequence_number[.server_name][@server]

The two jobs must be in the same location, so the server specification for the two jobs must agree.

Standard error

The qorder command will write diagnostic messages to standard error for each error occurrence.

Exit status

Upon successful processing of all the operands presented to the qorder command, the exit status will be a value of zero.

If the qorder command fails to process any operand, the command exits with a value greater than zero.

Related topics

- qsub on page 204
- qmove on page 189

Related topics (non-Adaptive Computing topics)

- pbs_orderjob(3B)
- pbs_movejob(3B)

qrerun

(Rerun a batch job)

Synopsis

qrerun [{-f}] <JOBID>[ <JOBID>] ...

Description

The qrerun command directs that the specified jobs are to be rerun if possible. To rerun a job is to terminate the session leader of the job and return the job to the queued state in the execution queue in which the job currently resides.

If a job is marked as not rerunnable then the rerun request will fail for that job. If the mini-server running the job is down, or it rejects the request, the Rerun Job batch request will return a failure unless -f is used.

Using -f violates IEEE Batch Processing Services Standard and should be handled with great care. It should only be used under exceptional circumstances. The best practice is to fix the problem mini-server host and let qrerun run normally. The nodes may need manual cleaning (see the -r option on the qsub and qalter commands).
Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-f</td>
<td>Force a rerun on a job</td>
</tr>
</tbody>
</table>

qrerun -f 15406

Operands

The qrerun command accepts one or more job_identifier operands of the form:

sequence_number[.server_name][@server]

Standard error

The qrerun command will write a diagnostic message to standard error for each error occurrence.

Exit status

Upon successful processing of all the operands presented to the qrerun command, the exit status will be a value of zero.

If the qrerun command fails to process any operand, the command exits with a value greater than zero.

Examples

> qrerun 3233

(Job 3233 will be re-run.)

Related topics

- qsub(1B)
- qalter(1B)

Non-Adaptive Computing topics

- pbs_alterjob(3B)
- pbs_rerunjob(3B)

qrls

(Release hold on PBS batch jobs)

Synopsis

qrls [{-h <HOLD LIST>| -t <array_range>}] <JOBID> [ <JOBID> ] ...
Description

The `qrls` command removes or releases holds which exist on batch jobs.

A job may have one or more types of holds which make the job ineligible for execution. The types of holds are USER, OTHER, and SYSTEM. The different types of holds may require that the user issuing the `qrls` command have special privileges. A user may always remove a USER hold on their own jobs, but only privileged users can remove OTHER or SYSTEM holds. An attempt to release a hold for which the user does not have the correct privilege is an error and no holds will be released for that job.

If no `-h` option is specified, the USER hold will be released.

If the job has no execution time pending, the job will change to the queued state. If an execution time is still pending, the job will change to the waiting state.

Options

<table>
<thead>
<tr>
<th>Command</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
</table>
| `-h`    | hold_list  | Defines the types of hold to be released from the jobs. The `hold_list` option argument is a string consisting of one or more of the letters "u", "o", and "s" in any combination. The hold type associated with each letter is:  
  - `u` – USER  
  - `o` – OTHER  
  - `s` – SYSTEM |
| `-t`    | array_range| The `array_range` argument is an integer id or a range of integers. Multiple ids or id ranges can be combined in a comma delimited list. Examples: `-t 1-100 or -t 1,10,50-100`  
  If an array range isn’t specified, the command tries to operate on the entire array. The command acts on the array (or specified range of the array) just as it would on an individual job. |

Operands

The `qrls` command accepts one or more `job_identifier` operands of the form:

`sequence_number[.server_name][@server]`

Examples

```
$ qrls -h u 3233 release user hold on job 3233
```

Standard error

The `qrls` command will write a diagnostic message to standard error for each error occurrence.
Exit status

Upon successful processing of all the operands presented to the qrls command, the exit status will be a value of zero.

If the qrls command fails to process any operand, the command exits with a value greater than zero.

Related topics

- qsub(1B)
- qalter(1B)
- qhold(1B)

Non-Adaptive Computing topics)

- pbs_alterjob(3B)
- pbs_holdjob(3B)
- pbs_rlsjob(3B)

qrun

(Run a batch job)

Synopsis

qrun [{-H <HOST>|-a} [<JOBID>][ <JOBID>] ...]

Overview

The qrun command runs a job.

Format

- **-H**

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt; Host Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>--</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies the host within the cluster on which the job(s) are to be run. The host argument is the name of a host that is a member of the cluster of hosts managed by the server. If the option is not specified, the server will select the &quot;worst possible&quot; host on which to execute the job.</td>
</tr>
<tr>
<td>Example</td>
<td><code>qrun -H hostname 15406</code></td>
</tr>
</tbody>
</table>
Command details

The `qrun` command is used to force a batch server to initiate the execution of a batch job. The job is run regardless of scheduling position or resource requirements.

In order to execute `qrun`, the user must have PBS Operation or Manager privileges.

Examples

```
> qrun 3233
```
(Run job 3233.)

### qsig

(Signal a job)

**Synopsis**

```
qsig [{-s <SIGNAL>}] <JOBID>[ <JOBID>] ...
[-a]
```

**Description**

The `qsig` command requests that a signal be sent to executing batch jobs. The signal is sent to the session leader of the job. If the `-s` option is not specified, SIGTERM is sent. The request to signal a batch job will be rejected if:

- The user is not authorized to signal the job.
- The job is not in the running state.
- The requested signal is not supported by the system upon which the job is executing.

The `qsig` command sends a Signal Job batch request to the server which owns the job.
Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-s</td>
<td>signal</td>
<td>Declares which signal is sent to the job. The signal argument is either a signal name, e.g. SIGKILL, the signal name without the SIG prefix, e.g. KILL, or an unsigned signal number, e.g. 9. The signal name SIGNULL is allowed; the server will send the signal 0 to the job which will have no effect on the job, but will cause an obituary to be sent if the job is no longer executing. Not all signal names will be recognized by qsig. If it doesn’t recognize the signal name, try issuing the signal number instead. Two special signal names, &quot;suspend&quot; and &quot;resume&quot;, are used to suspend and resume jobs. Cray systems use the Cray-specific suspend() / resume() calls. On non-Cray system, suspend causes a SIGTSTP to be sent to all processes in the job's top task, wait 5 seconds, and then send a SIGSTOP to all processes in all tasks on all nodes in the job. This differs from TORQUE 2.0.0 which did not have the ability to propagate signals to sister nodes. Resume sends a SIGCONT to all processes in all tasks on all nodes. When suspended, a job continues to occupy system resources but is not executing and is not charged for walltime. The job will be listed in the &quot;S&quot; state. Manager or operator privilege is required to suspend or resume a job.</td>
</tr>
<tr>
<td>-a</td>
<td>asynchronously</td>
<td>Makes the command run asynchronously.</td>
</tr>
</tbody>
</table>

Operands

The qsig command accepts one or more job_identifier operands of the form:

sequence_number[.server_name][@server]

Examples

```
> qsig -s SIGKILL 3233   send a SIGKILL to job 3233
> qsig -s KILL 3233     send a SIGKILL to job 3233
> qsig -s 9 3233        send a SIGKILL to job 3233
```

Standard error

The qsig command will write a diagnostic message to standard error for each error occurrence.

Exit status

Upon successful processing of all the operands presented to the qsig command, the exit status will be a value of zero.

If the qsig command fails to process any operand, the command exits with a value greater than zero.
Related topics

- `qsub(1B)`

Non-Adaptive Computing topics

- `pbs_sigjob(3B)`
- `pbs_resources_* (7B)` where * is system type
- PBS ERS

### qstat

Show status of PBS batch jobs.

**Synopsis**

```
qstat [-c on page 198] [-f [ -l]] [-W site_specific] [job_identifier... | destination...] [time]
[job_identifier... | destination...]
qstat -Q [-f [ -l]] [-c on page 198] [-W site_specific] [destination...]
qstat -q [-c on page 198] [-G|-M] [destination...]
qstat -B [-c on page 198] [-f [ -l]] [-W site_specific] [server_name...]
qstat -t [-c on page 198]
```

**Description**

The `qstat` command is used to request the status of jobs, queues, or a batch server. The requested status is written to standard out.

When requesting job status, synopsis format 1 or 2, `qstat` will output information about each job_identifier or all jobs at each destination. Jobs for which the user does not have status privilege are not displayed.

When requesting queue or server status, synopsis format 3 through 5, `qstat` will output information about each destination.

> You can configure TORQUE with `CFLAGS='DTXT'` to change the alignment of text in `qstat` output. This noticeably improves `qstat -r` output.
<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-c</td>
<td>Completed jobs are not displayed in the output. If desired, you can set the <code>PBS_QSTAT_NO_COMPLETE</code> environment variable to cause all <code>qstat</code> requests to not show completed jobs by default.</td>
</tr>
<tr>
<td>-f</td>
<td>Specifies that a full status display be written to standard out. The [time] value is the amount of wall-time, in seconds, remaining for the job. [time] does not account for walltime multipliers.</td>
</tr>
<tr>
<td>-a</td>
<td>All jobs are displayed in the alternative format (see Standard output on page 200). If the operand is a destination id, all jobs at that destination are displayed. If the operand is a job id, information about that job is displayed.</td>
</tr>
<tr>
<td>-e</td>
<td>If the operand is a job id or not specified, only jobs in executable queues are displayed. Setting the <code>PBS_QSTAT_EXECONLY</code> environment variable will also enable this option.</td>
</tr>
<tr>
<td>-i</td>
<td>Job status is displayed in the alternative format. For a destination id operand, statuses for jobs at that destination which are not running are displayed. This includes jobs which are queued, held or waiting. If an operand is a job id, status for that job is displayed regardless of its state.</td>
</tr>
<tr>
<td>-r</td>
<td>If an operand is a job id, status for that job is displayed. For a destination id operand, statuses for jobs at that destination which are running are displayed, this includes jobs which are suspended.</td>
</tr>
<tr>
<td>-n</td>
<td>In addition to the basic information, nodes allocated to a job are listed.</td>
</tr>
<tr>
<td>-l</td>
<td>In combination with -n, the -l option puts all of the nodes on the same line as the job ID. In combination with -f, attributes are not folded to fit in a terminal window. This is intended to ease the parsing of the <code>qstat</code> output.</td>
</tr>
<tr>
<td>-s</td>
<td>In addition to the basic information, any comment provided by the batch administrator or scheduler is shown.</td>
</tr>
<tr>
<td>-G</td>
<td>Show size information in giga-bytes.</td>
</tr>
<tr>
<td>-M</td>
<td>Show size information, disk or memory in mega-words. A word is considered to be 8 bytes.</td>
</tr>
<tr>
<td>-R</td>
<td>In addition to other information, disk reservation information is shown. Not applicable to all systems.</td>
</tr>
<tr>
<td>Option</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
</tbody>
</table>
| -t     | Normal `qstat` output displays a summary of the array instead of the entire array. Job for job. `qstat -t` expands the output to display the entire array. Note that arrays are now named with brackets following the array name; for example:  
    `dbeer@napali:~/.dev/torque/array_changes$ echo sleep 20 | qsub -t 0-299 189
    [0].napali
    Individual jobs in the array are now also noted using square brackets instead of dashes; for example, here is part of the output of `qstat -t` for the preceding array:  
    189[299].napali STDIN[299] dbeer 0 Q batch |
| -u     | Job status is displayed in the alternative format. If an operand is a job id, status for that job is displayed. For a destination id operand, statuses for jobs at that destination which are owned by the user(s) listed in `user_list` are displayed. The syntax of the `user_list` is:  
    `user_name[@host][,user_name[@host],...]`
    Host names may be wildcarded on the left end, e.g. "*.nasa.gov". `User_name` without a "@host" is equivalent to "user_name@*", that is at any host. |
| -Q     | Specifies that the request is for queue status and that the operands are destination identifiers. |
| -q     | Specifies that the request is for queue status which should be shown in the alternative format. |
| -B     | Specifies that the request is for batch server status and that the operands are the names of servers. |

**Operands**

If neither the `-Q` nor the `-B` option is given, the operands on the `qstat` command must be either job identifiers or destination identifiers.

If the operand is a job identifier, it must be in the following form:

```
sequence_number[.server_name]@[server]
```

where `sequence_number.server_name` is the job identifier assigned at submittal time (see `qsub`). If the `.server_name` is omitted, the name of the default server will be used. If `@server` is supplied, the request will be for the job identifier currently at that server.

If the operand is a destination identifier, it is one of the following three forms:

- `queue`
- `@server`
- `queue@server`

If `queue` is specified, the request is for status of all jobs in that queue at the default server. If the `@server` form is given, the request is for status of all jobs at that server. If a full destination identifier, `queue@server`, is given, the request is for status of all jobs in the named queue at the named server.
If the -Q option is given, the operands are destination identifiers as specified above. If queue is specified, the status of that queue at the default server will be given. If queue@server is specified, the status of the named queue at the named server will be given. If @server is specified, the status of all queues at the named server will be given. If no destination is specified, the status of all queues at the default server will be given.

If the -B option is given, the operand is the name of a server.

Standard output

Displaying job status

If job status is being displayed in the default format and the -f option is not specified, the following items are displayed on a single line, in the specified order, separated by white space:

- the job identifier assigned by PBS.
- the job name given by the submitter.
- the job owner.
- the CPU time used.
- the job state:

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>Job is completed after having run.</td>
</tr>
<tr>
<td>E</td>
<td>Job is exiting after having run.</td>
</tr>
<tr>
<td>H</td>
<td>Job is held.</td>
</tr>
<tr>
<td>Q</td>
<td>Job is queued, eligible to run or routed.</td>
</tr>
<tr>
<td>R</td>
<td>Job is running.</td>
</tr>
<tr>
<td>T</td>
<td>Job is being moved to new location.</td>
</tr>
<tr>
<td>W</td>
<td>Job is waiting for its execution time (-a option) to be reached.</td>
</tr>
<tr>
<td>S</td>
<td>(Unicos only) Job is suspended.</td>
</tr>
</tbody>
</table>

- the queue in which the job resides.

If job status is being displayed and the -f option is specified, the output will depend on whether qstat was compiled to use a Tcl interpreter. See Configuration on page 203 for details. If Tcl is not being used, full display for each job consists of the header line:

Job Id: job identifier
Followed by one line per job attribute of the form:

```
attribute_name = value
```

If any of the options `-a`, `-i`, `-r`, `-u`, `-n`, `-s`, `-G`, or `-M` are provided, the alternative display format for jobs is used. The following items are displayed on a single line, in the specified order, separated by white space:

- the job identifier assigned by PBS
- the job owner
- the queue in which the job currently resides
- the job name given by the submitter
- the session id (if the job is running)
- the number of nodes requested by the job
- the number of cpus or tasks requested by the job
- the amount of memory requested by the job
- either the cpu time, if specified, or wall time requested by the job, (hh:mm)
- the jobs current state
- the amount of cpu time or wall time used by the job (hh:mm)

If the `-r` option is provided, the line contains:

- the job identifier assigned by PBS
- the job owner
- the queue in which the job currently resides
- the number of nodes requested by the job
- the number of cpus or tasks requested by the job
- the amount of memory requested by the job
- either the cpu time or wall time requested by the job
- the jobs current state
- the amount of cpu time or wall time used by the job
- the amount of SRFS space requested on the big file system
- the amount of SRFS space requested on the fast file system
- the amount of space requested on the parallel I/O file system

The last three fields may not contain useful information at all sites or on all systems.

**Displaying queue status**

If queue status is being displayed and the `-f` option was not specified, the following items are displayed on a single line, in the specified order, separated by white space:
- the queue name
- the maximum number of jobs that may be run in the queue concurrently
- the total number of jobs in the queue
- the enable or disabled status of the queue
- the started or stopped status of the queue
- for each job state, the name of the state and the number of jobs in the queue in that state
- the type of queue, execution or routing

If queue status is being displayed and the `-f` option is specified, the output will depend on whether `qstat` was compiled to use a Tcl interpreter. See the configuration section for details. If Tcl is not being used, the full display for each queue consists of the header line:

```
Queue: queue_name
```

Followed by one line per queue attribute of the form:

```
attribute_name = value
```

If the `-Q` option is specified, queue information is displayed in the alternative format: The following information is displayed on a single line:

- the queue name
- the maximum amount of memory a job in the queue may request
- the maximum amount of cpu time a job in the queue may request
- the maximum amount of wall time a job in the queue may request
- the maximum amount of nodes a job in the queue may request
- the number of jobs in the queue in the running state
- the number of jobs in the queue in the queued state
- the maximum number (limit) of jobs that may be run in the queue concurrently
- the state of the queue given by a pair of letters:
  - either the letter `E` if the queue is Enabled or `D` if Disabled
  - and
  - either the letter `R` if the queue is Running (started) or `S` if Stopped.

**Displaying server status**

If batch server status is being displayed and the `-f` option is not specified, the following items are displayed on a single line, in the specified order, separated by white space:

- the server name
- the maximum number of jobs that the server may run concurrently
- the total number of jobs currently managed by the server
• the status of the server
• for each job state, the name of the state and the number of jobs in the server in that state

If server status is being displayed and the -f option is specified, the output will depend on whether qstat was compiled to use a Tcl interpreter. See the configuration section for details. If Tcl is not being used, the full display for the server consists of the header line:

Server: server name

Followed by one line per server attribute of the form:

attribute_name = value

Standard error

The qstat command will write a diagnostic message to standard error for each error occurrence.

Configuration

If qstat is compiled with an option to include a Tcl interpreter, using the -f flag to get a full display causes a check to be made for a script file to use to output the requested information. The first location checked is $HOME/.qstatrc. If this does not exist, the next location checked is administrator configured. If one of these is found, a Tcl interpreter is started and the script file is passed to it along with three global variables. The command line arguments are split into two variable named flags and operands. The status information is passed in a variable named objects. All of these variables are Tcl lists. The flags list contains the name of the command (usually "qstat") as its first element. Any other elements are command line option flags with any options they use, presented in the order given on the command line. They are broken up individually so that if two flags are given together on the command line, they are separated in the list. For example, if the user typed:

qstat -QfWbigdisplay

the flags list would contain

qstat -Q -f -W bigdisplay

The operands list contains all other command line arguments following the flags. There will always be at least one element in operands because if no operands are typed by the user, the default destination or server name is used. The objects list contains all the information retrieved from the server(s) so the Tcl interpreter can run once to format the entire output. This list has the same number of elements as the operands list. Each element is another list with two elements.

The first element is a string giving the type of objects to be found in the second. The string can take the values "server", "queue", "job" or "error".

The second element will be a list in which each element is a single batch status object of the type given by the string discussed above. In the case of "error", the list will be empty. Each object is again a list. The first element is the name of the object. The second is a list of attributes.

The third element will be the object text.

All three of these object elements correspond with fields in the structure batch_status which is described in detail for each type of object by the man pages for pbs_statjob(3), pbs_statsqueue(3), and pbs_statserver (3). Each attribute in the second element list whose elements correspond with the attrlist structure. Each
will be a list with two elements. The first will be the attribute name and the second will be the attribute value.

Exit status

Upon successful processing of all the operands presented to the `qstat` command, the exit status will be a value of zero.

If the `qstat` command fails to process any operand, the command exits with a value greater than zero.

Related topics

- `qalter`(1B)
- `qsub`(1B)

Non-Adaptive Computing topics

- `pbs_alterjob`(3B)
- `pbs_statjob`(3B)
- `pbs_statqueue`(3B)
- `pbs_statserver`(3B)
- `pbs_submit`(3B)
- `pbs_job_attributes`(7B)
- `pbs_queue_attributes`(7B)
- `pbs_server_attributes`(7B)
- `qmgr query_other_jobs` parameter (allow non-admin users to see other users’ jobs)
- `pbs_resources_*`(7B) where * is system type
- PBS ERS

qsub

Submit PBS job.

Synopsis

```
```

Description

To create a job is to submit an executable script to a batch server. The batch server will be the default server unless the `-q` option is specified. The command parses a script prior to the actual script execution;
it does not execute a script itself. All script-writing rules remain in effect, including the "#!" at the head of the file (see discussion of PBS_DEFAULT under Environment variables on page 220). Typically, the script is a shell script which will be executed by a command shell such as sh or csh.

Options on the qsub command allow the specification of attributes which affect the behavior of the job. The qsub command will pass certain environment variables in the Variable_List attribute of the job. These variables will be available to the job. The value for the following variables will be taken from the environment of the qsub command: HOME, LANG, LOGNAME, PATH, MAIL, SHELL, and TZ. These values will be assigned to a new name which is the current name prefixed with the string "PBS_O_". For example, the job will have access to an environment variable named PBS_O_HOME which have the value of the variable HOME in the qsub command environment.

