

Torque Resource Manager

Administrator Guide 6.0.4

August 2017



© 2017 Adaptive Computing Enterprises, Inc. All rights reserved.

Distribution of this document for commercial purposes in either hard or soft copy form is strictly prohibited without prior written consent from Adaptive Computing Enterprises, Inc.

Adaptive Computing, Cluster Resources, Moab, Moab Workload Manager, Moab Viewpoint, Moab Cluster Manager, Moab Cluster Suite, Moab Grid Scheduler, Moab Grid Suite, Moab Access Portal, and other Adaptive Computing products are either registered trademarks or trademarks of Adaptive Computing Enterprises, Inc. The Adaptive Computing logo and the Cluster Resources logo are trademarks of Adaptive Computing Enterprises, Inc. All other company and product names may be trademarks of their respective companies.

Adaptive Computing Enterprises, Inc.

1712 S. East Bay Blvd., Suite 300

Provo, UT 84606

+1 (801) 717-3700

www.adaptivecomputing.com



Scan to open online help

Welcome	1
Torque Administrator Guide Overview	2
Chapter 1: Introduction	5
Chapter 2: Overview	9
Torque Installation Overview	9
Torque Architecture	9
Installing Torque Resource Manager	10
Requirements	11
Supported Operating Systems	11
Software Requirements	11
Prerequisites	12
Open Necessary Ports	12
Verify The Hostname	12
Install Dependencies, Packages, Or Clients	13
Install Packages	13
Install Torque Server	13
Install Torque MOMs	14
Install Torque Clients	16
Configure Data Management	17
Compute Nodes	17
Enabling Torque As A Service	19
Initializing/Configuring Torque On The Server (pbs_server)	20
./torque.setup	20
Pbs_server -t Create	21
Specifying Compute Nodes	22
Configuring Torque On Compute Nodes	23
Configuring Ports	24
Configuring The Pbs_server Listening Port	24
Configuring The Pbs_mom Listening Port	24
Configuring The Port Pbs_server Uses To Communicate With Pbs_mom	25
Configuring The Port Pbs_mom Uses To Communicate With Pbs_server	25
Configuring The Port Client Commands Use To Communicate With Pbs_server	25
Configuring The Port Trqauthd Uses To Communicate With Pbs_server	25
Changing Default Ports	26
MOM Service Port	26
Default Port On The Server	26
MOM Manager Port	26
Configuring Trqauthd For Client Commands	27
Running Trqauthd	27
Finalizing Configurations	28
Advanced Configuration	28

Customizing The Install	29
HAVE_WORDEXP	37
Server Configuration	37
Server Configuration Overview	37
Name Service Configuration	38
Configuring Job Submission Hosts	38
Configuring Torque On A Multi-Homed Server	39
Architecture Specific Notes	39
Specifying Non-Root Administrators	39
Setting Up Email	40
Using MUNGE Authentication	40
Setting Up The MOM Hierarchy (Optional)	41
MOM Hierarchy Example	42
Setting Up The MOM Hierarchy	43
Putting The MOM Hierarchy On The MOMs	43
Manual Setup Of Initial Server Configuration	44
Server Node File Configuration	45
Basic Node Specification	45
Specifying Virtual Processor Count For A Node	46
Specifying GPU Count For A Node	46
Specifying Node Features (Node Properties)	47
Testing Server Configuration	47
Configuring Torque For NUMA Systems	49
Torque NUMA-Aware Configuration	49
Multiple Cgroup Directory Configuration	52
Change Considerations For Pbs_mom	52
Torque NUMA-Support Configuration	53
Configure Torque For NUMA-Support	53
Create The Mom.layout File	53
Automatically Create Mom.layout (Recommended)	53
Manually Create Mom.layout	54
Configure The Server_priv/nodes File	56
Limit Memory Resources (Optional)	56
Torque Multi-MOM	57
Multi-MOM Configuration	57
Configure Server_priv/nodes	57
/etc/hosts File	58
Starting Pbs_mom With Multi-MOM Options	58
Stopping Pbs_mom In Multi-MOM Mode	59
Chapter 3: Submitting And Managing Jobs	61
Job Submission	61
Multiple Job Submission	62

Slot Limit	63
Managing Multi-Node Jobs	64
Requesting Resources	65
Native Torque Resources	65
Interpreting Resource Requests	72
Interpreting Node Requests	72
Moab Job Extensions	73
Requesting NUMA-Aware Resources	74
Requesting Generic Resources	75
Requesting Floating Resources	75
Requesting Other Resources	76
Exported Batch Environment Variables	76
Enabling Trusted Submit Hosts	78
Example Submit Scripts	78
Job Files	79
Monitoring Jobs	81
Monitoring NUMA Job Task Placement	81
Canceling Jobs	83
Job Preemption	83
Configuring A Checkpoint Script On A MOM	83
Keeping Completed Jobs	84
Job Checkpoint And Restart	85
Introduction To BLCR	85
Configuration Files And Scripts	86
Starting A Checkpointable Job	93
Checkpointing A Job	94
Restarting A Job	95
Restarting A Job In The Held State	95
Restarting A Job In The Completed State	95
Acceptance Tests	95
Job Exit Status	95
Service Jobs	99
Submitting Service Jobs	100
Submitting Service Jobs In MCM	100
Managing Service Jobs	100
Chapter 4: Managing Nodes	103
Adding Nodes	103
Run-time Node Changes	103
Node Properties	105
Run-time Node Changes	105
Manual Node Changes	105
Changing Node State	105

Node Recovery	106
Changing Node Power States	106
Host Security	109
Enabling PAM With Torque	109
Legacy Torque PAM Configuration	110
Linux Cpuset Support	111
Cpuset Overview	111
Cpuset Support	111
Configuring Cpuset	112
Cpuset Advantages/Disadvantages	112
Scheduling Cores	112
Geometry Request Configuration	113
Geometry Request Usage	113
Geometry Request Considerations	114
Scheduling Accelerator Hardware	114
Chapter 5: Setting Server Policies	115
Queue Configuration	115
Example Queue Configuration	115
Setting Queue Resource Controls With Resource Request Syntax 2.0	116
Setting A Default Queue	117
Mapping A Queue To Subset Of Resources	117
Creating A Routing Queue	118
Server High Availability	119
Redundant Server Host Machines	120
Enabling High Availability	121
Enhanced High Availability With Moab	121
How Commands Select The Correct Server Host	122
Shutting Down The Local Or Active Pbs_server	122
Job Names	123
Persistence Of The Pbs_server Process	123
High Availability Of The NFS Server	123
Installing Torque In High Availability Mode	124
Installing Torque In High Availability Mode On Headless Nodes	129
Example Setup Of High Availability	135
Setting Min_threads And Max_threads	136
Chapter 6: Integrating Schedulers For Torque	137
Chapter 7: Configuring Data Management	139
SCP Setup	139
Generating SSH Key On Source Host	139
Copying Public SSH Key To Each Destination Host	140
Configuring The SSH Daemon On Each Destination Host	140

Validating Correct SSH Configuration	141
Enabling Bi-Directional SCP Access	141
Compiling Torque To Support SCP	141
Troubleshooting	142
NFS And Other Networked Filesystems	142
File Stage-in/stage-out	143
Examples	143
Chapter 8: MPI (Message Passing Interface) Support	145
MPICH	145
MPIExec Overview	145
MPIExec Troubleshooting	145
General MPI Troubleshooting	146
Open MPI	146
TM Aware	146
Chapter 9: Resources	149
Chapter 10: Accounting Records	153
Chapter 11: Job Logging	155
Job Log Location And Name	155
Enabling Job Logs	155
Chapter 12: NUMA And Torque	157
NUMA-Aware Systems	157
About NUMA-Aware Systems	158
Installation And Configuration	159
Job Resource Requests	159
Job Monitoring	160
Moab/Torque NUMA Configuration	160
Considerations When Upgrading Versions Or Changing Hardware	160
NUMA Tutorials	160
NUMA Primer	161
Torque Cgroup Hierarchy	161
Cpuset Subsystem	162
Cpuacct Subsystem	162
Devices Subsystem	162
Memory Subsystem	163
Resource Request 2.0	163
Single Resource Request With Two Tasks And Default Settings	163
Multiple Lprocs	165
Usecores	165
Usethreads	166

Multiple Resource Requests	166
Place Directives	167
Pbsnodes And Dedicated Resources	168
How NUMA Places Jobs	169
The Placement Algorithm	169
NUMA Discovery And Persistence	172
Initial Discovery	172
Job Placement Decisions	172
Persistence Across Restarts.	172
-L NUMA Resource Request	173
Syntax	173
Allocation Options	174
Pbsnodes With NUMA-Awareness	182
NUMA-Support Systems	184
About NUMA-Supported Systems	185
Torque Installation And Configuration	185
Moab/Torque NUMA Configuration	185
Chapter 13: Troubleshooting	187
Automatic Queue And Job Recovery	187
Host Resolution	187
Firewall Configuration	188
Torque Log Files	188
Pbs_server And Pbs_mom Log Files	188
Trqauthd Log Files	189
Using "tracejob" To Locate Job Failures	190
Overview	190
Syntax	190
Example	191
Using GDB To Locate Job Failures	192
Other Diagnostic Options	193
Stuck Jobs	193
Frequently Asked Questions (FAQ)	194
Cannot Connect To Server: Error=15034	195
Deleting 'stuck' Jobs	195
Which User Must Run Torque?	196
Scheduler Cannot Run Jobs - Rc: 15003	196
PBS_Server: Pbsd_init, Unable To Read Server Database	196
Qsub Will Not Allow The Submission Of Jobs Requesting Many Processors	198
Qsub Reports 'Bad UID For Job Execution'	198
Why Does My Job Keep Bouncing From Running To Queued?	198
How Do I Use PVM With Torque?	199
My Build Fails Attempting To Use The TCL Library	199

My Job Will Not Start, Failing With The Message 'cannot Send Job To Mom, State=PRERUN'	199
How Do I Determine What Version Of Torque I Am Using?	199
How Do I Resolve Autogen.sh Errors That Contain "error: Possibly Undefined Macro: AC_MSG_ERROR"?	200
How Do I Resolve Compile Errors With Libssl Or Libcrypto For Torque 4.0 On Ubuntu 10.04?	200
Why Are There So Many Error Messages In The Client Logs (trqauthd Logs) When I Don't Notice Client Commands Failing?	200
Compute Node Health Check	201
Configuring MOMs To Launch A Health Check	201
Creating The Health Check Script	202
Adjusting Node State Based On The Health Check Output	203
Example Health Check Script	203
Debugging	203
Torque Error Codes	204

Appendices	211
Commands Overview	213
Torque Services	213
Client Commands	213
Momctl	215
Synopsis	215
Overview	215
Format	215
Query Attributes	220
Diagnose Detail	221
Pbs_mom	222
Synopsis	222
Description	222
Options	223
Configuration File	225
Resources	226
Health Check	227
Files	227
Signal Handling	228
Exit Status	228
Pbs_server	229
Synopsis	229
Description	229
Options	229
Files	231
Signal Handling	231
Exit Status	232
Pbs_track	232

Synopsis	233
Description	233
Options	233
Operands	233
Exit Status	233
Pbsdsh	234
Synopsis	234
Description	234
Options	235
Operands	235
Standard Error	235
Exit Status	235
Pbsnodes	236
Synopsis	236
Description	236
NUMA-Awareness	236
Options	236
Examples	238
Qalter	239
Synopsis	239
Description	240
Options	240
Operands	247
Standard Error	247
Exit Status	247
Copyright	247
Qchkpt	248
Synopsis	248
Description	248
Options	248
Operands	248
Examples	248
Standard Error	248
Exit Status	249
Qdel	249
Synopsis	249
Description	249
Options	250
Operands	250
Examples	251
Standard Error	251
Exit Status	251
Qgpumode	251

Synopsis	252
Description	252
Options	252
Qgpureset	253
Synopsis	253
Description	253
Options	253
Qhold	253
Synopsis	253
Description	253
Options	254
Operands	254
Example	254
Standard Error	255
Exit Status	255
Qmgr	255
Synopsis	255
Description	255
Options	256
Operands	256
Standard Input	257
Standard Output	257
Standard Error	257
Directive Syntax	257
Exit Status	258
Qmove	259
Synopsis	259
Description	259
Operands	259
Standard Error	259
Exit Status	260
Qorder	260
Synopsis	260
Description	260
Operands	260
Standard Error	260
Exit Status	261
Qrerun	261
Synopsis	261
Description	261
Options	261
Operands	262
Standard Error	262

Exit Status	262
Examples	262
Qrls	263
Synopsis	263
Description	263
Options	263
Operands	264
Examples	264
Standard Error	264
Exit Status	264
Qrun	264
Synopsis	264
Overview	264
Format	265
Command Details	265
Examples	265
Qsig	265
Synopsis	266
Description	266
Options	266
Operands	267
Examples	267
Standard Error	267
Exit Status	267
Qstat	267
Synopsis	267
Options	268
Operands	270
Standard Output	270
Standard Error	274
Configuration	274
Exit Status	275
Qsub	275
Synopsis	275
Description	276
Options	277
Operands	291
Standard Input	291
Input Files	291
Standard Output	292
Standard Error	292
Environment Variables	292
Torque.cfg	292

Extended Description	292
Exit Status	294
Qterm	295
Synopsis	295
Description	295
Options	295
Operands	295
Standard Error	295
Exit Status	296
Trqauthd	296
Synopsis	296
Description	296
Options	297
Server Parameters	299
Parameters	300
Node Manager (MOM) Configuration	327
Parameters	327
Node Features And Generic Consumable Resource Specification	349
Command-line Arguments	350
Diagnostics And Error Codes	351
Considerations Before Upgrading	359
Rolling Upgrade	359
Large Cluster Considerations	361
Scalability Guidelines	361
End-User Command Caching	362
Qstat	362
Moab And Torque Configuration For Large Clusters	364
Starting Torque In Large Environments	365
Other Considerations	366
Job_stat_rate	366
Poll_jobs	366
Scheduler Settings	366
File System	366
Network ARP Cache	366
Prologue And Epilogue Scripts	369
Script Order Of Execution	370
Script Environment	370
Prologue Environment	371
Epilogue Environment	371
Per Job Prologue And Epilogue Scripts	372
Prologue And Epilogue Scripts Time Out	373
Prologue Error Processing	373
Running Multiple Torque Servers And MOMs On The Same Node	377

Configuring The First Torque	377
Configuring The Second Torque	377
Bringing The First Torque Server Online	377
Bringing The Second Torque Server Online	378
Security Overview	379
Trqauthd And Pbs_iff Authorization Theory	379
Job Submission Filter ("qsub Wrapper")	381
"torque.cfg" Configuration File	383
Configuration Parameters	383
Torque Quick Start Guide	389
Initial Installation	389
Initialize/Configure Torque On The Server (pbs_server)	389
Install Torque On The Compute Nodes	390
Configure Torque On The Compute Nodes	390
Configure Data Management On The Compute Nodes	390
Update Torque Server Configuration	391
Start The Pbs_mom Daemons On Compute Nodes	391
Verify Correct Torque Installation	391
Enable The Scheduler	392
Startup/Shutdown Service Script For Torque/Moab (OPTIONAL)	392
BLCR Acceptance Tests	393
Test Environment	393
Test 1 - Basic Operation	393
Introduction	393
Test Steps	394
Possible Failures	394
Successful Results	394
Test 2 - Persistence Of Checkpoint Images	396
Introduction	396
Test Steps	396
Possible Failures	396
Successful Results	396
Test 3 - Restart After Checkpoint	396
Introduction	396
Test Steps	397
Successful Results	397
Test 4 - Multiple Checkpoint/Restart	397
Introduction	397
Test Steps	397
Test 5 - Periodic Checkpoint	397
Introduction	397
Test Steps	398
Successful Results	398

Test 6 - Restart From Previous Image	398
Introduction	398
Test Steps	398
Successful Results	398
Queue Attributes	399
Attributes	399
Assigning Queue Resource Limits	409

Welcome

Welcome to the Torque 6.0.4 Administrator Guide.

This guide is intended as a reference for system administrators.

For more information about this guide, see these topics:

- ["Torque Administrator Guide Overview" on the next page](#)
- ["Introduction" on page 5](#)

Torque Administrator Guide Overview

[Chapter 1: Introduction on page 5](#) provides basic introduction information to help you get started using Torque.

[Chapter 2: Overview on page 9](#) 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 3: Submitting and Managing Jobs on page 61](#) 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 4: Managing Nodes on page 103](#) 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 109](#).

[Chapter 5: Setting Server Policies on page 115](#) details server-side configurations of queue and high availability.

[Chapter 6: Integrating Schedulers for Torque on page 137](#) offers information about using the native scheduler versus an advanced scheduler.

[Chapter 7: Configuring Data Management on page 139](#) deals with issues of data management. For non-network file systems, ["SCP Setup" on page 139](#) details setting up SSH keys and nodes to automate transferring data. ["NFS and Other Networked Filesystems" on page 142](#) 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 8: MPI \(Message Passing Interface\) Support on page 145](#) offers details supporting MPI.

[Chapter 9: Resources on page 149](#) covers configuration, utilization, and states of resources.

[Chapter 10: Accounting Records on page 153](#) explains how jobs are tracked by Torque for accounting purposes.

[Chapter 11: Job Logging on page 155](#) explains how to enable job logs that contain information for completed jobs.

[Chapter 12: NUMA and Torque on page 157](#) provides a centralized location for information on configuring Torque for NUMA systems.

[Chapter 13: Troubleshooting on page 187](#) 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 213](#)
- [Server Parameters on page 299](#)

- [Node Manager \(MOM\) Configuration on page 327](#)
- [Diagnostics and Error Codes on page 351](#)
- [Considerations Before Upgrading on page 359](#)
- [Large Cluster Considerations on page 361](#)
- [Prologue and Epilogue Scripts on page 369](#)
- [Running Multiple Torque Servers and MOMs on the Same Node on page 377](#)
- [Security Overview on page 379](#)
- [Job Submission Filter \("qsub Wrapper"\) on page 381](#)
- ["torque.cfg" Configuration File on page 383](#)
- [Torque Quick Start Guide on page 389](#)
- [BLCR Acceptance Tests on page 393](#)
- [Queue Attributes on page 399](#)

Related Topics

["Introduction" on page 5](#)

Chapter 1: Introduction

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

- ["What is a Resource Manager?" below](#)
- ["What are Batch Systems?" below](#)
- ["Basic Job Flow" on the next page](#)

1.0.1 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.

1.0.2 What are Batch Systems?

While Torque is flexible enough to handle scheduling a conference room, it is primarily used in batch systems. Torque is based on a job scheduler called Portable Batch System (PBS). 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.

Component	Description
Master Node	A batch system will have a master node where <i>pbs_server</i> 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.

Component	Description
Submit/Interactive Nodes	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 <i>qsub</i> and <i>qhold</i>).
Computer Nodes	Compute nodes are the workhorses of the system. Their role is to execute submitted jobs. On each compute node, <i>pbs_mom</i> runs to start, kill, and manage submitted jobs. It communicates with <i>pbs_server</i> on the master node. Depending on the needs of the systems, a compute node may double as the master node (or more).
Resources	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.

1.0.3 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	<p>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 (<i>walltime</i>), what resources are necessary to run, and what to execute. The following is an example submit file:</p> <pre>#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</pre> <p>This submit script specifies the name of the job (<i>localBlast</i>), what environment to use (<i>/bin/sh</i>), that it needs both processors on a single node (<i>nodes=1:ppn=2</i>), that it will run for at most 10 days, and that Torque should email "user@my.organization.com" when the job exits or aborts. Additionally, the user specifies where and what to execute.</p>
Submission	A job is submitted with the <i>qsub</i> 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 <i>qstat</i> .

Stage	Description
Finalilzation	When a job completes, by default, the <code>stdout</code> and <code>stderr</code> files are copied to the directory where the job was submitted.

Chapter 2: 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" below](#)
- ["Initializing/Configuring Torque on the Server \(pbs_server\)" on page 20](#)
- ["Advanced Configuration" on page 28](#)
- ["Manual Setup of Initial Server Configuration" on page 44](#)
- ["Server Node File Configuration" on page 45](#)
- ["Testing Server Configuration" on page 47](#)
- ["Configuring Torque for NUMA Systems" on page 49](#)
- ["Torque Multi-MOM" on page 57](#)

Torque Installation Overview

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

For details, see these topics:

- ["Torque Architecture" below](#)
- ["Installing Torque Resource Manager" on the next page](#)
- ["Compute Nodes" on page 17](#)
- ["Enabling Torque as a Service" on page 19](#)

Related Topics

["Troubleshooting" on page 187](#)

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 the previous page](#)

["Installing Torque Resource Manager" below](#)

Installing Torque Resource Manager



If you intend to use Torque Resource Manager 6.0.4 with Moab Workload Manager, you must run Moab version 8.0 or later. However, some Torque 6.0 functionality requires Moab 9.0 or later.

This topic contains instructions on how to install and start Torque Resource Manager (Torque).



For Cray systems, Adaptive Computing recommends that you install Moab and Torque Servers (head nodes) on commodity hardware (*not* on Cray compute/service/login nodes).

However, you must install the Torque `pbs_mom` daemon and Torque client commands on Cray login and "mom" service nodes since the `pbs_mom` must run on a Cray service node within the Cray system so it has access to the Cray ALPS subsystem.

See Installation Notes for Moab and Torque for Cray in the *Moab Workload Manager Administrator Guide* for instructions on installing Moab and Torque on a non-Cray server.



Adaptive Computing does not officially support the init scripts and `systemd` unit files; the community provides them as a starting point. You will want to direct questions about these scripts to the `torqueusers` mailing list.

In this topic:

- [Requirements on page 11](#)
- [Prerequisites on page 12](#)
- [Install Dependencies, Packages, or Clients on page 13](#)
- [Install Torque Server on page 13](#)
- [Install Torque MOMs on page 14](#)

- [Install Torque Clients on page 16](#)
- [Configure Data Management on page 17](#)

Requirements

In this section:

- [Supported Operating Systems on page 11](#)
- [Software Requirements on page 11](#)

Supported Operating Systems

- CentOS 6.x, 7.x
- RHEL 6.x, 7.x
- Scientific Linux 6.x, 7.x
- SUSE Linux Enterprise Server 11, 12

Software Requirements

- libxml2-devel package (package name may vary)
- openssl-devel package (package name may vary)
- Tcl/Tk version 8 or later if you plan to build the GUI portion of Torque, or use a Tcl-based scheduler
- cpusets and cgroups
 - NUMA-awareness uses cgroups, which include cpusets. Red Hat systems must use libcgrouop version 0.40.rc1-16.el6 or later; SUSE systems need to use a comparative libcgrouop version.
 - cpusets: libhwloc 1.9.1 is the minimum supported, however NVIDIA K80 requires libhwloc 1.11.0. If you need to install libhwloc and the corresponding hwloc-devel package, see "[Linux Cpuset Support](#)" on page 111.

i `--enable-geometry-requests` is *not* compatible with `--enable-cgroups`. In addition, if `--enable-cgroups` is specified, `--enable-cpuset` is ignored.

i If you are building with cgroups enabled, you must have boost version 1.41 or later.

- if you build Torque from source (i.e. clone from github), the following additional software is required:
 - gcc
 - gcc-c++
 - posix-compatible version of make
 - libtool 1.5.22 or later

- boost-devel 1.36.0 or later



Red Hat 6-based systems come packaged with 1.41.0 and Red Hat 7-based systems come packaged with 1.53.0. If needed, use the `--with-boost-path=DIR` option to change the packaged boost version. See ["Customizing the Install" on page 29](#) for more information.

Prerequisites

In this section:

- [Open Necessary Ports on page 12](#)
- [Verify the hostname on page 12](#)

Open Necessary Ports

Torque requires certain ports to be open for essential communication.

- For client and pbs_mom communication to pbs_server, the default port is 15001.
- For pbs_server communication to pbs_mom, the default port is 15002.
- For pbs_mom communication to pbs_mom, the default port is 15003.
- If using the MOM hierarchy (documented in [Setting Up the MOM Hierarchy \(Optional\) on page 41](#)) you must also open port 15003 from the server to the nodes.

For more information on how to configure the ports that Torque uses for communication, see ["Configuring Ports" on page 24](#) for more information.

If you have a firewall enabled, do the following:

1. On the Torque Server Host:
 - Red Hat 6-based systems using iptables
 - Red Hat 7-based systems using firewalld
 - SUSE 11-based systems using SuSEfirewall2
 - SUSE 12-based systems using SuSEfirewall2
2. On the Torque MOM Hosts (compute nodes):
 - Red Hat 6-based systems using iptables
 - Red Hat 7-based systems using firewalld
 - SUSE 11-based systems using SuSEfirewall2
 - SUSE 12-based systems using SuSEfirewall2

Verify the hostname

On the Torque Server Host, confirm your host (with the correct IP address) is in your `/etc/hosts` file. To verify that the hostname resolves correctly, make sure that `hostname` and `hostname -f` report the correct name for the host.

Install Dependencies, Packages, or Clients

Install Packages

On the Torque Server Host, use the following commands to install the `libxml2-devel`, `openssl-devel`, and `boost-devel` packages.