In addition to the above, the following environment variables will be available to the batch job:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBS_O_HOST</td>
<td>The name of the host upon which the qsub command is running.</td>
</tr>
<tr>
<td>PBS_SERVER</td>
<td>The hostname of the pbs_server which qsub submits the job to.</td>
</tr>
<tr>
<td>PBS_O_QUEUE</td>
<td>The name of the original queue to which the job was submitted.</td>
</tr>
<tr>
<td>PBS_O_WORKDIR</td>
<td>The absolute path of the current working directory of the qsub command.</td>
</tr>
<tr>
<td>PBS_ARRAYID</td>
<td>Each member of a job array is assigned a unique identifier (see -l option).</td>
</tr>
<tr>
<td>PBS_ENVIRONMENT</td>
<td>Set to PBS_BATCH to indicate the job is a batch job, or to PBS_INTERACTIVE to indicate the job is a PBS interactive job (see -I option).</td>
</tr>
<tr>
<td>PBS_GPUFILE</td>
<td>The name of the file containing the list of assigned GPUs. For more information about how to set up TORQUE with GPUs, see the Moab Workload Manager accelerators documentation.</td>
</tr>
<tr>
<td>PBS_JOBID</td>
<td>The job identifier assigned to the job by the batch system. It can be used in the stdout and stderr paths. TORQUE replaces $PBS_JOBID with the job's jobid (for example, #PBS -o /tmp/$PBS_JOBID.output).</td>
</tr>
<tr>
<td>PBS_JOBNAME</td>
<td>The job name supplied by the user.</td>
</tr>
<tr>
<td>PBS_NODEFILE</td>
<td>The name of the file contains the list of nodes assigned to the job (for parallel and cluster systems).</td>
</tr>
<tr>
<td>PBS_QUEUE</td>
<td>The name of the queue from which the job is executed.</td>
</tr>
</tbody>
</table>
Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
</table>
| -a     | date_time    | Declares the time after which the job is eligible for execution. The date_time argument is in the form:  
|        |              | [[[[CC]YY]MM]DD]hhmm[.SS] where CC is the first two digits of the year (the century), YY is the second two digits of the year, MM is the two digits for the month, DD is the day of the month, hh is the hour, mm is the minute, and the optional SS is the seconds.  
|        |              | If the month (MM) is not specified, it will default to the current month if the specified day (DD) is in the future. Otherwise, the month will be set to next month. Likewise, if the day (DD) is not specified, it will default to today if the time (hhmm) is in the future. Otherwise, the day will be set to tomorrow.  
|        |              | For example, if you submit a job at 11:15 am with a time of -a 1110, the job will be eligible to run at 11:10 am tomorrow. |
| -A     | account_string | Defines the account string associated with the job. The account_string is an undefined string of characters and is interpreted by the server which executes the job. See section 2.7.1 of the PBS ERS. |
| -b     | seconds      | Defines the maximum number of seconds qsub will block attempting to contact pbs_server. If pbs_server is down, or for a variety of communication failures, qsub will continually retry connecting to pbs_server for job submission.  
<p>|        |              | This value overrides the CLIENTRETRY parameter in torque.cfg. This is a non-portable TORQUE extension. Portability-minded users can use the PBS_CLIENTRETRY environmental variable. A negative value is interpreted as infinity. The default is 0. |</p>
<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
</table>
| -c     | checkpoint_options    | Defines the options that will apply to the job. If the job executes upon a host which does not support checkpoint, these options will be ignored. Valid checkpoint options are:  
  - *none* – No checkpointing is to be performed.  
  - *enabled* – Specify that checkpointing is allowed but must be explicitly invoked by either the `qhold` or `qchkpt` commands.  
  - *shutdown* – Specify that checkpointing is to be done on a job at `pbs_mom` shutdown.  
  - *periodic* – Specify that periodic checkpointing is enabled. The default interval is 10 minutes and can be changed by the `$checkpoint_interval` option in the MOM config file or by specifying an interval when the job is submitted  
  - *interval=minutes* – Checkpointing is to be performed at an interval of minutes, which is the integer number of minutes of wall time used by the job. This value must be greater than zero.  
  - *depth=number* – Specify a number (depth) of checkpoint images to be kept in the checkpoint directory.  
  - *dir=path* – Specify a checkpoint directory (default is `/var/spool/torque/checkpoint`). |
<p>| -C     | directive_prefix      | Defines the prefix that declares a directive to the <code>qsub</code> command within the script file. (See the paragraph on script directives under <em>Extended description</em> on page 221.) If the <code>-C</code> option is presented with a <code>directive_prefix</code> argument that is the null string, <code>qsub</code> will not scan the script file for directives. |
| -d     | path                  | Defines the working directory path to be used for the job. If the <code>-d</code> option is not specified, the default working directory is the home directory. This option sets the environment variable <code>PBS_O_INITDIR</code>. |
| -D     | path                  | Defines the root directory to be used for the job. This option sets the environment variable <code>PBS_O_ROOTDIR</code>. |</p>
<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
</table>
| -e     | path | Defines the path to be used for the standard error stream of the batch job. The path argument is of the form:  
\[hostname:]path_name  
where hostname is the name of a host to which the file will be returned, and path_name is the path name on that host in the syntax recognized by POSIX.  

**Note:** When specifying a directory for the location, you need to include a trailing slash.  
The argument will be interpreted as follows:  
- *path_name* – where path_name is not an absolute path name, then the qsub command will expand the path name relative to the current working directory of the command. The command will supply the name of the host upon which it is executing for the hostname component.  
- *hostname:path_name* – where path_name is not an absolute path name, then the qsub command will not expand the path name relative to the current working directory of the command. On delivery of the standard error, the path name will be expanded relative to the user’s home directory on the hostname system.  
- *path_name* – where path_name specifies an absolute path name, then the qsub command will supply the name of the host on which it is executing for the hostname.  
- *hostname:path_name* – where path_name specifies an absolute path name, the path will be used as specified.  
If the -e option is not specified, the default file name for the standard error stream will be used. The default name has the following form:  
- *job_name:sequence_number* – where job_name is the name of the job (see the -n name option) and sequence_number is the job number assigned when the job is submitted.  

| -f     | ---  | Job is made fault tolerant. Jobs running on multiple nodes are periodically polled by mother superior. If one of the nodes fails to report, the job is canceled by mother superior and a failure is reported. If a job is fault tolerant, it will not be canceled based on failed polling (no matter how many nodes fail to report). This may be desirable if transient network failures are causing large jobs not to complete, where ignoring one failed polling attempt can be corrected at the next polling attempt.  

**Note:** If TORQUE is compiled with PBS_NO_POSIX_VIOLATION (there is no config option for this), you have to use -W fault_tolerant=true to mark the job as fault tolerant.
<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
</table>
| `-F`  |      | Specifies the arguments that will be passed to the job script when the script is launched. The accepted syntax is:  
  
  qsub `-F "myarg1 myarg2 myarg3=myarg3value" myscript2.sh`  
  
  Quotation marks are required. `qsub` will fail with an error message if the argument following `-F` is not a quoted value. The pbs_mom server will pass the quoted value as arguments to the job script when it launches the script. |
| `-h`  |      | Specifies that a user hold be applied to the job at submission time. |
| `-I`  |      | Declares that the job is to be run "interactively". The job will be queued and scheduled as any PBS batch job, but when executed, the standard input, output, and error streams of the job are connected through `qsub` to the terminal session in which `qsub` is running. Interactive jobs are forced to not rerunable. See [Extended description on page 221](#) for additional information of interactive jobs. |
| `-j`  | join | Declares if the standard error stream of the job will be merged with the standard output stream of the job.  
  
  An option argument value of `oe` directs that the two streams will be merged, intermixed, as standard output. An option argument value of `eo` directs that the two streams will be merged, intermixed, as standard error.  
  
  If the join argument is `n` or the option is not specified, the two streams will be two separate files.  
  
  If using either the `-e` or the `-o` option and the `-j eo|oe` option, the `-j` option takes precedence and all standard error and output messages go to the chosen output file. |
### Option `-k` - keep

**Description**

Defines which (if either) of standard output or standard error will be retained on the execution host. If set for a stream, this option overrides the path name for that stream. If not set, neither stream is retained on the execution host.

The argument is either the single letter "e" or "o", or the letters "e" and "o" combined in either order. Or the argument is the letter "n".

- **e** - The standard error stream is to be retained on the execution host. The stream will be placed in the home directory of the user under whose user id the job executed. The file name will be the default file name given by:

  \[
  \text{job\_name.osequence}
  \]

  where `job_name` is the name specified for the job, and `sequence` is the sequence number component of the job identifier.

- **o** - The standard output stream is to be retained on the execution host. The stream will be placed in the home directory of the user under whose user id the job executed. The file name will be the default file name given by:

  \[
  \text{job\_name.osequence}
  \]

  where `job_name` is the name specified for the job, and `sequence` is the sequence number component of the job identifier.

- **eo** - Both the standard output and standard error streams will be retained.

- **oe** - Both the standard output and standard error streams will be retained.

- **n** - Neither stream is retained.

### Option `-l` - resource_list

**Description**

Defines the resources that are required by the job and establishes a limit to the amount of resource that can be consumed. If not set for a generally available resource, such as CPU time, the limit is infinite. The `resource_list` argument is of the form:

\[
\text{resource\_name [=value]}[,\text{resource\_name [=value]}]...\]

In this situation, you should request the more inclusive resource first. For example, a request for procs should come before a gres request.

In TORQUE 3.0.2 or later, `qsub` supports the mapping of `-l gpus=X` to `-l gres=gpus:X`. This allows users who are using NUMA systems to make requests such as `-l ncpus=20:gpus=5` indicating they are not concerned with the GPUs in relation to the NUMA nodes they request, they only want a total of 20 cores and 5 GPUs.

For more information, see Requesting resources on page 42.

For information on specifying multiple types of resources for allocation, see "Multi-Req Support" under "General Job Policies" in the Moab Workload Manager documentation.
<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
</table>
| -m     | mail_options       | Defines the set of conditions under which the execution server will send a mail message about the job. The mail_options argument is a string which consists of either the single character "n", or one or more of the characters "a", "b", and "e". If the character "n" is specified, no normal mail is sent. Mail for job cancels and other events outside of normal job processing are still sent. For the letters "a", "b", and "e":  
  - a – Mail is sent when the job is aborted by the batch system.  
  - b – Mail is sent when the job begins execution.  
  - e – Mail is sent when the job terminates.  
If the -m option is not specified, mail will be sent if the job is aborted. |
| -M     | user_list          | Declares the list of users to whom mail is sent by the execution server when it sends mail about the job. The user_list argument is of the form:  
  user[@host],[user[@host],...]  
If unset, the list defaults to the submitting user at the qsub host, i.e. the job owner. |
<p>| -n     | node-exclusive     | Allows a user to specify an exclusive-node access/allocation request for the job. This affects only cpusets and compatible schedulers (see Linux cpuset support on page 72). |
| -N     | name               | Declares a name for the job. The name specified may be an unlimited number of characters in length. It must consist of printable, nonwhite space characters with the first character alphabetic. If the -N option is not specified, the job name will be the base name of the job script file specified on the command line. If no script file name was specified and the script was read from the standard input, then the job name will be set to STDIN. |</p>
<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
</table>
| -o     | path     | Defines the path to be used for the standard output stream of the batch job. The path argument is of the form:  
|        |          | [hostname:]path_name  
|        |          | where hostname is the name of a host to which the file will be returned, and path_name is the path name on that host in the syntax recognized by POSIX.  
|        |          | **i** When specifying a directory for the location you need to include a trailing slash. |
|        |          | The argument will be interpreted as follows:  
|        |          | - *path_name* – where *path_name* is not an absolute path name, then the qsub command will expand the path name relative to the current working directory of the command. The command will supply the name of the host upon which it is executing for the hostname component.  
|        |          | - *hostname:path_name* – where *path_name* is not an absolute path name, then the qsub command will not expand the path name relative to the current working directory of the command. On delivery of the standard output, the path name will be expanded relative to the user’s home directory on the hostname system.  
|        |          | - *path_name* – where *path_name* specifies an absolute path name, then the qsub command will supply the name of the host on which it is executing for the hostname.  
|        |          | - *hostname:path_name* where *path_name* specifies an absolute path name, the path will be used as specified.  
|        |          | If the -o option is not specified, the default file name for the standard output stream will be used. The default name has the following form:  
<p>|        |          | - <em>job_name</em>:<em>sequence_number</em> – where <em>job_name</em> is the name of the job (see the -n name option) and <em>sequence_number</em> is the job number assigned when the job is submitted. |
| -p     | priority | Defines the priority of the job. The priority argument must be an integer between -1024 and +1023 inclusive. The default is no priority which is equivalent to a priority of zero. |
| -P     | user [group] | Allows a root user or manager to submit a job as another user. TORQUE treats proxy jobs as though the jobs were submitted by the supplied username. This feature is available in TORQUE 2.4.7 and later, however, TORQUE 2.4.7 does not have the ability to supply the [:group] option; it is available in TORQUE 2.4.8 and later. |</p>
<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-q</td>
<td>destination</td>
<td>Defines the destination of the job. The destination names a queue, a server, or a queue at a server.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The <code>qsub</code> command will submit the script to the server defined by the destination argument. If the destination is a routing queue, the job may be routed by the server to a new destination.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If the <code>-q</code> option is not specified, the <code>qsub</code> command will submit the script to the default server. (See <code>Environment variables</code> on page 220 and the PBS ERS section 2.7.4, &quot;Default Server&quot;).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If the <code>-q</code> option is specified, it is in one of the following three forms:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- queue</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- @server</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- queue@server</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If the destination argument names a queue and does not name a server, the job will be submitted to the named queue at the default server.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If the destination argument names a server and does not name a queue, the job will be submitted to the default queue at the named server.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If the destination argument names both a queue and a server, the job will be submitted to the named queue at the named server.</td>
</tr>
<tr>
<td>-r</td>
<td>y/n</td>
<td>Declares whether the job is rerunnable (see the <code>qrerun</code> command). The option argument is a single character, either y or n.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If the argument is &quot;y&quot;, the job is rerunnable. If the argument is &quot;n&quot;, the job is not rerunnable. The default value is y, rerunnable.</td>
</tr>
<tr>
<td>-S</td>
<td>path_list</td>
<td>Declares the path to the desires shell for this job.</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>qsub script.sh -S /bin/tcsh</code></td>
</tr>
<tr>
<td></td>
<td></td>
<td>If the shell path is different on different compute nodes, use the following syntax:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>path[@host],[path[@host]],...</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>qsub script.sh -S /bin/tcsh@node1,/usr/bin/tcsh@node2</code></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Only one path may be specified for any host named. Only one path may be specified without the corresponding host name. The path selected will be the one with the host name that matched the name of the execution host. If no matching host is found, then the path specified without a host will be selected, if present.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If the <code>-S</code> option is not specified, the option argument is the null string or no entry from the <code>path_list</code> is selected, the execution will use the user’s login shell on the execution host.</td>
</tr>
<tr>
<td>Option</td>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| -t     | array_request     | Specifies the task ids of a job array. Single task arrays are allowed. The array_request argument is an integer id or a range of integers. Multiple ids or id ranges can be combined in a comma delimited list. Examples: `-t 1-100` or `-t 1,10,50-100`  
An optional slot limit can be specified to limit the amount of jobs that can run concurrently in the job array. The default value is unlimited. The slot limit must be the last thing specified in the array_request and is delimited from the array by a percent sign (%).  
`qsub script.sh -t 0-299%5`  
This sets the slot limit to 5. Only 5 jobs from this array can run at the same time. You can use `qalter` to modify slot limits on an array. The server parameter `max_slot_limit` can be used to set a global slot limit policy. |
| -u     | user_list         | Defines the user name under which the job is to run on the execution system. The user_list argument is of the form:  
`user[@host],[user[@host],...].`  
Only one user name may be given per specified host. Only one of the user specifications may be supplied without the corresponding host specification. That user name will used for execution on any host not named in the argument list. If unset, the user list defaults to the user who is running `qsub`. |
<p>| -v     | variable_list     | Expands the list of environment variables that are exported to the job. In addition to the variables described in the &quot;Description&quot; section above, variable_list names environment variables from the <code>qsub</code> command environment which are made available to the job when it executes. The variable_list is a comma separated list of strings of the form <code>variable</code> or <code>variable=value</code>. These variables and their values are passed to the job. Note that <code>-v</code> has a higher precedence than <code>-V</code>, so identically named variables specified via <code>-v</code> will provide the final value for an environment variable in the job. |
| -V     | ---               | Declares that all environment variables in the <code>qsub</code> commands environment are to be exported to the batch job. |</p>
<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-W</td>
<td>additional_attributes</td>
<td>The -W option allows for the specification of additional job attributes. The general syntax of -W is in the form: -W attr_name=attr_value. You can use multiple -W options with this syntax: -W attr_name1=attr_value1 -W attr_name2=attr_value2.</td>
</tr>
</tbody>
</table>

If white space occurs anywhere within the option argument string or the equal sign, "=", occurs within an attribute_value string, then the string must be enclosed with either single or double quote marks.

PBS currently supports the following attributes within the -W option:

- **depend=dependency_list** - Defines the dependency between this and other jobs. The dependency_list is in the form:
  
  `type[:argument[:argument...][,type:argument...]]`

  The argument is either a numeric count or a PBS job id according to type. If argument is a count, it must be greater than 0. If it is a job id and not fully specified in the form `seq_number.server.name`, it will be expanded according to the default server rules which apply to job IDs on most commands. If argument is null (the preceding colon need not be specified), the dependency of the corresponding type is cleared (unset). For more information, see `depend=dependency_list valid dependencies` on page 216.

- **group_list=g_list** - Defines the group name under which the job is to run on the execution system. The g_list argument is of the form:
  
  `group[@host][,group[@host],...]`

  Only one group name may be given per specified host. Only one of the group specifications may be supplied without the corresponding host specification. That group name will be used for execution on any host not named in the argument list. If not set, the group_list defaults to the primary group of the user under which the job will be run.

- **interactive=true** - If the interactive attribute is specified, the job is an interactive job. The -I option is an alternative method of specifying this attribute.

- **job_radix=<int>** - To be used with parallel jobs. It directs the Mother Superior of the job to create a distribution radix of size `<int>` between sisters. See `Managing multi-node jobs` on page 42.

- **stagein=file_list**

- **stageout=file_list** - Specifies which files are staged (copied) in before job start or staged out after the job completes execution. On completion of the job, all staged-in and staged-out files are removed from the execution system. The file_list is in the form:
  
  `local_file@hostname:remote_file[,...]`

  regardless of the direction of the copy. The name `local_file` is the name of the file on the system where the job executed. It may be an absolute path or relative to the home directory of the user. The name `remote_file` is the destination name on the host specified by hostname. The name may be
<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>absolute or relative to the user’s home directory on the destination host. The use of wildcards in the file name is not recommended. The file names map to a remote copy program (rcp) call on the execution system in the follow manner:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>○ For stagein: rcp hostname:remote_file local_file</td>
</tr>
<tr>
<td></td>
<td></td>
<td>○ For stageout: rcp local_file hostname:remote_file</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data staging examples:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-W stagein=/tmp/input.txt@headnode:/home/user/input.txt</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-W stageout=/tmp/output.txt@headnode:/home/user/output.txt</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If TORQUE has been compiled with wordexp support, then variables can be used in the specified paths. Currently only $PBS_JOBID, $HOME, and $TMPDIR are supported for stagein.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• umask=XXX – Sets umask used to create stdout and stderr spool files in pbs_mom spool directory. Values starting with 0 are treated as octal values, otherwise the value is treated as a decimal umask value.</td>
</tr>
<tr>
<td>-x</td>
<td>---</td>
<td>By default, if you submit an interactive job with a script, the script will be parsed for PBS directives but the rest of the script will be ignored since it’s an interactive job. The -x option allows the script to be executed in the interactive job and then the job completes. For example:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>script.sh</td>
</tr>
<tr>
<td></td>
<td></td>
<td>#!/bin/bash</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ls</td>
</tr>
<tr>
<td></td>
<td></td>
<td>---end script---</td>
</tr>
<tr>
<td></td>
<td></td>
<td>qsub -I script.sh</td>
</tr>
<tr>
<td></td>
<td></td>
<td>qsub: waiting for job 5.napali to start</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dbeer@napali:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;displays the contents of the directory, because of the ls command&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>qsub: job 5.napali completed</td>
</tr>
<tr>
<td>-X</td>
<td>---</td>
<td>Enables X11 forwarding. The DISPLAY environment variable must be set.</td>
</tr>
<tr>
<td>-z</td>
<td>---</td>
<td>Directs that the qsub command is not to write the job identifier assigned to the job to the commands standard output.</td>
</tr>
</tbody>
</table>

**depend=dependency_list valid dependencies**

For job dependencies to work correctly, you must set the keep_completed on page 233 server parameter.
<table>
<thead>
<tr>
<th>Dependency</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>synccount:count</td>
<td>This job is the first in a set of jobs to be executed at the same time. Count is the number of additional jobs in the set.</td>
</tr>
<tr>
<td>syncwith:jobid</td>
<td>This job is an additional member of a set of jobs to be executed at the same time. In the above and following dependency types, jobid is the job identifier of the first job in the set.</td>
</tr>
<tr>
<td>after:jobid[:jobid...]</td>
<td>This job may be scheduled for execution at any point after jobs jobid have started execution.</td>
</tr>
<tr>
<td>afterok:jobid[:jobid...]</td>
<td>This job may be scheduled for execution only after jobs jobid have terminated with no errors. See the csh warning under <a href="#">Extended description on page 221</a>.</td>
</tr>
<tr>
<td>afternotok:jobid[:jobid...]</td>
<td>This job may be scheduled for execution only after jobs jobid have terminated with errors. See the csh warning under <a href="#">Extended description on page 221</a>.</td>
</tr>
<tr>
<td>afterany:jobid[:jobid...]</td>
<td>This job may be scheduled for execution after jobs jobid have terminated, with or without errors.</td>
</tr>
<tr>
<td>on:count</td>
<td>This job may be scheduled for execution after count dependencies on other jobs have been satisfied. This form is used in conjunction with one of the &quot;before&quot; forms (see below).</td>
</tr>
<tr>
<td>before:jobid[:jobid...]</td>
<td>When this job has begun execution, then jobs jobid... may begin.</td>
</tr>
<tr>
<td>beforeok:jobid[:jobid...]</td>
<td>If this job terminates execution without errors, then jobs jobid... may begin. See the csh warning under <a href="#">Extended description on page 221</a>.</td>
</tr>
<tr>
<td>beforenotok:jobid[:jobid...]</td>
<td>If this job terminates execution with errors, then jobs jobid... may begin. See the csh warning under <a href="#">Extended description on page 221</a>.</td>
</tr>
<tr>
<td>Dependency</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>beforeany:jobid[:jobid...]</td>
<td>When this job terminates execution, jobs jobid... may begin. If any of the before forms are used, the jobs referenced by jobid must have been submitted with a dependency type of on. If any of the before forms are used, the jobs referenced by jobid must have the same owner as the job being submitted. Otherwise, the dependency is ignored.</td>
</tr>
</tbody>
</table>

**Array dependencies** make a job depend on an array or part of an array. If no count is given, then the entire array is assumed. For examples, see Dependency examples on page 219.

<table>
<thead>
<tr>
<th>afterstartarray:arrayid[count]</th>
<th>After this many jobs have started from arrayid, this job may start.</th>
</tr>
</thead>
<tbody>
<tr>
<td>afterokarray:arrayid[count]</td>
<td>This job may be scheduled for execution only after jobs in arrayid have terminated with no errors.</td>
</tr>
<tr>
<td>afternotokarray:arrayid[count]</td>
<td>This job may be scheduled for execution only after jobs in arrayid have terminated with errors.</td>
</tr>
<tr>
<td>afteranyarray:arrayid[count]</td>
<td>This job may be scheduled for execution after jobs in arrayid have terminated, with or without errors.</td>
</tr>
<tr>
<td>beforestartarray:arrayid[count]</td>
<td>Before this many jobs have started from arrayid, this job may start.</td>
</tr>
<tr>
<td>beforeokarray:arrayid[count]</td>
<td>If this job terminates execution without errors, then jobs in arrayid may begin.</td>
</tr>
<tr>
<td>beforenotokarray:arrayid[count]</td>
<td>If this job terminates execution with errors, then jobs in arrayid may begin.</td>
</tr>
<tr>
<td>beforeanyarray:arrayid[count]</td>
<td>When this job terminates execution, jobs in arrayid may begin. If any of the before forms are used, the jobs referenced by arrayid must have been submitted with a dependency type of on. If any of the before forms are used, the jobs referenced by arrayid must have the same owner as the job being submitted. Otherwise, the dependency is ignored.</td>
</tr>
</tbody>
</table>
### Dependency examples

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>qsub -W depend=afterok:123.big.iron.com /tmp/script</code></td>
<td>Error processing of the existence, state, or condition of the job on which the newly submitted job is a deferred service, i.e. the check is performed after the job is queued. If an error is detected, the new job will be deleted by the server. Mail will be sent to the job submitter stating the error.</td>
</tr>
<tr>
<td><code>qsub -W depend=before:234.hunk1.com:235.hunk1.com /tmp/script</code></td>
<td>(This assumes every job in array 427 has to finish successfully for the dependency to be satisfied.)</td>
</tr>
<tr>
<td><code>qsub_script.sh -W depend=afterokarray:427[]</code></td>
<td>(This means that 5 of the jobs in array 427 have to successfully finish in order for the dependency to be satisfied.)</td>
</tr>
</tbody>
</table>

### Operands

The `qsub` command accepts a script operand that is the path to the script of the job. If the path is relative, it will be expanded relative to the working directory of the `qsub` command.

If the script operand is not provided or the operand is the single character "-", the `qsub` command reads the script from standard input. When the script is being read from Standard Input, `qsub` will copy the file to a temporary file. This temporary file is passed to the library interface routine `pbs_submit`. The temporary file is removed by `qsub` after `pbs_submit` returns or upon the receipt of a signal which would cause `qsub` to terminate.

### Standard input

The `qsub` command reads the script for the job from standard input if the script operand is missing or is the single character "-".

### Input files

The script file is read by the `qsub` command. `qsub` acts upon any directives found in the script.

When the job is created, a copy of the script file is made and that copy cannot be modified.

### Standard output

Unless the `-z` option is set, the job identifier assigned to the job will be written to standard output if the job is successfully created.
Standard error

The `qsub` command will write a diagnostic message to standard error for each error occurrence.

Environment variables

The values of some or all of the variables in the `qsub` commands environment are exported with the job (see the `-y` and `-y` options).

The environment variable PBS_DEFAULT defines the name of the default server. Typically, it corresponds to the system name of the host on which the server is running. If PBS_DEFAULT is not set, the default is defined by an administrator established file.

The environment variable PBS_DPREFIX determines the prefix string which identifies directives in the script.

The environment variable PBS_CLIENTRETRY defines the maximum number of seconds `qsub` will block (see the `-b` option). Despite the name, currently `qsub` is the only client that supports this option.

**torque.cfg**

The `torque.cfg` file, located in PBS_SERVER_HOME (`/var/spool/torque` by default) controls the behavior of the `qsub` command. This file contains a list of parameters and values separated by whitespace.

- `QSUBSLEEP` – takes an integer operand which specifies time to sleep when running `qsub` command. Used to prevent users from overwhelming the scheduler.
- `SUBMITFILTER` – specifies the path to the submit filter used to pre-process job submission. The default path is `libexecdir/qsub_filter`, which falls back to `/usr/local/sbin/torque_submitfilter` for backwards compatibility. This `torque.cfg` parameter overrides this default.
- `SERVERHOST`
- `QSUBHOST`
- `QSUBSENDUID`
- `XAUTHPATH`
- `CLIENTRETRY`
- `VALIDATEGROUP`
- `DEFAULTCKPT`
- `VALIDATEPATH`
- `RERUNNABLEBYDEFAULT`

For example:

```
QSUBSLEEP 2
RERUNNABLEBYDEFAULT false
```
Extended description

**Script Processing:**

A job script may consist of PBS directives, comments and executable statements. A PBS directive provides a way of specifying job attributes in addition to the command line options. For example:

```plaintext
#PBS -N Job_name
#PBS -l walltime=10:30,mem=320kb
#PBS -m be
#
step1 arg1 arg2
step2 arg3 arg4
```

The `qsub` command scans the lines of the script file for directives. An initial line in the script that begins with the characters "#" or the character ":" will be ignored and scanning will start with the next line. Scanning will continue until the first executable line, that is a line that is not blank, not a directive line, nor a line whose first nonwhite space character is "#". If directives occur on subsequent lines, they will be ignored.

A line in the script file will be processed as a directive to `qsub` if and only if the string of characters starting with the first nonwhite space character on the line and of the same length as the directive prefix matches the directive prefix.

The remainder of the directive line consists of the options to `qsub` in the same syntax as they appear on the command line. The option character is to be preceded with the "-" character.

If an option is present in both a directive and on the command line, that option and its argument, if any, will be ignored in the directive. The command line takes precedence.

If an option is present in a directive and not on the command line, that option and its argument, if any, will be processed as if it had occurred on the command line.

The directive prefix string will be determined in order of preference from:

- The value of the `-c` option argument if the option is specified on the command line.
- The value of the environment variable PBS_DPREFIX if it is defined.
- The four character string #PBS.

If the `-c` option is found in a directive in the script file, it will be ignored.

**User Authorization:**

When the user submits a job from a system other than the one on which the PBS Server is running, the name under which the job is to be executed is selected according to the rules listed under the `-u` option. The user submitting the job must be authorized to run the job under the execution user name. This authorization is provided if:

- The host on which `qsub` is run is trusted by the execution host (see `/etc/hosts.equiv`).
- The execution user has an `.rhosts` file naming the submitting user on the submitting host.
C-Shell .logout File:
The following warning applies for users of the c-shell, csh. If the job is executed under the csh and a .logout file exists in the home directory in which the job executes, the exit status of the job is that of the .logout script, not the job script. This may impact any inter-job dependencies. To preserve the job exit status, either remove the .logout file or place the following line as the first line in the .logout file:

```bash
set EXITVAL = $status
```

and the following line as the last executable line in .logout:

```bash
exit $EXITVAL
```

Interactive Jobs:
If the -I option is specified on the command line or in a script directive, or if the "interactive" job attribute declared true via the -W option, -W interactive=true, either on the command line or in a script directive, the job is an interactive job. The script will be processed for directives, but will not be included with the job. When the job begins execution, all input to the job is from the terminal session in which qsub is running.

When an interactive job is submitted, the qsub command will not terminate when the job is submitted. qsub will remain running until the job terminates, is aborted, or the user interrupts qsub with an SIGINT (the control-C key). If qsub is interrupted prior to job start, it will query if the user wishes to exit. If the user response "yes", qsub exits and the job is aborted.

One the interactive job has started execution, input to and output from the job pass through qsub. Keyboard generated interrupts are passed to the job. Lines entered that begin with the tilde (~) character and contain special sequences are escaped by qsub. The recognized escape sequences are:

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>~</td>
<td>qsub terminates execution. The batch job is also terminated.</td>
</tr>
<tr>
<td>~susp</td>
<td>Suspend the qsub program if running under the C shell. &quot;susp&quot; is the suspend character (usually CNTL-Z).</td>
</tr>
<tr>
<td>~asusp</td>
<td>Suspend the input half of qsub (terminal to job), but allow output to continue to be displayed. Only works under the C shell. &quot;asusp&quot; is the auxiliary suspend character, usually CNTL-Y.</td>
</tr>
</tbody>
</table>

Exit status
Upon successful processing, the qsub exit status will be a value of zero.

If the qsub command fails, the command exits with a value greater than zero.

Related topics
- qalter(1B)
- qdel(1B)
- `qhold`(1B)
- `qrls`(1B)
- `qsig`(1B)
- `qstat`(1B)
- `pbs_server`(8B)

**Non-Adaptive Computing topics**

- `pbs_connect`(3B)
- `pbs_job_attributes`(7B)
- `pbs_queue_attributes`(7B)
- `pbs_resources_irix5`(7B)
- `pbs_resources_sp2`(7B)
- `pbs_resources_sunos4`(7B)
- `pbs_resources_unixos8`(7B)
- `pbs_server_attributes`(7B)
- `qselect`(1B)
- `qmove`(1B)
- `qmsg`(1B)
- `qrerun`(1B)

**qterm**

Terminate processing by a PBS batch server.

**Synopsis**

```
qterm [-t type] [server...]
```

**Description**

The `qterm` command terminates a batch server. When a server receives a terminate command, the server will go into the "Terminating" state. No new jobs will be allowed to be started into execution or enqueued into the server. The impact on jobs currently being run by the server depends on the state of the server.

In order to execute `qterm`, the user must have PBS Operation or Manager privileges.
Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-t</td>
<td>type</td>
<td>Specifies the type of shut down. The types are:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- <em>immediate</em> - If checkpointing is supported, all running jobs are to immediately stop execution. If checkpointing is supported, running jobs that can be checkpointed are checkpointed, terminated, and requeued. If checkpointing is not supported or the job cannot be checkpointed, running jobs are requeued if the rerunable attribute is true. Otherwise, jobs are killed.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- <em>delay</em> - If checkpointing is supported, running jobs that can be checkpointed are checkpointed, terminated, and requeued. If a job cannot be checkpointed, but can be rerun, the job is terminated and requeued. Otherwise, running jobs are allowed to continue to run.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- <em>quick</em> - This is the default action if the -t option is not specified. This option is used when you wish that running jobs be left running when the server shuts down. The server will cleanly shutdown and can be restarted when desired. Upon restart of the server, jobs that continue to run are shown as running; jobs that terminated during the server’s absence will be placed into the exiting state.</td>
</tr>
</tbody>
</table>

Operands

The server operand specifies which servers are to shut down. If no servers are given, then the default server will be terminated.

Standard error

The qterm command will write a diagnostic message to standard error for each error occurrence.

Exit status

Upon successful processing of all the operands presented to the qterm command, the exit status will be a value of zero.

If the qterm command fails to process any operand, the command exits with a value greater than zero.

Related topics (non-Adaptive Computing topics)

- pbs_server(8B)
- qmgr(8B)
- pbs_resources_aix4(7B)
- pbs_resources_irix5(7B)
- pbs_resources_sp2(7B)
- pbs_resources_sunos4(7B)
- pbs_resources_unicos8(7B)
trqauthd

(TORQUE authorization daemon)

Synopsis

trqauthd -D
trqauthd -d

Description

The trqauthd daemon, introduced in TORQUE 4.0.0, replaced the pbs_iff authentication process. When
users connect to pbs_server by calling one of the TORQUE utilities or by using the TORQUE APIs, the new
user connection must be authorized by a trusted entity which runs as root. The advantage of trqauthd's
doing this rather than pbs_iff is that trqauthd is resident, meaning you do not need to be loaded every
time a connection is made; multi-threaded; scalable; and more easily adapted to new functionality than
pbs_iff.

Beginning in TORQUE 4.2.6, trqauthd can remember the currently active pbs_server host, enhancing high
availability functionality. Previously, trqauthd tried to connect to each host in the $TORQUE_HOME/<server_name> file until it could successfully connect. Because it now remembers the active
server, it tries to connect to that server first. If it fails to connect, it will go through the <server_name> file and try to connect to a host where an active pbs_server is running.

Options

- **-D — Debug**

  | Format | --- |
  | Default | --- |
  | Description | Run trqauthd in debug mode. |
  | Example | trqauthd -D |

- **-d — Terminate**

  | Format | --- |
  | Default | --- |
  | Description | Terminate trqauthd. |
  | Example | trqauthd -d |
Server parameters

TORQUE server parameters are specified using the `qmgr` command. The `set` subcommand is used to modify the `server` object. For example:

```bash
> qmgr -c 'set server default_queue=batch'
```

Parameters

<table>
<thead>
<tr>
<th>acl_hosts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td><strong>Default</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
</tbody>
</table>

In version 2.5 and later, the wildcard (*) character can appear anywhere in the host name, and ranges are supported; these specifications also work for managers and operators:

```bash
Qmgr: set queue batch acl_hosts = "hostA,hostB"
Qmgr: set queue batch acl_hosts += "hostE,hostF,hostG"
```

<table>
<thead>
<tr>
<th>acl_host_enable</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td><strong>Default</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
</tbody>
</table>
### acl_logic_or

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;BOOLEAN&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>FALSE</td>
</tr>
<tr>
<td>Description</td>
<td>When set to TRUE, the user and group queue ACL’s are logically OR’d. When set to FALSE, they are AND’d.</td>
</tr>
</tbody>
</table>

### allow_node_submit

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;BOOLEAN&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>FALSE</td>
</tr>
<tr>
<td>Description</td>
<td>When set to TRUE, specifies that users can submit jobs directly from any trusted compute host directly or from within batch jobs (see Configuring job submission hosts on page 22). When you enable allow_node_submit you must also enable the allow_proxy_user on page 228 parameter to allow user proxying when submitting and running jobs.</td>
</tr>
</tbody>
</table>

### allow_proxy_user

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;BOOLEAN&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>FALSE</td>
</tr>
<tr>
<td>Description</td>
<td>When set to TRUE, specifies that users can proxy from one user to another. Proxy requests will be either validated by ruserok() or by the scheduler (see Job submission on page 39).</td>
</tr>
</tbody>
</table>

### auto_node_np

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;BOOLEAN&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>DISABLED</td>
</tr>
<tr>
<td>Description</td>
<td>When set to TRUE, automatically configures a node’s np (number of processors) value based on the ncpus value from the status update. Requires full manager privilege to set or alter.</td>
</tr>
</tbody>
</table>
### automatic_requeue_exit_code

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;LONG&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>---</td>
</tr>
<tr>
<td>Description</td>
<td>This is an exit code, defined by the admin, that tells pbs_server to requeue the job instead of considering it as completed. This allows the user to add some additional checks that the job can run meaningfully, and if not, then the job script exits with the specified code to be requeued.</td>
</tr>
</tbody>
</table>

### checkpoint_defaults

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>---</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies for a queue the default checkpoint values for a job that does not have checkpointing specified. The checkpoint_defaults parameter only takes effect on execution queues.</td>
</tr>
</tbody>
</table>

### clone_batch_delay

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>1</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies the delay (in seconds) between clone batches (see <code>clone_batch_size</code>).</td>
</tr>
</tbody>
</table>

### clone_batch_size

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>256</td>
</tr>
<tr>
<td>Description</td>
<td>Job arrays are created in batches of size X. X jobs are created, and after the <code>clone_batch_delay</code>, X more are created. This repeats until all are created.</td>
</tr>
</tbody>
</table>
### cray_enabled

**Format**  
<BOOLEAN>

**Default**  
FALSE

**Description**  
When set to TRUE, specifies that this instance of pbs_server has Cray hardware that reports to it. See **Installation Notes for Moab and TORQUE for Cray** in the Moab Workload Manager documentation.

### default_queue

**Format**  
<String>

**Default**  
---

**Description**  
Indicates the queue to assign to a job if no queue is explicitly specified by the submitter.

### disable_server_id_check

**Format**  
<BOOLEAN>

**Default**  
FALSE

**Description**  
When set to TRUE, makes it so the user for the job doesn’t have to exist on the server. The user must still exist on all the compute nodes or the job will fail when it tries to execute.

> !important If you have **disable_server_id_check** set to TRUE, a user could request a group to which they do not belong. Setting **VALIDATEGROUP** to TRUE in the torque.cfg file prevents such a scenario (see “torque.cfg” configuration file on page 295).

### display_job_server_suffix

**Format**  
<BOOLEAN>

**Default**  
TRUE

**Description**  
When set to TRUE, TORQUE will display both the job ID and the host name. When set to FALSE, only the job ID will be displayed.

> tip If set to **FALSE**, the environment variable **NO_SERVER_SUFFIX** must be set to **TRUE** for pbs_track to work as expected.
<table>
<thead>
<tr>
<th><strong>interactive_jobs_can_roam</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td><strong>Default</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>job_force_cancel_time</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td><strong>Default</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>job_log_file_max_size</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td><strong>Default</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>job_log_file_roll_depth</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td><strong>Default</strong></td>
</tr>
<tr>
<td><strong>job_log_file_roll_depth</strong></td>
</tr>
<tr>
<td>-----------------------------</td>
</tr>
</tbody>
</table>
| **Description**             | This sets the maximum number of new log files that are kept in a day if the *job_log_file_max_size* parameter is set. For example, if the roll depth is set to 3, no file can roll higher than `<filename.3>`. If a file is already at the specified depth, such as `<filename.3>`, the file is deleted so it can be replaced by the incoming file roll, `<filename.2>`.

<table>
<thead>
<tr>
<th><strong>job_log_keep_days</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td><strong>Default</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>job_nanny</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td><strong>Default</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>job_stat_rate</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td><strong>Default</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
</tbody>
</table>
## job_start_timeout

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>---</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies the pbs_server to pbs_mom TCP socket timeout in seconds that is used when the pbs_server sends a job start to the pbs_mom. It is useful when the MOM has extra overhead involved in starting jobs. If not specified, then the tcp_timeout parameter is used.</td>
</tr>
</tbody>
</table>

## job_sync_timeout

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>60</td>
</tr>
<tr>
<td>Description</td>
<td>When a stray job is reported on multiple nodes, the server sends a kill signal to one node at a time. This timeout determines how long the server waits between kills if the job is still being reported on any nodes.</td>
</tr>
</tbody>
</table>

## keep_completed

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>---</td>
</tr>
<tr>
<td>Description</td>
<td>The amount of time a job will be kept in the queue after it has entered the completed state. keep_completed must be set for job dependencies to work. For more information, see Keeping completed jobs on page 52.</td>
</tr>
</tbody>
</table>

If you ran `torque.setup` on TORQUE installation, the default is 300.

## lock_file

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td><code>torque/server_priv/server.lock</code></td>
</tr>
</tbody>
</table>
## lock_file

| Description | Specifies the name and location of the lock file used to determine which high availability server should be active. If a full path is specified, it is used verbatim by TORQUE. If a relative path is specified, TORQUE will prefix it with `torque/server_priv`. |

## lock_file_update_time

| Format | <INTEGER> |
| Default | 3 |
| Description | Specifies how often (in seconds) the thread will update the lock file. |

## lock_file_check_time

| Format | <INTEGER> |
| Default | 9 |
| Description | Specifies how often (in seconds) a high availability server will check to see if it should become active. |

## log_events

| Format | Bitmap |
| Default | --- |
### log_events

**Description**

By default, all events are logged. However, you can customize things so that only certain events show up in the log file. These are the bitmaps for the different kinds of logs:

```c
#define PBSEVENT_ERROR 0x0001 /* internal errors */
#define PBSEVENT_SYSTEM 0x0002 /* system (server) events */
#define PBSEVENT_ADMIN 0x0004 /* admin events */
#define PBSEVENT_JOB 0x0008 /* job related events */
#define PBSEVENT_JOB_USAGE 0x0010 /* End of Job accounting */
#define PBSEVENT_SECURITY 0x0020 /* security violation events */
#define PBSEVENT_SCHED 0x0040 /* scheduler events */
#define PBSEVENT_DEBUG 0x0080 /* common debug messages */
#define PBSEVENT_DEBUG2 0x0100 /* less needed debug messages */
#define PBSEVENT_FORCE 0x8000 /* set to force a message */
```

If you want to log only error, system, and job information, use `qmgr` to set `log_events` to 11:

```bash
set server log_events = 11
```

### log_file_max_size

**Format**

`<INTEGER>`

**Default**

0

**Description**

Specifies a soft limit, in kilobytes, for the server's log file. The file size is checked every 5 minutes, and if the current day file size is greater than or equal to this value then it will be rolled from X to X.1 and a new empty log will be opened. Any value less than or equal to 0 will be ignored by `pbs_server` (the log will not be rolled).