- Red Hat 6-based or Red Hat 7-based systems
- SUSE 11-based or SUSE 12-based systems

Install Torque Server

i You must complete the prerequisite tasks and the tasks to install the dependencies, packages, or clients before installing Torque Server. See [Prerequisites on page 12](#) and [Install Dependencies, Packages, or Clients on page 13](#).

On the Torque Server Host, do the following:

1. Download the latest 6.0.4 build from the [Adaptive Computing](#) website. It can also be downloaded via command line for the tarball distribution (recommended) or the github method.
 - Get the tarball source distribution.
 - Red Hat 6-based or Red Hat 7-based systems
 - SUSE 11-based or SUSE 12-based systems
 - Clone the source from github.

i If git is not installed:

```
# Red Hat 6-based or Red Hat 7-based systems
# SUSE 11-based or SUSE 12-based systems
```

```
[root]# git clone https://github.com/adaptivecomputing/torque.git -b 6.0.4 6.0.4
[root]# cd 6.0.4
[root]# ./autogen.sh
```

2. Run each of the following commands in order.

```
[root]# ./configure
[root]# make
[root]# make install
```

See ["Customizing the Install" on page 29](#) for information on which options are available to customize the `./configure` command.

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

```
[root]# echo <torque_server_hostname> > /var/spool/torque/server_name
```

4. Configure the `trqauthd` daemon to start automatically at system boot.
 - Red Hat 6-based systems
 - Red Hat 7-based systems
 - SUSE 11-based systems
 - SUSE 12-based systems
5. By default, Torque installs all binary files to `/usr/local/bin` and `/usr/local/sbin`. Make sure the path environment variable includes these directories for both the installation user and the root user.

```
[root]# export PATH=/usr/local/bin:/usr/local/sbin:$PATH
```

6. Initialize `serverdb` by executing the `torque.setup` script.

```
[root]# ./torque.setup root
```

7. Add nodes to the `/var/spool/torque/server_priv/nodes` file. See ["Specifying Compute Nodes" on page 22](#) for information on syntax and options for specifying compute nodes.
8. Configure `pbs_server` to start automatically at system boot, and then start the daemon.
 - Red Hat 6-based systems
 - Red Hat 7-based systems
 - SUSE 11-based systems
 - SUSE 12-based systems

Install Torque MOMs

In most installations, you will install a Torque MOM on each of your compute nodes.

 See ["Specifying Compute Nodes" on page 22](#) or ["Configuring Torque on Compute Nodes" on page 23](#) for more information.

Do the following:

1. On the Torque Server Host, do the following:

a. Create the self-extracting packages that are copied and executed on your nodes.

```
[root]# make packages
Building ./torque-package-clients-linux-x86_64.sh ...
Building ./torque-package-mom-linux-x86_64.sh ...
Building ./torque-package-server-linux-x86_64.sh ...
Building ./torque-package-gui-linux-x86_64.sh ...
Building ./torque-package-devel-linux-x86_64.sh ...
Done.
```

The package files are self-extracting packages that can be copied and executed on your production machines. Use `--help` for options.

b. Copy the self-extracting packages to each Torque MOM Host.

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

i The only required package for the compute node is `mom-linux`. Additional packages are recommended so you can use client commands and submit jobs from compute nodes.

```
[root]# scp torque-package-mom-linux-x86_64.sh <mom-node>:
[root]# scp torque-package-clients-linux-x86_64.sh <mom-node>:
```

c. Copy the `pbs_mom` startup script to each Torque MOM Host.

i If you have already installed init scripts, you may want to skip this step after comparing your existing scripts to the ones provided in the `contrib` directory. If you decide to use the new ones, be sure to check for manual alterations (e.g., to `ulimits`) in the old ones. The script includes a check for the presence of `/etc/sysconfig/pbs_mom`, so the natural place to put `ulimits` would be there, if not directly in the init script itself.

- Red Hat 6-based systems
- Red Hat 7-based systems
- SUSE 11-based systems
- SUSE 12-based systems

i Not all sites see an inherited `ulimit`, but those that do can change it in the `pbs_mom` init script, which is responsible for starting and stopping the `pbs_mom` process. The script includes a check for the presence of `/etc/sysconfig/pbs_mom`, so the natural place to put `ulimits` would be there (or directly in the init script itself).

2. On each Torque MOM Host, do the following:
 - a. Install the self-extracting packages and run `ldconfig`.

```
[root]# ssh root@<mom-node>
[root]# ./torque-package-mom-linux-x86_64.sh --install
[root]# ./torque-package-clients-linux-x86_64.sh --install
[root]# echo /usr/local/lib > /etc/ld.so.conf.d/torque.conf
[root]# ldconfig
```

- b. Configure `pbs_mom` to start at system boot, and then start the daemon.

- Red Hat 6-based systems
- Red Hat 7-based systems
- SUSE 11-based systems
- SUSE 12-based systems

Install Torque Clients

If you want to have the Torque client commands installed on hosts other than the Torque Server Host (such as the compute nodes or separate login nodes), do the following:

1. On the Torque Server Host, do the following:
 - a. Copy the self-extracting client package to each Torque Client Host.

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

```
[root]# scp torque-package-clients-linux-x86_64.sh <torque-client-host>:
```

- b. Copy the `trqauthd` startup script to each Torque Client Host.

- Red Hat 6-based systems
- Red Hat 7-based systems
- SUSE 11-based systems
- SUSE 12-based systems

2. On each Torque Client Host, do the following:

i Many of these steps can be done from the Torque server from a remote shell, such as SSH. Set up shared SSH keys if you do not want to supply a password for each Torque Client Host.

- a. Install the self-extracting client package.

```
[root]# ./torque-package-clients-linux-x86_64.sh --install
[root]# echo /usr/local/lib > /etc/ld.so.conf.d/torque.conf
[root]# ldconfig
```

- b. Enable and start the trqauthd service.

- Red Hat 6-based systems
- Red Hat 7-based systems
- SUSE 11-based systems
- SUSE-12 based systems

Configure Data Management

When a batch job completes, stdout and stderr files are generated and placed in the spool directory on the master Torque MOM Host for the job instead of the submit host. You can configure the Torque batch environment to copy the stdout and stderr files back to the submit host. See ["Configuring Data Management" on page 139](#) for more information.

Compute Nodes

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



If you installed Torque using the RPMs, you must install and configure your nodes manually by modifying the `/var/spool/torque/mom_priv/config` file of each one. This file is identical for all compute nodes and can be created on the head node and distributed in parallel to all systems.

```
[root]# vi /var/spool/torque/mom_priv/config

$pbsserver      headnode      # hostname running pbs server
$logevent       225           # bitmap of which events to log

[root]# service pbs_mom restart
```

To create Torque packages

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

```
> make packages
Building ./torque-package-clients-linux-x86_64.sh ...
Building ./torque-package-mom-linux-x86_64.sh ...
Building ./torque-package-server-linux-x86_64.sh ...
Building ./torque-package-gui-linux-x86_64.sh ...
Building ./torque-package-devel-linux-x86_64.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.

```
> cp torque-package-mom-linux-x86_64.sh /shared/storage/
> cp torque-package-clients-linux-x86_64.sh /shared/storage/
```

3. Install the Torque packages on the compute nodes.

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

i The only required package for the compute node is `mom-linux`. Additional packages are recommended so you can use client commands and submit jobs from compute nodes.

The following is an example of how to copy and install `mom-linux` in a distributed fashion.

```
> for i in node01 node02 node03 node04 ; do scp torque-package-mom-linux-x86_64.sh
${i}:/tmp/. ; done
> for i in node01 node02 node03 node04 ; do scp torque-package-clients-linux-x86_
64.sh ${i}:/tmp/. ; done
> for i in node01 node02 node03 node04 ; do ssh ${i} /tmp/torque-package-mom-linux-
x86_64.sh --install ; done
> for i in node01 node02 node03 node04 ; do ssh ${i} /tmp/torque-package-clients-
linux-x86_64.sh --install ; done
> for i in node01 node02 node03 node04 ; do ssh ${i} ldconfig ; done
```

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

To use a tool like xCAT

1. Copy the Torque package to the nodes.

```
> prcp torque-package-linux-x86_64.sh noderange:/destinationdirectory/
```

2. Install the Torque package.

```
> psh noderange /tmp/torque-package-linux-x86_64.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.

Related Topics

["Installing Torque Resource Manager" on page 10](#)

["Torque Installation Overview" on page 9](#)

Enabling Torque as a Service

i Enabling Torque as a service is optional. In order to run Torque as a service, you must enable `trqauthd`. (see ["Configuring trqauthd for Client Commands" on page 27](#)).

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

- SUSE (as root)

```
> cp contrib/init.d/suse.pbs_mom /etc/init.d/pbs_mom
> insserv -d pbs_mom
> 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_mom /etc/init.d/pbs_mom
> update-rc.d pbs_mom defaults
> cp contrib/init.d/debian.pbs_server /etc/init.d/pbs_server
> update-rc.d pbs_server defaults
```

i 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 9](#)

["Installing Torque Resource Manager" on page 10](#)

["Configuring trqauthd for Client Commands" on page 27](#)

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 (See ["Configuring Torque on Compute Nodes" on page 23](#)). It also maintains the status of each MOM node through updates from the MOMs in the cluster (see ["pbsnodes" on page 236](#)). 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" below](#)).
- Use `pbs_server -t create` (see ["pbs_server -t create" on the facing page](#)).

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'

#
# 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

```

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

This section contains these topics:

- ["Specifying Compute Nodes" on the next page](#)
- ["Configuring Torque on Compute Nodes" on page 23](#)
- ["Finalizing Configurations" on page 28](#)

Related Topics

["Node Manager \(MOM\) Configuration" on page 327](#)

["Advanced Configuration" on page 28](#)

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.

i 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.

i You can specify a compute node's hostname by starting each `pbs_mom` with the "[-H](#)" [on page 224](#) `hostname` option, or by adding a line for `$mom_host` in `TORQUE_HOME/mom_priv/config` on the `pbs_mom` host(s). (You can run `hostname -f` to obtain a node's hostname.)

- 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 20](#)

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:`

```
$pbsserver      headnode      # hostname running pbs server
$logevent       1039           # bitmap of which events to log
```

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 20](#)

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`

i 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.

```
$ 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.

```
$ 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.

```
$ pbs_mom -M port_num
```

- Edit the `pbs_server nodes` file to 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.

```
$ pbs_server -M 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 port_num/tcp.
- Start pbs_mom with the -S option.

```
$ pbs_mom -p port_num
```

- 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 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>

Changing Default Ports

This section provides examples of changing the default ports (using non-standard ports).

MOM Service Port

The MOM service port is the port number on which MOMs are listening. This example shows how to change the default MOM service port (15002) to port 30001.

Do the following:

- On the server, for the `server_priv/nodes` file, change the node entry.

```
nodename np=4 mom_service_port=30001
```

- On the MOM

```
pbs_mom -M 30001
```

Default Port on the Server

Do the following:

- Set the `$(TORQUE_HOME)/server_name` file.

```
hostname:newport  
numa3.ac:45001
```

- On the MOM, start `pbs_mom` with the `-S` option and the new port value for the server.

```
pbs_mom -S 45001
```

MOM Manager Port

The MOM manager port tell MOMs which ports on which other MOMs are listening for MOM-to-MOM communication. This example shows how to change the default MOM manager port (15003) to port 30002.

Do the following:

- On the server nodes file.

```
nodename np=4 mom_manager_port=30002
```

- Start the MOM.

```
pbs_mom -R 30002
```

Related Topics

["Initializing/Configuring Torque on the Server \(pbs_server\)" on page 20](#)

[pbs_server](#)

[pbs_mom](#)

[trqauthd](#)

client commands

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`.

i If you run trqauthd before starting pbs_server, you will receive a warning that no servers are available. To avoid this message, start pbs_server before running trqauthd.

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:

Script	Description
debian.trqauthd	Used for apt-based systems (debian, ubuntu are the most common variations of this)
suse.trqauthd	Used for suse-based systems
trqauthd	An example for other package managers (Redhat, Scientific, CentOS, and Fedora are some common examples)

i 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
```

i 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 20](#)

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 20](#)

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 the facing page](#)
- ["Server Configuration" on page 37](#)

Related Topics

["Server Parameters" on page 299](#)

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 2-1: Optional Features

Option	Description
<code>--disable-clients</code>	Directs Torque not to build and install the Torque client utilities such as <code>qsub</code> , <code>qstat</code> , <code>qdel</code> , etc.
<code>--disable-FEATURE</code>	Do not include FEATURE (same as <code>--enable-FEATURE=no</code>).
<code>--disable-lib-tool-lock</code>	Avoid locking (might break parallel builds).
<code>--disable-mom</code>	Do not include the MOM daemon.
<code>--disable-mom-check-spool</code>	Don't check free space on spool directory and set an error.
<code>--disable-posixmemlock</code>	Disable the MOM's use of <code>mlockall</code> . Some versions of OSs seem to have buggy POSIX <code>MEMLOCK</code> .
<code>--disable-priv-ports</code>	Disable the use of privileged ports for authentication. Some versions of OSX have a buggy <code>bind()</code> and cannot bind to privileged ports.
<code>--disable-qsub-keep-override</code>	Do not allow the <code>qsub -k</code> flag to override <code>-o -e</code> .

Option	Description
--disable-server	Do not include server and scheduler.
--disable-shell-pipe	Give the job script file as standard input to the shell instead of passing its name via a pipe.
--disable-spool	If disabled, Torque will create output and error files directly in TORQUE_HOME/.pbs_spool if it exists or in TORQUE_HOME otherwise. By default, Torque will spool files in TORQUE_HOME/spool and copy them to the users home directory when the job completes.
--disable-xopen-net-working	With HPUX and GCC, don't force usage of XOPEN and libxnet.
--enable-acct-x	Enable adding x attributes to accounting log.
--enable-array	Setting this under IRIX enables the SGI Origin 2000 parallel support. Normally autodetected from the /etc/config/array file.
--enable-autorun	Turn on the AUTORUN_JOBS flag. When enabled, Torque runs the jobs as soon as they are submitted (destroys Moab compatibly). This option is not supported.
--enable-blcr	Enable BLCR support.
--enable-cgroups	<p>Enable cgroups for NUMA-aware configurations.</p> <div> <i>i</i> If you are building with cgroups enabled, you must have boost version 1.41 or later. </div> <div> <i>i</i> --enable-geometry-requests is <i>not</i> compatible with --enable-cgroups. </div>
--enable-cpa	Enable Cray's CPA support.
--enable-cpu-set	<p>Enable Linux 2.6 kernel cpusets.</p> <div> <i>i</i> It is recommended that you turn on this feature to prevent a job from expanding across more CPU cores than it is assigned. </div> <div> <i>i</i> If using NUMA-awareness, cgroups are supported and cpusets are handled by the cgroup cpuset subsystem. If --enable-cgroups is specified, --enable-cpuset is ignored. </div>

Option	Description
--enable-debug	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.)
--enable-dependency-tracking	Do not reject slow dependency extractors.
--enable-fast-install[=PKGS]	Optimize for fast installation [default=yes].
--enable-FEATURE [ARG]	Include FEATURE [ARG=yes].
--enable-file-sync	Open files with sync on each write operation. This has a negative impact on Torque performance. This is disabled by default.
--enable-force-nodefile	Forces creation of nodefile regardless of job submission parameters. Not on by default.
--enable-gcc-warnings	Enable gcc strictness and warnings. If using gcc, default is to error on any warning.
--enable-geometry-requests	<p>Torque is compiled to use procs_bitmap during job submission.</p> <div>  --enable-geometry-requests is <i>not</i> compatible with --enable-cgroups. </div>
--enable-gui	Include the GUI-clients.
--enable-main-tainer-mode	This is for the autoconf utility and tells autoconf to enable so called rebuild rules. See main-tainer mode for more information.

Option	Description
--enable-maxdefault	<p>Turn on the RESOURCEMAXDEFAULT flag.</p> <div> <p>i Versions 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).</p> <p>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 <code>--enable-maxdefault</code>.</p> </div>
--enable-nochildsignal	Turn on the NO_SIGCHLD flag.
--enable-nodemask	Enable nodemask-based scheduling on the Origin 2000.
--enable-pemask	Enable pemask-based scheduling on the Cray T3e.
--enable-plock-daemons[=ARG]	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).
--enable-quick-commit	Turn on the QUICKCOMMIT flag. When enabled, adds a check to make sure the job is in an expected state and does some bookkeeping for array jobs. This option is not supported.
--enable-shared[=PKGS]	Build shared libraries [default=yes].
--enable-shell-use-argv	Enable this to put the job script name on the command line that invokes the shell. Not on by default. Ignores --enable-shell-pipe setting.
--enable-sp2	Build PBS for an IBM SP2.
--enable-srfs	Enable support for SRFS on Cray.
--enable-static[=PKGS]	Build static libraries [default=yes].

Option	Description
--enable-syslog	Enable (default) the use of syslog for error reporting.
--enable-tcl-qstat	Setting this builds qstat with Tcl interpreter features. This is enabled if Tcl is enabled.
--enable-unix-sockets	Enable the use of Unix Domain sockets for authentication.

Table 2-2: Optional Packages

Option	Description
--with-bcr=DIR	BLCR installation prefix (Available in versions 2.5.6 and 3.0.2 and later).
--with-bcr-include=DIR	Include path for <code>libcr.h</code> (Available in versions 2.5.6 and 3.0.2 and later).
--with-bcr-lib=DIR	Lib path for <code>libcr</code> (Available in versions 2.5.6 and 3.0.2 and later).
--with-bcr-bin=DIR	Bin path for BLCR utilities (Available in versions 2.5.6 and 3.0.2 and later).

Option	Description
--with-boost-path=DIR	<div>  Boost version 1.36.0 or later is supported. Red Hat 6-based systems come packaged with 1.41.0 and Red Hat 7-based systems come packaged with 1.53.0. </div> <p>Set the path to the Boost header files to be used during make. This option does not require Boost to be built or installed.</p> <p>The <code>--with-boost-path</code> value must be a directory containing a subdirectory called <code>boost</code> that contains the <code>boost .hpp</code> files.</p> <p>For example, if downloading the boost 1.55.0 source tarball to the adaptive user's home directory:</p> <pre>[adaptive]\$ cd ~ [adaptive]\$ wget http://sourceforge.net/projects/boost/files/boost/1.55.0/boost_1_55_0.tar.gz/download [adaptive]\$ tar xzf boost_1_55_0.tar.gz [adaptive]\$ ls boost_1_55_0 boost boost-build.jam ...</pre> <p>In this case use <code>--with-boost-path=/home/adaptive/boost_1_55_0</code> during configure.</p> <p>Another example would be to use an installed version of Boost. If the installed Boost header files exist in <code>/usr/include/boost/*.hpp</code>, use <code>--with-boost-path=/usr/include</code>.</p>
--with-cpa-include=DIR	Include path for <code>cpalib.h</code> .
--with-cpa-lib=DIR	Lib path for <code>libcpalib</code> .
--with-debug=no	Do not compile with debugging symbols.
--with-default-server-r=HOSTNAME	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.
--with-envir=PATH	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 "pbs_environment" in the server home. Relative paths are interpreted within the context of the server home.
--with-gnu-ld	Assume the C compiler uses GNU <code>ld</code> [default=no].

Option	Description
--with-mail-domain=MAILDOMAIN	Override the default domain for outgoing mail messages, i.e. "user@maildomain". The default maildomain is the hostname where the job was submitted from.
--with-modulefiles[=DIR]	Use module files in specified directory [/etc/modulefiles].
--with-momlogdir	Use this directory for MOM logs.
--with-momlogsuffix	Use this suffix for MOM logs.
--without-PACKAGE	Do not use PACKAGE (same as --with-PACKAGE=no).
--without-readline	Do not include readline support (default: included if found).
--with-PACKAGE[=ARG]	Use PACKAGE [ARG=yes].
--with-pam=DIR	Directory that holds the system PAM modules. Defaults to /lib(64)/security on Linux.
--with-pic	Try to use only PIC/non-PIC objects [default=use both].
--with-qstatrc-file=FILE	Set the name of the file that qstat will use if there is no .qstatrc 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 qstatrc (no dot) in the server home directory.
--with-rcp	One of scp, rcp, mom_rcp, 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.
--with-reserved-port-start=t=PORT	Set the lower bound of the reserved port range that Torque will use when opening a reserved port. PORT must be between 144 and 823, inclusive. <div>  Setting this parameter reduces the number of privileged ports available to the system. This could affect performance, because it limits the number of concurrent reserved ports pbs_server can open. </div>

Option	Description
--with-sched=TYPE	Sets the scheduler type. If TYPE is <code>c</code> , the scheduler will be written in C. If TYPE is <code>tcl</code> the server will use a Tcl based scheduler. If TYPE is <code>basl</code> , Torque will use the rule based scheduler. If TYPE is <code>no</code> , then no scheduling is done. <code>c</code> is the default.
--with-sched-code=PATH	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 <code>.basl</code> . It will be interpreted relative to <code>src/tree/src/schedulers.SCHD_TYPE/samples</code> . As an example, an appropriate BASL scheduler relative path would be <code>nas.basl</code> . The default scheduler code for "C" schedulers is <code>fifo</code> .
--with-scp	In Torque 2.1 and later, SCP is the default remote copy protocol. See " --with-rcp " on the previous page if a different protocol is desired.
--with-sendmail[=PATH_TO_EXECUTABLE]	Sendmail executable to use. If <code>=PATH_TO_EXECUTABLE</code> is not specified or if <code>--with-sendmail</code> is not used at all, configure will attempt to find sendmail.
--with-server-home=DIR	Set the server <code>home/spool</code> directory for PBS use. Defaults to <code>/var/spool/torque</code> .
--with-server-name-file=FILE	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 <code>/var/spool/torque</code> or is set using the <code>--with-server-home</code> option to configure. If this option is not specified, the default name for this file will be set to <code>"server_name"</code> .
--with-tcl	Directory containing tcl configuration (<code>tclConfig.sh</code>).
--with-tclatrsep=CHAR	Set the Tcl attribute separator character this will default to <code>"."</code> if unspecified.
--with-tclinclude	Directory containing the public Tcl header files.
--with-tclx	Directory containing <code>tclx</code> configuration (<code>tclxConfig.sh</code>).
--with-tk	Directory containing <code>tk</code> configuration (<code>tkConfig.sh</code>).
--with-tkinclude	Directory containing the public Tk header files.
--with-tkx	Directory containing <code>tkx</code> configuration (<code>tkxConfig.sh</code>).
--with-xauth=PATH	Specify path to <code>xauth</code> program.

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 28](#)

["Server Configuration" below](#)

Server Configuration

This topic contains information and instructions to configure your server.

In this topic:

- ["Server Configuration Overview" below](#)
- ["Name Service Configuration" on the next page](#)
- ["Configuring Job Submission Hosts" on the next page](#)
- ["Configuring Torque on a Multi-Homed Server" on page 39](#)
- ["Architecture Specific Notes" on page 39](#)
- ["Specifying Non-Root Administrators" on page 39](#)
- ["Setting Up Email" on page 40](#)
- ["Using MUNGE Authentication" on page 40](#)

Also see ["Setting Up the MOM Hierarchy \(Optional\)" on page 41](#)

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 399](#)).

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'
```

i 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.

i 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. Specify the location of the sendmail executable. You can do this using the `sendmail_path` server attribute.

```
qmgr -c 'set server sendmail_path = <path_to_executable>'
```

If this server option is not set, you can set a default location during the build.

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

If a location for the sendmail executable is not specified, Torque will attempt to find it when you run `configure`. If you installed Torque using RPMs from Adaptive Computing, the default path will be `/usr/sbin/sendmail`.

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

```
> 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**:

```
> qmgr -c 'set server mail_body_fmt=Job: %i \n Name: %j \n On host: %h \n \n %m \n \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 381).

Using MUNGE Authentication



The same version of MUNGE must be installed on all of your Torque Hosts (Server, Client, MOM).

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 <https://dun.github.io/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`.

```
> ./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 MUNGЕ library into its executables. It calls the MUNGЕ and UNMUNGЕ utilities which are part of the MUNGЕ daemon. The MUNGЕ daemon must be running on the server and all submission hosts. The Torque client utilities call MUNGЕ 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:

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

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:

```
> 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.

Related Topics

["Setting Up the MOM Hierarchy \(Optional\)" below](#)

["Advanced Configuration" on page 28](#)

Setting Up the MOM Hierarchy (Optional)

i Mom hierarchy is designed for large systems to configure how information is passed directly to the *pbs_server*.

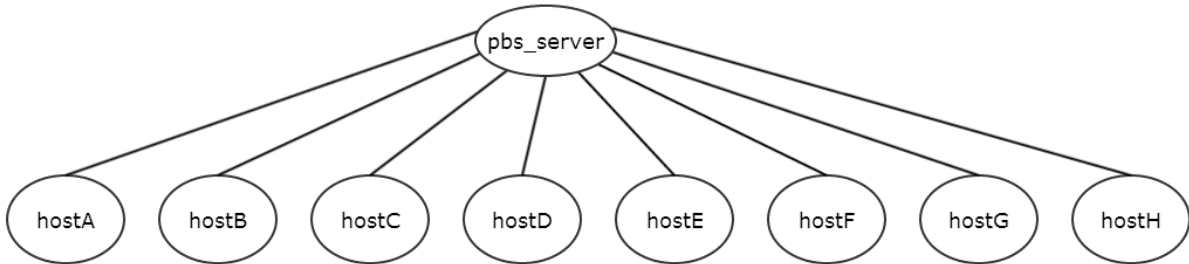
The MOM hierarchy allows you to override the compute nodes' default behavior of reporting status updates directly to the *pbs_server*. Instead, you configure compute nodes so that each node sends its status update information to another compute node. The compute nodes pass the information up a tree or hierarchy until eventually the information reaches a node that will pass the information directly to *pbs_server*. This can significantly reduce network traffic and ease the load on the *pbs_server* in a large system.

i Adaptive Computing recommends approximately 25 nodes per path. Numbers larger than this may reduce the system performance.