### log_file_roll_depth

**Format**

`<INTEGER>`

**Default**

1

**Description**

Controls how deep the current day log files will be rolled, if `log_file_max_size` is set, before they are deleted.

### log_keep_days

**Format**

`<INTEGER>`
### log_keep_days

<table>
<thead>
<tr>
<th>Default</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>Specifies how long (in days) a server or MOM log should be kept.</td>
</tr>
</tbody>
</table>

### log_level

<table>
<thead>
<tr>
<th><strong>Format</strong></th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Default</strong></td>
<td>0</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>Specifies the pbs_server logging verbosity. Maximum value is 7.</td>
</tr>
</tbody>
</table>

### mail_body_fmt

<table>
<thead>
<tr>
<th><strong>Format</strong></th>
<th>A printf-like format string</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Default</strong></td>
<td>PBS Job Id: %i Job Name: %j Exec host: %h %m %d</td>
</tr>
</tbody>
</table>
| **Description** | Override the default format for the body of outgoing mail messages. A number of printf-like format specifiers and escape sequences can be used:  
\n new line  
\t tab  
\ \ backslash  
\ single quote  
\" double quote  
%d details concerning the message  
%h PBS host name  
%i PBS job identifier  
%j PBS job name  
%m long reason for message  
%r short reason for message  
%% a single % |

### mail_domain

<table>
<thead>
<tr>
<th><strong>Format</strong></th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Default</strong></td>
<td>---</td>
</tr>
</tbody>
</table>
### mail_domain

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Override the default domain for outgoing mail messages. If set, emails will be addressed to &lt;user-@&lt;hostdomain&gt;. If unset, the job’s Job_Owner attribute will be used. If set to never, TORQUE will never send emails.</td>
</tr>
</tbody>
</table>

### mail_from

<table>
<thead>
<tr>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;STRING&gt;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>adm</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specify the name of the sender when TORQUE sends emails.</td>
</tr>
</tbody>
</table>

### mail_subject_fmt

<table>
<thead>
<tr>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>A printf-like format string</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBS JOB %i</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Override the default format for the subject of outgoing mail messages. A number of printf-like format specifiers and escape sequences can be used: \n new line \t tab \ backslash ' single quote &quot; double quote %d details concerning the message %h PBS host name %i PBS job identifier %j PBS job name %m long reason for message %r short reason for message %% a single %</td>
</tr>
</tbody>
</table>

### managers

<table>
<thead>
<tr>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;user&gt;@&lt;host.sub.domain&gt; [,&lt;user&gt;@&lt;host.sub.domain&gt;...]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>root@localhost</td>
</tr>
</tbody>
</table>
### managers

**Description** List of users granted batch administrator privileges. The host, sub-domain, or domain name may be wildcarded by the use of an asterisk character (*). Requires full manager privilege to set or alter.

### max_job_array_size

**Format** `<INTEGER>`

**Default** Unlimited

**Description** Sets the maximum number of jobs that can be in a single job array.

### max_slot_limit

**Format** `<INTEGER>`

**Default** Unlimited

**Description** This is the maximum number of jobs that can run concurrently in any job array. Slot limits can be applied at submission time with `qsub`, or it can be modified with `qalter`.

```
qmgr -c 'set server max_slot_limit=10'
```

No array can request a slot limit greater than 10. Any array that does not request a slot limit receives a slot limit of 10. Using the example above, slot requests greater than 10 are rejected with the message: "Requested slot limit is too large, limit is 10."

### max_threads

**Format** `<INTEGER>`

**Default** The value of `min_threads` `(2 * the number of procs listed in /proc/cpuinfo) + 1) * 10`

**Description** This is the maximum number of threads that should exist in the thread pool at any time.

### max_user_queuable

**Format** `<INTEGER>`
**max_user_queuable**

<table>
<thead>
<tr>
<th>Default</th>
<th>Unlimited</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>When set, max_user_queuable places a system-wide limit on the amount of jobs that an individual user can queue.</td>
</tr>
<tr>
<td></td>
<td>qmgr -c 'set server max_user_queuable=500'</td>
</tr>
</tbody>
</table>

**min_threads**

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Default</strong></td>
<td>(2 * the number of pros listed in <code>/proc/cpuinfo</code>) + 1. If TORQUE is unable to read <code>/proc/cpuinfo</code>, the default is 10.</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>This is the minimum number of threads that should exist in the thread pool at any time.</td>
</tr>
</tbody>
</table>

**moab_array_compatible**

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;BOOLEAN&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Default</strong></td>
<td>TRUE</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>This parameter places a hold on jobs that exceed the slot limit in a job array. When one of the active jobs is completed or deleted, one of the held jobs goes to a queued state.</td>
</tr>
</tbody>
</table>

**mom_job_sync**

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;BOOLEAN&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Default</strong></td>
<td>TRUE</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>When set to TRUE, specifies that the pbs_server will synchronize its view of the job queue and resource allocation with compute nodes as they come online. If a job exists on a compute node, it will be automatically cleaned up and purged. (Enabled by default in TORQUE 2.2.0 and higher.) Jobs that are no longer reported by the mother superior are automatically purged by pbs_server. Jobs that pbs_server instructs the MOM to cancel have their processes killed in addition to being deleted (instead of leaving them running as in versions of TORQUE prior to 4.1.1).</td>
</tr>
</tbody>
</table>
### next_job_number

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>---</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies the ID number of the next job. If you set your job number too low and TORQUE repeats a job number that it has already used, the job will fail. Before setting next_job_number to a number lower than any number that TORQUE has already used, you must clear out your .e and .o files.</td>
</tr>
</tbody>
</table>

If you use Moab Workload Manager and have configured it to synchronize job IDs with TORQUE (See Synchronizing Job IDs in TORQUE and Moab in the Moab Workload Manager Administrator Guide for more information.), then Moab will generate the job ID and next_job_number will have no effect on the job ID.

### node_check_rate

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>600</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies the minimum duration (in seconds) that a node can be unresponsive to server queries before being marked down by the pbs_server daemon.</td>
</tr>
</tbody>
</table>

### node_pack

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;BOOLEAN&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>---</td>
</tr>
<tr>
<td>Description</td>
<td>Controls how multiple processor nodes are allocated to jobs. If this attribute is set to TRUE, jobs will be assigned to the multiple processor nodes with the fewest free processors. This packs jobs into the fewest possible nodes leaving multiple processor nodes free for jobs which need many processors on a node. If set to false, jobs will be scattered across nodes reducing conflicts over memory between jobs. If unset, the jobs are packed on nodes in the order that the nodes are declared to the server (in the nodes file). Default value: unset - assigned to nodes as nodes in order that were declared.</td>
</tr>
</tbody>
</table>

### node_ping_rate

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>Server parameters</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>---</td>
</tr>
<tr>
<td><strong>node_ping_rate</strong></td>
<td></td>
</tr>
<tr>
<td>Default</td>
<td>300</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies the maximum interval (in seconds) between successive &quot;pings&quot; sent from the pbs_server daemon to the pbs_mom daemon to determine node/daemon health.</td>
</tr>
</tbody>
</table>

| **no_mail_force** |  |
| Format | <BOOLEAN> |
| Default | FALSE |
| Description | When set to TRUE, eliminates all e-mails when mail_options (see qsub on page 204) is set to "n". The job owner won't receive e-mails when a job is deleted by a different user or a job failure occurs. If no_mail_force is unset or is FALSE, then the job owner receives e-mails when a job is deleted by a different user or a job failure occurs. |

| **np_default** |  |
| Format | <INTEGER> |
| Default | --- |
| Description | Allows the administrator to unify the number of processors (np) on all nodes. The value can be dynamically changed. A value of 0 tells pbs_server to use the value of np found in the nodes file. The maximum value is 32767. |

| **operators** |  |
| Format | <user>@<host.sub.domain>,<user>@<host.sub.domain>... |
| Default | root@localhost |
| Description | List of users granted batch operator privileges. Requires full manager privilege to set or alter. |

<p>| <strong>poll_jobs</strong> |  |
| Format | &lt;BOOLEAN&gt; |</p>
<table>
<thead>
<tr>
<th>poll_jobs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>TRUE (FALSE in TORQUE 1.2.0p5 and earlier)</td>
</tr>
<tr>
<td>Description</td>
<td>If set to TRUE, pbs_server will poll job info from MOMs over time and will not block on handling requests which require this job information. If set to FALSE, no polling will occur and if requested job information is stale, pbs_server may block while it attempts to update this information. For large systems, this value should be set to TRUE.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>query_other_jobs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>&lt;BOOLEAN&gt;</td>
</tr>
<tr>
<td>Default</td>
<td>FALSE</td>
</tr>
<tr>
<td>Description</td>
<td>When set to TRUE, specifies whether or not non-admin users may view jobs they do not own.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>record_job_info</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>&lt;BOOLEAN&gt;</td>
</tr>
<tr>
<td>Default</td>
<td>FALSE</td>
</tr>
<tr>
<td>Description</td>
<td>This must be set to TRUE in order for job logging to be enabled.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>record_job_script</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>&lt;BOOLEAN&gt;</td>
</tr>
<tr>
<td>Default</td>
<td>FALSE</td>
</tr>
<tr>
<td>Description</td>
<td>If set to TRUE, this adds the contents of the script executed by a job to the log. For record_job_script to take effect, record_job_info on page 242 must be set to TRUE.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>resources_available</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>&lt;STRING&gt;</td>
</tr>
</tbody>
</table>
### resources_available

<table>
<thead>
<tr>
<th>Default</th>
<th>---</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>Allows overriding of detected resource quantity limits (see <a href="#">Assigning queue resource limits</a> on page 87). pbs_server must be restarted for changes to take effect. Also, resources_available is constrained by the smallest of queue.resources_available and the server.resources_available.</td>
</tr>
</tbody>
</table>

### scheduling

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;BOOLEAN&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Default</strong></td>
<td>---</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>Allows pbs_server to be scheduled. When FALSE, pbs_server is a resource manager that works on its own. When TRUE, TORQUE allows a scheduler, such as Moab or Maui, to dictate what pbs_server should do.</td>
</tr>
</tbody>
</table>

### submit_hosts

<table>
<thead>
<tr>
<th>Format</th>
<th>&quot;&lt;HOSTNAME&gt;[,&lt;HOSTNAME&gt;]...&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Default</strong></td>
<td>---</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>Indicates which hosts included in the server nodes file located at $TORQUE/server_priv/nodes (see <a href="#">Server node file configuration</a> on page 27) can submit batch or interactive jobs (see <a href="#">Configuring job submission hosts</a> on page 22). For more information on adding hosts that are not included in the first nodes file, see the acl_hosts parameter.</td>
</tr>
</tbody>
</table>

### tcp_timeout

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Default</strong></td>
<td>300</td>
</tr>
</tbody>
</table>
### tcp_timeout

**Description**
Specifies the timeout for idle TCP connections. If no communication is received by the server on the connection after the timeout, the server closes the connection. There is an exception for connections made to the server on port 15001 (default); timeout events are ignored on the server for such connections established by a client utility or scheduler. Responsibility rests with the client to close the connection first (See Large cluster considerations on page 273 for additional information.).

<table>
<thead>
<tr>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>If you use Moab Workload Manager, prevent communication errors by giving tcp_timeout at least twice the value of the Moab RMPOLIINTERVAL.</td>
</tr>
</tbody>
</table>

### thread_idle_seconds

**Format**

```<INTEGER>```

**Default**

300

**Description**

This is the number of seconds a thread can be idle in the thread pool before it is deleted. If threads should not be deleted, set to -1. TORQUE will always maintain at least min_threads number of threads, even if all are idle.
Node manager (MOM) configuration

Under TORQUE, MOM configuration is accomplished using the mom_priv/config file located in the PBS directory on each execution server. You must create this file and insert any desired lines in a text editor (blank lines are allowed). When you modify the mom_priv/config file, you must restart pbs_mom.

The following examples demonstrate two methods of modifying the mom_priv/config file:

```
> echo "$loglevel 3" >> /var/spool/torque/mom_priv/config
> vim /var/spool/torque/mom_priv/config

...$loglevel 3
```

For details, see these topics:

- Parameters on page 245
- Node features and generic consumable resource specification on page 260
- Command-line arguments on page 261

Related topics

- Commands overview on page 145
- Prologue and epilogue scripts on page 281

Parameters

<table>
<thead>
<tr>
<th>arch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
</tr>
<tr>
<td>Description</td>
</tr>
<tr>
<td>Example</td>
</tr>
</tbody>
</table>

$attempt_to_make_dir

| Format | <BOOLEAN> |
### $attempt_to_make_dir

**Description**
When set to *TRUE*, specifies that you want TORQUE to attempt to create the output directories for jobs if they do not already exist.
Default is *FALSE*.

TORQUE uses this parameter to make the directory as the *user* and not as *root*. TORQUE will create the directory (or directories) **ONLY** if the user has permissions to do so.

**Example**
```bash
$attempt_to_make_dir true
```

### $clienthost

**Format**
<STRING>

**Description**
Specifies the machine running `pbs_server`.

This parameter is deprecated. Use `$pbsserver`.

**Example**
```bash
$clienthost node01.teracluster.org
```

### $check_poll_time

**Format**
<STRING>

**Description**
Amount of time between checking running jobs, polling jobs, and trying to resend obituaries for jobs that haven’t sent successfully. Default is 45 seconds.

**Example**
```bash
$check_poll_time 90
```

### $configversion

**Format**
<STRING>

**Description**
Specifies the version of the config file data.

**Example**
```bash
$configversion 113
```
### $cputmult

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;FLOAT&gt;</th>
</tr>
</thead>
</table>
| Description| CPU time multiplier.  

<table>
<thead>
<tr>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>$cputmult 2.2</td>
</tr>
</tbody>
</table>

**Example**

If set to 0.0, MOM level cputime enforcement is disabled.

### $exec_with_exec

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;BOOLEAN&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>PBS mom uses the exec command to start the job script rather than the TORQUE default method, which is to pass the script's contents as the input to the shell. This means that if you trap signals in the job script, they will be trapped for the job. Using the default method, you would need to configure the shell to also trap the signals. Default is FALSE.</td>
</tr>
<tr>
<td>Example</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>$exec_with_exec true</td>
<td></td>
</tr>
</tbody>
</table>

### $ext_pwd_retry

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>(Available in TORQUE 2.5.10, 3.0.4, and later.) Specifies the number of times to retry checking the password. Useful in cases where external password validation is used, such as with LDAP. The default value is 3 retries.</td>
</tr>
<tr>
<td>Example</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>$ext_pwd_retry = 5</td>
<td></td>
</tr>
</tbody>
</table>

### $ideal_load

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;FLOAT&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Ideal processor load.</td>
</tr>
<tr>
<td>Example</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>$ideal_load 4.0</td>
<td></td>
</tr>
<tr>
<td>$igncput</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Format</strong></td>
<td>&lt;BOOLEAN&gt;</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>Ignores limit violation pertaining to CPU time. Default is FALSE.</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>$igncput true</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$ignmem</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
<td>&lt;BOOLEAN&gt;</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>Ignores limit violations pertaining to physical memory. Default is FALSE.</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>$ignmem true</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$ignvmem</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
<td>&lt;BOOLEAN&gt;</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>Ignores limit violations pertaining to virtual memory. Default is FALSE.</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>$ignvmem true</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$ignwalltime</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
<td>&lt;BOOLEAN&gt;</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>Ignore walltime (do not enable MOM based walltime limit enforcement).</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>$ignwalltime true</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$job_exit_wait_time</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
<td>&lt;INTEGER&gt;</td>
</tr>
<tr>
<td><strong>$job_exit_wait_time</strong></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>This is the timeout to clean up parallel jobs after one of the sister nodes for the parallel job goes down or is otherwise unresponsive. The MOM sends out all of its kill job requests to sisters and marks the time. Additionally, the job is placed in the substate JOB_SUBSTATE_EXIT_WAIT. The MOM then periodically checks jobs in this state and if they are in this state for more than the specified time, death is assumed and the job gets cleaned up. Default is 10 minutes.</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>$job_exit_wait_time 300</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>$job_output_file_umask</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
<td>&lt;STRING&gt;</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>Uses the specified umask when creating job output and error files. Values can be specified in base 8, 10, or 16; leading 0 implies octal and leading 0x or 0X hexadecimal. A value of &quot;userdefault&quot; will use the user’s default umask. This parameter is in version 2.3.0 and later.</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>$job_output_file_umask 027</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>$job_starter</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
<td>&lt;STRING&gt;</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>Specifies the fully qualified pathname of the job starter. If this parameter is specified, instead of executing the job command and job arguments directly, the MOM will execute the job starter, passing the job command and job arguments to it as its arguments. The job starter can be used to launch jobs within a desired environment.</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>$job_starter /var/torque/mom_priv/job_starter.sh &gt; cat /var/torque/mom_priv/job_starter.sh #!/bin/bash export FOOHOME=/home/foo ulimit -n 314 *</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>$log_directory</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
<td>&lt;STRING&gt;</td>
</tr>
</tbody>
</table>
### $log_directory

| Description | Changes the log directory. Default is `TORQUE_HOME/mom_logs/`. `TORQUE_HOME` default is `/var/spool/torque/` but can be changed in the `./configure` script. The value is a string and should be the full path to the desired MOM log directory. |
| Example | `$log_directory /opt/torque/mom_logs/` |

### $log_file_suffix

| Format | <STRING> |
| Description | Optional suffix to append to log file names. If `%h` is the suffix, `pbs_mom` appends the hostname for where the log files are stored if it knows it, otherwise it will append the hostname where the MOM is running. |
| Example | `$log_file_suffix %h = 20100223.mybox`  
            `$log_file_suffix foo = 20100223.foo` |

### $logevent

| Format | <STRING> |
| Description | Specifies a bitmap for event types to log. |
| Example | `$logevent 255` |

### $loglevel

| Format | <INTEGER> |
| Description | Specifies the verbosity of logging with higher numbers specifying more verbose logging. Values may range between 0 and 7. |
| Example | `$loglevel 4` |

### $log_file_max_size

| Format | <INTEGER> |
### $log_file_max_size

<table>
<thead>
<tr>
<th>Description</th>
<th>Soft limit for log file size in kilobytes. Checked every 5 minutes. If the log file is found to be greater than or equal to $log_file_max_size the current log file will be moved from X to X.1 and a new empty file will be opened.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>$log_file_max_size = 100</td>
</tr>
</tbody>
</table>

### $log_file_roll_depth

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Specifies how many times a log fill will be rolled before it is deleted.</td>
</tr>
<tr>
<td>Example</td>
<td>$log_file_roll_depth = 7</td>
</tr>
</tbody>
</table>

### $log_keep_days

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Specifies how many days to keep log files. pbs_mom deletes log files older than the specified number of days. If not specified, pbs_mom won't delete log files based on their age.</td>
</tr>
<tr>
<td>Example</td>
<td>$log_keep_days 10</td>
</tr>
</tbody>
</table>

### $max_conn_timeout_micro_sec

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Specifies how long pbs_mom should wait for a connection to be made. Default value is 10000 (1 sec).</td>
</tr>
<tr>
<td>Example</td>
<td>$max_conn_timeout_micro_sec 30000</td>
</tr>
<tr>
<td></td>
<td>This sets the connection timeout on the MOM to .3 seconds.</td>
</tr>
</tbody>
</table>

### $max_join_job_wait_time

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
</table>
### $max\_join\_job\_wait\_time$

<table>
<thead>
<tr>
<th>Description</th>
<th>The interval to wait for jobs stuck in a prerun state before deleting them from the MOMs and requeueing them on the server. Default is 10 minutes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>$max_join_job_wait_time\ 300$</td>
</tr>
</tbody>
</table>

### $max\_load$

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;FLOAT&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Maximum processor load.</td>
</tr>
<tr>
<td>Example</td>
<td>$max_load\ 4.0$</td>
</tr>
</tbody>
</table>

### $memory\_pressure\_duration$

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td><em>(Applicable in version 3.0 and later)</em> Memory pressure duration sets a limit to the number of times the value of memory_pressure_threshold can be exceeded before a process is terminated. This can only be used with $memory_pressure_threshold.*</td>
</tr>
<tr>
<td>Example</td>
<td>$memory_pressure_duration\ 5$</td>
</tr>
</tbody>
</table>

### $memory\_pressure\_threshold$

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INTEGER&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td><em>(Applicable in version 3.0 and later)</em> The memory_pressure of a cpuset provides a simple per-cpuset running average of the rate that the processes in a cpuset are attempting to free up in-use memory on the nodes of the cpuset to satisfy additional memory requests. The memory_pressure_threshold is an integer number used to compare against the reclaim rate provided by the memory_pressure file. If the threshold is exceeded and memory_pressure_duration is set, then the process terminates after exceeding the threshold by the number of times set in memory_pressure_duration. If memory_pressure duration is not set, then a warning is logged and the process continues. Memory_pressure_threshold is only valid with memory_pressure enabled in the root cpuset.*</td>
</tr>
<tr>
<td>Example</td>
<td>$memory_pressure_threshold\ 1000$</td>
</tr>
</tbody>
</table>
### $mom_hierarchy_retry_time

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;SECONDS&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Specifies the amount of time that a MOM waits to retry a node in the hierarchy path after a failed connection to that node. The default is 90 seconds.</td>
</tr>
<tr>
<td>Example</td>
<td>$mom_hierarchy_retry_time 30</td>
</tr>
</tbody>
</table>

### $node_check_script

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Specifies the fully qualified pathname of the health check script to run (see Compute node health check on page 134 for more information).</td>
</tr>
<tr>
<td>Example</td>
<td>$node_check_script /opt/batch_tools/nodecheck.pl</td>
</tr>
</tbody>
</table>

### $node_check_interval

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Specifies the number of MOM intervals between subsequent executions of the specified health check. This value default to 1 indicating the check is run every MOM interval (see Compute node health check on page 134 for more information).</td>
</tr>
<tr>
<td>Example</td>
<td>$node_check_interval 5</td>
</tr>
</tbody>
</table>

### $nodefile_suffix

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Specifies the suffix to append to a host names to denote the data channel network adapter in a multi-homed compute node.</td>
</tr>
<tr>
<td>Example</td>
<td>$nodefile_suffix i with the suffix of &quot;i&quot; and the control channel adapter with the name node01, the data channel would have a hostname of node01i.</td>
</tr>
</tbody>
</table>
### $nospool_dir_list

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>If this is configured, the job’s output is spooled in the working directory of the job or the specified output directory. Specify the list in full paths, delimited by commas. If the job’s working directory (or specified output directory) is in one of the paths in the list (or a subdirectory of one of the paths in the list), the job is spooled directly to the output location. $nospool_dir_list * is accepted. The user that submits the job must have write permission on the folder where the job is written, and read permission on the folder where the file is spooled. Alternatively, you can use the $spool_as_final_name parameter to force the job to spool directly to the final output.</td>
</tr>
<tr>
<td>Example</td>
<td>$nospool_dir_list /home/mike/jobs/,,/var/tmp/spool/</td>
</tr>
</tbody>
</table>

### opsys

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Specifies the operating system of the local machine. This information is used by the scheduler only.</td>
</tr>
<tr>
<td>Example</td>
<td>opsys RHEL3</td>
</tr>
</tbody>
</table>

### $pbsclient

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Specifies machines which the MOM daemon will trust to run resource manager commands via momctl. This may include machines where monitors, schedulers, or admins require the use of this command.</td>
</tr>
<tr>
<td>Example</td>
<td>$pbsclient node01.teracluster.org</td>
</tr>
</tbody>
</table>

### $pbsserver

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
</table>
### $pbsserver

**Description**
Specifies the machine running pbs_server.

- **Note:** This parameter replaces the deprecated parameter $clienthost.

**Example**
$pbsserver node01.teracluster.org

### $prologalarm

**Format**
<INTEGER>

**Description**
Specifies maximum duration (in seconds) which the MOM will wait for the job prologue or job epilogue to complete. The default value is 300 seconds (5 minutes). The maximum value is 300 and when set to anything higher than that, it is treated as 300.

**Example**
$prologalarm 60

### $rcpcmd

**Format**
<STRING>

**Description**
Specifies the full path and optional additional command line args to use to perform remote copies.

**Example**
mom_priv/config:
$rcpcmd /usr/local/bin/scp -i /etc/sshauth.dat

### $remote_reconfig

**Format**
<STRING>

**Description**
Enables the ability to remotely reconfigure pbs_mom with a new config file. Default is disabled. This parameter accepts various forms of true, yes, and 1. For more information on how to reconfigure MOMs, see `momctl -r`.

**Example**
$remote_reconfig true
| **$reduce_prolog_checks** |  |
| Format | <STRING> |
| Description | If enabled, TORQUE will only check if the file is a regular file and is executable, instead of the normal checks listed on the prologue and epilogue page. Default is FALSE. |
| Example | $reduce_prolog_checks true |

| **$reject_job_submission** |  |
| Format | <BOOLEAN> |
| Description | If set to TRUE, jobs will be rejected and the user will receive the message, “Jobs cannot be run on mom %s.” Default is FALSE. |
| Example | $reject_job_submission job01 |

| **$resend_join_job_wait_time** |  |
| Format | <INTEGER> |
| Description | This is the timeout for the Mother Superior to re-send the join job request if it didn’t get a reply from all the sister MOMs. The resend happens only once. Default is 5 minutes. |
| Example | $resend_join_job_wait_time 120 |

| **$restricted** |  |
| Format | <STRING> |
| Description | Specifies hosts which can be trusted to access MOM services as non-root. By default, no hosts are trusted to access MOM services as non-root. |
| Example | $restricted *.teracluster.org |

| **$rpp_throttle** |  |
| Format | <INTEGER> |
### $rpp\_throttle

<table>
<thead>
<tr>
<th>Description</th>
<th>This integer is in microseconds and causes a sleep after every RPP packet is sent. It is for systems that experience job failures because of incomplete data.</th>
</tr>
</thead>
</table>
| Example     | `$rpp\_throttle 100`  
(will cause a 100 microsecond sleep) |

#### size[fs=\(<FS>\)]

<table>
<thead>
<tr>
<th>Format</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Specifies that the available and configured disk space in the (&lt;FS&gt;) filesystem is to be reported to the pbs_server and scheduler.</td>
</tr>
</tbody>
</table>
| Example | size[fs=/localscratch]  
The available and configured disk space in the /localscratch filesystem will be reported. |

#### $source\_login\_batch

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Specifies whether or not MOM will source the /etc/profile, etc. type files for batch jobs. Parameter accepts various forms of true, false, yes, no, 1 and 0. Default is TRUE. This parameter is in version 2.3.1 and later.</td>
</tr>
</tbody>
</table>
| Example | $source\_login\_batch False  
MOM will bypass the sourcing of /etc/profile, etc. type files. |

#### $source\_login\_interactive

<p>| Format | &lt;STRING&gt; |</p>
<table>
<thead>
<tr>
<th>$source_login_interactive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
</tr>
</tbody>
</table>
| **Example** | `$source_login_interactive False`  
MOM will bypass the sourcing of `/etc/profile, etc.` type files. |

<table>
<thead>
<tr>
<th>$spool_as_final_name</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
<tr>
<td><strong>Example</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$status_update_time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
</tbody>
</table>
| **Example** | `status_update_time:  
$status_update_time 120  
MOM will send server update reports every 120 seconds.` |

<table>
<thead>
<tr>
<th>$thread_unlink_calls</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
</tbody>
</table>
| **Example** | `thread_unlink_calls:  
$thread_unlink_calls true` |
### $timeout

**Format**

<INTEGER>

**Description**

Specifies the number of seconds before a TCP connection on the MOM will timeout. Default is 300 seconds.

In version 3.x and earlier, this specifies the number of seconds before MOM-to-MOM messages will timeout if RPP is disabled. Default is 60 seconds.

**Example**

$timeout 120

A TCP connection will wait up to 120 seconds before timing out.

For 3.x and earlier, MOM-to-MOM communication will allow up to 120 seconds before timing out.

### $tmpdir

**Format**

<STRING>

**Description**

Specifies a directory to create job-specific scratch space (see Creating Per-Job Temporary Directories).

**Example**

$tmpdir /localscratch

### $usecp

**Format**

<HOST>:<SRCDIR> <DSTDIR>

**Description**

Specifies which directories should be staged (see NFS and other networked filesystems on page 110).

**Example**

$usecp *.fte.com:/data /usr/local/data

### $use_smt

**Format**

<BOOLEAN>

**Description**

Indicates that the user would like to use SMT. If set, each logical core inside of a physical core will be used as a normal core for cpusets. This parameter is on by default.

If SMT is used, you will need to set the np attribute so that each logical processor is counted.

**Example**

$use_smt false
$\textbf{varattr}$

<table>
<thead>
<tr>
<th><strong>Format</strong></th>
<th>&lt;INTEGER&gt; &lt;STRING&gt;</th>
</tr>
</thead>
</table>
| **Description** | Provides a way to keep track of dynamic attributes on nodes. <INTEGER> is how many seconds should go by between calls to the script to update the dynamic values. If set to -1, the script is read only one time. <STRING> is the script path. This script should check for whatever dynamic attributes are desired, and then output lines in this format: \begin{verbatim}
name=value
\end{verbatim}
Include any arguments after the script's full path. These features are visible in the output of \texttt{pbsnodes}:
\begin{verbatim}
varattr=Matlab=7.1;Octave=1.0.
\end{verbatim}
For information about using \texttt{varattr} to request dynamic features in Moab, see \url{Configuring dynamic features in TORQUE and Moab} in the \textit{Moab Workload Manager Administrator Guide}.

| **Example** | $\texttt{varattr 25 /usr/local/scripts/nodeProperties.pl arg1 arg2 arg3}$ |

$\textbf{wallmult}$

<table>
<thead>
<tr>
<th><strong>Format</strong></th>
<th>&lt;FLOAT&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>Sets a factor to adjust walltime usage by multiplying a default job time to a common reference system. It modifies real walltime on a per-MOM basis (MOM configuration parameters). The factor is used for walltime calculations and limits in the same way that \texttt{cputmult} is used for cpu time. If set to 0.0, MOM level walltime enforcement is disabled.</td>
</tr>
</tbody>
</table>

| **Example** | $\texttt{wallmult 2.2}$ |

Related topics
- \url{Node manager (MOM) configuration} on page 245

**Node features and generic consumable resource specification**

Node features (a.k.a. "node properties") are opaque labels which can be applied to a node. They are not consumable and cannot be associated with a value. (Use generic resources described below for these purposes). Node features are configured within the nodes file on the \texttt{pbs_server} head node. This file can be used to specify an arbitrary number of node features.
Additionally, per node consumable generic resources may be specified using the format "<ATTR> <VAL>" with no leading dollar ("$") character. When specified, this information is routed to the scheduler and can be used in scheduling decisions. For example, to indicate that a given host has two tape drives and one node-locked matlab license available for batch jobs, the following could be specified:

```
mom_priv/config:
$clienthost 241.13.153.7
tape 2
matlab 1
```

Dynamic consumable resource information can be routed in by specifying a path preceded by an exclamation point (!) as in the example below. If the resource value is configured in this manner, the specified file will be periodically executed to load the effective resource value.