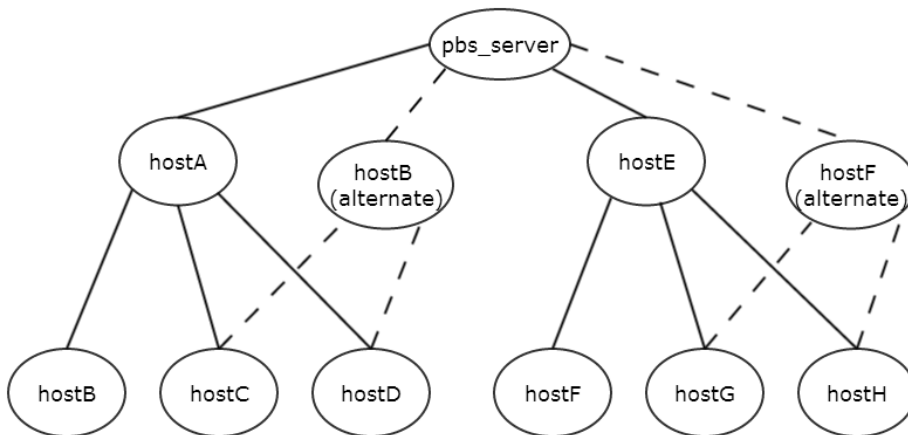
MOM Hierarchy Example

The following example illustrates how information is passed to the *pbs_server* without and with *mom_hierarchy*.

Without *mom_hierarchy*



With *mom_hierarchy*



i The dotted lines indicates an alternate path if the hierarchy-designated node goes down.

The following is the *mom_hierarchy_file* for the with *mom_hierarchy* example:

```

<path>
  <level>hostA,hostB</level>
  <level>hostB,hostC,hostD</level>
</path>
<path>
  <level>hostE,hostF</level>
  <level>hostE,hostF,hostG</level>
</path>
  
```

Setting Up the MOM Hierarchy

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

Within a `<path></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.

i When setting up the MOM hierarchy, you must open port 15003 for communication from `pbs_server` to `pbs_mom`.

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

i 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>
...
```

Putting the MOM Hierarchy on the MOMs

You can put the MOM hierarchy file directly on the MOMs. The default location is `/var/spool/torque/mom_priv/mom_hierarchy`. 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. When using a global file system accessible from both the MOMs and the `pbs_server`, it is recommended that the hierarchy file be symbolically linked to the MOMs.

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.

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`:

```
qmgr -c 'p s'
#
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.

```
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"
```

i 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 323](#) 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 9](#)

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. See these topics for details:

- ["Basic Node Specification" below](#)
- ["Specifying Virtual Processor Count for a Node" on the next page](#)
- ["Specifying GPU Count for a Node" on the next page](#)
- ["Specifying Node Features \(Node Properties\)" on page 47](#)

Related Topics

["Torque Installation Overview" on page 9](#)

["Server Parameters" on page 299](#)

Node Features/Node Properties

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
```



The `server nodes` file also displays the parameters applied to the node. See [Adding nodes](#) for more information on the parameters.

Related Topics

["Server Node File Configuration" above](#)

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.

`server_priv/nodes:`

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

Related Topics

["Server Node File Configuration" on the previous page](#)

Specifying GPU Count for a Node



This section describes a rudimentary method for configuring GPUs manually. Administrators can configure the MOMs to automatically detect the number of NVIDIA GPUs and get detailed GPU reporting on each node (the recommended method). Combining this with `cgroups` will also prevent unauthorized access to resources. See Scheduling GPUs in the *Moab Workload Manager Administrator Guide* for details on this automated method.

When using this method, `pbs_server` automatically appends "`gpus=<count>`" to the end of the line in `TORQUE_HOME/server_priv/nodes` for any node with a GPU, overriding any such manual configuration.

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 GPU 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.

`server_priv/nodes:`

```
node001 gpus=2
node002 gpus=4
...
```

Related Topics

["Server Node File Configuration" on page 45](#)

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:

```
qsub -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 45](#)

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

Queue      Memory      CPU Time      Walltime      Node      Run      Que      Lm      State
-----
batch      --           --           --           --           0       0       --      ER
                                     0       0

# 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 = 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 = user1@kmn
set server operators = user1@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 = 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,varattr=,jobs=,state=free,netload=6814326158,gres=,loadave
=0.21,ncpus=6,physmem=8193724kb,
availmem=13922548kb,totmem=16581304kb,idletime=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
  gpus = 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
-----
```



```
0.kmn      STDIN      knielson      0  Q  batch
```

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 Torque" on page 137](#) for details on setting up an advanced scheduler.

Related Topics

["Torque Installation Overview" on page 9](#)

Configuring Torque for NUMA Systems

Torque supports these two types of Non-Uniform Memory Architecture (NUMA) systems:

- **NUMA-Aware** – For Torque 6.0 and later, supports multi-req jobs and jobs that span hosts. Requires the `--enable-cgroups` configuration command to support cgroups. See [Torque NUMA-Aware Configuration on page 49](#) for instructions and additional information.
- **NUMA-Support** – For Torque version 3.0 and later; *only* for large-scale SLES systems (SGI Altix and UV hardware). Requires the `--enable-numa-support` configuration command. See [Torque NUMA-Support Configuration on page 53](#) for instructions and additional information.



Torque cannot be configured for both systems at the same.

Related Topics

- [Torque NUMA-Aware Configuration on page 49](#)
- [Torque NUMA-Support Configuration on page 53](#)

Torque NUMA-Aware Configuration

This topic provides instructions for enabling NUMA-aware, including cgroups, and requires Torque 6.0 or later. For instructions on NUMA-support configurations, see [Torque NUMA-Support Configuration on page 53](#). This topic assumes you have a basic understanding of cgroups. See [RedHat Resource Management Guide](#)

(https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/6/html/Resource_Management_Guide/ch01.html) or [cgroups on kernel.org](#)

(<https://www.kernel.org/doc/Documentation/cgroup-v1/cgroups.txt>) for basic information on cgroups.

Torque uses cgroups to better manage cpu and memory accounting, memory enforcement, cpuset management, and binding jobs to devices such as MICs and GPUs.

i Be aware of the following:

- If you are building with cgroups enabled, you must have boost version 1.41 or later.
- The pbs_mom daemon is the binary that interacts cgroups, but both the server and the MOM must be built with `--enable-cgroups` to understand all of the new structures.
- Beginning with Torque 6.0.2, Cray-enabled Torque may be configured with cgroups. On the login node, each job will have all of the cpus and all of the memory controllers in it's cgroup.

Prerequisites

1. Install the prerequisites found in ["Installing Torque Resource Manager" on page 10](#).
2. hwloc version 1.9 or later is required. Version 1.11 is needed if installing with NVIDIA K80 or newer GPU hardware

- download hwloc-1.9.tar.gz from: <https://www.open-mpi.org/software/hwloc/v1.9>
- perform the following command line actions:

```
$ tar -xzf hwloc-1.9.tar.gz
$ cd hwloc-1.9.tar.gz
$ sudo ./configure
```

- You do not need to overwrite the default installation of hwloc. By default hwloc will install to the /usr/local directory. You can also configure hwloc with the `--prefix` option to have it install to a location of your choosing. If you do not install hwloc to /usr directory you can tell Torque where to find the version you want it to use at configure time using the `--with-hwloc-path` option. For example:

```
./configure --enable-cgroups --with-hwloc-path=/usr/local
```

- Run make
- sudo make install

Installation Instructions

Do the following:

1. Install the libcgroup package.

i Red Hat-based Systems must use libcgroup version 0.40.rc1-16.el6 or later; SUSE-based systems need to use a comparative libcgroup version.

- Red Hat-based systems

```
yum install libcgroup-tools libcgroup
```

- SUSE-based systems

```
zypper install libcgroup-tools
```

2. Enable Torque to access cgroups.

```
$ ./configure --enable-cgroups
```

3. Run `lssubsys -am` to determine whether your system has mounted cgroups.

- a. If cgroups are *not* mounted, you will see:

```
$ lssubsys -am

nsperf_event
net_prio
cpuset
cpu
cpuacct
memory
devices
freezer
net_cls blkio
```

- b. If cgroups are mounted, you will see:

```
$ lssubsys -am

ns
perf_event
net_prio
cpuset,cpu,cpuacct /cgroup/cpu
memory /cgroup/memory
devices /cgroup/devices
freezer /cgroup/freezer
net_cls /cgroup/net_cls
blkio /cgroup/blkio
```

4. If you determined that cgroups are not mounted, take one of the following actions; otherwise, proceed to the next step.

- Follow the cgroup mounting instructions for your Red Hat operating system.
- Manually mount cgroups from the command line.

```
mount -t cgroup -o <subsystem>[,<subsystem>,...] name <dir path>/name
```

The name parameter will be the name of the hierarchy.



The following commands create five hierarchies, one for each subsystem.

```
mount -t cgroup -o cpuset cpuset /var/spool/torque/cgroup/cpuset
mount -t cgroup -o cpu cpu /var/spool/torque/cgroup/cpu
mount -t cgroup -o cpuacct cpuacct /var/spool/torque/cgroup/cpuacct
mount -t cgroup -o memory memory /var/spool/torque/cgroup/memory
mount -t cgroup -o devices devices /var/spool/torque/cgroup/devices
```

Once you have mounted the cgroups, run `lssubsys -am` again. You should now see:

```
cpuset /var/spool/torque/cgroup/cpuset
cpu /var/spool/torque/cgroup/cpu
cpuacct /var/spool/torque/cgroup/cpuacct
memory /var/spool/torque/cgroup/memory
devices /var/spool/torque/cgroup/devices
freezer
blkio
perf_event
```

After successful configuration, the `pbsnodes` command should show these additional fields for each node:

```
total_sockets=1
total_numa_nodes = 1
total_cores = 8
total_threads = 8
dedicated_sockets = 0
dedicated_numa_nodes = 0
dedicated_cores = 0
dedicated_threads = 0
```

Multiple cgroup Directory Configuration

If your system has more than one cgroup directory configured, you must create the `trq-cgroup-paths` file in the `$TORQUE_HOME` directory. This file has a list of the cgroup subsystems and the mount points for each subsystem in the syntax of `<subsystem> <mount point>`.

All five subsystems used by `pbs_mom` must be in the `trq-cgroup-paths` file. In the example that follows, a directory exists at `/cgroup` with subdirectories for each subsystem. Torque uses this file first to configure where it will look for cgroups.

```
cpuset /cgroup/cpuset
cpuacct /cgroup/cpuacct
cpu /cgroup/cpu
memory /cgroup/memory
devices /cgroup/devices
```

Change Considerations for `pbs_mom`

In order to improve performance when removing cgroup hierarchies and job files Torque 6.0.0 added a new MOM configuration parameter `$thread_unlink_calls`. This parameter puts the job file cleanup processes on their own thread which increases the performance of the MOM. However, the addition of new threads also increases the size of `pbs_mom` from around 50 mb to 100mb.

`$thread_unlink_calls` is true by default which will thread job deletion. If `pbs_mom` is too large for your configuration set `$thread_unlink_calls` to false and jobs will be deleted within the main `pbs_mom` thread.

Torque NUMA-Support Configuration

i This topic provides instructions for enabling NUMA-support on large-scale SLES systems using SGI Altix and UV hardware and requires Torque 3.0 or later. For instructions on enabling NUMA-aware, see [Torque NUMA-Aware Configuration on page 49](#).

Do the following in order:

- ["Configure Torque for NUMA-Support" below](#)
- ["Create the mom.layout File" below](#)
- ["Configure the server_priv/nodes File" on page 56](#)
- ["Limit Memory Resources \(Optional\)" on page 56](#)

Configure Torque for 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
```

i 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.

Create the mom.layout File

This section provides instructions to create the *mom.layout* file.

Do *one* of the following:

- [Automatically Create mom.layout \(Recommended\) on page 53](#)
- [Manually Create mom.layout on page 54](#)

Automatically Create mom.layout (Recommended)

A perl script named *mom_genclfg* 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, do the following:

1. Verify `hwloc` library and corresponding `hwloc-devel` package are installed. See ["Installing Torque Resource Manager" on page 10](#) for more information.
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.

```
$ ./mom_gencfg
```

Manually Create 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:

```
Partition number: 0
6 Blades
72 CPUs
378.43 Gb Memory Total
```

Blade	ID	asic	NASID	Memory
0	r001i01b00	UVHub 1.0	0	67089152 kB
1	r001i01b01	UVHub 1.0	2	67092480 kB
2	r001i01b02	UVHub 1.0	4	67092480 kB
3	r001i01b03	UVHub 1.0	6	67092480 kB
4	r001i01b04	UVHub 1.0	8	67092480 kB
5	r001i01b05	UVHub 1.0	10	67092480 kB

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

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 root 0 Dec 3 12:14 .
drwxr-xr-x 14 root root 0 Dec 3 12:13 ..
-r--r--r-- 1 root root 4096 Dec 3 14:58 has_cpu
-r--r--r-- 1 root root 4096 Dec 3 14:58 has_normal_memory
drwxr-xr-x 2 root root 0 Dec 3 12:14 node0
drwxr-xr-x 2 root root 0 Dec 3 12:14 node1
drwxr-xr-x 2 root root 0 Dec 3 12:14 node10
drwxr-xr-x 2 root root 0 Dec 3 12:14 node11
drwxr-xr-x 2 root root 0 Dec 3 12:14 node2
drwxr-xr-x 2 root root 0 Dec 3 12:14 node3
drwxr-xr-x 2 root root 0 Dec 3 12:14 node4
drwxr-xr-x 2 root root 0 Dec 3 12:14 node5
drwxr-xr-x 2 root root 0 Dec 3 12:14 node6
drwxr-xr-x 2 root root 0 Dec 3 12:14 node7
drwxr-xr-x 2 root root 0 Dec 3 12:14 node8
drwxr-xr-x 2 root root 0 Dec 3 12:14 node9
-r--r--r-- 1 root root 4096 Dec 3 14:58 online
-r--r--r-- 1 root root 4096 Dec 3 14:58 possible
```

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.

Configure the `server_priv/nodes` File

The `pbs_server` requires awareness of how the MOM is reporting nodes since there is only one MOM daemon and multiple MOM nodes.

You need to 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.

The following is an example of how to configure the nodes file with `num_node_boards`.

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

In this example, 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.

The previous example showed that the NUMA system is uniform in its configuration of CPUs per node board. However, 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 example, `pbs_server` knows it has 3 MOM nodes and the nodes have 6, 8, and 12 CPUs respectively. Notice that the attribute `np` is not used. The `np` attribute is ignored because the number of CPUs per node is expressly given.

Limit Memory Resources (Optional)

Torque can better enforce memory limits with the use of the `memacctd` utility. The `memacctd` utility is a daemon that caches memory footprints when it is queried. When configured to use the memory monitor, Torque queries `memacctd`.



The `memacctd` utility is provided by SGI for SLES systems only. It is up to the user to make sure `memacctd` is installed.

To configure Torque to use `memacctd` for memory enforcement, do the following:

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.

i You use the `qsub` filter to include a default memory limit for all jobs that are not submitted with memory limit.

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" below](#)
- ["Stopping pbs_mom in Multi-MOM Mode" on page 59](#)

Multi-MOM Configuration

There are three steps to setting up multi-MOM capability:

1. ["Configure server_priv/nodes" below](#)
2. ["/etc/hosts file" on the next page](#)
3. ["Starting pbs_mom with Multi-MOM Options" on the next page](#)

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 222](#) 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 the previous page](#)

["Stopping pbs_mom in Multi-MOM Mode" on the facing page](#)

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 57](#)

["Multi-MOM Configuration" on page 57](#)

Chapter 3: Submitting and Managing Jobs

This section contains information about how you can submit and manage jobs with Torque. See the following topics for details:

- ["Job Submission" below](#)
- ["Monitoring Jobs" on page 81](#)
- ["Canceling Jobs" on page 83](#)
- ["Job Preemption" on page 83](#)
- ["Keeping Completed Jobs" on page 84](#)
- ["Job Checkpoint and Restart" on page 85](#)
- ["Job Exit Status" on page 95](#)
- ["Service Jobs" on page 99](#)

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.

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

This is an example of properly declared PBS directives.

```
#PBS -S /bin/bash
SOMEVARIABLE=42
#PBS -m abe
#PBS -M <yourEmail@company.com>
echo sleep 300
```

This is an example of improperly declared PBS directives. PBS directives below "SOMEVARIABLE=42" are ignored.

i By default, job submission is allowed only on the Torque server host (host on which `pbs_server` is running). Enablement of job submission from other hosts is documented in ["Configuring Job Submission Hosts" on page 38](#).

i Versions 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.

This section contains these topics:

- ["Multiple Job Submission" below](#)
- ["Requesting Resources" on page 65](#)
- ["Requesting Generic Resources" on page 75](#)
- ["Requesting Floating Resources" on page 75](#)
- ["Requesting Other Resources" on page 76](#)
- ["Exported Batch Environment Variables" on page 76](#)
- ["Enabling Trusted Submit Hosts" on page 78](#)
- ["Example Submit Scripts" on page 78](#)

Related Topics

- [Maui Documentation](#)
- ["Job Submission Filter \("qsub Wrapper"\)" on page 381](#) – Allow local checking and modification of submitted job

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 ...
```

i 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`.

Related Topics

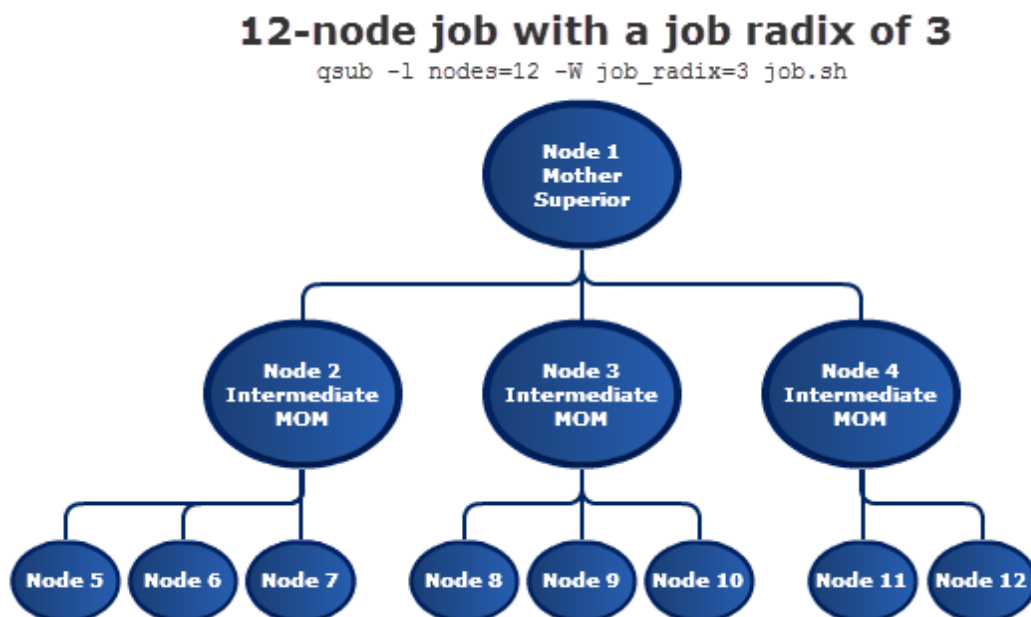
["Job Submission" on page 61](#)

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 ["qsub" on page 275](#) 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 3-1: 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.

i 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 (with the `-l <resource>` syntax) or external scheduler resource extensions (with `-W x=`) may be specified.

`qsub -l` supports:

- All the native Torque resources. See [Native Torque Resources on page 65](#) for a list of resources.
- Some Moab scheduler job extensions (for legacy support). See [Moab Job Extensions on page 73](#) for a list of resources.

i For Moab resource extensions, `qsub -W x=` is recommended instead of `qsub -l`. See Resource Manager Extensions in the *Moab Workload Manager Administrator Guide* for a complete list of scheduler-only job extensions.


Native Torque Resources

The native Torque resources are listed in the following table.



Resource	Format	Description
arch	string	Specifies the administrator defined system architecture required. This defaults to whatever the <code>PBS_MACH</code> string is set to in <code>local.mk</code> .
cput	seconds, or [[HH:]MM:]SS	Maximum amount of CPU time used by all processes in the job.

Resource	Format	Description
cpuclock	string	<p>Specify the CPU clock frequency for each node requested for this job. A <code>cpuclock</code> request applies to every processor on every node in the request. Specifying varying CPU frequencies for different nodes or different processors on nodes in a single job request is not supported.</p> <p>Not all processors support all possible frequencies or ACPI states. If the requested frequency is not supported by the CPU, the nearest frequency is used.</p> <p>ALPS 1.4 or later is required when using <code>cpuclock</code> on Cray.</p> <p>The clock frequency can be specified via:</p> <ul style="list-style-type: none"> a number that indicates the clock frequency (with or without the SI unit suffix). <div data-bbox="695 709 1409 873" data-label="Text"> <pre>qsub -l cpuclock=1800,nodes=2 script.sh qsub -l cpuclock=1800mhz,nodes=2 script.sh</pre> <p><i>This job requests 2 nodes and specifies their CPU frequencies should be set to 1800 MHz.</i></p> </div> a Linux power governor policy name. The governor names are: <ul style="list-style-type: none"> <code>performance</code>: This governor instructs Linux to operate each logical processor at its maximum clock frequency. This setting consumes the most power and workload executes at the fastest possible speed. <code>powersave</code>: This governor instructs Linux to operate each logical processor at its minimum clock frequency. This setting executes workload at the slowest possible speed. This setting does not necessarily consume the least amount of power since applications execute slower, and may actually consume more energy because of the additional time needed to complete the workload's execution. <code>ondemand</code>: This governor dynamically switches the logical processor's clock frequency to the maximum value when system load is high and to the minimum value when the system load is low. This setting causes workload to execute at the fastest possible speed or the slowest possible speed, depending on OS load. The system switches between consuming the most power and the least power.

Resource	Format	Description
		<div> <div> <i>i</i> <p>The power saving benefits of <code>ondemand</code> might be non-existent due to frequency switching latency if the system load causes clock frequency changes too often. This has been true for older processors since changing the clock frequency required putting the processor into the C3 "sleep" state, changing its clock frequency, and then waking it up, all of which required a significant amount of time.</p> <p>Newer processors, such as the Intel Xeon E5-2600 Sandy Bridge processors, can change clock frequency dynamically and much faster.</p> </div> <div> <ul style="list-style-type: none"> ◦ <code>conservative</code>: This governor operates like the <code>ondemand</code> governor but is more conservative in switching between frequencies. It switches more gradually and uses all possible clock frequencies. <p>This governor can switch to an intermediate clock frequency if it seems appropriate to the system load and usage, which the <code>ondemand</code> governor does not do.</p> </div> </div> <div> <pre>qsub -l cpuclock=performance,nodes=2 script.sh</pre> <div> <p><i>This job requests 2 nodes and specifies their CPU frequencies should be set to the performance power governor policy.</i></p> </div> </div> <div> <ul style="list-style-type: none"> • an ACPI performance state (or P-state) with or without the P prefix. P-states are a special range of values (0-15) that map to specific frequencies. Not all processors support all 16 states, however, they all start at P0. P0 sets the CPU clock frequency to the highest performance state which runs at the maximum frequency. P15 sets the CPU clock frequency to the lowest performance state which runs at the lowest frequency. </div> <div> <pre>qsub -l cpuclock=3,nodes=2 script.sh qsub -l cpuclock=p3,nodes=2 script.sh</pre> <div> <p><i>This job requests 2 nodes and specifies their CPU frequencies should be set to a performance state of 3.</i></p> </div> </div>
		<p>When reviewing job or node properties when <code>cpuclock</code> was used, be mindful of unit conversion. The OS reports frequency in Hz, not MHz or GHz.</p>

Resource	Format	Description
epilogue	string	Specifies a user owned epilogue script which will be run before the system epilogue and epilogue.user scripts at the completion of a job. The syntax is <code>epilogue=<file></code> . The file can be designated with an absolute or relative path. For more information, see " Prologue and Epilogue Scripts " on page 369.
feature	string	Specifies a property or feature for the job. Feature corresponds to Torque node properties and Moab features. <pre>qsub script.sh -l procs=10,feature=bigmem</pre>
file	size	Sets RLIMIT_FSIZE for each process launched through the TM interface. See <code>FILEREQUESTISJOBCENTRIC</code> for information on how Moab schedules.
host	string	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.
mem	size	Maximum amount of physical memory used by the job. Ignored on Darwin, Digital Unix, Free BSD, HPUX 11, IRIX, NetBSD, and SunOS. Not implemented on AIX and HPUX 10. The <code>mem</code> resource will only work for single-node jobs. If your job requires multiple nodes, use <code>pmem</code> instead.
ncpus	integer	The number of processors in one task where a task cannot span nodes. <div> You cannot request both <code>ncpus</code> and <code>nodes</code> in the same job.</div>
nice	integer	Number between -20 (highest priority) and 19 (lowest priority). Adjust the process execution priority.