```
mom_priv/config:
$clienthost 241.13.153.7
tape !/opt/rm/gettapecount.pl
matlab !/opt/tools/getlicensecount.pl
```

Related topics

- Node manager (MOM) configuration on page 245

**Command-line arguments**

Below is a table of `pbs_mom` command-line startup flags.

<table>
<thead>
<tr>
<th>Flag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>a &lt;integer&gt;</td>
<td>Alarm time in seconds.</td>
</tr>
<tr>
<td>c &lt;file&gt;</td>
<td>Config file path.</td>
</tr>
<tr>
<td>C &lt;directory&gt;</td>
<td>Checkpoint path.</td>
</tr>
<tr>
<td>d &lt;directory&gt;</td>
<td>Home directory.</td>
</tr>
<tr>
<td>L &lt;file&gt;</td>
<td>Log file.</td>
</tr>
<tr>
<td>M &lt;integer&gt;</td>
<td>MOM port to listen on.</td>
</tr>
<tr>
<td>p</td>
<td>Perform 'poll' based job recovery on restart (jobs persist until associated processes terminate).</td>
</tr>
<tr>
<td>Flag</td>
<td>Description</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
</tr>
<tr>
<td>P</td>
<td>On restart, deletes all jobs that were running on MOM (Available in 2.4.X and later).</td>
</tr>
<tr>
<td>q</td>
<td>On restart, requeues all jobs that were running on MOM (Available in 2.4.X and later).</td>
</tr>
<tr>
<td>r</td>
<td>On restart, kills all processes associated with jobs that were running on MOM, and then requeues the jobs.</td>
</tr>
<tr>
<td>R &lt;integer&gt;</td>
<td>MOM 'RM' port to listen on.</td>
</tr>
<tr>
<td>S &lt;integer&gt;</td>
<td>pbs_server port to connect to.</td>
</tr>
<tr>
<td>v</td>
<td>Display version information and exit.</td>
</tr>
<tr>
<td>x</td>
<td>Disable use of privileged port.</td>
</tr>
<tr>
<td>?</td>
<td>Show usage information and exit.</td>
</tr>
</tbody>
</table>

For more details on these command-line options, see **pbs_mom on page 151**.

**Related topics**
- **Node manager (MOM) configuration** on page 245
Diagnostics and error codes

TORQUE has a diagnostic script to assist you in giving TORQUE Support the files they need to support issues. It should be run by a user that has access to run all TORQUE commands and access to all TORQUE directories (this is usually root).

The script (contrib/diag/tdiag.sh) is available in TORQUE 2.3.8, TORQUE 2.4.3, and later. The script grabs the node file, server and MOM log files, and captures the output of qmgr -c 'ps'. These are put in a tar file.

The script also has the following options (this can be shown in the command line by entering ./tdiag.sh -h):

Usage: ./torque_diag [-d DATE] [-h] [-o OUTPUT_FILE] [-t TORQUE_HOME]

- **DATE** should be in the format YYYYmmdd. For example, "20091130" would be the date for November 30th, 2009. If no date is specified, today's date is used.

- **OUTPUT_FILE** is the optional name of the output file. The default output file is torque_diag<today's_date>.tar.gz. TORQUE_HOME should be the path to your TORQUE directory. If no directory is specified, /var/spool/torque is the default.

<table>
<thead>
<tr>
<th>Error code name</th>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBSE_FLOOR</td>
<td>15000</td>
<td>No error</td>
</tr>
<tr>
<td>PBSE_UNKJOBID</td>
<td>15001</td>
<td>Unknown job ID error</td>
</tr>
<tr>
<td>PBSE_NOATTR</td>
<td>15002</td>
<td>Undefined attribute</td>
</tr>
<tr>
<td>PBSE_ATTRIBRO</td>
<td>15003</td>
<td>Cannot set attribute, read only or insufficient permission</td>
</tr>
<tr>
<td>PBSE_IVALREQ</td>
<td>15004</td>
<td>Invalid request</td>
</tr>
<tr>
<td>PBSE_UNKREQ</td>
<td>15005</td>
<td>Unknown request</td>
</tr>
<tr>
<td>PBSE_TOOMANY</td>
<td>15006</td>
<td>Too many submit retries</td>
</tr>
<tr>
<td>Error code name</td>
<td>Number</td>
<td>Description</td>
</tr>
<tr>
<td>-------------------------</td>
<td>--------</td>
<td>-----------------------------------------------------</td>
</tr>
<tr>
<td>PBSE_PERM</td>
<td>15007</td>
<td>Unauthorized Request</td>
</tr>
<tr>
<td>PBSE_IFF_NOT_FOUND</td>
<td>15008</td>
<td>trqauthd unable to authenticate</td>
</tr>
<tr>
<td>PBSE_MUNGE_NOT_FOUND</td>
<td>15009</td>
<td>Munge executable not found, unable to authenticate</td>
</tr>
<tr>
<td>PBSE_BADHOST</td>
<td>15010</td>
<td>Access from host not allowed, or unknown host</td>
</tr>
<tr>
<td>PBSE_JOBEXIST</td>
<td>15011</td>
<td>Job with requested ID already exists</td>
</tr>
<tr>
<td>PBSE_SYSTEM</td>
<td>15012</td>
<td>System error</td>
</tr>
<tr>
<td>PBSE_INTERNAL</td>
<td>15013</td>
<td>PBS server internal error</td>
</tr>
<tr>
<td>PBSE_REGROUTE</td>
<td>15014</td>
<td>Dependent parent job currently in routing queue</td>
</tr>
<tr>
<td>PBSE_UNKSIG</td>
<td>15015</td>
<td>Unknown/illegal signal name</td>
</tr>
<tr>
<td>PBSE_BADATVAL</td>
<td>15016</td>
<td>Illegal attribute or resource value for</td>
</tr>
<tr>
<td>PBSE_MODATRRUN</td>
<td>15017</td>
<td>Cannot modify attribute while job running</td>
</tr>
<tr>
<td>PBSE_BADSTATE</td>
<td>15018</td>
<td>Request invalid for state of job</td>
</tr>
<tr>
<td>PBSE_UNKQUE</td>
<td>15020</td>
<td>Unknown queue</td>
</tr>
<tr>
<td>PBSE_BADCRED</td>
<td>15021</td>
<td>Invalid credential</td>
</tr>
<tr>
<td>PBSE_EXPIRED</td>
<td>15022</td>
<td>Expired credential</td>
</tr>
<tr>
<td>PBSE_QUNOENB</td>
<td>15023</td>
<td>Queue is not enabled</td>
</tr>
<tr>
<td>PBSE_QACCESS</td>
<td>15024</td>
<td>Access to queue is denied</td>
</tr>
<tr>
<td>PBSE_BADUSER</td>
<td>15025</td>
<td>Bad UID for job execution</td>
</tr>
<tr>
<td>PBSE_HOPCOUNT</td>
<td>15026</td>
<td>Job routing over too many hops</td>
</tr>
<tr>
<td>PBSE_QUEUEEXIST</td>
<td>15027</td>
<td>Queue already exists</td>
</tr>
<tr>
<td>Error code name</td>
<td>Number</td>
<td>Description</td>
</tr>
<tr>
<td>------------------</td>
<td>--------</td>
<td>-----------------------------------------------------</td>
</tr>
<tr>
<td>PBSE_ATTRTYPE</td>
<td>15028</td>
<td>Incompatible type</td>
</tr>
<tr>
<td>PBSE_QUEUEBUSY</td>
<td>15029</td>
<td>Cannot delete busy queue</td>
</tr>
<tr>
<td>PBSE_QUENBIG</td>
<td>15030</td>
<td>Queue name too long</td>
</tr>
<tr>
<td>PBSE_NOSUP</td>
<td>15031</td>
<td>No support for requested service</td>
</tr>
<tr>
<td>PBSE_QUENOEN</td>
<td>15032</td>
<td>Cannot enable queue, incomplete definition</td>
</tr>
<tr>
<td>PBSE_PROTOCOL</td>
<td>15033</td>
<td>Batch protocol error</td>
</tr>
<tr>
<td>PBSE_BADATLST</td>
<td>15034</td>
<td>Bad attribute list structure</td>
</tr>
<tr>
<td>PBSE_NOCONNECTS</td>
<td>15035</td>
<td>No free connections</td>
</tr>
<tr>
<td>PBSE_NOSERVER</td>
<td>15036</td>
<td>No server specified</td>
</tr>
<tr>
<td>PBSE_UNKRESC</td>
<td>15037</td>
<td>Unknown resource type</td>
</tr>
<tr>
<td>PBSE_EXCQRESC</td>
<td>15038</td>
<td>Job exceeds queue resource limits</td>
</tr>
<tr>
<td>PBSE_QUENODFLT</td>
<td>15039</td>
<td>No default queue specified</td>
</tr>
<tr>
<td>PBSE_NORERUN</td>
<td>15040</td>
<td>Job is not rerunnable</td>
</tr>
<tr>
<td>PBSE_ROUTEREJ</td>
<td>15041</td>
<td>Job rejected by all possible destinations (check syntax, queue resources, ...)</td>
</tr>
<tr>
<td>PBSE_ROUTEEXPD</td>
<td>15042</td>
<td>Time in Route Queue Expired</td>
</tr>
<tr>
<td>PBSE_MOMREJECT</td>
<td>15043</td>
<td>Execution server rejected request</td>
</tr>
<tr>
<td>PBSE_BADSCRIPT</td>
<td>15044</td>
<td>(qsub) cannot access script file</td>
</tr>
<tr>
<td>PBSE_STAGEIN</td>
<td>15045</td>
<td>Stage-in of files failed</td>
</tr>
<tr>
<td>PBSE_RESCUNAV</td>
<td>15046</td>
<td>Resource temporarily unavailable</td>
</tr>
<tr>
<td>Error code name</td>
<td>Number</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------</td>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>PBSE_BADGRP</td>
<td>15047</td>
<td>Bad GID for job execution</td>
</tr>
<tr>
<td>PBSE_MAXQUED</td>
<td>15048</td>
<td>Maximum number of jobs already in queue</td>
</tr>
<tr>
<td>PBSE_CKPBSY</td>
<td>15049</td>
<td>Checkpoint busy, may retry</td>
</tr>
<tr>
<td>PBSE_EXLIMIT</td>
<td>15050</td>
<td>Resource limit exceeds allowable</td>
</tr>
<tr>
<td>PBSE_BADACCT</td>
<td>15051</td>
<td>Invalid Account</td>
</tr>
<tr>
<td>PBSE_ALRDYEXIT</td>
<td>15052</td>
<td>Job already in exit state</td>
</tr>
<tr>
<td>PBSE_NOCOPYFILE</td>
<td>15053</td>
<td>Job files not copied</td>
</tr>
<tr>
<td>PBSE_CLEANEDOUT</td>
<td>15054</td>
<td>Unknown job id after clean init</td>
</tr>
<tr>
<td>PBSE_NOSYNCMSTR</td>
<td>15055</td>
<td>No master found for sync job set</td>
</tr>
<tr>
<td>PBSE_BADDEPEND</td>
<td>15056</td>
<td>Invalid Job Dependency</td>
</tr>
<tr>
<td>PBSE_DUPLIST</td>
<td>15057</td>
<td>Duplicate entry in list</td>
</tr>
<tr>
<td>PBSE_DISPROTO</td>
<td>15058</td>
<td>Bad DIS based Request Protocol</td>
</tr>
<tr>
<td>PBSE_EXECTHERE</td>
<td>15059</td>
<td>Cannot execute at specified host because of checkpoint or stagein files</td>
</tr>
<tr>
<td>PBSE_SISREJECT</td>
<td>15060</td>
<td>Sister rejected</td>
</tr>
<tr>
<td>PBSE_SISCOMM</td>
<td>15061</td>
<td>Sister could not communicate</td>
</tr>
<tr>
<td>PBSE_SVRDOWN</td>
<td>15062</td>
<td>Request not allowed: Server shutting down</td>
</tr>
<tr>
<td>PBSE_CKPSHORT</td>
<td>15063</td>
<td>Not all tasks could checkpoint</td>
</tr>
<tr>
<td>PBSE_UNKNODE</td>
<td>15064</td>
<td>Unknown node</td>
</tr>
<tr>
<td>PBSE_UNKNODEATR</td>
<td>15065</td>
<td>Unknown node-attribute</td>
</tr>
<tr>
<td>Error code name</td>
<td>Number</td>
<td>Description</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>PBSE_NONODES</td>
<td>15066</td>
<td>Server has no node list</td>
</tr>
<tr>
<td>PBSE_NODENBIG</td>
<td>15067</td>
<td>Node name is too big</td>
</tr>
<tr>
<td>PBSE_NODEEXIST</td>
<td>15068</td>
<td>Node name already exists</td>
</tr>
<tr>
<td>PBSE_BADNDATVAL</td>
<td>15069</td>
<td>Illegal value for</td>
</tr>
<tr>
<td>PBSE_MUTUALEX</td>
<td>15070</td>
<td>Mutually exclusive values for</td>
</tr>
<tr>
<td>PBSE_GMODERR</td>
<td>15071</td>
<td>Modification failed for</td>
</tr>
<tr>
<td>PBSE_NORELYMOM</td>
<td>15072</td>
<td>Server could not connect to MOM</td>
</tr>
<tr>
<td>PBSE_NOTSNODE</td>
<td>15073</td>
<td>No time-share node available</td>
</tr>
<tr>
<td>PBSE_JOBTYPE</td>
<td>15074</td>
<td>Wrong job type</td>
</tr>
<tr>
<td>PBSE_BADAACLHOST</td>
<td>15075</td>
<td>Bad ACL entry in host list</td>
</tr>
<tr>
<td>PBSE_MAXUSERQUED</td>
<td>15076</td>
<td>Maximum number of jobs already in queue for user</td>
</tr>
<tr>
<td>PBSE_BADDISALLOWTYPE</td>
<td>15077</td>
<td>Bad type in disallowed_types list</td>
</tr>
<tr>
<td>PBSE_NOINTERACTIVE</td>
<td>15078</td>
<td>Queue does not allow interactive jobs</td>
</tr>
<tr>
<td>PBSE_NOBATCH</td>
<td>15079</td>
<td>Queue does not allow batch jobs</td>
</tr>
<tr>
<td>PBSE_NORERUNABLE</td>
<td>15080</td>
<td>Queue does not allow rerunable jobs</td>
</tr>
<tr>
<td>PBSE_NONONRERUNABLE</td>
<td>15081</td>
<td>Queue does not allow nonrerunable jobs</td>
</tr>
<tr>
<td>PBSE_UNKARRAYID</td>
<td>15082</td>
<td>Unknown Array ID</td>
</tr>
<tr>
<td>PBSE_BAD_ARRAY_REQ</td>
<td>15083</td>
<td>Bad Job Array Request</td>
</tr>
<tr>
<td>PBSE_BAD_ARRAY_DATA</td>
<td>15084</td>
<td>Bad data reading job array from file</td>
</tr>
<tr>
<td>PBSE_TIMEOUT</td>
<td>15085</td>
<td>Time out</td>
</tr>
<tr>
<td>Error code name</td>
<td>Number</td>
<td>Description</td>
</tr>
<tr>
<td>-------------------------</td>
<td>--------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>PBSE_JOBNOTFOUND</td>
<td>15086</td>
<td>Job not found</td>
</tr>
<tr>
<td>PBSE_NOFAULTTOLERANT</td>
<td>15087</td>
<td>Queue does not allow fault tolerant jobs</td>
</tr>
<tr>
<td>PBSE_NOFAULTINTOLERANT</td>
<td>15088</td>
<td>Queue does not allow fault intolerant jobs</td>
</tr>
<tr>
<td>PBSE_NOJOBARRAYS</td>
<td>15089</td>
<td>Queue does not allow job arrays</td>
</tr>
<tr>
<td>PBSE_RELAYED_TO_MOM</td>
<td>15090</td>
<td>Request was relayed to a MOM</td>
</tr>
<tr>
<td>PBSE_MEM_MALLOC</td>
<td>15091</td>
<td>Error allocating memory - out of memory</td>
</tr>
<tr>
<td>PBSE_MUTEX</td>
<td>15092</td>
<td>Error allocating controling mutex (lock/unlock)</td>
</tr>
<tr>
<td>PBSE_THREADATTR</td>
<td>15093</td>
<td>Error setting thread attributes</td>
</tr>
<tr>
<td>PBSE_THREAD</td>
<td>15094</td>
<td>Error creating thread</td>
</tr>
<tr>
<td>PBSE_SELECT</td>
<td>15095</td>
<td>Error in socket select</td>
</tr>
<tr>
<td>PBSE_SOCKET_FAULT</td>
<td>15096</td>
<td>Unable to get connection to socket</td>
</tr>
<tr>
<td>PBSE_SOCKET_WRITE</td>
<td>15097</td>
<td>Error writing data to socket</td>
</tr>
<tr>
<td>PBSE_SOCKET_READ</td>
<td>15098</td>
<td>Error reading data from socket</td>
</tr>
<tr>
<td>PBSE_SOCKET_CLOSE</td>
<td>15099</td>
<td>Socket close detected</td>
</tr>
<tr>
<td>PBSE_SOCKET_LISTEN</td>
<td>15100</td>
<td>Error listening on socket</td>
</tr>
<tr>
<td>PBSE_AUTH_INVALID</td>
<td>15101</td>
<td>Invalid auth type in request</td>
</tr>
<tr>
<td>PBSE_NOT_IMPLEMENTED</td>
<td>15102</td>
<td>This functionality is not yet implemented</td>
</tr>
<tr>
<td>PBSE_QUEUEAVAILABLE</td>
<td>15103</td>
<td>Queue is currently not available</td>
</tr>
<tr>
<td>PBSE_TMPDIFFOWNER</td>
<td>15104</td>
<td>tmpdir owned by another user</td>
</tr>
<tr>
<td>PBSE_TMPNOTDIR</td>
<td>15105</td>
<td>tmpdir exists but is not a directory</td>
</tr>
<tr>
<td>Error code name</td>
<td>Number</td>
<td>Description</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>--------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>PBSE_TMPNONAME</td>
<td>15106</td>
<td>tmpdir cannot be named for job</td>
</tr>
<tr>
<td>PBSE_CANTOPENSOCKET</td>
<td>15107</td>
<td>Cannot open demux sockets</td>
</tr>
<tr>
<td>PBSE_CANTCONTACTSISTERS</td>
<td>15108</td>
<td>Cannot send join job to all sisters</td>
</tr>
<tr>
<td>PBSE_CANTCREATETMPDIR</td>
<td>15109</td>
<td>Cannot create tmpdir for job</td>
</tr>
<tr>
<td>PBSE_BADMOMSTATE</td>
<td>15110</td>
<td>Mom is down, cannot run job</td>
</tr>
<tr>
<td>PBSE_SOCKET_INFORMATION</td>
<td>15111</td>
<td>Socket information is not accessible</td>
</tr>
<tr>
<td>PBSE_SOCKET_DATA</td>
<td>15112</td>
<td>Data on socket does not process correctly</td>
</tr>
<tr>
<td>PBSE_CLIENT_INVALID</td>
<td>15113</td>
<td>Client is not allowed/trusted</td>
</tr>
<tr>
<td>PBSE_PREMATURE_EOF</td>
<td>15114</td>
<td>Premature End of File</td>
</tr>
<tr>
<td>PBSE_CAN_NOT_SAVE_FILE</td>
<td>15115</td>
<td>Error saving file</td>
</tr>
<tr>
<td>PBSE_CAN_NOT_OPEN_FILE</td>
<td>15116</td>
<td>Error opening file</td>
</tr>
<tr>
<td>PBSE_CAN_NOT_WRITE_FILE</td>
<td>15117</td>
<td>Error writing file</td>
</tr>
<tr>
<td>PBSE_JOB_FILE_CORRUPT</td>
<td>15118</td>
<td>Job file corrupt</td>
</tr>
<tr>
<td>PBSE_JOB_RERUN</td>
<td>15119</td>
<td>Job can not be rerun</td>
</tr>
<tr>
<td>PBSE_CONNECT</td>
<td>15120</td>
<td>Can not establish connection</td>
</tr>
<tr>
<td>PBSE_JOBWORKDELAY</td>
<td>15121</td>
<td>Job function must be temporarily delayed</td>
</tr>
<tr>
<td>PBSE_BAD_PARAMETER</td>
<td>15122</td>
<td>Parameter of function was invalid</td>
</tr>
<tr>
<td>PBSE_CONTINUE</td>
<td>15123</td>
<td>Continue processing on job. (Not an error)</td>
</tr>
<tr>
<td>PBSE_JOBSUBSTATE</td>
<td>15124</td>
<td>Current sub state does not allow trasaction.</td>
</tr>
<tr>
<td>PBSE_CAN_NOT_MOVE_FILE</td>
<td>15125</td>
<td>Error moving file</td>
</tr>
<tr>
<td>Error code name</td>
<td>Number</td>
<td>Description</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>--------</td>
<td>-------------------------------------------------------</td>
</tr>
<tr>
<td>PBSE_JOB_RECYCLED</td>
<td>15126</td>
<td>Job is being recycled</td>
</tr>
<tr>
<td>PBSE_JOB_ALREADY_IN_QUEUE</td>
<td>15127</td>
<td>Job is already in destination queue.</td>
</tr>
<tr>
<td>PBSE_INVALID_MUTEX</td>
<td>15128</td>
<td>Mutex is NULL or otherwise invalid</td>
</tr>
<tr>
<td>PBSE_MUTEX_ALREADY_LOCKED</td>
<td>15129</td>
<td>The mutex is already locked by this object</td>
</tr>
<tr>
<td>PBSE_MUTEX_ALREADY_UNLOCKED</td>
<td>15130</td>
<td>The mutex has already been unlocked by this object</td>
</tr>
<tr>
<td>PBSE_INVALID_SYNTAX</td>
<td>15131</td>
<td>Command syntax invalid</td>
</tr>
<tr>
<td>PBSE_NODE_DOWN</td>
<td>15132</td>
<td>A node is down. Check the MOM and host</td>
</tr>
<tr>
<td>PBSE_SERVER_NOT_FOUND</td>
<td>15133</td>
<td>Could not connect to batch server</td>
</tr>
<tr>
<td>PBSE_SERVER_BUSY</td>
<td>15134</td>
<td>Server busy. Currently no available threads</td>
</tr>
</tbody>
</table>
Considerations before upgrading

TORQUE is flexible in regards to how it can be upgraded. In most cases, a TORQUE "shutdown" followed by a configure, make, make install procedure as documented in this guide is all that is required (see Installing TORQUE on page 2). This process will preserve existing configuration and in most cases, existing workload.

A few considerations are included below:

- If upgrading from OpenPBS, PBSPro, or TORQUE 1.0.3 or earlier, queued jobs whether active or idle will be lost. In such situations, job queues should be completely drained of all jobs.

- If not using the pbs_mom -r or -p flag (see Command-line arguments on page 261), running jobs may be lost. In such cases, running jobs should be allowed to be completed or should be requeued before upgrading TORQUE.

- pbs_mom and pbs_server daemons of differing versions may be run together. However, not all combinations have been tested and unexpected failures may occur.

- trqauthd is an intermediary between client commands and pbs_server. It is recommended that when you upgrade pbs_server you also upgrade the client utilities and trqauthd to prevent unexpected failures when you execute client commands. Because no direct relationship exists between the MOMs and trqauthd, you can upgrade trqauthd without upgrading the MOMs.

- When upgrading from early versions of TORQUE (pre-4.0) and Moab, you may encounter a problem where Moab core files are regularly created in /opt/moab. This can be caused by old TORQUE library files used by Moab that try to authorize with the old TORQUE pbs_iff authorization daemon. You can resolve the problem by removing the old version library files from /usr/local/lib.

To upgrade

1. Build new release (do not install).
2. Stop all TORQUE daemons (see qterm and momctl –s).
3. Install new TORQUE (use make install).
4. Start all TORQUE daemons.

Rolling upgrade

If you are upgrading to a new point release of your current version (for example, from 4.2.2 to 4.2.3) and not to a new major release from your current version (for example, from 4.1 to 4.2), you can use the following procedure to upgrade TORQUE without taking your nodes offline.
Because TORQUE version 4.1.4 changed the way that pbs_server communicates with the MOMs, it is not recommended that you perform a rolling upgrade of TORQUE from version 4.1.3 to 4.1.4.

To perform a rolling upgrade in TORQUE

1. Enable the **enablemomrestart** on page 157 flag on the MOMs you want to upgrade. The **enablemomrestart** option causes a MOM to check if its binary has been updated and restart itself at a safe point when no jobs are running. You can enable this in the MOM configuration file, but it is recommended that you use **momctl** instead.

   ```bash
   momctl -q enablemomrestart=1 -h :ALL
   # The enablemomrestart flag is enabled on all nodes.
   ```

2. Replace the **pbs_mom** binary, located in */usr/local/bin* by default. **pbs_mom** will continue to run uninterrupted because the *pbs_mom* binary has already been loaded in RAM.

   ```bash
   torque-package-mom-linux-x86_64.sh --install
   # The next time pbs_mom is in an idle state, it will check for changes in the binary. If pbs_mom detects that the binary on disk has changed, it will restart automatically, causing the new pbs_mom version to load.
   ```

   After the **pbs_mom** restarts on each node, the **enablemomrestart** parameter will be set back to false (0) for that node.

   If you have a cluster with high utilization, you may find that the nodes never enter an idle state so **pbs_mom** never restarts. When this occurs, you must manually take the nodes offline and wait for the running jobs to complete before restarting **pbs_mom**. To set the node to an offline state, which will allow running jobs to complete but will not allow any new jobs to be scheduled on that node, use **pbsnodes -o <nodeName>**. After the new MOM has started, you must make the node active again by running **pbsnodes -c <nodeName>**.
Large cluster considerations

TORQUE has enhanced much of the communication found in the original OpenPBS project. This has resulted in a number of key advantages including support for:

- larger clusters.
- more jobs.
- larger jobs.
- larger messages.

In most cases, enhancements made apply to all systems and no tuning is required. However, some changes have been made configurable to allow site specific modification. The configurable communication parameters are: node_check_rate, node_ping_rate, and tcp_timeout.

For details, see these topics:

- Scalability guidelines on page 273
- End user command caching on page 274
- Moab and TORQUE configuration for large clusters on page 276
- Starting TORQUE in large environments on page 277
- Other considerations on page 278

Scalability guidelines

In very large clusters (in excess of 1,000 nodes), it may be advisable to tune a number of communication layer timeouts. By default, PBS MOM daemons timeout on inter-MOM messages after 60 seconds. In TORQUE 1.1.0p5 and higher, this can be adjusted by setting the timeout parameter in the mom_priv/config file (see, Node manager (MOM) configuration on page 245). If 15059 errors (cannot receive message from sisters) are seen in the MOM logs, it may be necessary to increase this value.

Client-to-server communication timeouts are specified via the tcp_timeout server option using the qmgr command.

On some systems, ulimit values may prevent large jobs from running. In particular, the open file descriptor limit (i.e., ulimit -n) should be set to at least the maximum job size in procs + 20. Further, there may be value in setting the fs.file-max in sysctl.conf to a high value, such as:

/etc/sysctl.conf:
f$s.file-max = 65536

Large cluster considerations
End user command caching

qstat

In a large system, users may tend to place excessive load on the system by manual or automated use of resource manager end user client commands. A simple way of reducing this load is through the use of client command wrappers which cache data. The example script below will cache the output of the command `qstat -e` for 60 seconds and report this info to end users.
#!/bin/sh

# USAGE: qstat @

CMDPATH=/usr/local/bin/qstat
CACHETIME=60
TMPFILE=/tmp/qstat.f.tmp

if [ "$1" != "-f" ]; then
    echo "direct check (arg1=$1)"
    CMDPATH $1 $2 $3 $4
    exit $?;
fi

if [ -n "$2" ]; then
    echo "direct check (arg2=$2)"
    CMDPATH $1 $2 $3 $4
    exit $?;
fi

if [ -f $TMPFILE ]; then
    TMPFILEMTIME=`stat -c %Z $TMPFILE`
else
    TMPFILEMTIME=0
fi

NOW=`date +%s`
AGE=$((NOW - $TMPFILEMTIME))

#echo AGE=$AGE

for i in 1 2 3;do
    if [ "$AGE" -gt $CACHETIME ]; then
        #echo "cache is stale"
        if [ -f $TMPFILE.1 ]; then
            #echo someone else is updating cache
            sleep 5
            NOW=`date +%s`
            TMPFILEMTIME=`stat -c %Z $TMPFILE`
        else
            break;
        fi
    fi
done

if [ -f $TMPFILE.1 ]; then
    #echo someone else is hung
    rm $TMPFILE.1
fi

if [ "$AGE" -gt $CACHETIME ]; then
    #echo updating cache
    CMDPATH -f > $TMPFILE.1
    mv $TMPFILE.1 $TMPFILE
fi

#echo "using cache"
The above script can easily be modified to cache any command and any combination of arguments by changing one or more of the following attributes:

- script name
- value of $CMDPATH
- value of $CACHETIME
- value of $TMPFILE

For example, to cache the command `pbsnodes -a`, make the following changes:

- Move original `pbsnodes` command to `pbsnodes.orig`.
- Save the script as `pbsnodes`.
- Change $CMDPATH to `pbsnodes.orig`.
- Change $TMPFILE to `/tmp/pbsnodes.a.tmp`.

**Related topics**

- [Large cluster considerations](#) on page 273

---

**Moab and TORQUE configuration for large clusters**

There are a few basic configurations for Moab and TORQUE that can potentially improve performance on large clusters.

**Moab configuration**

In the `moab.cfg` file, add:

1. `RMPOLLINTERVAL 30,30` - This sets the minimum and maximum poll interval to 30 seconds.
2. `RMCFG[<name>] FLAGS=ASYNCSTART` - This tells Moab not to block until it receives a confirmation that the job starts.
3. `RMCFG[<name>] FLAGS=ASYNCDELETE` - This tells Moab not to block until it receives a confirmation that the job was deleted.

**TORQUE configuration**

1. Follow the [Starting TORQUE in large environments](#) recommendations.
2. Increase `job_start_timeout` on `pbs_server`. The default is 300 (5 minutes), but for large clusters the value should be changed to something like 600 (10 minutes). Sites running very large parallel jobs might want to set this value even higher.
3. Use a node health check script on all MOM nodes. This helps prevent jobs from being scheduled on bad nodes and is especially helpful for large parallel jobs.

4. Make sure that `ulimit -n` (maximum file descriptors) is set to `unlimited`, or a very large number, and not the default.

5. For clusters with a high job throughput it is recommended that the server parameter `max_threads` be increased from the default. The default is `(2 * number of cores + 1) * 10`.

Related topics

- [Large cluster considerations on page 273](#)

### Starting TORQUE in large environments

If running TORQUE in a large environment, use these tips to help TORQUE start up faster.

**Fastest possible start up**

1. Create a [MOM hierarchy](#), even if your environment has a one-level MOM hierarchy (meaning all MOMs report directly to `pbs_server`), and copy the file to the `mom_priv` directory on the MOMs.

2. Start `pbs_server` with the `-n` option. This specifies that `pbs_server` won’t send the hierarchy to the MOMs unless a MOM requests it.

3. Start the MOMs normally.

**If no daemons are running**

1. Start `pbs_server` with the `-c` option.

2. Start the MOMs without the `-w` option.

**If MOMs are running and just restarting `pbs_server`**

1. Start `pbs_server` without the `-c` option.

**If restarting a MOM or all MOMs**

1. Start `pbs_server` without the `-w` option. Starting it with `-w` causes the MOMs to appear to be down.

Related topics

- [Large cluster considerations on page 273](#)
Other considerations

job_stat_rate

In a large system, there may be many users, many jobs, and many requests for information. To speed up response time for users and for programs using the API the job_stat_rate can be used to tweak when the pbs_server daemon will query MOMs for job information. By increasing this number, a system will not be constantly querying job information and causing other commands to block.

poll_jobs

The poll_jobs parameter allows a site to configure how the pbs_server daemon will poll for job information. When set to TRUE, the pbs_server will poll job information in the background and not block on user requests. When set to FALSE, the pbs_server may block on user requests when it has stale job information data. Large clusters should set this parameter to TRUE.

Internal settings

On large, slow, and/or heavily loaded systems, it may be desirable to increase the pbs_tcp_timeout setting used by the pbs_mom daemon in MOM-to-MOM communication. This setting defaults to 20 seconds and requires rebuilding code to adjust. For client-server based communication, this attribute can be set using the qmgr command. For MOM-to-MOM communication, a source code modification is required. To make this change, edit the $TORQUEBUILDDIR/src/lib/Libifl/tcp_dis.c file and set pbs_tcp_timeout to the desired maximum number of seconds allowed for a MOM-to-MOM request to be serviced.

A system may be heavily loaded if it reports multiple 'End of File from addr' or 'Premature end of message' failures in the pbs_mom or pbs_server logs.

Scheduler settings

If using Moab, there are a number of parameters which can be set on the scheduler which may improve TORQUE performance. In an environment containing a large number of short-running jobs, the JOBAGGREGATIONTIME parameter (see the "Parameters" section of the Moab Workload Manager Administrator Guide) can be set to reduce the number of workload and resource queries performed by the scheduler when an event based interface is enabled. If the pbs_server daemon is heavily loaded and PBS API timeout errors (i.e. "Premature end of message") are reported within the scheduler, the "TIMEOUT" attribute of the RMCFG parameter may be set with a value of between 30 and 90 seconds.

File system

TORQUE can be configured to disable file system blocking until data is physically written to the disk by using the --disable-fsync argument with configure. While having fsync enabled is more reliable, it may lead to server delays for sites with either a larger number of nodes, or a large number of jobs. Fsync is enabled by default.
Network ARP cache

For networks with more than 512 nodes it is mandatory to increase the kernel’s internal ARP cache size. For a network of ~1000 nodes, we use these values in /etc/sysctl.conf on all nodes and servers:

```
/etc/sysctl.conf
net.ipv4.neigh.default.gc_thresh3 = 4096
# Don't allow the arp table to become bigger than this
net.ipv4.neigh.default.gc_thresh2 = 2048
# Tell the gc when to become aggressive with arp table cleaning.
net.ipv4.neigh.default.gc_thresh1 = 1024
# Adjust this based on size of the LAN.
net.ipv4.neigh.default.gc_interval = 3600
# Adjust where the gc will leave arp table alone
net.ipv4.neigh.default.gc_stale_time = 3600
# ARP cache entry timeout
```

Use `sysctl -p` to reload this file.

The ARP cache size on other Unixes can presumably be modified in a similar way.

An alternative approach is to have a static /etc/ethers file with all hostnames and MAC addresses and load this by `arp -f /etc/ethers`. However, maintaining this approach is quite cumbersome when nodes get new MAC addresses (due to repairs, for example).

Related topics

- [Large cluster considerations](#) on page 273
Prologue and epilogue scripts

TORQUE provides administrators the ability to run scripts before and/or after each job executes. With such a script, a site can prepare systems, perform node health checks, prepend and append text to output and error log files, cleanup systems, and so forth.

The following table shows which MOM runs which script. All scripts must be in the TORQUE_HOME/mom_priv/ directory and be available on every compute node. "Mother Superior," as referenced in the following table, is the pbs_mom on the first node allocated, and the term "Sisters" refers to pbs_moms, although note that a Mother Superior is also a sister node.

<table>
<thead>
<tr>
<th>Script</th>
<th>Execution location</th>
<th>Execute as</th>
<th>File permissions</th>
</tr>
</thead>
<tbody>
<tr>
<td>prologue</td>
<td>Mother Superior</td>
<td>root</td>
<td>Readable and executable by root and NOT writable by anyone but root (e.g., -r-x-----)</td>
</tr>
<tr>
<td>epilogue</td>
<td>root</td>
<td>root</td>
<td>Readable and executable by root and other (e.g., -r-x-- -r-x)</td>
</tr>
<tr>
<td>prologue.user</td>
<td>user</td>
<td>user</td>
<td>Readable and executable by root and other (e.g., -r-x-- -r-x)</td>
</tr>
<tr>
<td>epilogue.user</td>
<td>root</td>
<td>user</td>
<td>Readable and executable by root and other (e.g., -r-x-- -r-x)</td>
</tr>
<tr>
<td>prologue.parallel</td>
<td>Sister</td>
<td>root</td>
<td>Readable and executable by root and NOT writable by anyone but root (e.g., -r-x-----)</td>
</tr>
<tr>
<td>epilogue.parallel</td>
<td>root</td>
<td>root</td>
<td>Readable and executable by root and other (e.g., -r-x-- -r-x)</td>
</tr>
<tr>
<td>prologue.user.parallel</td>
<td>user</td>
<td>user</td>
<td>Readable and executable by root and other (e.g., -r-x-- -r-x)</td>
</tr>
<tr>
<td>epilogue.user.parallel</td>
<td>user</td>
<td>user</td>
<td>Readable and executable by root and other (e.g., -r-x-- -r-x)</td>
</tr>
<tr>
<td>epilogue.precancel</td>
<td>Mother Superior</td>
<td>user</td>
<td>Readable and executable by root and other (e.g., -r-x-- -r-x)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The execution directory for each script is TORQUE_HOME/mom_priv/.
This section contains these topics:

- **Script order of execution** on page 282
- **Script environment** on page 282
- **Per job prologue and epilogue scripts** on page 284
- **Prologue and epilogue scripts time out** on page 285
- **Prologue error processing** on page 285

### Script order of execution

When jobs start, the order of script execution is prologue followed by prologue.user. On job exit, the order of execution is epilogue.user followed by epilogue unless a job is canceled. In that case, epilogue.precancel is executed first. epilogue.parallel is executed only on the Sister nodes when the job is completed.

- The epilogue and prologue scripts are controlled by the system administrator. However, beginning in TORQUE version 2.4 a user epilogue and prologue script can be used on a per job basis. (See **Per job prologue and epilogue scripts** on page 284 for more information.)

- Root squashing is now supported for epilogue and prologue scripts.

**Related topics**

- **Prologue and epilogue scripts** on page 281

### Script environment

The prologue and epilogue scripts can be very simple. On most systems, the script must declare the execution shell using the `#!<SHELL>` syntax (for example, `#!/bin/sh`). In addition, the script may want to process context sensitive arguments passed by TORQUE to the script.

**Prologue Environment**

The following arguments are passed to the prologue, prologue.user, and prologue.parallel scripts:

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>argv[1]</td>
<td>job id</td>
</tr>
<tr>
<td>Argument</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>argv[2]</td>
<td>job execution user name</td>
</tr>
<tr>
<td>argv[3]</td>
<td>job execution group name</td>
</tr>
<tr>
<td>argv[4]</td>
<td>job name (TORQUE 1.2.0p4 and higher only)</td>
</tr>
<tr>
<td>argv[5]</td>
<td>list of requested resource limits (TORQUE 1.2.0p4 and higher only)</td>
</tr>
<tr>
<td>argv[6]</td>
<td>job execution queue (TORQUE 1.2.0p4 and higher only)</td>
</tr>
<tr>
<td>argv[7]</td>
<td>job account (TORQUE 1.2.0p4 and higher only)</td>
</tr>
</tbody>
</table>

**Epilogue Environment**

TORQUE supplies the following arguments to the `epilogue`, `epilogue.user`, `epilogue.precancel`, and `epilogue.parallel` scripts:

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>argv[1]</td>
<td>job id</td>
</tr>
<tr>
<td>argv[2]</td>
<td>job execution user name</td>
</tr>
<tr>
<td>argv[3]</td>
<td>job execution group name</td>
</tr>
<tr>
<td>argv[4]</td>
<td>job name</td>
</tr>
<tr>
<td>argv[5]</td>
<td>session id</td>
</tr>
<tr>
<td>argv[6]</td>
<td>list of requested resource limits</td>
</tr>
<tr>
<td>argv[7]</td>
<td>list of resources used by job</td>
</tr>
<tr>
<td>argv[8]</td>
<td>job execution queue</td>
</tr>
<tr>
<td>argv[9]</td>
<td>job account</td>
</tr>
<tr>
<td>argv[10]</td>
<td>job exit code</td>
</tr>
</tbody>
</table>
The `epilogue.precancel` script is run after a job cancel request is received by the MOM and before any signals are sent to job processes. If this script exists, it is run whether the canceled job was active or idle.

### Related topics
- [Prologue and epilogue scripts](#) on page 281

## Per job prologue and epilogue scripts

TORQUE supports per job prologue and epilogue scripts when using the `qsub` `-l` option. The syntax is:

```
qsub -l prologue=<prologue_script_path> epilogue=<epilogue_script_path> <script>.
```

The path can be either relative (from the directory where the job is submitted) or absolute. The files must be owned by the user with at least execute and read privileges, and the permissions must not be writeable by group or other.

```
/home/usertom/dev/  
-r-x------ 1 usertom usertom 2009-11-09 16:11 prologue_script.sh  
-r-x------ 1 usertom usertom 2009-11-09 16:11 epilogue_script.sh
```

### Example G-1:

```
$ qsub -l prologue=/home/usertom/dev/prologue_  
  script.sh, epilogue=/home/usertom/dev/epilogue_script.sh job14.pl
```

This job submission executes the `prologue` script first. When the `prologue` script is complete, `job14.pl` runs. When `job14.pl` completes, the `epilogue` script is executed.
Prologue and epilogue scripts time out

TORQUE takes preventative measures against prologue and epilogue scripts by placing an alarm around the scripts execution. By default, TORQUE sets the alarm to go off after 5 minutes of execution. If the script exceeds this time, it will be terminated and the node will be marked down. This timeout can be adjusted by setting the \texttt{prologalarm} parameter in the \texttt{mom\_priv/config} file.

\begin{itemize}
  \item While TORQUE is executing the \texttt{epilogue}, \texttt{epilogue.user}, or \texttt{epilogue.precancel} scripts, the job will be in the \texttt{E} (exiting) state.
\end{itemize}

If an \texttt{epilogue.parallel} script cannot open the \texttt{.O}U or \texttt{.E}R files, an error is logged but the script is continued.

Related topics

- \texttt{Prologue and epilogue scripts} on page 281

Prologue error processing

If the \texttt{prologue} script executes successfully, it should exit with a zero status. Otherwise, the script should return the appropriate error code as defined in the table below. The \texttt{pbs\_mom} will report the script’s exit status to \texttt{pbs\_server} which will in turn take the associated action. The following table describes each exit code for the prologue scripts and the action taken.

\begin{table}[h]
\centering
\begin{tabular}{|c|p{8cm}|c|}
\hline
Error & Description & Action \\
\hline
-4 & The script timed out & Job will be requeued \\
\hline
-3 & The wait(2) call returned an error & Job will be requeued \\
\hline
-2 & Input file could not be opened & Job will be requeued \\
\hline
-1 & Permission error \newline (script is not owned by root, or is writable by others) & Job will be requeued \\
\hline
0 & Successful completion & Job will run \\
\hline
1 & Abort exit code & Job will be aborted \\
\hline
>1 & other & Job will be requeued \\
\hline
\end{tabular}
\end{table}
Example G-2:

Following are example prologue and epilogue scripts that write the arguments passed to them in the job's standard out file:

<table>
<thead>
<tr>
<th>prologue</th>
<th>Script</th>
<th>stdout</th>
</tr>
</thead>
</table>
|          | `#!/bin/sh` | Prologue Args:  
|          | `echo "Prologue Args:"` | Job ID: 13724.node01  
|          | `echo "Job ID: $1"` | User ID: user1  
|          | `echo "User ID: $2"` | Group ID: user1  
|          | `echo "Group ID: $3"` |  
|          | `echo ""` |  
|          | `exit 0` |  

<table>
<thead>
<tr>
<th>epilogue</th>
<th>Script</th>
<th>stdout</th>
</tr>
</thead>
</table>
|          | `#!/bin/sh` | Prologue Args:  
|          | `echo "Epilogue Args:"` | Job ID: 13724.node01  
|          | `echo "Job ID: $1"` | User ID: user1  
|          | `echo "User ID: $2"` | Group ID: user1  
|          | `echo "Group ID: $3"` | Job Name: script.sh  
|          | `echo "Job Name: $4"` | Session ID: 28244  
|          | `echo "Session ID: $5"` | Resource List: neednodes=node01,nodes=1,walltime=00:01:00  
|          | `echo "Resource List: $6"` | Resources Used: cput=00:00:00,mem=0kb,vmem=0kb,walltime=00:00:07  
|          | `echo "Resources Used: $7"` | Queue Name: batch  
|          | `echo "Queue Name: $8"` | Account String:  
|          | `echo """"""` |  
|          | `exit 0` |  

Prologue and epilogue scripts
Example G-3:

The Ohio Supercomputer Center contributed the following scripts:

"prologue creates a unique temporary directory on each node assigned to a job before the job begins to run, and epilogue deletes that directory after the job completes.

```
prologue
#!/bin/sh
# Create TMPDIR on all the nodes
# Copyright 1999, 2000, 2001 Ohio Supercomputer Center
# prologue gets 3 arguments:
# 1 -- jobid
# 2 -- userid
# 3 -- grpdir
# jobid=$1
# user=$2
# group=$3
nodefile=/var/spool/pbs/aux/$jobid
if [ -r $nodefile ]; then
    nodes=$(sort $nodefile | uniq)
else
    nodes=localhost
fi
tmp=/tmp/pbstmp.$jobid
for i in $nodes ; do
    ssh $i mkdir -m 700 $tmp && chown $user.$group $tmp
done
exit 0

epilogue
#!/bin/sh
# Clear out TMPDIR
# Copyright 1999, 2000, 2001 Ohio Supercomputer Center
# epilogue gets 9 arguments:
# 1 -- jobid
# 2 -- userid
# 3 -- grpdir
# 4 -- job name
# 5 -- sessionid
# 6 -- resource limits
# 7 -- resources used
# 8 -- queue
# 9 -- account
# jobid=$1
nodefile=/var/spool/pbs/aux/$jobid
if [ -r $nodefile ]; then
    nodes=$(sort $nodefile | uniq)
else
    nodes=localhost
fi
tmp=/tmp/pbstmp.$jobid
for i in $nodes ; do
    ssh $i rm -rf $tmp
done
exit 0
```
prologue, prologue.user, and prologue.parallel scripts can have dramatic effects on job scheduling if written improperly.

Related topics

- Prologue and epilogue scripts on page 281
Running multiple TORQUE servers and MOMs on the same node

TORQUE can be configured to allow multiple servers and MOMs to run on the same node. This example will show how to configure, compile and install two different TORQUE servers and MOMs on the same node. For details, see these topics:

- **Configuring the first TORQUE** on page 289
- **Configuring the second TORQUE** on page 289
- **Bringing the first TORQUE server online** on page 289
- **Bringing the second TORQUE server online** on page 290

Configuring the first TORQUE

```
./configure --with-server-home=/usr/spool/PBS1 --bindir=/usr/spool/PBS1/bin --
            sbindir=/usr/spool/PBS1/sbin
```

Then make and make install will place the first TORQUE into `/usr/spool/PBS1` with the executables in their corresponding directories.

Configuring the second TORQUE

```
./configure --with-server-home=/usr/spool/PBS2 --bindir=/usr/spool/PBS2/bin --
            sbindir=/usr/spool/PBS2/sbin
```

Then make and make install will place the second TORQUE into `/usr/spool/PBS2` with the executables in their corresponding directories.

Bringing the first TORQUE server online

Each command, including pbs_server and pbs_mom, takes parameters indicating which servers and ports to connect to or listen on (when appropriate). Each of these is documented in their corresponding man pages (configure with --enable-docs).

In this example the first TORQUE server will accept batch requests on port 35000, communicate with the MOMs on port 35001, and communicate via RPP on port 35002. The first TORQUE MOM will try to connect to the server on port 35000, it will listen for requests from the server on port 35001 and will communicate via RPP on port 35002. (Each of these command arguments is discussed in further details on the corresponding man page. In particular, -t create is only used the first time a server is run.)

```
> pbs_server -p 35000 -M 35001 -R 35002 -t create
> pbs_mom -S 35000 -M 35001 -R 35002
```
Afterwards, when using a client command to make a batch request it is necessary to specify the server name and server port (35000):

```bash
> pbsnodes -a -s node01:35000
```

Submitting jobs can be accomplished using the `-q` option `[@queue][@host[@port]]`:

```bash
> qsub -q @node01:35000 /tmp/script.pbs
```

Bringing the second TORQUE server online

In this example the second TORQUE server will accept batch requests on port 36000, communicate with the MOMS on port 36002, and communicate via RPP on port 36002. The second TORQUE MOM will try to connect to the server on port 36000, it will listen for requests from the server on port 36001 and will communicate via RPP on port 36002.

```bash
> pbs_server -p 36000 -M 36001 -R 36002 -t create
> pbs_mom -S 36000 -M 36001 -R 36002
```

Afterward, when using a client command to make a batch request it is necessary to specify the server name and server port (36002):

```bash
> pbsnodes -a -s node01:36000
> qsub -q @node01:36000 /tmp/script.pbs
```
Security overview

The authorization model for TORQUE changed in version 4.0.0 from pbs_iff to a daemon called trqauthd. The job of the trqauthd daemon is the same as pbs_iff. The difference is that trqauthd is a resident daemon whereas pbs_iff is invoked by each client command. pbs_iff is not scalable and is prone to failure under even small loads. trqauthd is very scalable and creates the possibility for better security measures in the future.

trqauthd and pbs_iff authorization theory

The key to security of both trqauthd and pbs_iff is the assumption that any host which has been added to the TORQUE cluster has been secured by the administrator. Neither trqauthd nor pbs_iff do authentication. They only do authorization of users. Given that the host system is secure the following is the procedure by which trqauthd and pbs_iff authorize users to pbs_server.

1. Client utility makes a connection to pbs_server on a dynamic port.
2. Client utility sends a request to trqauthd with the user name and port.
3. trqauthd verifies the user ID and then sends a request to pbs_server on a privileged port with the user ID and dynamic port to authorize the connection.
4. trqauthd reports results of the server to client utility.

Both trqauthd and pbs_iff use Unix domain sockets for communication from the client utility. Unix domain sockets have the ability to verify that a user is who they say they are by using security features that are part of the file system.
When a "submit filter" exists, TORQUE will send the command file (or contents of STDIN if piped to qsub) to that script/executable and allow it to evaluate the submitted request based on specific site policies. The resulting file is then handed back to qsub and processing continues. Submit filters can check user jobs for correctness based on site policies. They can also modify user jobs as they are submitted. Some examples of what a submit filter might evaluate and check for are:

- Memory Request - Verify that the job requests memory and rejects if it does not.
- Job event notifications - Check if the job does one of the following and rejects it if it:
  o explicitly requests no notification.
  o requests notifications but does not provide an email address.
- Walltime specified - Verify that the walltime is specified.
- Global Walltime Limit - Verify that the walltime is below the global max walltime.
- Test Walltime Limit - If the job is a test job, this check rejects the job if it requests a walltime longer than the testing maximum.

The script below reads the original submission request from STDIN and shows how you could insert parameters into a job submit request:

```
#!/bin/sh
# add default memory constraints to all requests
# that did not specify it in user's script or on command line
echo "#PBS -l mem=16MB"
while read i
do
echo $i
done
```

The same command line arguments passed to qsub will be passed to the submit filter and in the same order. It should be noted that as of TORQUE 2.2.0 extended attributes are not passed to the filter. Exit status of -1 will cause qsub to reject the submission with a message stating that it failed due to administrative policies.

The "submit filter" must be executable, must be available on each of the nodes where users may submit jobs, and by default must be located at ${libexecdir}/qsub_filter (for version 2.1 and older: /usr/local/sbin/torque_submitfilter). At run time, if the file does not exist at this new preferred path then qsub will fall back to the old hard-coded path. The submit filter location can be customized by setting the SUBMITFILTER parameter inside the file (see "torque.cfg" configuration file on page 295), as in the following example:

```
torque.cfg:
```

Job submission filter ("qsub wrapper")

When a "submit filter" exists, TORQUE will send the command file (or contents of STDIN if piped to qsub) to that script/executable and allow it to evaluate the submitted request based on specific site policies. The resulting file is then handed back to qsub and processing continues. Submit filters can check user jobs for correctness based on site policies. They can also modify user jobs as they are submitted. Some examples of what a submit filter might evaluate and check for are:

- Memory Request - Verify that the job requests memory and rejects if it does not.
- Job event notifications - Check if the job does one of the following and rejects it if it:
  - explicitly requests no notification.
  - requests notifications but does not provide an email address.
- Walltime specified - Verify that the walltime is specified.
- Global Walltime Limit - Verify that the walltime is below the global max walltime.
- Test Walltime Limit - If the job is a test job, this check rejects the job if it requests a walltime longer than the testing maximum.

The script below reads the original submission request from STDIN and shows how you could insert parameters into a job submit request:

```
#!/bin/sh
# add default memory constraints to all requests
# that did not specify it in user's script or on command line
echo "#PBS -l mem=16MB"
while read i
do
echo $i
done
```

The same command line arguments passed to qsub will be passed to the submit filter and in the same order. It should be noted that as of TORQUE 2.2.0 extended attributes are not passed to the filter. Exit status of -1 will cause qsub to reject the submission with a message stating that it failed due to administrative policies.

The "submit filter" must be executable, must be available on each of the nodes where users may submit jobs, and by default must be located at ${libexecdir}/qsub_filter (for version 2.1 and older: /usr/local/sbin/torque_submitfilter). At run time, if the file does not exist at this new preferred path then qsub will fall back to the old hard-coded path. The submit filter location can be customized by setting the SUBMITFILTER parameter inside the file (see "torque.cfg" configuration file on page 295), as in the following example:

```
torque.cfg:
```
SUBMITFILTER /opt/torque/submit.pl
...

Initial development courtesy of Oak Ridge National Laboratories.
## "torque.cfg" configuration file

### CLIENTRETRY

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INT&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>0</td>
</tr>
<tr>
<td>Description</td>
<td>Seconds between retry attempts to talk to pbs_server.</td>
</tr>
</tbody>
</table>
| Example  | CLIENTRETRY 10

TORQUE waits 10 seconds after a failed attempt before it attempts to talk to pbs_server again.

### DEFAULTCKPT

<table>
<thead>
<tr>
<th>Format</th>
<th>One of None, Enabled, Shutdown, Periodic, Interval=minutes, depth=number, or dir=path</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>None</td>
</tr>
<tr>
<td>Description</td>
<td>Default value for job’s checkpoint attribute. For a description of all possible values, see the qsub -c documentation.</td>
</tr>
</tbody>
</table>

This default setting can be overridden at job submission with the qsub -c option.

### FAULT_TOLERANT_BY_DEFAULT

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;BOOLEAN&gt;</th>
</tr>
</thead>
</table>
### FAULT_TOLERANT_BY_DEFAULT

<table>
<thead>
<tr>
<th>Default</th>
<th>FALSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Sets all jobs to fault tolerant by default. (See <code>qsub -h</code> for more information on fault tolerance.)</td>
</tr>
<tr>
<td>Example</td>
<td>FAULT_TOLERANT_BY_DEFAULT TRUE</td>
</tr>
</tbody>
</table>

**Example:**

```
FAULT_TOLERANT_BY_DEFAULT TRUE
```

Jobs are fault tolerant by default. They will not be canceled based on failed polling, no matter how many nodes fail to report.

### HOST_NAME_SUFFIX

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>---</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies a hostname suffix. When <code>qsub</code> submits a job, it also submits the username of the submitter and the name of the host from which the user submitted the job. TORQUE appends the value of HOST_NAME_SUFFIX to the hostname. This is useful for multi-homed systems that may have more than one name for a host.</td>
</tr>
<tr>
<td>Example</td>
<td>HOST_NAME_SUFFIX -ib</td>
</tr>
</tbody>
</table>

**Example:**

```
HOST_NAME_SUFFIX -ib
```

When a job is submitted, the `-ib` suffix is appended to the host name.

### QSUBHOST

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;HOSTNAME&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>---</td>
</tr>
<tr>
<td>Description</td>
<td>The hostname given as the argument of this option will be used as the PBS_O_HOST variable for job submissions. By default, PBS_O_HOST is the hostname of the submission host. This option allows administrators to override the default hostname and substitute a new name.</td>
</tr>
<tr>
<td>Example</td>
<td>QSUBHOST host1</td>
</tr>
</tbody>
</table>

**Example:**

```
QSUBHOST host1
```

The default hostname associated with a job is `host1`. 
### QSUBSENDUID

<table>
<thead>
<tr>
<th>Format</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>---</td>
</tr>
<tr>
<td>Description</td>
<td>Integer for job’s PBS_OUID variable. Specifying the parameter name anywhere in the config file enables the feature. Removing the parameter name disables the feature.</td>
</tr>
<tr>
<td>Example</td>
<td>QSUBSENDUID</td>
</tr>
</tbody>
</table>

TORQUE assigns a unique ID to a job when it is submitted by `qsub`.

### QSUBSLEEP

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;INT&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>0</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies time, in seconds, to sleep between a user’s submitting and TORQUE’s starting a <code>qsub</code> command. Used to prevent users from overwhelming the scheduler.</td>
</tr>
<tr>
<td>Example</td>
<td>QSUBSLEEP 2</td>
</tr>
</tbody>
</table>

When a job is submitted with `qsub`, it will sleep for 2 seconds.

### RERUNNABLEBYDEFAULT

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;BOOLEAN&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>TRUE</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies if a job is re-runnable by default. Setting this to false causes the re-runnable attribute value to be false unless the users specifies otherwise with the <code>qsub -r</code> option. (New in TORQUE 2.4.)</td>
</tr>
<tr>
<td>Example</td>
<td>RERUNNABLEBYDEFAULT FALSE</td>
</tr>
</tbody>
</table>

By default, `qsub` jobs cannot be rerun.
### SERVERHOST

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>localhost</td>
</tr>
<tr>
<td>Description</td>
<td>If set, the <code>qsub</code> on page 204 command will open a connection to the host specified by the <code>SERVERHOST</code> string.</td>
</tr>
</tbody>
</table>
| Example  | `SERVERHOST orion15`  
```bash
The server will open socket connections and and communicate using serverhost orion15.
``` |

### SUBMITFILTER

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td><code>${libexecdir}/qsub_filter</code> (for version 2.1 and older: <code>/usr/local/sbin/torque_submitfilter</code>)</td>
</tr>
<tr>
<td>Description</td>
<td>Specifies the location of the submit filter (see <code>Job submission filter (&quot;qsub wrapper&quot;)</code> on page 293) used to pre-process job submission.</td>
</tr>
</tbody>
</table>
| Example  | `SUBMITFILTER /usr/local/sbin/qsub_filter`  
```bash
The location of the submit filter is specified as /usr/local/sbin/qsub_filter.
``` |

### TRQ_IFNAME

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;STRING&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>null</td>
</tr>
<tr>
<td>Description</td>
<td>Allows you to specify a specific network interface to use for outbound TORQUE requests. The string is the name of a network interface, such as <code>eth0</code> or <code>eth1</code>, depending on which interface you want to use.</td>
</tr>
</tbody>
</table>
| Example  | `TRQ_IFNAME eth1`  
```bash
Outbound TORQUE requests are handled by eth1.
``` |
### VALIDATEGROUP

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;BOOLEAN&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>FALSE</td>
</tr>
<tr>
<td>Description</td>
<td>Validate submit user’s group on <code>qsub</code> commands. For TORQUE builds released after 2/8/2011, <code>VALIDATEGROUP</code> also checks any groups requested in <code>group_list</code> at the submit host. Set <code>VALIDATEGROUP</code> to &quot;TRUE&quot; if you set <code>disable_server_id_check</code> to TRUE.</td>
</tr>
</tbody>
</table>
| Example      | `VALIDATEGROUP TRUE

qsub verifies the submitter’s group ID.` |

### VALIDATEPATH

<table>
<thead>
<tr>
<th>Format</th>
<th>&lt;BOOLEAN&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>TRUE</td>
</tr>
<tr>
<td>Description</td>
<td>Validate local existence of ‘~d’ working directory.</td>
</tr>
</tbody>
</table>
| Example      | `VALIDATEPATH FALSE

qsub does not validate the path.` |
TORQUE Quick Start Guide

Initial installation

TORQUE is now hosted at https://github.com under the adaptivecomputing organization. To download source, you will need to use the git utility. For example:

```
[root]# git clone https://github.com/adaptivecomputing.com/torque.git -b 4.2.6 4.2.6
```

To download a different version, replace each 4.2.6 with the desired version. After downloading a copy of the repository, you can list the current branches by typing `git branch --a` from within the directory of the branch you cloned.

If you're checking source out from git, read the README.building-40 file in the repository.

Extract and build the distribution on the machine that will act as the "TORQUE server" - the machine that will monitor and control all compute nodes by running the pbs_server daemon. See the example below:

```
> tar -xzf torque.tar.gz
> cd torque
> ./configure
> make
> make install
```

OSX 10.4 users need to change the `#define _TDARWIN in src/include/pbs_config.h to #define _TDARWIN_8`.

After installation, verify you have PATH environment variables configured for /usr/local/bin/ and /usr/local/sbin/. Client commands are installed to /usr/local/bin and server binaries are installed to /usr/local/sbin.

In this document, TORQUE_HOME corresponds to where TORQUE stores its configuration files. The default is /var/spool/torque.

Initialize/Configure TORQUE on the server (pbs_server)

- Once installation on the TORQUE server is complete, configure the pbs_server daemon by executing the command `torque.setup <USER>` found packaged with the distribution source code, where `<USER>` is a username that will act as the TORQUE admin. This script will set up a
basic batch queue to get you started. If you experience problems, make sure that the most recent TORQUE executables are being executed, or that the executables are in your current PATH.

If you are upgrading from TORQUE 2.5.9, run pbs_server -u before running torque.setup.

```
[root]# pbs_server -u
```

- If doing this step manually, be certain to run the command `pbs_server -t create` to create the new batch database. If this step is not taken, the pbs_server daemon will be unable to start.
- Proper server configuration can be verified by following the steps listed in Testing server configuration.

Install TORQUE on the compute nodes

To configure a compute node do the following on each machine (see page 19, Section 3.2.1 of PBS Administrator's Manual for full details):

- Create the self-extracting, distributable packages with make packages (See the INSTALL file for additional options and features of the distributable packages) and use the parallel shell command from your cluster management suite to copy and execute the package on all nodes (i.e. xCAT users might do `prcp torque-package-linux-i686.sh main:/tmp/; psh main /tmp/torque-package-linux-i686.sh --install`). Optionally, distribute and install the clients package.

Configure TORQUE on the compute nodes

- For each compute host, the MOM daemon must be configured to trust the pbs_server daemon. In TORQUE 2.0.0p4 and earlier, this is done by creating the `<TorqueHome>/mom_priv/config` file and setting the `Sbpsserver` parameter. In TORQUE 2.0.0p5 and later, this can also be done by creating the `<TorqueHome>/server_name_name` file and placing the server hostname inside.
- Additional config parameters may be added to `<TorqueHome>/mom_priv/config` (see Node manager (MOM) configuration on page 245 for details).

Configure data management on the compute nodes

Data management allows jobs' data to be staged in/out or to and from the server and compute nodes.

- For shared filesystems (i.e., NFS, DFS, AFS, etc.) use the `$usecp` parameter in the `mom_priv/config` files to specify how to map a user's home directory.
  
  (Example: `$usecp gridmaster.tmx.com:/home /home`

- For local, non-shared filesystems, rcp or scp must be configured to allow direct copy without prompting for passwords (key authentication, etc.)

Update TORQUE server configuration

On the TORQUE server, append the list of newly configured compute nodes to the `<TorqueHome>/server_priv/nodes` file:
Start the \texttt{pbs\_mom} daemons on compute nodes

- Next start the \texttt{pbs\_mom} daemon on each compute node by running the \texttt{pbs\_mom} executable.

Run the \texttt{trqauthd} daemon to run client commands (see \texttt{Configuring trqauthd for client commands} on page 13). This enables running client commands.

Verifying correct TORQUE installation

The \texttt{pbs\_server} daemon was started on the TORQUE server when the \texttt{torque\_setup} file was executed or when it was manually configured. It must now be restarted so it can reload the updated configuration changes.

```
# shutdown server
> qterm # shutdown server
# start server
> pbs\_server
# verify all queues are properly configured
> qstat -q
# view additional server configuration
> qmgr -c 'p s'
# verify all nodes are correctly reporting
> pbsnodes -a
# submit a basic job
>echo "sleep 30" | qsub
# verify jobs display
> qstat
```

At this point, the job will not start because there is no scheduler running. The scheduler is enabled in the next step below.

Enabling the scheduler

Selecting the cluster scheduler is an important decision and significantly affects cluster utilization, responsiveness, availability, and intelligence. The default TORQUE scheduler, \texttt{pbs\_sched}, is very basic and will provide poor utilization of your cluster's resources. Other options, such as \texttt{Maui Scheduler} or \texttt{Moab Workload Manager} are highly recommended. If using Maui/Moab, refer to the Moab-PBS Integration Guide. If using \texttt{pbs\_sched}, start this daemon now.

If you are installing ClusterSuite, TORQUE and Moab were configured at installation for interoperability and no further action is required.
Startup/Shutdown service script for TORQUE/Moab (OPTIONAL)

Optional startup/shutdown service scripts are provided as an example of how to run TORQUE as an OS service that starts at bootup. The scripts are located in the contrib/init.d/ directory of the TORQUE tarball you downloaded. In order to use the script you must:

- Determine which init.d script suits your platform the best.
- Modify the script to point to TORQUE’s install location. This should only be necessary if you used a non-default install location for TORQUE (by using the --prefix option of ./configure).
- Place the script in the /etc/init.d/ directory.
- Use a tool like chkconfig to activate the start-up scripts or make symbolic links (S99moab and K15moab, for example) in desired runtimes (/etc/rc.d/rc3.d/ on Redhat, etc).

Related topics

- Advanced configuration on page 14
BLCR acceptance tests

This section contains a description of the testing done to verify the functionality of the BLCR implementation. For details, see these topics:

- **Test environment** on page 305
- **Test 1 - Basic operation** on page 305
- **Test 2 - Persistence of checkpoint images** on page 308
- **Test 3 - Restart after checkpoint** on page 308
- **Test 4 - Multiple checkpoint/restart** on page 309
- **Test 5 - Periodic checkpoint** on page 309
- **Test 6 - Restart from previous image** on page 310

Test environment

All these tests assume the following test program and shell script, test.sh.

```c
#include
int main( int argc, char *argv[] )
{
    int i;
    for (i=0; i<100; i++)
    {
        printf("i = %d\n", i);
        fflush(stdout);
        sleep(1);
    }
}
```

Related topics

- **BLCR acceptance tests** on page 305

Test 1 - Basic operation

Introduction

This test determines if the proper environment has been established.
Test steps

Submit a test job and the issue a hold on the job.

```
> qsub -c enabled test.sh
999.xxx.yyy
> qhold 999
```

Possible failures

Normally the result of `qhold` is nothing. If an error message is produced saying that `qhold` is not a supported feature then one of the following configuration errors might be present.

- The TORQUE images may have not be configured with `--enable-blcr`
- BLCR support may not be installed into the kernel with `insmod`.
- The config script in `mom_priv` may not exist with `$checkpoint_script` defined.
- The config script in `mom_priv` may not exist with `$restart_script` defined.
- The config script in `mom_priv` may not exist with `$checkpoint_run_exe` defined.
- The scripts referenced in the config file may not exist.
- The scripts referenced in the config file may not have the correct permissions.

Successful results

If no configuration was done to specify a specific directory location for the checkpoint file, the default location is off of the TORQUE directory, which in my case is `/var/spool/torque/checkpoint`.

Otherwise, go to the specified directory for the checkpoint image files. This was done by either specifying an option on job submission, i.e. `-c dir=/home/test` or by setting an attribute on the execution queue. This is done with the command `qmgr -c 'set queue batch checkpoint_dir=/home/test`.

Doing a directory listing shows the following.

```
# find /var/spool/torque/checkpoint
/var/spool/torque/checkpoint
/var/spool/torque/checkpoint/999.xxx.yyy.CK
/var/spool/torque/checkpoint/999.xxx.yyy.CK/ckpt.999.xxx.yyy.1205266630
# find /var/spool/torque/checkpoint |xargs ls -l
-r-------- 1 root root 543779 2008-03-11 14:17
/var/spool/torque/checkpoint/999.xxx.yyy.CK/ckpt.999.xxx.yyy.1205266630
/var/spool/torque/checkpoint:
  total 4
  drwxr-xr-x 2 root root 4096 2008-03-11 14:17 999.xxx.yyy.CK
/var/spool/torque/checkpoint/999.xxx.yyy.CK:
  total 536
  r-------- 1 root root 543779 2008-03-11 14:17 ckpt.999.xxx.yyy.1205266630
```

Doing a `qstat -f` command should show the job in a held state, `job_state = H`. Note that the attribute `checkpoint_name` is set to the name of the file seen above.
If a checkpoint directory has been specified, there will also be an attribute `checkpoint_dir` in the output of `qstat -f`.

```
$ qstat -f
Job Id: 999.xxx.yyy
  Job_Name = test.sh
  Job_Owner = test@xxx.yyy
  resources_used.cput = 00:00:00
  resources_used.mem = 0kB
  resources_used.vmem = 0kB
  resources_used.walltime = 00:00:06
  job_state = H
  queue = batch
  server = xxx.yyy
  Checkpoint = u
  ctime = Tue Mar 11 14:17:04 2008
  Error_Path = xxx.yyy:/home/test/test.sh.e999
  exec_host = test/0
  Hold_Types = u
  Join_Path = n
  Keep_Files = n
  Mail_Points = a
  mtime = Tue Mar 11 14:17:10 2008
  Output_Path = xxx.yyy:/home/test/test.sh.o999
  Priority = 0
  qtime = Tue Mar 11 14:17:04 2008
  Rerunable = True
  Resource_List.neednodes = 1
  Resource_List.nodect = 1
  Resource_List.nodes = 1
  Resource_List.walltime = 01:00:00
  session_id = 9402
  substate = 20
  Variable_List = PBS_O_HOME=/home/test,PBS_O_LANG=en_US.UTF-8,
                 PBS_O_LOGNAME=test,
                 PBS_O_PATH=/usr/local/perltests/bin:/home/test/bin:/usr/local/bin:/usr/local/sbin:/usr/bin:/sbin:/bin:/usr/games,
                 PBS_O_SHELL=/bin/bash,PBS_SERVER=xxx.yyy,
                 PBS_O_HOST=xxx.yyy,PBS_O_WORKDIR=/home/test,
                 PBS_O_QUEUE=batch
  euser = test
egroup = test
  hashname = 999.xxx.yyy
  queue_rank = 3
  queue_type = E comment = Job started on Tue Mar 11 at 14:17
  exit_status = 271
  submit_args = test.sh
  start_time = Tue Mar 11 14:17:04 2008
  start_count = 1
  checkpoint_dir = /var/spool/torque/checkpoint/999.xxx.yyy.CK
  checkpoint_name = ckpt.999.xxx.yyy.1205266630
```

The value of `Resource_List.*` is the amount of resources requested.

Related topics

- [BLCR acceptance tests](#) on page 305
Test 2 - Persistence of checkpoint images

Introduction
This test determines if the checkpoint files remain in the default directory after the job is removed from the TORQUE queue.

Note that this behavior was requested by a customer but in fact may not be the right thing to do as it leaves the checkpoint files on the execution node. These will gradually build up over time on the node being limited only by disk space. The right thing would seem to be that the checkpoint files are copied to the user's home directory after the job is purged from the execution node.

Test steps
Assuming the steps of Test 1 (see Test 1 - Basic operation on page 305), delete the job and then wait until the job leaves the queue after the completed job hold time. Then look at the contents of the default checkpoint directory to see if the files are still there.

```
> qsub -c enabled test.sh
999.xxx.yyy
> qhold 999
> qdel 999
> sleep 100
> qstat
>
> find /var/spool/torque/checkpoint
... files ...
```

Possible failures
The files are not there, did Test 1 actually pass?

Successful results
The files are there.

Related topics
- BLCR acceptance tests on page 305

Test 3 - Restart after checkpoint

Introduction
This test determines if the job can be restarted after a checkpoint hold.

Test steps
Assuming the steps of Test 1 (see Test 1 - Basic operation on page 305), issue a qrls command. Have another window open into the /var/spool/torque/spool directory and tail the job.
Successful results

After the qrls, the job's output should resume.

Related topics

- BLCR acceptance tests on page 305

Test 4 - Multiple checkpoint/restart

Introduction

This test determines if the checkpoint/restart cycle can be repeated multiple times.

Test steps

Start a job and then while tailing the job output, do multiple qhold/qrls operations.

```bash
> qsub -c enabled test.sh
999.xxx.yyy
> qhold 999
> qrls 999
> qhold 999
> qrls 999
> qhold 999
> qrls 999
```

Successful results

After each qrls, the job's output should resume. Also tried "while true; do qrls 999; qhold 999; done" and this seemed to work as well.

Related topics

- BLCR acceptance tests on page 305

Test 5 - Periodic checkpoint

Introduction

This test determines if automatic periodic checkpoint will work.

Test steps

Start the job with the option -c enabled,periodic,interval=1 and look in the checkpoint directory for checkpoint images to be generated about every minute.

```bash
> qsub -c enabled,periodic,interval=1 test.sh
999.xxx.yyy
```
Successful results

After each qrls, the job's output should resume. Also tried "while true; do qrls 999; qhold 999; done" and this seemed to work as well.

Related topics

- BLCR acceptance tests on page 305

Test 6 - Restart from previous image

Introduction

This test determines if the job can be restarted from a previous checkpoint image.

Test steps

Start the job with the option -c enabled,periodic,interval=1 and look in the checkpoint directory for checkpoint images to be generated about every minute. Do a qhold on the job to stop it. Change the attribute checkpoint_name with the qalter command. Then do a qrls to restart the job.

```
> qsub -c enabled,periodic,interval=1 test.sh
999.xxx.yyy
> qhold 999
> qalter -W checkpoint_name=ckpt.999.xxx.yyy.1234567
> qrls 999
```

Successful results

The job output file should be truncated back and the count should resume at an earlier number.

Related topics

- BLCR acceptance tests on page 305