Resource	Format	Description
nodes	<pre>{<node_count> <hostname>} [:ppn=<ppn>] [:gpus=<gpu>] [:<property>] [:<property>]... [+ ...]</pre>	<p>Number and/or type of nodes to be reserved for exclusive use by the job. The value is one or more <code>node_specs</code> joined with the + (plus) character: <code>node_spec[+node_spec...]</code>. Each <code>node_spec</code> is a number of nodes required of the type declared in the <code>node_spec</code> and a name of one or more properties desired for the nodes. The number, the name, and each property in the <code>node_spec</code> 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:</p> <ul style="list-style-type: none"> • <code>ppn=#</code> - 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 <code>TORQUE_HOME/server_priv/nodes</code> file using the <code>np</code> attribute (see "Server Node File Configuration" on page 45). 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 <code>np</code> value greater than the number of physical cores (or hyper-thread contexts). The <code>ppn</code> value is a characteristic of the hardware, system, and site, and its value is to be determined by the administrator. • <code>gpus=#</code> - Specify the number of GPUs per node requested for this job. The number of GPUs available on a node can be configured in the <code>TORQUE_HOME/server_priv/nodes</code> file using the <code>gpu</code> attribute (see "Server Node File Configuration" on page 45). The GPU value is a characteristic of the hardware, system, and site, and its value is to be determined by the administrator. • <code>property</code> - 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. <div> <p>i 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 <code>qsub</code>.</p> </div> <p>See "qsub -l nodes" on page 73 for examples.</p> <div> <p>i 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 <code>JOBNODEMATCHPOLICY</code> parameter. See Moab Parameters in the <i>Moab Workload Manager Administrator Guide</i> for more information.</p> </div>

Resource	Format	Description
		<p> All nodes in Torque have their own name as a property. You may request a specific node by using its name in the nodes request. Multiple nodes can be requested this way by using '+' as a delimiter. For example:</p> <pre>qsub -l nodes=node01:ppn=3+node02:ppn=6</pre> <p>See the HOSTLIST RM extension in the <i>Moab Workload Manager Administrator Guide</i> for more information.</p>
opsys	string	Specifies the administrator defined operating system as defined in the MOM configuration file.
other	string	<p>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.</p> <p> This does not work for <i>msub</i> using Moab and Maui.</p>
pcput	seconds, or [[HH:]MM:]SS	Maximum amount of CPU time used by any single process in the job.
pmem	size	Maximum amount of physical memory used by any single process of the job. (Ignored on Fujitsu. Not implemented on Digital Unix and HP-UX.)
procs	procs=<integer>	<p>(Applicable in version 2.5.0 and later.) 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 qsub command.</p> <pre>> qsub -l nodes=3 -l procs=2</pre>
procs_bitmap	string	<p>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.</p> <p>For more information, see "Scheduling Cores" on page 112.</p>
prologue	string	<p>Specifies a user owned prologue script which will be run after the system prologue and prologue.user scripts at the beginning of a job. The syntax is <code>prologue=<file></code>. The file can be designated with an absolute or relative path.</p> <p>For more information, see "Prologue and Epilogue Scripts" on page 369.</p>

Resource	Format	Description
pvmem	size	Maximum amount of virtual memory used by any single process in the job. (Ignored on Unicos.)
size	integer	For Torque, this resource has no meaning. It is passed on to the scheduler for interpretation. In the Moab scheduler, the <code>size</code> resource is intended for use in Cray installations only. See the note at the end of this table about the supported format.
software	string	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 License Management in the <i>Moab Workload Manager Administrator Guide</i> for more information.
vmem	size	Maximum amount of virtual memory used by all concurrent processes in the job. (Ignored on Unicos.)
walltime	seconds, or [[HH:]MM:]SS	Maximum amount of real time during which the job can be in the 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.

Suffix		Multiplier
b	w	1
kb	kw	1024
mb	mw	1,048,576
gb	gw	1,073,741,824
tb	tw	1,099,511,627,776

Interpreting Resource Requests

The table below shows how various requests are interpreted in the `qsub -l` syntax and corresponding cgroup usage.

Memory parameters (`mem`, `pmem`, `vmem`, `pvmem`) may specify units (examples: `mem=1024mb`, `mem=320kb`, `mem=1gb`). Recognized units are kb (kilobytes), mb (megabytes), gb (gigabytes), tb (terabyte), pb (petabytes), and eb (exabyte). If units are not specified, mb (megabytes) is assumed.

Example 3-1: Interpreting `qsub -l` requests

Usage	Description
<code>node=X:ppn=Y</code>	Creates <i>X</i> tasks that will use <i>Y</i> lprocs per task.
<code>procs=X</code>	Creates <i>X</i> tasks that will use 1 lproc each.
<code>ncpus=X</code>	Creates 1 task that will use <i>X</i> lprocs.
<code>mem=X</code>	The entire job will use <i>X</i> memory, divided evenly among the tasks.*
<code>pmem=X</code>	Each task will use <i>X</i> memory. In translation, <code>-l nodes=1:ppn=4,pmem=1gb</code> will use 4 GB of memory.*
<code>vmem=X</code>	The entire job will use <i>X</i> swap, divided evenly among the tasks. If <code>legacy_vmem</code> is set to true in the server, then the entire specified value will be given per host.**
<code>pvmem=X</code>	Each task will use <i>X</i> swap. In translation, <code>-l nodes=1:ppn=4,pvmem=1gb</code> will use 4 GB of swap.**

*If both `mem` and `pmem` are specified, the less restrictive of the two will be used as the limit for the job. For example, `qsub job.sh -l nodes=2:ppn=2,mem=4gb,pmem=1gb` will apply the `mem` requested instead of `pmem`, because it will allow 2 GB per task (4 GB/2 tasks) instead of 1 GB per task.

**If both `vmem` and `pvmem` are specified, the less restrictive of the two will be used as the limit for the job. For example, `qsub job.sh -l nodes=2:ppn=2,vmem=4gb,pvmem=1gb` will apply `pvmem` instead of `vmem`, because it will allow 2 GB swap per task (1 GB * 2 ppn) instead of .5 GB per task (1 GB/2 tasks).

Interpreting Node Requests

The table below shows how various `qsub -l nodes` requests are interpreted.

Example 3-2: qsub -l nodes

Usage	Description
> qsub -l nodes=12	Request 12 nodes of any type
> qsub -l nodes=2:server+14	Request 2 "server" nodes and 14 other nodes (a total of 16) - this specifies two node_specs, "2:server" and "14"
> qsub -l nodes=s=server:hippi+10:noserver+3:bigmem:hippi	Request (a) 1 node that is a "server" and has a "hippi" interface, (b) 10 nodes that are not servers, and (c) 3 nodes that have a large amount of memory and have hippy
> qsub -l nodes=b2005+b1803+b1813	Request 3 specific nodes by hostname
> qsub -l nodes=4:ppn=2	Request 2 processors on each of four nodes
> qsub -l nodes=1:ppn=4	Request 4 processors on one node
> qsub -l nodes=2:blue:ppn=2+red:ppn=3+b1014	Request 2 processors on each of two blue nodes, three processors on one red node, and the compute node "b1014"

Example 3-3:

This job requests a node with 200MB of available memory:

```
> qsub -l mem=200mb /home/user/script.sh
```

Example 3-4:

This job will wait until node01 is free with 200MB of available memory:

```
> qsub -l nodes=node01,mem=200mb /home/user/script.sh
```

Moab Job Extensions

qsub -l also supports some Moab resource extension values listed below, but be advised that this usage has been deprecated. The ones that currently work will remain for the purpose of legacy support, but additional ones will not be added. Instead, we recommend transitioning to the -W x=<resource> syntax mentioned at the top of this page.

advres	image	procs	subnode_list
cpuclock	jgroup	procs_bitmap	taskdistpolicy
deadline	jobflags	prologue	template
depend	latency	qos	termsig
ddisk	loglevel	queuejob	termtime
dmem	minprocspeed	reqattr	tid
energy_used	minpreempttime	retrycount	tpn
epilogue	minwclimit	retrycc	trig
feature	naccesspolicy	rmtype	trl
flags	nallocpolicy	select	var
gattr	nodeset	sid	vcores
geometry	opsys	signal	wcrequeue
gmetric	os	stagein	
gres	partition	spriority	
hostlist	pref	subnode	

Related Topics

["Job Submission" on page 61](#)

Requesting NUMA-Aware Resources

i This topic only applies for NUMA-aware systems and requires Torque Resource Manager 6.0 and later.

Various NUMA resources can be requested at the time of job submission.

The `qsub -L` option allows administrators the ability to place jobs at the "task" or "OS process" level to get maximum efficiency out of the available hardware. In addition, multiple, non-symmetric resource requests can be made for the same job using the `-L` job submission syntax. See [-L NUMA Resource Request on page 173](#) for a complete list of `-L` values.

For example:

```
qsub -L tasks=4:lprocs=2:usecores:memory=500mb -L tasks=8:lprocs=4:memory=2gb
```

Creates two requests. The first will create 4 tasks with two logical processors per task and 500 mb of memory per task. The logical processors will be placed on cores. The second request calls for 8 tasks with 4 logical processors per task and 2 gb of memory per task. The logical processors may be placed on cores or threads since the default placement is allowthreads.

i The queue attribute `resources_default` has several options that are not compatible with the `qsub -L` syntax. If a queue has any of the following `resources_default` options set (again, without a corresponding `req_information_default` setting), the job will be rejected from the queue:

`nodes`, `size`, `mppwidth`, `mem`, `hostlist`, `ncpus`, `procs`, `pvmem`, `pmem`, `vmem`, `reqattr`, `software`, `geometry`, `opsys`, `tpn`, and `trl`.

i See [Setting Queue Resource Controls with Resource Request Syntax 2.0 on page 116](#) for more information about setting queue resource requirements and the use of `-l` and `-L` job submission syntaxes.

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 **Managing Consumable Generic Resources** on page 1 in the *Moab Workload Manager Administrator Guide* for details on configuration within Moab.

Example 3-5: Generic

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

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

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

Related Topics

["Requesting Resources" on page 65](#)

["Job Submission" on page 61](#)

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 **Managing Shared Cluster Resources (Floating Resources)** on page 1 in the *Moab Workload Manager Administrator Guide* for details on configuration within Moab.

Example 3-6: Floating

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

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

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

Related Topics

["Requesting Resources" on page 65](#)

["Job Submission" on page 61](#)

Requesting Other Resources

Many other resources can be requested at the time of job submission using Moab Workload Manager (via the `qsub -W x=syntax` (or `qsub -l` in limited cases), or via `msub -l` or `msub -W x=`). 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 65](#)
- ["Job Submission" on page 61](#)

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:

Variable	Description
PBS_JOBNAME	User specified jobname
PBS_ARRAYID	Zero-based value of job array index for this job (in version 2.2.0 and later)
PBS_GPUFILE	Line-delimited list of GPUs allocated to the job located in <code>TORQUE_HOME/aux/jobidgpu</code> . Each line follows the following format: <code><host>-gpu<number></code> For example, <code>myhost-gpu1</code> .
PBS_O_WORKDIR	Job's submission directory

Variable	Description
PBS_ENVIRONMENT	N/A
PBS_TASKNUM	Number of tasks requested
PBS_O_HOME	Home directory of submitting user
PBS_MOMPORT	Active port for MOM daemon
PBS_O_LOGNAME	Name of submitting user
PBS_O_LANG	Language variable for job
PBS_JOBCOOKIE	Job cookie
PBS_JOBID	Unique pbs job id
PBS_NODENUM	Node offset number
PBS_NUM_NODES	Number of nodes allocated to the job
PBS_NUM_PPN	Number of procs per node allocated to the job
PBS_O_SHELL	Script shell
PBS_O_HOST	Host on which job script is currently running
PBS_QUEUE	Job queue
PBS_NODEFILE	File containing line delimited list of nodes allocated to the job
PBS_NP	Number of execution slots (cores) for the job
PBS_O_PATH	Path variable used to locate executables within job script

Related Topics

["Requesting Resources" on page 65](#)

["Job Submission" on page 61](#)

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 39](#)).

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 39](#)).

Allows specified hosts to be trusted as a submit host.

- Use `.rhosts` to enable `ruserok()` based authentication (see ["Using RCmd authentication" on page 38](#)).

See ["Configuring Job Submission Hosts" on page 38](#) for more information.



When you enable ["allow_node_submit" on page 302](#), you must also enable the ["allow_proxy_user" on page 302](#) parameter to allow user proxying when submitting and running jobs.

Related Topics

["Job Submission" on page 61](#)

Example Submit Scripts

The following is an example job test script:

```
#!/bin/sh
#
#This is an example script example.sh
#
#These commands set up the Grid Environment for your job:
#PBS -N ExampleJob
#PBS -l nodes=1,walltime=00:01:00
#PBS -q np_workq
#PBS -M YOURUNIQNAME@umich.edu
#PBS -m abe

#print the time and date
date

#wait 10 seconds
sleep 10

#print the time and date again
date
```

Related Topics

["Job Submission" on page 61](#)

Job Files

Torque 4.5.0 was updated to accept XML-based job files in addition to the binary job files. The change allows job files to be more human-readable and easier to parse. Below is a sample job file in the new XML format:

```

<?xml version="1.0"?>
<job>
  <version>131842</version>
  <state>1</state>
  <substate>10</substate>
  <server_flags>33</server_flags>
  <start_time>0</start_time>
  <jobid>340</jobid>
  <fileprefix>340</fileprefix>
  <queue>batch</queue>
  <destination_queue></destination_queue>
  <record_type>1</record_type>
  <mom_address>0</mom_address>
  <mom_port>11</mom_port>
  <mom_rmport>0</mom_rmport>
  <attributes>
    <Job_Name flags="1">job2.sh</Job_Name>
    <Job_Owner flags="1">echan@moabServer.cn</Job_Owner>
    <job_state flags="1">Q</job_state>
    <queue flags="3">batch</queue>
    <server flags="1">company.com</server>
    <Checkpoint flags="1">u</Checkpoint>
    <ctime flags="1">1384292754</ctime>
    <Error_Path flags="1">moabServer.cn:/home/echan/work/job2.sh.e340</Error_Path>
    <Hold_Types flags="1">n</Hold_Types>
    <Join_Path flags="1">n</Join_Path>
    <Keep_Files flags="1">n</Keep_Files>
    <Mail_Points flags="1">a</Mail_Points>
    <mtime flags="1">1384292754</mtime>
    <Output_Path flags="1">moabServer.cn:/home/echan/work/job2.sh.o340</Output_Path>
    <Priority flags="1">0</Priority>
    <qtime flags="1">1384292754</qtime>
    <Rerunable flags="1">True</Rerunable>
    <Resource_List>
      <epilogue flags="1">/tmp/epilogue.sh</epilogue>
      <neednodes flags="1">moabServer:ppn=1</neednodes>
      <nodect flags="1">1</nodect>
      <nodes flags="1">moabServer:ppn=1</nodes>
    </Resource_List>
    <substate flags="1">10</substate>
    <Variable_List flags="1">PBS_O_QUEUE=batch
PBS_O_HOME=/home/echan
PBS_O_LOGNAME=echan
PBS_O_
PATH=/home/echan/eclipse:/usr/lib/lightdm/lightdm:/usr/local/sbin:/usr/local/bin:/usr/
sbin:/usr/bin:/sbin:/bin:/usr/games:/opt/moab/bin:/opt/moab/sbin
PBS_O_SHELL=/bin/bash
PBS_O_LANG=en_US
PBS_O_WORKDIR=/home/echan/work
PBS_O_HOST=moabServer.cn
PBS_O_SERVER=moabServer
</Variable_List>
    <euser flags="1">echan</euser>
    <egroup flags="5">company</egroup>
    <hop_count flags="1">1</hop_count>
    <queue_rank flags="1">2</queue_rank>
    <queue_type flags="1">E</queue_type>
    <etime flags="1">1384292754</etime>
    <submit_args flags="1">-l nodes=lei:ppn=1 -l epilogue=/tmp/epilogue.sh
./job2.sh</submit_args>
    <fault_tolerant flags="1">False</fault_tolerant>
    <job_radix flags="1">0</job_radix>
  </attributes>
</job>

```



```
<submit_host flags="1">lei.ac</submit_host>
</attributes>
</job>
```

The above job was submitted with this submit command:

```
qsub -l nodes=moabServer:ppn=1 -l epilogue=/tmp/epilogue.sh ./job2.sh
```

Related Topics

["Job Submission" on page 61](#)

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
...
```

Monitoring NUMA Job Task Placement

 NUMA-aware job task placement is available with Torque Resource Manager 6.0 and later.

When using NUMA, job resources are tracked per task. To support this `qstat -f` produces a new category of information that begins with the "req_information" keyword. Following each "req_information keyword" is another keyword giving information about how the job was allocated. See [-L NUMA Resource Request on page 173](#) for available allocation keywords.

When the job has completed, the output will also include the per task resident memory used and per task cpu time used. The following is a sample `qstat -f` completed job output.

You will see that `req_information.task_usage.O.task.O.cpu_list` gives the cores to which the job is bound for the cpuset. The same for `mem_list`. The keywords `memory_used` and `cput_used` report the per task resident memory used and cpu time used respectively.

```

Job Id: 832.pv-knielson-dt
Job Name = bigmem.sh
Job_Owner = knielson@pv-knielson-dt
resources_used.cput = 00:00:00
resources_used.energy_used = 0
resources_used.mem = 3628kb
resources_used.vmem = 31688kb
resources_used.walltime = 00:00:00
job_state = C
queue = second
server = pv-knielson-dt
Checkpoint = u
ctime = Tue Jul 28 23:23:15 2015
Error Path = pv-knielson-dt:/home/knielson/jobs/bigmem.sh.e832
exec_host = pv-knielson-dt/0-3
Hold_Types = n
Join_Path = n
Keep_Files = n
Mail_Points = a
mtime = Tue Jul 28 23:23:18 2015
Output_Path = pv-knielson-dt:/home/knielson/jobs/bigmem.sh.o832
Priority = 0
qtime = Tue Jul 28 23:23:15 2015
Rerunable = True
Resource_List.walltime = 00:05:00
session_id = 2708
substate = 59
Variable_List = PBS_O_QUEUE=routeme,PBS_O_HOME=/home/knielson,
PBS_O_LOGNAME=knielson,
PBS_O_PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/b
in:/usr/games:/usr/local/games,PBS_O_SHELL=/bin/bash,PBS_O_LANG=en_US,
PBS_O_WORKDIR=/home/knielson/jobs,PBS_O_HOST=pv-knielson-dt,
PBS_O_SERVER=pv-knielson-dt
euser = knielson
egroup = company
hashname = 832.pv-knielson-dt
queue_rank = 391
queue_type = E
etime = Tue Jul 28 23:23:15 2015
exit_status = 0
submit_args = -L tasks=2:lprocs=2 ../scripts/bigmem.sh
start_time = Tue Jul 28 23:23:18 2015
start_count = 1
fault_tolerant = False
comp_time = Tue Jul 28 23:23:18 2015
job_radix = 0
total_runtime = 0.093262
submit_host = pv-knielson-dt
req_information.task_count.0 = 2
req_information.lprocs.0 = 2
req_information.thread_usage_policy.0 = allowthreads
req_information.hostlist.0 = pv-knielson-dt:ppn=4
req_information.task_usage.0.task.0.cpu_list = 2,6
req_information.task_usage.0.task.0.mem_list = 0
req_information.task_usage.0.task.0.memory_used = 258048
req_information.task_usage.0.task.0.cput_used = 18
req_information.task_usage.0.task.0.cores = 0
req_information.task_usage.0.task.0.threads = 0
req_information.task_usage.0.task.0.host =
req_information.task_usage.0.task.1.cpu_list = 3,7
req_information.task_usage.0.task.1.mem_list = 0
req_information.task_usage.0.task.1.memory_used = 258048
req_information.task_usage.0.task.1.cput_used = 18
req_information.task_usage.0.task.1.cores = 0
req_information.task_usage.0.task.1.threads = 2
req_information.task_usage.0.task.1.host = pv-knielson-dt

```

Related Topics

["Submitting and Managing Jobs" on page 61](#)

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.

```
$ qstat
Job id          Name          User          Time Use S Queue
-----
4807            scatter       user01        12:56:34 R batch
...
$ qdel -m "hey! Stop abusing the NFS servers" 4807
$
```

Related Topics

["Submitting and Managing Jobs" on page 61](#)

Job Preemption

Torque supports job preemption by allowing authorized users to suspend and resume jobs. This is supported using one of two methods. If the node supports OS-level preemption, Torque will recognize that during the configure process and enable it. Otherwise, the MOM may be configured to launch a custom *checkpoint script* in order to support preempting a job. Using a custom checkpoint script requires that the job understand how to resume itself from a checkpoint after the preemption occurs.

Configuring a Checkpoint Script on a MOM

To configure the MOM to support a checkpoint script, the `$checkpoint_script` parameter must be set in the MOM's configuration file found in `TORQUE_HOME/mom_priv/config`. The checkpoint script should have execute permissions set. A typical configuration file might look as follows:

`mom_priv/config`:

```
$pbsserver      node06
$logevent        255
$restricted      *.mycluster.org
$checkpoint_script /opt/moab/tools/mom-checkpoint.sh
```

The second thing that must be done to enable the checkpoint script is to change the value of `MOM_CHECKPOINT` to 1 in `/src/include/pbs_config.h`. (In some instances, `MOM_CHECKPOINT` may already be defined as 1.) The new line should be as follows:

```
/src/include/pbs_config.h:
```

```
#define MOM_CHECKPOINT 1
```

Related Topics

["Submitting and Managing Jobs" on page 61](#)

Keeping Completed Jobs

Torque provides the ability to report on the status of completed jobs for a configurable duration after the job has completed. This can be enabled by setting the ["keep_completed" on page 403](#) attribute on the job execution queue or the ["keep_completed" on page 311](#) 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.

i 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 Moab Parameters in the *Moab Workload Manager Administrator Guide* for more information.

Related Topics

["Submitting and Managing Jobs" on page 61](#)

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" below](#)
- ["Configuration Files and Scripts" on the next page](#)
- ["Starting a Checkpointable Job" on page 93](#)
- ["Checkpointing a Job" on page 94](#)
- ["Restarting a Job" on page 95](#)
- ["Acceptance Tests" on page 95](#)

Related Topics

["Submitting and Managing Jobs" on page 61](#)

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:

```
# /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 the previous page](#)

Configuration Files and Scripts

Configuring and Building Torque for BLCR:

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

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

Option	Description
<code>--with-bldr-include=DIR</code>	include path for <code>libcr.h</code>
<code>--with-bldr-lib=DIR</code>	lib path for <code>libcr</code>
<code>--with-bldr-bin=DIR</code>	bin path for BLCR utilities

The `pbs_mom` 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.

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

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

i The script files below must be executable by the user. Be sure to use `chmod` to set the permissions to 754.

Example 3-7: Script file permissions

```
# chmod 754 blcr*
# ls -l
total 20
-rwxr-xr-- 1 root root 2112 2008-03-11 13:14 blcr_checkpoint_script
-rwxr-xr-- 1 root root 1987 2008-03-11 13:14 blcr_restart_script
-rw-r--r-- 1 root 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 root 7 2008-03-11 13:15 mom.lock
```

Example 3-8: 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 3-9: mom_priv/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          \n";

# Note that depth is not used in this script but could control a limit to the number
of checkpoint
# image files that are preserved on the disk.
#
# Note also that a request was made to identify whether this script was invoked by the
job's
# owner or by a system administrator. While this information is known to pbs_server,
it
# is not propagated to pbs_mom and thus it is not possible to pass this to the script.

# Therefore, a workaround is to invoke qmgr and attempt to set a trivial variable.
# This will fail if the invoker is not a manager.

if (@ARGV == 7)
{
    ($sessionId, $jobId, $userId, $checkpointDir, $checkpointName, $signalNum $depth)
    =
        @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): $!\n")
    if $logLevel;

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
{

```

```

    my ($level, $message) = @_ ;
    my @severity = ('none', 'warning', 'info', 'debug');

    return if $level > $logLevel;

    openlog('checkpoint_script', '', 'user');
    syslog($severity[$level], $message);
    closelog();
}

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

```

Example 3-10: mom_priv/blcr_restart_script

```

#!/usr/bin/perl
#####
#
# 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): !\n")
    if $logLevel;

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);
}

```

```
    die($message);
}
```

Related Topics

["Job Checkpoint and Restart" on page 85](#)

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 show below.

Option	Description
none	No checkpointing (not highly useful, but included for completeness).
enabled	Specify that checkpointing is allowed, but must be explicitly invoked by either the <code>qhold</code> or <code>qchkpt</code> commands.
shutdown	Specify that checkpointing is to be done on a job at <code>pbs_mom</code> shutdown.
periodic	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.
interval=minutes	Specify the checkpoint interval in minutes.
depth=number	Specify a number (depth) of checkpoint images to be kept in the checkpoint directory.
dir=path	Specify a checkpoint directory (default is <code>/var/spool/torque/checkpoint</code>).

Example 3-11: Sample test program

```
#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 3-12: Instructions for building test program

```
> gcc -o test test.c
```

Example 3-13: Sample test script

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

Example 3-14: Starting the test job

```
> qstat
> qsub -c enabled,periodic,shutdown,interval=1 test.sh
77.jakaa.cridomain
> qstat
```

Job id	Name	User	Time Use	S	Queue
77.jakaa	test.sh	jsmith	0	Q	batch

```
>
```

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 85](#)

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 85](#)

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 85](#)

Acceptance Tests

A number of tests were made to verify the functioning of the BLCR implementation. See "[BLCR Acceptance Tests](#)" on page 393 for a description of the testing.

Related Topics

["Job Checkpoint and Restart" on page 85](#)

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 3-15: qstat -f (job failure)

```

Job Id: 179.host
Job_Name = STDIN
Job_Owner = user@host
job_state = C
queue = batchq server = host
Checkpoint = u ctime = Fri Aug 29 14:55:55 2008
Error_Path = host:/opt/moab/STDIN.e179
exec_host = node1/0
Hold_Types = n
Join_Path = n
Keep_Files = n
Mail_Points = a
mtime = Fri Aug 29 14:55:55 2008
Output_Path = host:/opt/moab/STDIN.o179
Priority = 0
qtime = Fri Aug 29 14:55:55 2008
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_hint = Post job file processing error; job 179.host on host node1/0Ba
d UID for job execution REJHOST=pala.cridomain MSG=cannot find user 'user' in
password file
etime = Fri Aug 29 14:55:55 2008
exit_status = -1

```

i 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 3-16: Torque supplied exit codes

Name	Value	Description
JOB_EXEC_OK	0	Job execution successful
JOB_EXEC_FAIL1	-1	Job execution failed, before files, no retry
JOB_EXEC_FAIL2	-2	Job execution failed, after files, no retry
JOB_EXEC_RETRY	-3	Job execution failed, do retry
JOB_EXEC_INITABT	-4	Job aborted on MOM initialization

Name	Value	Description
JOB_EXEC_INITRST	-5	Job aborted on MOM init, chkpt, no migrate
JOB_EXEC_INITRMG	-6	Job aborted on MOM init, chkpt, ok migrate
JOB_EXEC_BADRESRT	-7	Job restart failed
JOB_EXEC_CMDFAIL	-8	Exec() of user command failed
JOB_EXEC_STDOUTFAIL	-9	Could not create/open stdout stderr files
JOB_EXEC_OVERLIMIT_MEM	-10	Job exceeded a memory limit
JOB_EXEC_OVERLIMIT_WT	-11	Job exceeded a walltime limit
JOB_EXEC_OVERLIMIT_CPUT	-12	Job exceeded a CPU time limit

Example 3-17: Exit code from C program

```

$ cat error.c

#include
#include

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

$ gcc -o error error.c

$ echo ./error | qsub
180.xxx.yyy

$ 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
  ctime = Wed Apr 30 11:29:37 2008
  Error_Path = xxx.yyy:/home/test/STDIN.e180
  exec_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
  qtime = Wed Apr 30 11:29:37 2008
  Rerunnable = True
  Resource_List.needsnodes = 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/s
    bin:/usr/local/bin:/usr/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 = 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

```

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 85](#)

["Submitting and Managing Jobs" on page 61](#)

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 the next page](#)
- ["Submitting Service Jobs in MCM" on the next page](#)

- ["Managing Service Jobs" below](#)

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 y" 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 the previous page](#)

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 the previous page](#)

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 99](#)

Chapter 4: Managing Nodes

This chapter 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" below](#)
- ["Node Properties" on page 105](#)
- ["Changing Node State" on page 105](#)
- ["Host Security" on page 109](#)
- ["Linux Cpuset Support" on page 111](#)
- ["Scheduling Cores" on page 112](#)

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 20](#).



Be aware of the following:

- Nodes cannot be added or deleted dynamically if there is a `mom_hierarchy` file in the `server_priv` directory.
- When you make changes to nodes by directly editing the nodes file, you must restart `pbs_server` for those changes to take effect. Changes made using `qmgr` do not require a restart.
- When you make changes to a node's ip address, you must clear the `pbs_server` cache. Either restart `pbs_server` or delete the changed node and then re-add it.
- Before a newly added node is set to a free state, the cluster must be informed that the new node is valid and they can trust it for running jobs. Once this is done, the node will automatically transition to free.
- Adding or changing a hostname on a node requires a `pbs_server` restart in order to add the new hostname as a node.

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[,node004,node005...] [np=n,][TTL=yyyy-mm-ddThh:mm:ssZ,][acl="user==user1:user2:user3",][requestid=n]'
```

The optional parameters are used as follows:

- **np** – Number of virtual processors.
- **TTL** – (Time to Live) Specifies the time in UTC format that the node is supposed to be retired by Moab. Moab will not schedule any jobs on a node after its time to live has passed.
- **acl** – (Access control list) Can be used to control which users have access to the node in Moab.



Except for temporary nodes and/or the simplest of cluster configurations, Adaptive Computing recommends avoiding the use of the `acl` parameter, as this can lead to confusion about the root cause of jobs being unable to run. Use Moab reservations with user ACLs instead.

- **requestid** – An ID that can be used to track the request that created the node.

You can alter node parameters by following these examples:

```
qmgr -c 'set node node003 np=6'
qmgr -c 'set node node003 TTL=2020-12-31T23:59:59Z'
qmgr -c 'set node node003 requestid=23234'
qmgr -c 'set node node003 acl="user:user10:user11:user12"'
qmgr -c 'set node node003 acl=""'
```



Torque does not use the `TTL`, `acl`, and `requestid` parameters. Information for those parameters are simply passed to Moab.



The `set node` subcommand of `qmgr` supports the `+=` and `-=` syntax, but has known problems when used to alter the `acl` parameter. Do not use it for this. Instead, simply reset the full user list, as shown in the above example.

The `create node` and `set node` commands will append the following line(s) to the bottom of the `TORQUE_HOME/server_priv/nodes` file:

```
node003 np=6 TTL=2020-12-31T23:59:59Z acl=user1:user2:user3 requestid=3210
node004 ...
```

Nodes can also be removed with a similar command:

```
> qmgr -c 'delete node node003[,node004,node005...]'
```

Related Topics

["Changing Node State" on the facing page](#)

["Managing Nodes" on the previous page](#)

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
```

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

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

Related Topics

["Job Submission" on page 61](#)

["Managing Nodes" on page 103](#)

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.)

Node Recovery

When a mom gets behind on replying to requests, `pbs_server` has a failsafe to allow for node recovery in processing the backlog. After three failures without having two consecutive successes in servicing a request, `pbs_server` will mark that mom as offline for five minutes to allow the mom extra time to process the backlog before it resumes its normal activity. If the mom has two consecutive successes in responding to network requests before the timeout, then it will come back earlier.

Related Topics

[Chapter 4: Managing Nodes on page 103](#)

Changing Node Power States

Beginning with Torque 5.0.0, the `pbsnodes -m` command can modify the power state of nodes. Node cannot go from one low-power state to another low-power state. They must be brought up to the Running state and then moved to the new low-power state. The supported power states are:

State	Description
Running	<ul style="list-style-type: none">Physical machine is actively workingPower conservation is on a per-device basisProcessor power consumption controlled by P-states
Standby	<ul style="list-style-type: none">System appears offProcessor halted (OS executes a "halt" instruction)Processor maintains CPU and system cache stateRAM refreshed to maintain memory stateMachine in low-power modeRequires interrupt to exit stateLowest-latency sleep state - has no effect on software

State	Description
Suspend	<ul style="list-style-type: none"> • System appears off • Processor and support chipset have no power • OS maintains CPU, system cache, and support chipset state in memory • RAM in slow refresh • Machine in lowest-power state • Usually requires specific interrupt (keyboard, mouse) to exit state • Third lowest-latency sleep state - system must restore power to processor and support chipset
Hibernate	<ul style="list-style-type: none"> • System is off • Physical machine state and memory saved to disk • Requires restoration of power and machine state to exit state • Second highest-latency sleep state - system performs faster boot using saved machine state and copy of memory
Shutdown	<ul style="list-style-type: none"> • Equivalent to <code>shutdown now</code> command as root

In order to wake nodes and bring them up to a running state:

- the nodes must support, and be configured to use, Wake-on-LAN (WOL).
- the `pbsnodes` command must report the node's MAC address correctly.

To configure nodes to use Wake-on-LAN

1. Enable WOL in the BIOS for each node. If needed, contact your hardware manufacturer for details.
2. Use the `ethtool` command to determine what types of WOL packets your hardware supports. Torque uses the `g` packet. If the `g` packet is not listed, you cannot use WOL with Torque.

```
[root]# ethtool eth0
Settings for eth0:
    Supported ports: [ TP ]
    Supported link modes:   10baseT/Half 10baseT/Full
                           100baseT/Half 100baseT/Full
                           1000baseT/Full

    Supported pause frame use: No
    Supports auto-negotiation: Yes
    Advertised link modes:  10baseT/Half 10baseT/Full
                           100baseT/Half 100baseT/Full
                           1000baseT/Full

    Advertised pause frame use: No
    Advertised auto-negotiation: Yes
    Speed: 100Mb/s
    Duplex: Full
    Port: Twisted Pair
    PHYAD: 2
    Transceiver: internal
    Auto-negotiation: on
    MDI-X: off
    Supports Wake-on: pumbg
    Wake-on: p
    Current message level: 0x00000007 (7)
                           drv probe link

    Link detected: yes
```

This Ethernet interface supports the g WOL packet, but is currently set to use the p packet.

3. If your Ethernet interface supports the g packet, but is configured for a different packet, use `ethtool -s <interface> wol g` to configure it to use g.

```
[root]# ethtool -s eth0 wol g
[root]# ethtool eth0
Settings for eth0:
    Supported ports: [ TP ]
    Supported link modes:   10baseT/Half 10baseT/Full
                           100baseT/Half 100baseT/Full
                           1000baseT/Full

    Supported pause frame use: No
    Supports auto-negotiation: Yes
    Advertised link modes:  10baseT/Half 10baseT/Full
                           100baseT/Half 100baseT/Full
                           1000baseT/Full

    Advertised pause frame use: No
    Advertised auto-negotiation: Yes
    Speed: 100Mb/s
    Duplex: Full
    Port: Twisted Pair
    PHYAD: 2
    Transceiver: internal
    Auto-negotiation: on
    MDI-X: off
    Supports Wake-on: pumbg
    Wake-on: g
    Current message level: 0x00000007 (7)
                           drv probe link

    Link detected: yes
```

Now the power state of your nodes can be modified and they can be woken up from power-saving states.

Related Topics

["pbsnodes" on page 236](#)

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.

i Security Enhanced Linux (SELinux) must either be disabled or configured to properly work with PAM.

To enable Torque PAM on the compute nodes, include the `--with-pam[=<DIR>]` option when running `configure` before installing, and then make the required security configuration file changes described in this section.

The installation process (below) will place `pam_pbssimpleauth.la` and `pam_pbssimpleauth.so` in either `/lib/security` or `/lib64/security` (depending on your build architecture), or in the directory optionally designated with the `configure` option.

When installing with packages, be sure to also run `torque-package-pam-linux.sh` or `torque-package-pam-linux-x86_64.sh` on each `pbs_mom` host.

When installing from source with `"make install mom"`, be sure to also run `"make install_pam"` on each `pbs_mom` host. (Running `"make install"` will install everything.)

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.

i This example is not intended to be used as a general purpose solution, and must be modified for your configuration.

In this example, after installing Torque with PAM enabled, you would add the following two lines to `/etc/pam.d/sshd` (on the compute nodes):

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

Also, in `/etc/security/access.conf` on the compute nodes, make sure to add all users who access the node to the configuration. This example permits node access to users `root`, `george`, `allen`, and `michael`.

```
-: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 open sessions to the node. If `george` currently has no running jobs on a node, `ssh` will refuse login attempts to that node as that user, and will close the connection.

The Torque PAM module will keep users out unless they 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 369 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 103](#)


Linux Cpuset Support

- ["Cpuset Overview" below](#)
- ["Cpuset Support" below](#)
- ["Configuring Cpuset" on the next page](#)
- ["Cpuset Advantages/Disadvantages" on the next page](#)

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 and corresponding `hwloc-devel` package installed. See ["Installing Torque Resource Manager" on page 10](#) for more information.

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.

Configuring Cpuset

To configure cpuset

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

```
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:

```
./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 103](#)

["Geometry Request Configuration" on the facing page](#)

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.

See these topics for details:

- ["Geometry Request Configuration" on the facing page](#)
- ["Geometry Request Usage" on the facing page](#)

- ["Geometry Request Considerations" on the next page](#)

Geometry Request Configuration

A Linux kernel of 2.6 or later is required to use geometry requests, because this feature uses Linux cpusets 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 111](#). 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.

i The geometry request feature uses a subset of the cpusets 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 the previous page](#)

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.

i 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 openssl.sh
```

The job `openssl.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 the previous page](#)

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 112](#)

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. See **Accelerators** on page 1 in the *Moab Workload Manager Administrator Guide* for information on configuring accelerators in Torque.

Chapter 5: Setting Server Policies

This section explains how to set up and configure your queue. This section also explains how to set up Torque to run in high availability mode. For details, see these topics:

- ["Queue Configuration" below](#)
- ["Server High Availability" on page 119](#)

Queue Configuration

To initially define a queue, use the *create* subcommand of **qmgr**:

```
> 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*.

```
> qmgr -c "set queue batch started=true"
> qmgr -c "set queue batch enabled=true"
```

Further configuration is possible using any combination of the following attributes.

For Boolean attributes, *T*, *t*, *1*, *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:

- [Example Queue Configuration on page 115](#)
- [Setting Queue Resource Controls with Resource Request Syntax 2.0 on page 116](#)
- [Setting a Default Queue on page 117](#)
- [Mapping a Queue to Subset of Resources on page 117](#)
- [Creating a Routing Queue on page 118](#)

Related Topics

[Queue Attributes on page 399](#)

[Server Parameters on page 299](#)

["qalter" on page 239](#) - command which can move jobs from one queue to another

Example Queue Configuration

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

```
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.



See [Setting Queue Resource Controls with Resource Request Syntax 2.0](#) on page 116 for more information about setting queue resource requirements and the use of `-l` and `-L` job submission syntaxes.

Related Topics

["Queue Configuration" on the previous page](#)

Setting Queue Resource Controls with Resource Request Syntax 2.0

Using the `-L` syntax, you can set default, minimum, and maximum values for `lprocs`, `memory`, `swap`, `disk`, `sockets`, `numanode`, `cores` and `threads` with resource request 2.0.

These can be set in the general format:

```
qmgr -c "set queue <queue_name> req_information_[default|min|max].
[lprocs|memory|swap|disk|sockets|numanode|core|thread]"
```

Example 5-1: Jobs using `-L` syntax

```
qmgr -c "set queue q1 req_information_default.lprocs=2"
qmgr -c "set queue q1 req_information_minimum.memory=2gb"
qmgr -c "set queue q1 req_information_maximum.core=10"
```



Regarding queue defaults and the newer `-L` NUMA-aware syntax: with *only* the newer `req_information_default.<attribute>` configured on a queue, the queue will only accept submissions with the `-L` syntax. The same holds true for `resources_default.<attribute>` and `-l` submissions. Setting both on a queue (as in [Example 5-2](#)) enables the queue to accept job submissions with either syntax.

Example 5-2: Jobs using `-L` or `-l` syntax

This example shows how to enable a queue to be able to accept *both* kinds of jobs and still be able to enforce defaults.

```
qmgr -c "create queue batch"
qmgr -c "set queue batch queue_type = Execution"
qmgr -c "set queue batch resources_default.mem = 3gb"
qmgr -c "set queue batch enabled = True"
qmgr -c "set queue batch started = True"
qmgr -c "set queue batch req_information_default.memory = 3gb"
```

In this example, jobs submitted that explicitly use the `-L` syntax will have the `req_information_default.memory` setting applied. If the job does *not* explicitly use this syntax, then the `resources_default.mem` setting will be applied.

Related Topics

["Queue Configuration" on page 115](#)

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 115](#)

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 and Maui 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.

```
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 115](#)

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.

```
qmgr

# routing queue
create queue route
set queue route queue_type = Route
set queue route route_destinations = reg_64
set queue route route_destinations += reg_32
set queue route route_destinations += reg
set queue route enabled = True
set queue route started = True

# queue for jobs using 1-15 nodes
create queue reg
set queue reg queue_type = Execution
set queue reg resources_min.ncpus = 1
set queue reg resources_min.nodect = 1
set queue reg resources_default.ncpus = 1
set queue reg resources_default.nodes = 1
set queue reg enabled = True
set queue reg started = True

# queue for jobs using 16-31 nodes
create queue reg_32
set queue reg_32 queue_type = Execution
set queue reg_32 resources_min.ncpus = 31
set queue reg_32 resources_min.nodes = 16
set queue reg_32 resources_default.walltime = 12:00:00
set queue reg_32 enabled = True
set queue reg_32 started = True

# queue for jobs using 32+ nodes
create queue reg_64
set queue reg_64 queue_type = Execution
set queue reg_64 resources_min.ncpus = 63
set queue reg_64 resources_min.nodes = 32
set queue reg_64 resources_default.walltime = 06:00:00
set queue reg_64 enabled = True
set queue reg_64 started = True

# have all jobs go through the routing queue
set server default_queue = route
set server resources_default.ncpus = 1
set server resources_default.walltime = 24:00:00
...
```

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.



You should not use a Torque routing queue in conjunction with Moab remap classes to route jobs to queues/nodes. You should select only one of the two methods.

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
```

Torque waits to apply the server and queue defaults until the job is assigned to its final execution queue. Queue defaults override the server defaults. 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.

i Routing queue defaults are not applied to job attributes in versions 2.1.0 and before.

i 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 115](#)

["Queue Attributes" on page 399](#)

Server High Availability

Torque can run 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 primary server fails.

i The high availability feature is available in the 2.3 and later versions of Torque.

i The "native" high availability implementation, as described here, is only suitable for Moab Basic Edition. Contact Adaptive Computing for information on high availability for Enterprise Edition.

For more details, see these sections:

- "Redundant server host machines" below
- "Enabling High Availability" on the facing page
- "Enhanced High Availability with Moab" on the facing page
- "How Commands Select the Correct Server Host" on page 122
- "Shutting Down the Local or Active pbs_server" on page 122
- "Job Names" on page 123
- "Persistence of the pbs_server Process" on page 123
- "High Availability of the NFS Server" on page 123
- "Installing Torque in High Availability Mode" on page 124
- "Installing Torque in High Availability Mode on Headless Nodes" on page 129
- "Example Setup of High Availability" on page 135

Redundant server host machines

High availability enables Moab HPC Suite to continue running even if pbs_server is brought down. This is done by running multiple copies of pbs_server, each of which has its TORQUE_HOME/server_priv directory mounted on a shared file system.



Do not use symlinks when sharing the Torque home directory or server_priv directories. A workaround for this is to use `mount --rbind /path/to/share /var/spool/torque`. Also, it is highly recommended that you only share the server_priv directory, not the entire TORQUE_HOME directory.

The server_name file (for *all* servers and compute nodes) must include the names of every pbs_server host. The syntax of the server_name file is a comma-delimited list of host names.

For example:

```
host1,host2,host3
```



When configuring high availability, do not use \$pbsserver in the pbs_mom configuration file (TORQUE_HOME/mom_priv/config on the compute nodes) to specify the server host names. You must use the TORQUE_HOME/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. Each subsequent server to start will block very early upon startup when it tries to lock the TORQUE_HOME/server_priv/server.lock file. When the 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 any server starting after the first 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.

Three server options help manage high availability. The server parameters are "`lock_file`" on page 312, "`lock_file_update_time`" on page 313, and "`lock_file_check_time`" on page 313.

The `lock_file` option allows the administrator to change the location of the lock file (default location: `TORQUE_HOME/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` updates the lock file based on the `lock_file_update_time` (default value of 3 seconds). All backup `pbs_server`s 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`.

i 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.

i Do not use anything but a simple NFS fileshare that is not used by anything else (i.e., only Moab can use the fileshare).

i Do not use a general-purpose NAS, parallel file system, or company-wide shared infrastructure to set up Moab high availability using "native" high availability.

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, set the `PBS_ARGS` environment variable in the `/etc/sysconfig/pbs_server` file.

```
PBS_ARGS="-l <moabhost:port>"
```

`moabhost` is the name of the alternate server node and `port` is the port on which Moab on the alternate server node is listening (the port is required and the default is 15004).

If Moab is running in HA mode, use the the `PBS_ARGS` environment variable to set the `-l` option for each redundant server.

```
PBS_ARGS="-l <moabhost:port> -l <moabhost2:port>"
```

If `pbs_server` and Moab run on the same host, use the `PBS_ARGS` environment variable in the `/etc/sysconfig/pbs_server` file to set the `--ha` option.

```
PBS_ARGS="--ha
```

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 `PBS_DEFAULT` environment variable or from the `TORQUE_HOME/server_name`.



`PBS_DEFAULT` overrides the value in the `server_name` file.

Whenever a Torque client command executes with no explicit server mentioned, it attempts to connect to the first server name listed in `PBS_DEFAULT` or `TORQUE_HOME/server_name`. If this fails, the command tries the next server name. 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.

Shutting Down the Local or Active `pbs_server`

Different commands are available to shut down `pbs_server`, depending upon whether you are executing the commands on the active server host or a backup server host, and whether you want to shut down the active server or a local backup server.

- The `qterm` command shuts down the active `pbs_server`, regardless of whether the command is executed on the active server host or a backup server host.

- The `service` or `systemctl` commands (depending on host OS version) shut down the `pbs_server` running on the local host, whether it is the active `pbs_server` host or a backup server host. For example, executing `service pbs_server stop` on a backup server host shuts down `pbs_server` on the local backup server host and leaves the active `pbs_server` running.

Red Hat 6 or SUSE 11-based systems

```
> service pbs_server stop
```

Red Hat 7 or SUSE 12-based systems

```
> systemctl stop pbs_server
```

Job Names

Job names normally contain the name of the host machine where `pbs_server` is running. When job names are constructed, only the server name in `$PBS_DEFAULT` or the first name from the server specification list, `TORQUE_HOME/server_name`, 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_servers` in the process table and starts more if needed.

High Availability of the NFS Server

i Before installing a specific NFS HA solution please contact Adaptive Computing Support for a detailed discussion on NFS HA type and implementation path.

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](#)
- [Sourceforge Linux NFS FAQ](#)
- [Red Hat Enterprise Linux 3: Reference Guide: Chapter 9: Network File System \(NFS\)](#)
- [Red Hat Enterprise Linux Deployment Guide: Chapter 18. Network File System \(NFS\)](#)
- [SUSE Linux Enterprise Administration Guide: Sharing File Systems with NFS](#)

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

- [Red Hat Enterprise Linux: Chapter 1. High Availability Add-On Overview](#)

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 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

To install Torque in HA mode

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

- Red Hat 6 or SUSE 11-based systems

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

- Red Hat 7 or SUSE 12-based systems

```
> systemctl stop firewalld
> systemctl disable firewalld
```

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 TORQUE_HOME=/var/spool/torque

# Library Path

export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:${TORQUE_HOME}/lib

# Update system paths
export PATH=${TORQUE_HOME}/bin:${TORQUE_HOME}/sbin:${PATH}
```

4. Verify `server1` and `server2` are resolvable via either DNS or looking for an entry in the `/etc/hosts` file.
5. Make the entries in the `/etc/sysconfig/pbs_server` files that enable `server1` and `server2` to communicate.

```
server1# PBS_ARGS="--ha -l <server2>:<port>"
server2# PBS_ARGS="--ha -l <server1>:<port>"
```

6. 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
```

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

- Red Hat 6 or SUSE 11-based systems

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

- Red Hat 7 or SUSE 12-based systems

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

8. 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
rsiz= 8192,wsiz=8192,timeo=14,intr
```

Repeat this step for server2.

9. Install Torque by following these steps:

- a. Download and extract Torque 6.0.4 on server1.

```
server1# wget http://github.com/adaptivecomputing/torque/branches/6.0.4/torque-
6.0.4.tar.gz
server1# tar -xvzf torque-6.0.4.tar.gz
```

- b. Navigate to the Torque directory and compile Torque 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.

10. Start `trqauthd`.

i For `trqauthd` to start correctly, you must have completed the "Configure the `trqauthd` daemon to start automatically at system boot" step in [Installing Torque Resource Manager on page 10](#)

- Red Hat 6 or SUSE 11-based systems

```
server1# service trqauthd start
```

- Red Hat 7 or SUSE 12-based systems

```
server1# systemctl start trqauthd
```

11. Configure Torque for HA.

- a. List the host names of all nodes that run `pbs_server` in the `TORQUE_HOME/server_name` file. You must also include the host names of all nodes running `pbs_server` in the `TORQUE_HOME/server_name` file of each MOM node. The syntax of `TORQUE_HOME/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"
```

i 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"
```

i Because `server_priv/*` is a shared drive, you do not need to repeat this step on Server 2.

- 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"
```

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

i 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.

i For `pbs_server` to start correctly, you must have completed the "Configure `pbs_server` to start automatically at system boot" step in [Installing Torque Resource Manager on page 10](#)

- Red Hat 6 or SUSE 11-based systems

```
service pbs_server stop
```

- Red Hat 7 or SUSE 12-based systems

```
systemctl stop pbs_server
```

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

- Red Hat 6 or SUSE 11-based systems

```
service pbs_server start
```

- Red Hat 7 or SUSE 12-based systems

```
systemctl start pbs_server
```

12. 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.

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

- Red Hat 6 or SUSE 11-based systems

```
service pbs_server stop
```

- Red Hat 7 or SUSE 12-based systems

```
systemctl stop pbs_server
```

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

13. 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/spool/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
```

i 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.

- Red Hat 6 or SUSE 11-based systems

```
service pbs_server stop
```

- Red Hat 7 or SUSE 12-based systems

```
systemctl stop pbs_server
```

- e. Start the `pbs_mom` on each execution node.

- Red Hat 6 or SUSE 11-based systems

```
service pbs_mom start
```

- Red Hat 7 or SUSE 12-based systems

```
systemctl start 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.

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)

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.

- Red Hat 6 or SUSE 11-based systems

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

- Red Hat 7 or SUSE 12-based systems

```
> systemctl stop firewalld
> systemctl disable firewalld
```

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 TORQUE_HOME=/var/spool/torque

# Library Path

export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:${TORQUE_HOME}/lib

# Update system paths
export PATH=${TORQUE_HOME}/bin:${TORQUE_HOME}/sbin:${PATH}
```

4. Verify `server1` and `server2` are resolvable via either DNS or looking for an entry in the `/etc/hosts` file.
5. Make the entries in the `/etc/sysconfig/pbs_server` files that enable `server1` and `server2` to communicate.

```
server1# PBS_ARGS="--ha -l <server2>:<port>"
server2# PBS_ARGS="--ha -l <server1>:<port>"
```

6. 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
```

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

- Red Hat 6 or SUSE 11-based systems

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

- Red Hat 7 or SUSE 12-based systems

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

8. 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,wsiz=8192,timeo=14,intr
```

Repeat this step for server2.

9. Install Torque by following these steps:

- a. Download and extract Torque 6.0.4 on server1.

```
server1# wget http://github.com/adaptivecomputing/torque/branches/6.0.4/torque-
6.0.4.tar.gz
server1# tar -xvzf torque-6.0.4.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.

10. Start `trqauthd`.

i For `trqauthd` to start correctly, you must have completed the "Configure the `trqauthd` daemon to start automatically at system boot" step in [Installing Torque Resource Manager on page 10](#)

- Red Hat 6 or SUSE 11-based systems

```
server1# service trqauthd start
```

- Red Hat 7 or SUSE 12-based systems

```
server1# systemctl start trqauthd
```

11. Configure Torque for HA.

- a. List the host names of all nodes that run `pbs_server` in the `TORQUE_HOME/server_name` file. You must also include the host names of all nodes running `pbs_server` in the `TORQUE_HOME/server_name` file of each MOM node. The syntax of `TORQUE_HOME/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"
```

i 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.

```
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"
```

i 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.

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

i 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.

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

i Because TORQUE_HOME is a shared drive, you do not need to repeat this step on server2.

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

i For `pbs_server` to start correctly, you must have completed the "Configure `pbs_server` to start automatically at system boot" step in [Installing Torque Resource Manager on page 10](#)

i You can specify command line arguments for `pbs_server` using the `PBS_ARGS` environment variable in the `/etc/sysconfig/pbs_server` file. Set `PBS_ARGS=--ha -l <host>:<port>` where `<host>` is the name of the alternate server node and `<port>` is the port on which `pbs_server` on the alternate server node is listening (default 15004).

- Red Hat 6 or SUSE 11-based systems

```
service pbs_server stop
```

- Red Hat 7 or SUSE 12-based systems

```
systemctl stop pbs_server
```

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

- Red Hat 6 or SUSE 11-based systems

```
service pbs_server start
```

- Red Hat 7 or SUSE 12-based systems

```
systemctl start pbs_server
```

12. 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.

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

- Red Hat 6 or SUSE 11-based systems

```
service pbs_server stop
```

- Red Hat 7 or SUSE 12-based systems

```
systemctl stop pbs_server
```

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

13. 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.

- Red Hat 6 or SUSE 11-based systems

```
service pbs_server stop
```

- Red Hat 7 or SUSE 12-based systems

```
systemctl stop pbs_server
```

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

- Red Hat 6 or SUSE 11-based systems

```
service pbs_mom start
```

- Red Hat 7 or SUSE 12-based systems

```
systemctl start pbs_mom
```

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`.

```
server1,server2
```

3. The machines running `pbs_server` must be listed in [acl_hosts](#).

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

4. Make the entries in the `/etc/sysconfig/pbs_server` files that enable `server1` and `server2` to communicate.

```
server1# PBS_ARGS="--ha -l <server2>:<port>"
server2# PBS_ARGS="--ha -l <server1>:<port>"
```

5. Start `pbs_server` with the `--ha` option.

- Red Hat 6 or SUSE 11-based systems

```
[root@server1]$ service pbs_server start
[root@server2]$ service pbs_server start
```

- Red Hat 7 or SUSE 12-based systems

```
[root@server1]$ systemctl start pbs_server
[root@server2]$ systemctl start pbs_server
```

Related Topics

- ["Setting Server Policies" on page 115](#)
- ["Queue Configuration" on page 115](#)

Setting min_threads and max_threads

There are two threadpools in Torque, one for background tasks and one for incoming requests from the MOMs and through the API (client commands, Moab, and so forth). The "[min_threads](#)" on page 318 and "[max_threads](#)" on page 317 parameters control the number of total threads used for both, not for each individually. The incoming requests' threadpool has three-quarters of `min_threads` for its minimum, and three-quarters of `max_threads` for its maximum, with the background pool receiving the other one-quarter.

Additionally, `pbs_server` no longer allows incoming requests to pile up indefinitely. When the threadpool is too busy for incoming requests, it indicates such, returning `PBSE_SERVER_BUSY` with the accompanying message that "Pbs Server is currently too busy to service this request. Please retry this request." The threshold for this message, if the request is from a manager, is that at least two threads be available in the threadpool. If the request comes from a non-manager, 5% of the threadpool must be available for the request to be serviced. Note that availability is calculated based on the maximum threads and not based on the current number of threads allocated.

If an undesirably large number of requests are given a busy response, one option is to increase the number of maximum threads for the threadpool. If the load on the server is already very high, then this is probably not going to help, but if the CPU load is lower, then it may help. Remember that by default the threadpool shrinks down once the extra threads are no longer needed. This is controlled via the "[thread_idle_seconds](#)" on page 324 server parameter.



Any change in the `min_threads`, `max_threads`, or `thread_idle_seconds` parameters requires a restart of `pbs_server` to take effect.

Chapter 6: 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, see **Moab-Torque Integration Guide** on page 1 in the *Moab Workload Manager Administrator Guide*. If using `pbs_sched`, simply start the `pbs_sched` daemon.

 If you are installing Moab Cluster Manager, Torque and Moab were configured at installation for interoperability and no further action is required.

Chapter 7: Configuring Data Management


This chapter provides instructions to configure Torque for data management purposes. For example, if you want to copy stdout and stderr files back to the submit host.

In this chapters:

- ["SCP Setup" below](#)
- ["NFS and Other Networked Filesystems" on page 142](#)
- ["File stage-in/stage-out" on page 143](#)

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" below](#)
- ["Copying Public SSH Key to Each Destination Host" on the next page](#)
- ["Configuring the SSH Daemon on Each Destination Host" on the next page](#)
- ["Validating Correct SSH Configuration" on page 141](#)
- ["Enabling Bi-Directional SCP Access" on page 141](#)
- ["Compiling Torque to Support SCP" on page 141](#)
- ["Troubleshooting" on page 142](#)

Related Topics

["Configuring Data Management" above](#)

Generating SSH Key on Source Host

On the source host as the transfer user, execute the following:

```
> ssh-keygen -t rsa
```

This will prompt for a passphrase (optional) and create two files (`id_rsa` and `id_rsa.pub`) inside `~/.ssh/`.

Related Topics

["SCP Setup" on the previous page](#)

["Copying Public SSH Key to Each Destination Host" below](#)

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 the previous page](#)

["SCP Setup" on the previous page](#)

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 the facing page](#) 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
```

Related Topics

["SCP Setup" on page 139](#)

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:

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

i 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 139](#)

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 139](#)

Compiling Torque to Support SCP

i 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-configured (and then rebuilt) to use SCP by passing in the `--with-scp` flag to the configure script:

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

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

Related Topics

["SCP Setup" on page 139](#)

Troubleshooting

If, after following all of the instructions in this section (see ["SCP Setup" on page 139](#)), 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 139](#)

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 to the `TORQUE_HOME/spool` directory (e.g., `/var/spool/torque/spool` on the execution host (or "mother superior").

When the job completes, the MOM copies the files back to the submission directory on the remote submit host. The file copy happens using a remote copy facility such as `rcp` (default) 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 `TORQUE_HOME/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 cannot copy the output or error files to the submission directory, it copies them to `TORQUE_HOME/undelivered` on that host.

Related Topics

- ["Configuring Data Management" on page 139](#)

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

```
# 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 139](#)

Chapter 8: 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" below](#)
- ["Open MPI" on the next page](#)

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](#) to determine if the reported issue has already been corrected.
- Send email to the mpiexec mailing list at 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 `$loglevel` to 5 or higher in `$TORQUE_HOME/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 `$TORQUE_HOME/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 369).

Related Topics

["MPI \(Message Passing Interface\) Support" on the previous page](#)

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](#).

TM Aware

To make use of Moab HPC Suite's TM interface, MPI must be configured to be TM aware.

Use these guidelines:

1. If you have installed from source, you need to use `./configure --with-tm` when you configure and make `openmpi`.
2. Run `mpirun` *without* the `-machinefile`. Moab HPC Suite will copy down the environment `PATH` and Library path down to each sister MOM. If `-machinefile` is used, `mpirun` will bypass the TM interface.

Example 8-1: Without TM aware

```
[jbooth@support-mpi1 ~]$ /usr/lib64/openmpi/bin/mpirun -np 4 -machinefile $PBS_
NODEFILE echo.sh
=====
support-mpi1
=====
/usr/lib64/openmpi/bin:/usr/lib64/openmpi/bin:/usr/lib64/qt-
3.3/bin:/usr/local/sbin:/usr/local/bin:/sbin:/bin:/usr/sbin:/usr/bin:/root/bin:/opt/mo
ab/bin:/opt/moab/sbin:/home/jbooth/bin

/usr/lib64/openmpi/lib:/usr/lib64/openmpi/lib

=====
support-mpi1
=====
/usr/lib64/openmpi/bin:/usr/lib64/openmpi/bin:/usr/lib64/qt-
3.3/bin:/usr/local/sbin:/usr/local/bin:/sbin:/bin:/usr/sbin:/usr/bin:/root/bin:/opt/mo
ab/bin:/opt/moab/sbin:/home/jbooth/bin

/usr/lib64/openmpi/lib:/usr/lib64/openmpi/lib

=====
support-mpi2
=====
/usr/lib64/openmpi/bin:/usr/lib64/openmpi/bin:/usr/lib64/qt-
3.3/bin:/usr/local/bin:/bin:/usr/bin

/usr/lib64/openmpi/lib:/usr/lib64/openmpi/lib:

=====
support-mpi2
=====
/usr/lib64/openmpi/bin:/usr/lib64/openmpi/bin:/usr/lib64/qt-
3.3/bin:/usr/local/bin:/bin:/usr/bin

/usr/lib64/openmpi/lib:/usr/lib64/openmpi/lib:
```

The paths, /opt/moab/bin and /opt/moab/sbin, were not passed down to the sister MOMs.

Example 8-2: With TM aware

```
[jbooth@support-mpi1 ~]$ /usr/local/bin/mpirun -np 4 echo.sh
=====
support-mpi1
=====
/usr/local/bin:/usr/local/bin:/usr/lib64/qt-
3.3/bin:/usr/local/sbin:/usr/local/bin:/sbin:/bin:/usr/sbin:/usr/bin:/root/bin:/opt/mo
ab/bin:/opt/roab/sbin:/home/jbooth/bin

/usr/local/lib:/usr/local/lib

=====
support-mpi1
=====
/usr/local/bin:/usr/local/bin:/usr/lib64/qt-
3.3/bin:/usr/local/sbin:/usr/local/bin:/sbin:/bin:/usr/sbin:/usr/bin:/root/bin:/opt/mo
ab/bin:/opt/roab/sbin:/home/jbooth/bin

/usr/local/lib:/usr/local/lib

=====
support-mpi2
=====
/usr/local/bin:/usr/local/bin:/usr/local/bin:/usr/lib64/qt-
3.3/bin:/usr/local/sbin:/usr/local/bin:/sbin:/bin:/usr/sbin:/usr/bin:/root/bin:/opt/mo
ab/bin:/opt/roab/sbin:/home/jbooth/bin

/usr/local/lib:/usr/local/lib:/usr/local/lib

=====
support-mpi2
=====
/usr/local/bin:/usr/local/bin:/usr/local/bin:/usr/lib64/qt-
3.3/bin:/usr/local/sbin:/usr/local/bin:/sbin:/bin:/usr/sbin:/usr/bin:/root/bin:/opt/mo
ab/bin:/opt/roab/sbin:/home/jbooth/bin

/usr/local/lib:/usr/local/lib:/usr/local/lib
```

The paths, /opt/roab/bin and /opt/roab/sbin, were passed down to the sister MOMs.

Related Topics

["MPI \(Message Passing Interface\) Support" on page 145](#)

Chapter 9: 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" below](#)
- ["Utilization" on the next page](#)
- ["Node States" on the next page](#)

9.0.1 Configuration

Configuration includes both detected hardware configuration and specified batch attributes.

Attribute	Description	Details
Architecture (arch)	operating system of the node	The value reported is a derivative of the operating system installed.
Node Features (properties)	arbitrary string attributes associated with the node	No node features are specified by default. If required, they are set using the <code>nodes</code> file located in the <code>TORQUE_HOME/server_priv</code> directory. They may specify any string and are most commonly used to allow users to request certain subsets of nodes when submitting jobs.
Local Disk (size)	configured local disk	By default, local disk space is not monitored. If the MOM configuration size [fs=<FS>] parameter is set, Torque will report, in kilobytes, configured disk space within the specified directory.
Memory (physmem)	local memory/RAM	Local memory/RAM is monitored and reported in kilobytes.

Attribute	Description	Details
Processors (ncpus/np)	real/virtual processors	The number of processors detected by Torque is reported via the <i>ncpus</i> attribute. However, for scheduling purposes, other factors are taken into account. In its default configuration, Torque operates in "dedicated" 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 <i>nodes</i> file located in the <code>TORQUE_HOME/server_priv</code> directory. An alternative to dedicated mode is "timeshared" mode. If Torque's time-shared mode is enabled, Torque will accept additional workload on each node until the node's <i>maxload</i> limit is reached.
Swap (totmem)	virtual memory/Swap	Virtual memory/Swap is monitored and reported in kilobytes.

9.0.2 Utilization

Utilization includes information regarding the amount of node resources currently in use as well as information about who or what is consuming it.

Attribute	Description	Details
Disk (size)	local disk availability	By default, local disk space is not monitored. If the MOM configuration size [fs=<FS>] parameter is set, Torque will report configured and currently available disk space within the specified directory in kilobytes.
Memory (availmem)	real memory/RAM	Available real memory/RAM is monitored and reported in kilobytes.
Network (netload)	local network adapter usage	Reports total number of bytes transferred in or out by the network adapter.
Processor Utilization (loadave)	node's cpu load average	Reports the node's 1 minute bsd load average.

9.0.3 Node States

State information includes administrative status, general node health information, and general usage status.

Attribute	Description	Details
Idle Time (idletime)	time since local keyboard/mouse activity has been detected	Time in seconds since local keyboard/mouse activity has been detected.
State (state)	monitored/admin node state	<p>A node can be in one or more of the following states:</p> <ul style="list-style-type: none"> • <i>busy</i> - node is full and will not accept additional work • <i>down</i> - node is failing to report, is detecting local failures with node • <i>free</i> - node is ready to accept additional work • <i>job-exclusive</i> - all available virtual processors are assigned to jobs • <i>job-sharing</i> - node has been allocated to run multiple shared jobs and will remain in this state until jobs are complete • <i>offline</i> - node has been instructed by an admin to no longer accept work • <i>reserve</i> - node has been reserved by the server • <i>time-shared</i> - node always allows multiple jobs to run concurrently • <i>unknown</i> - node has not been detected

Chapter 10: Accounting Records

Torque maintains accounting records for batch jobs in the following directory:

```
$TORQUE_HOME/server_priv/accounting/<TIMESTAMP>
```

`$TORQUE_HOME` defaults to `/var/spool/torque` 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:

Record marker	Record type	Description
A	abort	Job has been aborted by the server
C	checkpoint	Job has been checkpointed and held
D	delete	Job has been deleted
E	exit	Job has exited (either successfully or unsuccessfully)
Q	queue	Job has been submitted/queued
R	rerun	Attempt to rerun the job has been made
S	start	Attempt to start the job has been made (if the job fails to properly start, it may have multiple job start records)
T	restart	Attempt to restart the job (from checkpoint) has been made (if the job fails to properly start, it may have multiple job start records)

10.0.1 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 61](#).



Jobs submitted with the `-L` request syntax will have the `-L` submission recorded in the accounting log.

Variable	Description
ctime	Time job was created
etime	Time job became eligible to run
qtime	Time job was queued
start	Time job started to run

A sample record in this file can look like the following:

```
08/26/2014 17:07:44;Q;11923.napali;queue=batch
08/26/2014 17:07:50;S;11923.napali;user=dbeer group=company jobname=STDIN queue=batch
ctime=1409094464 qtime=1409094464 etime=1409094464 start=1409094470 owner=dbeer@napali
exec_host=napali/0+napali/1+napali/2+napali/3+napali/4+napali/5+torque-devtest-
03/0+torque-devtest-03/1+torque-devtest-03/2+torque-devtest-03/3+torque-devtest-
03/4+torque-devtest-03/5 Resource_List.nodes=2:ppn=6 Resource_List.nodect=2
Resource_List.nodes=2:ppn=6
08/26/2014 17:08:04;E;11923.napali;user=dbeer group=company jobname=STDIN queue=batch
ctime=1409094464 qtime=1409094464 etime=1409094464 start=1409094470 owner=dbeer@napali
exec_host=napali/0+napali/1+napali/2+napali/3+napali/4+napali/5+torque-devtest-
03/0+torque-devtest-03/1+torque-devtest-03/2+torque-devtest-03/3+torque-devtest-
03/4+torque-devtest-03/5 Resource_List.nodes=2:ppn=6 Resource_List.nodect=2
Resource_List.nodes=2:ppn=6 session=11352 total_execution_slots=12 unique_node_count=2
end=1409094484 Exit_status=265 resources_used.cput=00:00:00 resources_used.mem=82700kb
resources_used.vmem=208960kb resources_used.walltime=00:00:14 Error_Path=/dev/pts/11
Output_Path=/dev/pts/11
```

i The value of `Resource_List.*` is the amount of resources requested, and the value of `resources_used.*` is the amount of resources actually used.

i `total_execution_slots` and `unique_node_count` display additional information regarding the job resource usage.

Chapter 11: 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" below](#)
- ["Enabling Job Logs" below](#)

Job Log Location and Name

When job logging is enabled (see ["Enabling Job Logs" below](#)), 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" below](#)

["Job Logging" above](#)

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.

Parameter	Description
record_job_info	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. Changing this parameter requires a restart of <code>pbs_server</code> to take effect.
record_job_script	If set to true, this adds the contents of the script executed by a job to the log.

Parameter	Description
job_log_file_max_size	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 <i>current day</i> file size is greater than or equal to this value, it is rolled from <i><filename></i> to <i><filename.1></i> and a new empty log is opened. If the current day file size exceeds the maximum size a second time, the <i><filename.1></i> log file is rolled to <i><filename.2></i> , the current log is rolled to <i><filename.1></i> , 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).
job_log_file_roll_depth	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 <i><filename.3></i> . If a file is already at the specified depth, such as <i><filename.3></i> , the file is deleted so it can be replaced by the incoming file roll, <i><filename.2></i> .
job_log_keep_days	This maintains logs for the number of days designated. If set to 4, any log file older than 4 days old is deleted.

Related Topics

["Job Log Location and Name" on the previous page](#)

["Job Logging" on the previous page](#)

Chapter 12: NUMA and Torque

Torque supports two types of Non-Uniform Memory Architecture (NUMA) systems. This chapter serves as a central information repository for the various configuration settings involved when using either NUMA system configuration.



Torque cannot be configured for both NUMA types simultaneously.

In this chapter:

- [12.0.1 Supported NUMA Systems on page 157](#)
- [NUMA-Aware Systems on page 157](#)
- [NUMA-Support Systems on page 184](#)

12.0.1 Supported NUMA Systems

Torque supports these two NUMA system configurations:

- **NUMA-Aware** – Introduced with Torque version 6.0, this configuration supports multi-req jobs and jobs that span hosts. Moab version 9.0 and later is also required.
- **NUMA-Support** – Introduced with Torque version 3.0, this configuration supports only a single instance for pbs_mom that is treated as if there were multiple nodes running in the cluster. This configuration is only for large-scale SLES systems using SGI Altix and UV hardware.

NUMA-Aware Systems

This topic serves as a central information repository for NUMA-aware systems. This topic provides basic information and contains links to the various NUMA-aware topics found throughout the documentation.



Support for NUMA-aware systems is available only with Torque Resource Manager 6.0 and later and Moab Workload Manager 9.0 and later.

In this topic:

- [About NUMA-Aware Systems on page 158](#)
- [Installation and Configuration on page 159](#)
- [Job Resource Requests on page 159](#)
- [Job Monitoring on page 160](#)

- [Moab/Torque NUMA Configuration on page 160](#)
- [Considerations When Upgrading Versions or Changing Hardware on page 160](#)

About NUMA-Aware Systems

The NUMA-aware architecture is a hardware design which separates its cores into multiple clusters where each cluster has its own local memory region and still allows cores from one cluster to access all memory in the system. However, if a processor needs to use memory that is not its own memory region, it will take longer to access that (remote) memory. For applications where performance is crucial, preventing the need to access memory from other clusters is critical.

Torque uses cgroups to better manage cpu and memory accounting, memory enforcement, cpuset management, and binding jobs to devices such as MICs and GPUs. Torque will try to place jobs which request GPUs or MICs on NUMA nodes next to the GPU or MIC device to be used.

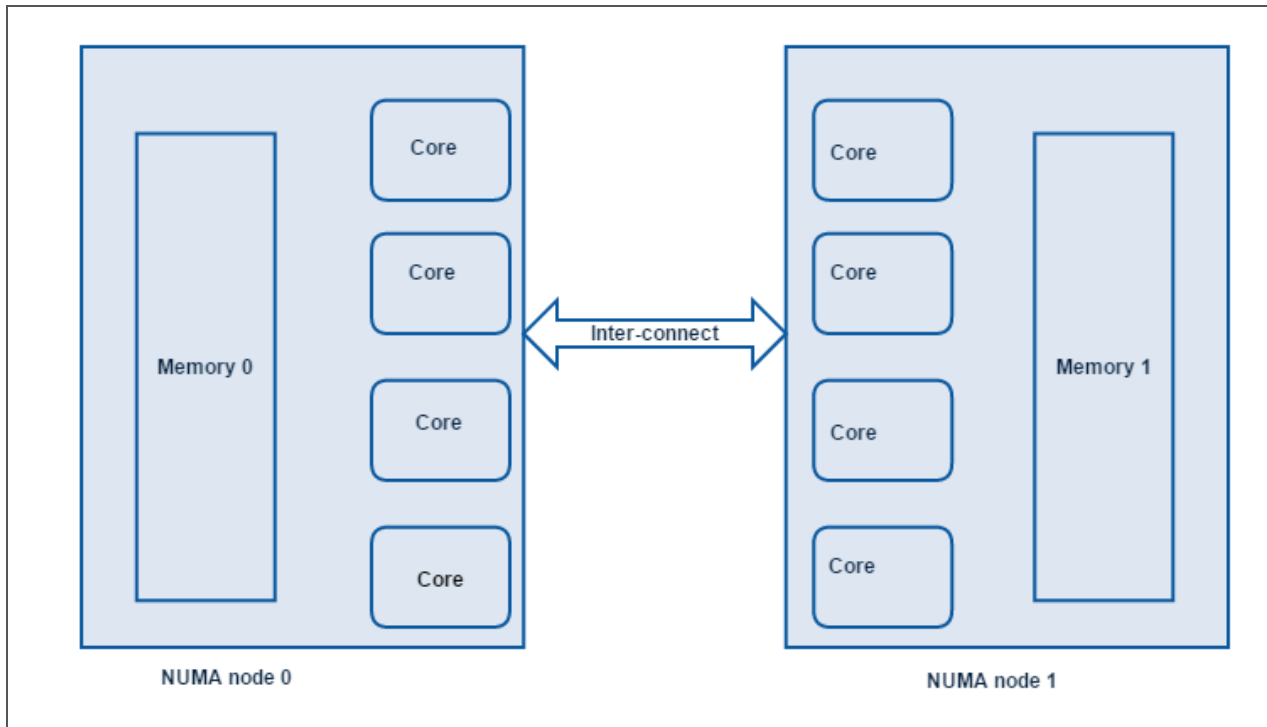
PCIe devices are similar to cores in that these devices will be closer to the memory of one NUMA node than another. Examples of PCIe devices are GPUs, NICs, disks, etc.

The resources of a processor chip have a hierarchy. The largest unit is a socket. A socket can contain one or more NUMA nodes with its cores and memory. A NUMA node will contain a set of cores and threads and memory which is local to the NUMA node. A core may have 0 or more threads.

- A socket refers to the physical location where a processor package plugs into a motherboard. The processor that plugs into the motherboard is also known as a socket. The socket can contain one or more NUMA nodes.
- A core is an individual execution unit within a processor that can independently execute a software execution thread and maintains its execution state separate from the execution state of any other cores within a processor.
- A thread refers to a hardware-based thread execution capability. For example, the Intel Xeon 7560 has eight cores, each of which has hardware that can effectively execute two software execution threads simultaneously, yielding 16 threads.

The following image is a simple depiction of a NUMA-aware architecture. In this example, the system has two NUMA nodes with four cores per NUMA node. The cores in each NUMA node have access to their own memory region but they can also access the memory region of the other NUMA node through the inter-connect.

i If the cores from NUMA chip 0 need to get memory from NUMA chip 1 there will be a greater latency to fetch the memory.



Installation and Configuration

Once Torque is first installed, you need to perform configuration steps.

See:

- [Torque NUMA-Aware Configuration on page 49](#)

Job Resource Requests

NUMA-aware resources can be requested at the time of job submission using the `qsub/msub -L` parameter. In addition, the `req_information_max` and `req_information_min` queue attributes let you specify the maximum and minimum resource limits allowed for jobs submitted to a queue.



Jobs requesting resources with `-L` can be run via `qrun` without a hostlist.

See:

- [Requesting NUMA-Aware Resources on page 74](#)
- [-L NUMA Resource Request on page 173](#)
- [Queue Attributes on page 399](#)

Job Monitoring

When using NUMA-aware, job resources are tracked per task. `qstat -f` produces a new category of information that begins with the "req_information" keyword. Following each "req_information keyword" is another keyword giving information about how the job was allocated. When the job has completed, the output will also include the per task resident memory used and per task cpu time used.

See

- [Monitoring NUMA Job Task Placement on page 81](#)

Moab/Torque NUMA Configuration

Moab does not require special configuration to support this NUMA-aware system. However, there are a few Moab-specific things that would be helpful to know and understand.

See

- [Using NUMA with Moab](#) in the *Moab Workload Manager Administrator Guide*

Considerations When Upgrading Versions or Changing Hardware

After upgrading server software or updating the hardware for a compute node, you should start the `pbs_mom` daemon with the `-f` flag to force the server to recognize the new configuration.

See

- Documentation for the [pbs_mom on page 222](#) `-f` flag

NUMA Tutorials

This section contains links to tutorials and other documentation useful in understanding NUMA-Aware systems.

In section:

- [NUMA Primer on page 161](#)
- [How NUMA Places Jobs on page 169](#)
- [NUMA Discovery and Persistence on page 172](#)

Related Topics

[NUMA-Aware Systems on page 157](#)

NUMA Primer

Torque 6.0 provides users with two major new features. First, cgroups are now used to manage each job. Second, resource request syntax has been updated, which gives users the ability to request resources on a per task basis, have multiple asymmetric resource requests in the same job, and control where jobs execute on cores and memory within NUMA hardware.

Control groups (cgroups) provide the ability to partition sets of tasks and their children into hierarchical groups with specialized behavior. In RHEL 7, the default directory for cgroups became `/sys/fs/cgroup`. The following examples use this standard.

i If you are building with cgroups enabled, you must have boost version 1.41 or later.

i For historical reasons, there are significant differences between how cgroups are administered between Red Hat 6 and Red Hat 7 and their corresponding equivalents in other Linux distributions. See [Red Hat Enterprise 6 Resource Management Guide](#) and [Red Hat Enterprise 7 Resource Management Guide](#) for more information about how cgroups work.

Torque cgroup Hierarchy

Torque 6.0 only uses the `cpu`, `devices`, `cpuacct`, `cpuset`, and `memory` subsystems. While `cpu` and `devices` subsystems are required for Torque 6.0 they are not yet fully implemented. When `pbs_mom` is initialized it creates a sub-directory named `torque` in each of the five subsystem directories. When a job is started on the MOM a directory with the name of the full job id is created under each `torque` directory. The following is an "ls -al" submission command example from the `cpuset/torque` hierarchy:

```
total 0
drwxr-xr-x 3 root root 0 Aug 28 13:36 .
drwxr-xr-x 4 root root 0 Aug 28 13:35 ..
drwx----- 4 root root 0 Aug 31 10:20 1301.hosta
-rw-r--r-- 1 root root 0 Aug 28 13:35 cgroup.clone_children
--w--w--w- 1 root root 0 Aug 28 13:35 cgroup.event_control
-rw-r--r-- 1 root root 0 Aug 28 13:35 cgroup.procs
-rw-r--r-- 1 root root 0 Aug 28 13:35 cpuset.cpu_exclusive
-rw-r--r-- 1 root root 0 Aug 28 13:35 cpuset.cpus
-rw-r--r-- 1 root root 0 Aug 28 13:35 cpuset.mem_exclusive
-rw-r--r-- 1 root root 0 Aug 28 13:35 cpuset.mem_hardwall
-rw-r--r-- 1 root root 0 Aug 28 13:35 cpuset.memory_migrate
-r--r--r-- 1 root root 0 Aug 28 13:35 cpuset.memory_pressure
-rw-r--r-- 1 root root 0 Aug 28 13:35 cpuset.memory_spread_page
-rw-r--r-- 1 root root 0 Aug 28 13:35 cpuset.memory_spread_slab
-rw-r--r-- 1 root root 0 Aug 28 13:35 cpuset.mems
-rw-r--r-- 1 root root 0 Aug 28 13:35 cpuset.sched_load_balance
-rw-r--r-- 1 root root 0 Aug 28 13:35 cpuset.sched_relax_domain_level
-rw-r--r-- 1 root root 0 Aug 28 13:35 notify_on_release
-rw-r--r-- 1 root root 0 Aug 28 13:35 tasks
```

Line 4 shows that the subdirectory is "1301.hosta". This is the `cpuset` cgroup for job "1301.hosta". If you were to issue an `ls` command on the "1301.hosta" subdirectory in this example, you would see the following:

```
total 0
drwx----- 4 root root 0 Aug 31 10:24 .
drwxr-xr-x 3 root root 0 Aug 31 10:22 ..
-rw-r--r-- 1 root root 0 Aug 31 10:24 cgroup.clone_children
--w--w--w- 1 root root 0 Aug 31 10:24 cgroup.event_control
-rw-r--r-- 1 root root 0 Aug 31 10:24 cgroup.procs
-rw-r--r-- 1 root root 0 Aug 31 10:24 cpuset.cpu_exclusive
-rw-r--r-- 1 root root 0 Aug 31 10:24 cpuset.cpus
-rw-r--r-- 1 root root 0 Aug 31 10:24 cpuset.mem_exclusive
-rw-r--r-- 1 root root 0 Aug 31 10:24 cpuset.mem_hardwall
-rw-r--r-- 1 root root 0 Aug 31 10:24 cpuset.memory_migrate
-r--r--r-- 1 root root 0 Aug 31 10:24 cpuset.memory_pressure
-rw-r--r-- 1 root root 0 Aug 31 10:24 cpuset.memory_spread_page
-rw-r--r-- 1 root root 0 Aug 31 10:24 cpuset.memory_spread_slab
-rw-r--r-- 1 root root 0 Aug 31 10:24 cpuset.mems
-rw-r--r-- 1 root root 0 Aug 31 10:24 cpuset.sched_load_balance
-rw-r--r-- 1 root root 0 Aug 31 10:24 cpuset.sched_relax_domain_level
-rw-r--r-- 1 root root 0 Aug 31 10:24 notify_on_release
drwx----- 2 root root 0 Aug 31 10:24 R0.t0
drwx----- 2 root root 0 Aug 31 10:24 R0.t1
-rw-r--r-- 1 root root 0 Aug 31 10:24 tasks
-rw-r--r-- 1 root root 0 Aug 28 13:35 tasks
```

For this job the `-L` resource request was:

```
qsub -L tasks=2:lprocs=2
```

This job has a single request and two tasks. The `R0` represents request 0 and the `t0` and `t1` represent the two tasks. In this case, `cpuset` information would be set for each task in their respective subdirectories. The `cpu`, `cpuacct`, `memory` and `devices` subsystems also utilize the same subdirectory structure.

cpuset Subsystem

The Linux `cpuset` functionality was integrated into `cgroups` so that when Torque is configured with the `--enable-cgroups` option, `cpuset` functionality is also included. When jobs are submitted using the `-L` resource request syntax, Torque allocates a `cpu` set and `memory` set for each task in the job request. Examples of how `cpusets` and `memory` sets are allocated will be shown in the examples at the end of this primer.

cpuacct Subsystem

The `cpuacct` subsystem keeps track of `cpu` time used for a `cgroup`. Torque now uses the `cpuacct` data to calculate `cpu` time used for a job. Also when using the `-L` resource request, `cpu` time per task is also recorded. Another advantage of `cgroups` is that the accounting information of a job does not disappear when the job process exits. So if `pbs_mom` goes down for any reason while running jobs the `cpu` time and `memory` used can still be tracked when `pbs_mom` is restarted.

devices Subsystem

The `cgroup` `devices` subsystem is used to restrict access to system devices such as disks, GPUs and MICs. In Torque 6.0.1, the `devices` subsystem was utilized to restrict jobs to use specific GPUs or MICs. Torque uses the `devices` subsystem to restrict access to GPUs and MICs in the `devices.deny` file. Take the following example:

```
qsub -L tasks=1:gpus=2
```

Assume a host with four GPUs indexed 0-3, and that when the job was started it was allocated GPUs 0 and 1. Torque would then add GPUs 2 and 3 to the `devices.deny` file of the jobs cgroup hierarchy, which would restrict the job to run only on GPUs 0 and 1.

If a job does not request any GPUs or MICs, then all GPUs and MICs on the MOM are added to the `devices.deny` file, thus restricting the job from using any GPUs or MICs.

memory Subsystem

The memory subsystem keeps track of the maximum memory used by a cgroup and also can be used to limit the maximum amount of resident memory a task can use or the maximum amount of swap a task can use. The `-L` resource request syntax has a memory and a swap option. Following are examples of how to request memory restrictions with the `-L` resource request.

```
qsub -L tasks=2:memory=300mb
```

Two tasks are created. The `memory=300mb` option restricts each task to a maximum of 300 megabytes of resident memory. If a task exceeds 300 mb, then the excess memory is sent to swap.

```
qsub -L tasks=2:swap=1Gb
```

Two tasks are created. The swap limit for each task is set to 1 GB.

i In order to be able to set swap and memory limits the Linux kernel must be built using the options `CONFIG_MEMCG=y`, `CONFIG_MEMCG_SWAP=y` and `CONFIG_MEMCG_SWAP_ENABLED=y`. For Red Hat 7-based systems, these options are set by default.

For SUSE 12-based systems, you will also need to modify the `/etc/default/grub` file. Do the following:

1. Edit `/etc/default/grub`.
2. Add the following inside of the `GRUB_CMDLINE_LINUX_DEFAULT` variable:

```
cgroup_enable=memory swapaccount=1
```

3. Run the following:

```
root# update-bootloader --refresh
```

4. Reboot your machine.

Resource Request 2.0

Following are several different types of `-L` resource requests. The examples show how to use the syntax to be able to have resources allocated which can best fit your job needs.

Single Resource Request With Two Tasks and Default settings

```
qsub -L tasks=2:lprocs=1
```

After this job runs, the summarized `qstat -f` output is shown:

```
Job Id: 1306.hosta
Job_Name = bigmem.sh
Job_Owner = knielson@hosta
resources_used.cput = 00:00:01
resources_used.energy_used = 0
resources_used.mem = 1956984kb
resources_used.vmem = 2672652kb
resources_used.walltime = 00:00:10
job_state = C
. . .
exit_status = 0
submit_args = -L tasks=2:lprocs=1 ../scripts/bigmem.sh
. . .
req_information.task_count.0 = 2
req_information.lprocs.0 = 1
req_information.thread_usage_policy.0 = allowthreads
req_information.hostlist.0 = hosta:ppn=2
req_information.task_usage.0.task.0.cpu_list = 0
req_information.task_usage.0.task.0.mem_list = 0
req_information.task_usage.0.task.0.memory_used = 976816kb
req_information.task_usage.0.task.0.cores = 0
req_information.task_usage.0.task.0.threads = 1
req_information.task_usage.0.task.0.host = hosta
req_information.task_usage.0.task.1.cpu_list = 4
req_information.task_usage.0.task.1.mem_list = 0
req_information.task_usage.0.task.1.memory_used = 976752kb
req_information.task_usage.0.task.1.cores = 0
req_information.task_usage.0.task.1.threads = 1
req_information.task_usage.0.task.1.host = hosta
```

resources_used is the same as in previous versions of Torque. In this job, 1 second of cpu time was used. 1956984kb of resident memory was used, but with the new -L syntax there is a new set of information which starts with req_information. This is the per task information of the job.

Output	Description
req_information.task_count.0 = 2	Two tasks are requested for this resource request; named tasks 0 and 1 respectively.
req_information.lprocs.0 = 1	One logical processor is requested per task. The lprocs value becomes the number of processing units per task allocated in the cpuset.
req_information.thread_usage_policy.0 = allow-threads	The processing unit allocation policy for the task. allow-threads is the user-specified default policy. allowthreads uses the first available core or thread. Processing units allocated in the cpuset are adjacent to each other unless other processors are also allocated.
req_information.hostlist.0 = hosta:ppn=2	On hostname hosta, two processing units are necessary . A single resource request can run on more than one host.

Output	Description
<code>req_information.task_usage.0.task.0.cpu_list = 0</code>	The <code>task_usage</code> keyword refers to the per task resource usage. 0 is the processing unit assigned to this task. In <code>req_information.task_usage.0.cpu_list.1</code> , the processing unit assigned is 4. This particular hardware is a 4 core system with hyper threading. So the core numbering is (0,4), (1,5), (2,6) and (3,7). Because the <code>thread_usage_policy</code> is <code>allow-threads</code> , the first two processing units are taken by default.
<code>req_information.task_usage.0.task.0.mem_list = 0</code>	Memory location 0 is allocated to this task.
<code>req_information.task_usage.0.task.0.memory_used = 976816kb</code>	The amount of resident memory used at task 0 is 976816kb.
<code>req_information.task_usage.0.task.0.cores = 0</code>	This is the number of cores used by the task. In this case no cores were used because the <code>allowthreads</code> policy uses only threads and not discrete cores.
<code>req_information.task_usage.0.task.0.host = hosta</code>	The task was run on hostname <code>hosta</code> .



The information for `req_information.task_usage.0.task.1` as opposed to `task.0`, means that the information displayed is referring to what was performed on task 1, rather than task 0.

Multiple lprocs

```
qsub -L tasks=1:lprocs=2
```

Two logical processors are specified with one task. The output of this job is as follows:

```
req_information.task_count.0 = 1
req_information.lprocs.0 = 2
req_information.thread_usage_policy.0 = allowthreads
req_information.hostlist.0 = hosta:ppn=2
req_information.task_usage.0.task.0.cpu_list = 0,4
req_information.task_usage.0.task.0.mem_list = 0
req_information.task_usage.0.task.0.cores = 0
req_information.task_usage.0.task.0.threads = 2
req_information.task_usage.0.task.0.host = hosta
```

The `req_information` for this syntax shows a `cpu_list` with two processing units. 0 and 4 are the first two processing units available so they are in the `cpu_list`. Notice that now there are two threads running.

usecores

The following example shows how a request to use cores changes the `cpu_list` allocation.

```
qsub -L tasks=1:lprocs=4:usecores
```

```
req_information.task_count.0 = 1
req_information.lprocs.0 = 4
req_information.thread_usage_policy.0 = usecores
req_information.hostlist.0 = hosta:ppn=4
req_information.task_usage.0.task.0.cpu_list = 0-3
req_information.task_usage.0.task.0.mem_list = 0
req_information.task_usage.0.task.0.cores = 4
req_information.task_usage.0.task.0.threads = 8
req_information.task_usage.0.task.0.host = hosta
```

Output	Description
req_information.task_usage.0.task.0.cores = 4	Four cores are used for task 0.
req_information.task_usage.0.task.0.threads = 8	When a core is requested, any threads for that core are no longer available for use by another task. In this case, each core has two threads. As a result, when one core is used two threads are also used. In this case, 8 threads are used in total.

usethreads

```
qsub -L tasks=1:lprocs=4:usethreads
```

The output of this job is as follows:

```
req_information.task_count.0 = 1
req_information.lprocs.0 = 4
req_information.thread_usage_policy.0 = usecores
req_information.hostlist.0 = hosta:ppn=4
req_information.task_usage.0.task.0.cpu_list = 0,4,1,5
req_information.task_usage.0.task.0.mem_list = 0
req_information.task_usage.0.task.0.cores = 0
req_information.task_usage.0.task.0.threads = 4
req_information.task_usage.0.task.0.host = hosta
```

Requesting usethreads gives adjacent processing units 0,4,1,5 and uses only 4 threads as indicated by req_information.task_usage.0.task.0.threads = 4.

Multiple Resource Requests

The -L resource requests makes it easier to request asymmetric resources for a single job. For example, you might have a job which needs several processors on a host to do work but only one or two processors on another host. The -L syntax easily accommodates this.

```
qsub -L tasks=2:lprocs=6:usecores -L tasks=1:lprocs=1:place=socket
```

```

req_information.task_count.0 = 2
req_information.lprocs.0 = 6
req_information.thread_usage_policy.0 = usecores
req_information.hostlist.0 = hostb:ppn=12
req_information.task_usage.0.task.0.cpu_list = 0-5
req_information.task_usage.0.task.0.mem_list = 0
req_information.task_usage.0.task.0.cores = 6
req_information.task_usage.0.task.0.threads = 12
req_information.task_usage.0.task.0.host = hostb
req_information.task_usage.0.task.1.cpu_list = 6-11
req_information.task_usage.0.task.1.mem_list = 1
req_information.task_usage.0.task.1.cores = 6
req_information.task_usage.0.task.1.threads = 12
req_information.task_usage.0.task.1.host = hostb
req_information.task_count.1 = 1
req_information.lprocs.1 = 1
req_information.socket.1 = 1
req_information.thread_usage_policy.1 = allowthreads
req_information.hostlist.1 = hostb:ppn=1
req_information.task_usage.1.task.0.cpu_list = 0
req_information.task_usage.1.task.0.mem_list = 0
req_information.task_usage.1.task.0.cores = 1
req_information.task_usage.1.task.0.threads = 1
req_information.task_usage.1.task.0.host = hostb

```

Output	Description
req_information.task_count.1=1	Only one task on request 1.
req_information.socket.1 = 1	One socket is requested and then allocated for use.

place Directives

The place directive takes one of five arguments: node, socket, numanode, core, and thread. The node, core, and thread arguments do not take an assignment, however socket and numanode can be assigned a number value requesting the number of sockets or numanodes per task. The use of "place=core" or "place=thread" is the equivalent of using the usecores or usethreads syntax.

When processes share the same memory cache and are run on adjacent cores or threads, the likelihood of swapping out a cache line is high. When memory needs to be fetched from primary memory instead of the cache processing execution times are increased and become less predictable. In these examples, Torque disables linearly allocating cores. To help ensure best performance by avoiding the sharing of caches between processors, cores are spread as far apart as possible.

The following examples show the results of each directive:

```
place=socket
```

If a socket is not given a number, it defaults to the number 1.

```
qsub -L tasks=2:lprocs=2:place=socket
```

This request allocates two tasks with two logical processors each. Each task is placed on its own socket.

```
req_information.task_count.0 = 2
req_information.lprocs.0 = 2
req_information.socket.0 = 1
req_information.thread_usage_policy.0 = allowthreads
req_information.hostlist.0 = hosta:ppn=4
req_information.task_usage.0.task.0.cpu_list = 0,3
req_information.task_usage.0.task.0.mem_list = 0
req_information.task_usage.0.task.0.cores = 2
req_information.task_usage.0.task.0.threads = 4
req_information.task_usage.0.task.0.host = hosta
req_information.task_usage.0.task.1.cpu_list = 6,9
req_information.task_usage.0.task.1.mem_list = 1
req_information.task_usage.0.task.1.cores = 2
req_information.task_usage.0.task.1.threads = 4
req_information.task_usage.0.task.1.host = hosta
cpuset_string = hosta:0,3,6,9
memset_string = hosta:0-1
```

For the last example the job was run on a dual socket host with 12 cores. Each core has two threads for a total of 24 processing units. Each socket has 6 cores and 12 threads. The cores for socket 0 are numbered 0, 1, 2, 3, 4, 5. The cores for socket 1 are numbered 6, 7, 8, 9, 10, 11. Task.0 uses cores 0 and 3 and task.1 uses cores 6 and 9.

```
place=numanode=2
```

```
qsub -L tasks=2:lprocs=2:place=numanode=2
```

This request allocates two numanodes, one for each task.

```
req_information.task_count.0 = 2
req_information.lprocs.0 = 2
req_information.numanode.0 = 2
req_information.thread_usage_policy.0 = allowthreads
req_information.hostlist.0 = hostb:ppn=2
req_information.task_usage.0.task.0.cpu_list = 0, 3
req_information.task_usage.0.task.0.mem_list = 0
req_information.task_usage.0.task.0.cores = 0
req_information.task_usage.0.task.0.threads = 0
req_information.task_usage.0.task.0.host = hostb
req_information.task_usage.0.task.1.cpu_list = 6, 9
req_information.task_usage.0.task.1.mem_list = 1
req_information.task_usage.0.task.1.cores = 2
req_information.task_usage.0.task.1.threads = 4
req_information.task_usage.0.task.1.host = hostb
```

pbsnodes and Dedicated Resources

When a resource is requested (core, numanode, socket, etc.), the entire resource is no longer available for other jobs to use, and enters a dedicated state. Starting with Torque 6.0, pbsnodes tracks total sockets, numanodes, cores and threads per node. pbsnodes also tracks dedicated sockets, numanodes, cores, and threads. Following is an example of node output in pbsnodes for Torque 6.0


```

state = free
power_state = Running
np = 12
ntype = cluster
status =
rectime=1441054213,macaddr=78:e3:b5:0a:c0:58,cpuclock=Fixed,varattr=,jobs=,state=fre
e...
mom_service_port = 15002
mom_manager_port = 15003
total_sockets = 2
total_numa_nodes = 2
total_cores = 12
total_threads = 12
dedicated_sockets = 0
dedicated_numa_nodes = 0
dedicated_cores = 0
dedicated_threads = 0

```

This node has a total of 2 sockets, 2 numanodes, 12 cores and 12 threads. The number of dedicated sockets, numanodes, cores, and threads are set to 0 indicating there are currently no jobs running on this nodes. If a job is run with a syntax of:

```
qsub -L tasks=2:lprocs=2
```

the number of dedicated threads becomes four.

Once a job is completed, dedicated_threads returns to 0.

Related Topics

[NUMA Tutorials on page 160](#)

How NUMA Places Jobs

This topic discusses how jobs are placed on their specific NUMA resources.

i In this topic, placing is defined as determining where on the node the job will go.

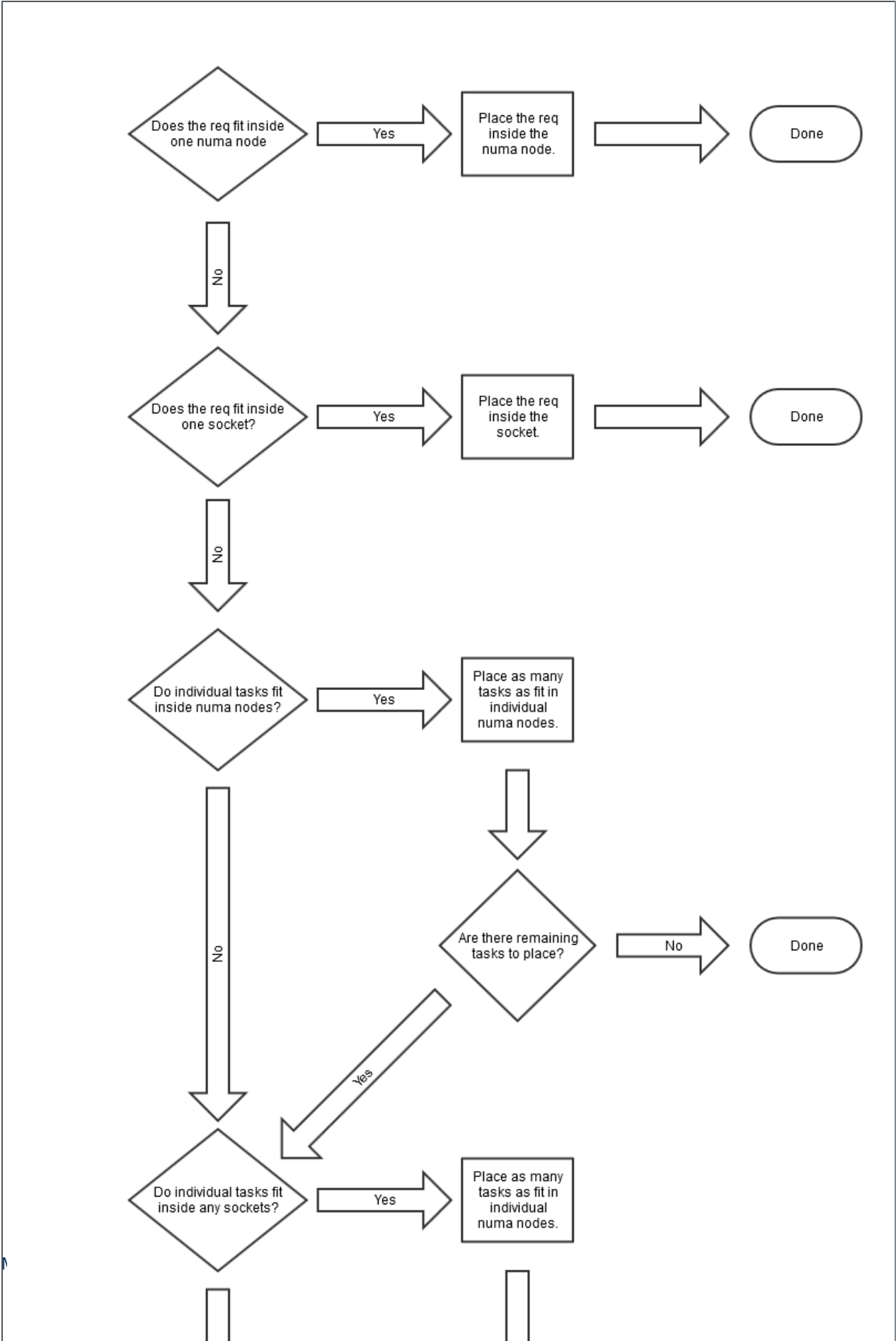
Moab determines where to place a job on a node and pbs_server places that job inside the node. pbs_server decides where to place the job based on the parameters specified by the job itself and optimal locality.

The Placement Algorithm

The following diagram shows the decision making process for each request not using the "place=" syntax.

Whenever possible, jobs are given placement preference next to the GPUs that they will use. If the job's tasks' memory and CPUs are available on the socket that has the GPUs, that job will be placed at that location.

i This placement fails if either there is no socket that contains all of the required resources, or if jobs are and the GPUs are on socket - but all of the cores are used by another job.



For jobs using the "place=" syntax, the decision making algorithm is much simpler. When the place is specified, it will become reserved and the job will be placed at that location.

i For a job to occupy the user-requested place option, that option must be completely available for use.

The following is an example of a job submitted using the -L option.

```
-L tasks=1:lprocs=2:sockets=2
```

This job placed on two sockets with one core reserved per socket.

i The use of the word "core" is intentional. If a "place=socket" or "place=numanode" is requested and the lprocs request is less than the number of cores inside the socket or NUMA node, then the job is given only cores.

Once pbs_server has determined where each task should run, that decision is stored in attributes on the job itself. The complete_req attribute stores where each task is allocated, and the mom reads that information to create appropriate cgroups for the job and for the entire task. This information is available to the user via qstat.

Related Topics

[NUMA Tutorials on page 160](#)

NUMA Discovery and Persistence

Initial Discovery

First, The mom performs the initial discovery of the host's layout, including the number of sockets, numa nodes, pci devices, cores, and threads. This is done using the hwloc library. Next, the mom sends that information to pbs_server, which notes it. Last, pbs_server writes files with the node's layout information locally. Following restarts, node information is gathered from these files.

Job Placement Decisions

Job placement decisions are done by pbs_server so that it immediately knows which NUMA resources have been used by which job. As a result, the second a job starts or finishes the information for available numa resources is updated and accurate. This information is then communicated to the mom daemon.

For more information on how jobs are placed, see [How NUMA Places Jobs on page 169](#).

Persistence Across Restarts.

To maintain correct usage information pbs_server writes files to a new directory in /server_priv/node_usage. The files are written in a json format. The following is a representation of what these files look like:

```
# Simple Node
{"node":{"socket":{"os_index":0,"numanode":{"os_index":0,"cores":"0-5","threads":"","mem":16435852}}}}

# More Complicated Node
"node":{"socket":{"os_index":0,"numanode":{"os_index":0,"cores":"0-7","threads":"","mem":16775316}, "numanode":{"os_index":1,"cores":"8-15","threads":"","mem":16777216}}, "socket": {"os_index":1,"numanode":{"os_index":2,"cores":"16-23","threads":"","mem":8388608}, "numanode":{"os_index":3,"cores":"24-31","threads":"","mem":8388608}}}}
```

When jobs are present, an allocation object will also be there to record what resources are being used by the job. The allocation object will be beneath the numanode object, so it is possible to have more than one per job.

Related Topics

[NUMA Tutorials on page 160](#)

-L NUMA Resource Request

The -L option is available in the qsub and msub commands to allow administrators the ability to place jobs at the "task" or "OS process" level to get maximum efficiency out of the available hardware.

Using the -L option requires a basic knowledge of the topologies of the available hardware where jobs will run. You will need to know how many cores, numanodes, sockets, etc. are available on the hosts within the cluster. The -L syntax is designed to allow for a wide variety of requests. However, if requests do not match the available hardware, you may have unexpected results.

In addition, multiple, non-symmetric resource requests can be made for the same job using the -L job submission syntax.

For example, the following command:

```
qsub -L tasks=4:lprocs=2:usecores:memory=500mb -L tasks=8:lprocs=4:memory=2gb
```

Creates two requests. The first request creates 4 tasks with two logical processors and 500 mb of memory per task. The logical processors are placed on cores. The second request calls for 8 tasks with 4 logical processors and 2 gb of memory per task. Logical processors may be placed on cores or threads since the default placement is allowthreads.


This topic provides the -L option syntax and a description of the allocation options.



Syntax

```
-L tasks=#[:lprocs=#|all]
[:{usecores|usethreads|allowthreads}]
[:place={socket|numanode|core|thread}[=#]{node}][:memory=#][:swap=#]
[:maxtpn=#][:gpus=#[:<mode>]][:mics=#][:gres=<gres>]
[:feature=<feature>]
[[:{cpt|cgroup_per_task}]][:{cph|cgroup_per_host}]
```

Allocation Options

The following table identifies the various allocation options you can specify per task.


**tasks=#** specifies the number of job tasks for which the resource request is to be applied. It is the only required element for the -L resource request. The remainder of the -L syntax allocates resources per task.

Value	Description
cpt, cgroup_ per_task, cph, cgroup_ per_host	<p>Specifies whether cgroups are created per-task or per-host. If submitting using <code>msub</code>, this information is passed through to Torque; there is no effect on Moab operations.</p> <div> This option lets you specify how cgroups are created during job submission. This option can be used to override the Torque <code>cgroup_per_task</code> server parameter. If this option is not specified, the server parameter value is used. See "Server Parameters" on page 299 in the <i>Torque Administrator Guide</i> for more information.</div> <ul style="list-style-type: none">• :cpt, :cgroup_per_task – Job request will have one cgroup created per task; all the processes on that host will be placed in the <i>first</i> task's cgroup.• :cph, :cgroup_per_host – Job request will have one cgroup created per host; this is similar to pre-6.0 cpuset implementations. <div> Some MPI implementations only launch one process through the TM API, and then fork each subsequent process that should be launched on that host. If the job is set to have one cgroup per task, this means that all of the processes on that host will be placed in the <i>first</i> task's cgroup. Confirm that the <code>cgroup_per_task</code> Torque server parameter is set to <code>FALSE</code> (default) or specify <code>:cph</code> or <code>:cgroup_per_host</code> at job submission. If you know that your MPI will communicate each process launch to the mom individually, then set the <code>cgroup_per_task</code> Torque server parameter is set to <code>TRUE</code> or specify <code>:cpt</code> or <code>:cgroup_per_task</code> at job submission.</div>
feature	<p>Specifies one or more node feature names used to qualify compute nodes for task resources; i.e., a compute node must have all ("&") or and (" ") of the specified feature name(s) assigned or the compute node's resources are ineligible for allocation to a job task.</p> <div><pre>:feature=bigmem :feature='bmem&fio' :feature='bmem fio'</pre></div>



Value	Description
gpus	<p>Specifies the quantity of GPU accelerators to allocate to a task, which requires placement at the locality-level to which an accelerator is connected or higher. <MODE> can be <code>exclusive_process</code>, <code>exclusive_thread</code>, or <code>reseterr</code>.</p> <div data-bbox="407 422 1433 573"> <p>i If you are using CUDA 8 or newer, the default of <code>exclusive_thread</code> is no longer supported. If the server specifies an <code>exclusive_thread</code> setting, the MOM will substitute an <code>exclusive_process</code> mode setting. We recommend that you set the default to <code>exclusive_process</code>.</p> </div> <p>The task resource request must specify placement at the numanode- (AMD only), socket-, or node-level. <code>place=core</code> and <code>place=thread</code> are invalid placement options when a task requests a PCIe-based accelerator device, since allowing other tasks to use cores and threads on the same NUMA chip or socket as the task with the PCIe device(s) would violate the consistent job execution time principle since these other tasks would likely interfere with the data transfers between the task's logical processors and its allocated accelerator(s).</p> <div data-bbox="407 793 1433 1041"> <pre>:gpus=1</pre> <p><i>Allocates one GPU per task.</i></p> <pre>:gpus=2:exclusive_process:reseterr</pre> <p><i>Allocates two GPUs per task with exclusive access by process and resets error counters.</i></p> </div>
gres	<p>Specifies the quantity of a specific generic resource <gres> to allocate to a task. If a quantity is not given, it defaults to one.</p> <div data-bbox="407 1157 1433 1245"> <p>i Specify multiple GRES by separating them with commas and enclosing all the GRES names, their quantities, and the commas within single quotation marks.</p> </div> <div data-bbox="407 1266 1433 1549"> <pre>:gres=matlab=1</pre> <p><i>Allocates one Matlab license per task.</i></p> <pre>:gres='dvd,blu=2'</pre> <p><i>Allocates one DVD drive and two Blu-ray drives per task, represented by the "dvd" and "blu" generic resource names, respectively.</i></p> </div> <div data-bbox="407 1570 1433 1696"> <p>i When scheduling, if a generic resource is node-locked, only compute nodes with the generic resource are eligible for allocation to a job task. If a generic resource is floating, it does not qualify or disqualify compute node resources from allocation to a job task.</p> </div>

Value	Description
lprocs	<p>Specifies the quantity of "logical processors" required by a single task to which it will be pinned by its control-group (cgroup).</p> <div> <p>i The "place" value specifies the total number of physical cores/threads to which a single task has exclusive access. The lprocs= keyword indicates the <i>actual</i> number of cores/threads to which the task has exclusive access for the task's cgroup to pin to the task.</p> <ul style="list-style-type: none"> When :lprocs is specified, and nothing is specified for #, the default is 1. When :lprocs=all is specified, all cores or threads in any compute node/server's available resource locality placement specified by the "place" option is eligible for task placement (the user has not specified a quantity, other than "give me all logical processors within the resource locality or localities"), which allows a user application to take whatever it can get and adapt to whatever it receives, which cannot exceed one node. <pre>qsub -L tasks=1:lprocs=4</pre> <p><i>One task is created which allocates four logical processors to the task. When the job is executed, the pbs_mom where the job is running will create a cpuset with four processors in the set. Torque will make a best effort to allocate the four processors next to each other but the placement is not guaranteed.</i></p> <pre>qsub -L tasks=1:lprocs=all:place=node</pre> <p><i>Places one task on a single node, and places all processing units in the cpuset of the task. The "lprocs=all" parameter specifies that the task will use all cores and/or threads available on the resource level requested.</i></p> </div>
maxtpn	<p>Specifies the maximum tasks per node; where "#" is the maximum tasks allocated per physical compute node. This restricts a task type to no more than "#" tasks per compute node and allows it to share a node with other task types or jobs. For example, a communication-intensive task may share a compute node with computation-intensive tasks.</p> <div> <p>i The number of nodes and tasks per node will not be known until the job is run.</p> <pre>qsub -L tasks=7:maxtpn=4</pre> <p><i>Allocates seven tasks but a maximum of four tasks can run on a single node.</i></p> </div>

Value	Description
memory	<div data-bbox="407 321 1433 443"> <p>i "memory" is roughly equivalent to the mem request for the qsub/msub -l resource request. However, with the -L qsub syntax, cgroups monitors the job memory usage and puts a ceiling on resident memory for each task of the job.</p> </div> <p>Specifies the maximum resident memory allocated per task. Allowable suffixes are kb (kilobytes), mb (megabytes), gb (gigabytes), tb (terabyte), pb (petabytes), and eb (exabyte). If a suffix is not provided by the user, kb (kilobytes) is default.</p> <div data-bbox="407 554 1433 615"> <p>i If a task uses more resident memory than specified the excess memory is moved to swap.</p> </div> <div data-bbox="407 636 1433 779"> <pre>qsub -L tasks=4:lprocs=2:usecores:memory=1gb</pre> <p><i>Allocates four tasks with two logical processors each. Each task is given a limit of 1 gb of resident memory.</i></p> </div> <div data-bbox="407 800 1433 921"> <pre>qsub -L tasks=2:memory=3500</pre> <p><i>Allocates two tasks with 3500 kb (the suffix was not specified so kilobytes is assumed).</i></p> </div>
mics	<p>Specifies the quantity of Intel MIC accelerators to allocate to a task, which requires placement at the locality-level to which a MIC is connected or higher.</p> <p>The task resource request must specify placement at the NUMA chip- (makes sense for AMD only), socket-, or node-level. place=core and place=thread are invalid placement options when a task requests a PCIe-based accelerator device since allowing other tasks to use cores and threads on the same NUMA chip or socket as the task with the PCIe device(s) would violate the consistent job execution time principle since these other tasks would likely interfere with the data transfers between the task's logical processors and its allocated accelerator(s).</p> <div data-bbox="407 1226 1433 1318"> <p>i Allocating resources for MICs operates in the exact same manner as for GPUs. See "gpus" on page 175.</p> </div> <div data-bbox="407 1339 1433 1451"> <pre>:mics=1</pre> <p><i>Allocates on MIC per task.</i></p> </div> <div data-bbox="407 1472 1433 1583"> <pre>:mics=2</pre> <p><i>Allocates two MICs per task.</i></p> </div>

Value	Description
place	<p>Specifies placement of a single task on the hardware. Specifically, this designates what hardware resource locality level and identifies the quantity of locality-level resources. Placement at a specific locality level is always exclusive, meaning a job task has exclusive use of all logical processor and physical memory resources at the specified level of resource locality, even if it does not use them.</p> <p>Valid Options:</p> <div><p> If a valid option is not specified, the usecores, usethreads, and allowthreads parameters are used.</p><ul style="list-style-type: none">socket[=#] – Refers to a socket within a compute node/server and specifies that each task is placed at the socket level with exclusive use of all logical processors and memory resources of the socket(s) allocated to a task. If a count is not specified, the default setting is 1.<div><pre>qsub -l tasks=2:lprocs=4:place=socket</pre><p><i>Two tasks are allocated with four logical processors each. Each task is placed on a socket where it will have exclusive access to all of the cores and memory of the socket. Although the socket may have more cores/threads than four, only four cores/threads will be bound in a cpuset per task per socket as indicated by "lprocs=4".</i></p></div><ul style="list-style-type: none">numanode[=#] – Refers to the numanode within a socket and specifies that each task is placed at the NUMA node level within a socket with exclusive use of all logical processor and memory resources of the NUMA node(s) allocated to the task. If a count is not given, the default value is 1. If a socket does not contain multiple numanodes, by default the socket contains one numanode.<p>To illustrate the locality level to which this option refers, the following examples are provided:</p><p>First, a Haswell-based Intel Xeon v3 processor with 10 or more cores is divided internally into two separate "nodes", each with an equal quantity of cores and its own local memory (referred to as a "numanode" in this topic).</p><p>Second, an AMD Opteron 6xxx processor is a "multi-chip module" that contains two separate physical silicon chips each with its own local memory (referred to as a "numanode" in this topic).</p><p>In both of the previous examples, a core in one "node" of the processor can access its own local memory faster than it can access the remote memory of the other "node" in the processor, which results in NUMA behavior.</p><div><pre>qsub -l tasks=2:lprocs=4:place=numanode</pre><p><i>Places a single task on a single numanode and the task has exclusive use of all the logical processors and memory of the numanode.</i></p></div></div>

Value	Description
	<pre>qsub -L tasks=2:lprocs=all:place=numanode=2</pre> <p><i>Allocates two tasks with each task getting two numanodes each. The "lprocs=all" specification indicates all of the cores/threads of each numanode will be bound in the cpuset for the task.</i></p> <ul style="list-style-type: none"> • core[=#] – Refers to a core within a numanode or socket and specifies each task is placed at the core level within the numanode or socket and has exclusive use of all logical processor and memory resources of the core(s) allocated to the task. Whether a core has SMT/hyper-threading enabled or not is irrelevant to this locality level. If a number of cores is not specified, it will default to the number of lprocs specified. <div> <p>i The amount of cores specified must be greater than or equal to the number of lprocs available, otherwise the job submission will be rejected.</p> </div> <pre>qsub -L tasks=2:place=core=2</pre> <p><i>Two tasks with one logical processor each will be placed on two cores per task.</i></p> <pre>qsub -L tasks=2:lprocs=2:place=core</pre> <p><i>Two tasks are allocated with two logical processors per task. Each logical process will be assigned to one core each (two cores total, the same as the number of lprocs). Torque will attempt to place the logical processors on non-adjacent cores.</i></p> <ul style="list-style-type: none"> • thread[=#] – Specifies each task is placed at the thread level within a core and has exclusive use of all logical processor and memory resources of the thread(s) allocated to a task. This affinity level refers to threads within a core and is applicable only to nodes with SMT or hyper-threading enabled. If a node does not have SMT or hyper-threading enabled, Moab will consider the node ineligible when allocating resources for a task. If a specific number of threads is not specified, it will default the the number of lprocs specified. <pre>qsub -L tasks=2:lprocs=4:place=thread</pre> <p><i>Allocates two tasks, each with four logical processors, which can be bound to any thread. Torque will make a best effort to bind the threads on the same numanode but placement is not guaranteed. Because the amount of threads is not specified, Torque will place the number of lprocs requested.</i></p> <ul style="list-style-type: none"> • node – Specifies that each task is placed at the node level and has exclusive use of all the resources of the node(s) allocated to a task. This locality level usually refers to a physical compute node, blade, or server within a cluster.

Value	Description
	<div><pre>qsub -L tasks=2:lprocs=all:place=node</pre></div> <div><p><i>Two tasks are allocated with one task per node, where the task has exclusive access to all the resources on the node. The "lprocs=all" specification directs Torque to create a cpuset with all of the processing units on the node. The "place=node" specification also claims all of the memory for the node/server.</i></p></div>
swap	<p>Specifies the maximum allocated resident memory <i>and</i> swap space allowed per task. Allowable suffixes are kb (kilobytes), mb (megabytes), gb (gigabytes), tb (terabyte), pb (petabytes), and eb (exabyte). If a suffix is not given, kb (kilobytes) is assumed.</p> <p>If a task exceeds the specified limit, the task will be killed; the associated job will be terminated.</p> <div><div></div><div><p>If the swap limit is unable to be set, the job will still be allowed to run. All other cgroup-related failures will cause the job to be rejected.</p></div></div> <p>When requesting swap, it is not required that you give a value for the :memory option.</p> <ul style="list-style-type: none">If using :swap <i>without</i> a specified :memory value, Torque will supply a memory value up to the value of :swap; but not larger than available physical memory. <div><pre>qsub -L tasks=4:lprocs=2:swap=4gb</pre></div> <div><p><i>Allocates four tasks with two logical processors each. Each task is given a combined limit of 4 gb of resident memory and swap space. If a task exceeds the limit, the task is terminated.</i></p></div> <ul style="list-style-type: none">If using :swap <i>with</i> a specified :memory value, Torque will only supply resident memory up to the :memory value. The rest of the swap can <i>only</i> be supplied from the swap space. <div><div></div><div><p>The :memory value <i>must</i> be smaller than or equal to the :swap value.</p></div></div> <div><pre>qsub -L tasks=2:memory=3.5gb:swap=5gb</pre></div> <div><p><i>Allocates two tasks and each task has up to 3.5 gb of resident memory and a maximum of 5 gb of swap. If a task exceed 3.5 gb of resident memory, the excess will be moved to the swap space. However, if the task exceed 5 gb of total swap, the task and job will be terminated.</i></p></div>

Value	Description
tasks	<p>Specifies the quantity of job tasks for which the resource request describes the resources needed by a single task.</p> <ul style="list-style-type: none">• Distributed memory systems - A single task must run within a single compute node/server; i.e., the task's resources must all come from the same compute node/server.• Shared memory systems - A single task may run on multiple compute nodes; i.e., the task's resources may come from multiple compute nodes. <p>This option is <i>required</i> for task-based resource allocation and placement.</p> <div><pre>qsub -L tasks=4</pre><p><i>Creates four tasks, each with one logical process. The tasks can be run on a core or thread (default allowthreads).</i></p></div>

Value	Description
usecores, usethreads, allow threads	<p>The usecores, usethreads, and allowthreads parameters are used to indicate whether the cgroup pins cores, threads, or either to a task, respectively. If no logical processor definition is given, the default is allowthreads for backward-compatible Moab scheduler and Torque resource manager behavior.</p> <p>In this context, "cores" means an AMD Opteron core, a hyperthread-disabled Intel Xeon core, or thread 0 <i>and only thread 0</i> of a hyperthread-enabled Intel Xeon core. The term "threads" refers to a hyperthread-enabled Intel Xeon thread. Likewise, "either" refers to an AMD Opteron core, a hyperthread-disabled Intel Xeon core, or any thread of a hyperthread-enabled Intel Xeon.</p> <ul style="list-style-type: none">• :usecores – Denotes that the logical processor definition for a task resource request is a physical core. This means if a core has hyper-threading enabled, the task will use only thread 0 of the core. <div><pre>qsub -L tasks=2:lprocs=2:usecores</pre><p><i>Two tasks are allocated with two logical processors per task. The usecores parameter indicates the processor types must be a core or thread 0 of a hyper-threaded core.</i></p></div> <ul style="list-style-type: none">• :usethreads – Specifies the logical processor definition for a task resource request is a hardware-based thread or virtual core. <div><pre>qsub -L tasks=2:lprocs=2:usethreads</pre><p><i>Two tasks are allocated with two logical processors per task. The usethreads parameter indicates that any type of hardware-based thread or virtual core can be used.</i></p></div> <ul style="list-style-type: none">• :allowthreads – Specifies that the logical processor definition for a task resource request can be either a physical core (e.g. AMD Opteron), or hardware-based thread of a core (hyperthread-enabled Intel Xeon). <div><pre>qsub -L tasks=2:lprocs=2:allowthreads</pre><p><i>Two tasks are allocated with two logical processors per task. The allowthreads parameter indicates hardware threads or cores can be used.</i></p></div>

Related Topics

- [qsub on page 275](#) in the *Torque Administrator Guide*
- [Requesting NUMA-Aware Resources on page 74](#) in the *Torque Administrator Guide*

pbsnodes with NUMA-Awareness

When Torque is configured with NUMA-awareness and configured with `--enable-cgroups`, the number of total *and* the number of available sockets, numachips (numa nodes), cores, and threads are returned when the status of nodes are queried by Moab (a call is made to `pbsnodes`).

The example output that follows shows a node with two sockets, four numachips, 16 cores and 32 threads. In this example, no jobs are currently running on this node; therefore, the available resources are the same as the total resources.

```
torque-devtest-01
  state = free
  power_state = Running
  np = 16
  ntype = cluster
  status =
rectime=1412732948,macaddr=00:26:6c:f4:66:a0,cpuclock=Fixed,varattr=,jobs=,state=free,
netload=17080856592,gres=,loadave=10.74,ncpus=16,physmem=49416100kb,availmem=50056608k
b,totmem=51480476kb,idletime=29,nusers=2,nsessions=3,sessions=8665
8671 1994,uname=Linux torque-devtest-01 2.6.32-358.el6.x86_64 #1 SMP
Fri Feb 22 00:31:26 UTC 2013 x86_64,opsys=linux
  mom_service_port = 15002
  mom_manager_port = 15003
  total_sockets = 2
  total_numachips = 4
  total_cores = 16
  total_threads = 32
  available_sockets = 2
  available_numachips = 4
  available_cores = 16
  available_threads = 32
```

However, if a job requesting only a single core was started on this node, the `pbsnodes` output will look like:

```
torque-devtest-01
  state = free
  power_state = Running
  np = 16
  ntype = cluster
  jobs = 0/112.torque-devtest-01
  status =
rectime=1412732948,macaddr=00:26:6c:f4:66:a0,cpuclock=Fixed,varattr=,jobs=,state=free,
netload=17080856592,gres=,loadave=10.74,ncpus=16,physmem=49416100kb,availmem=50056608k
b,totmem=51480476kb,idletime=29,nusers=2,nsessions=3,sessions=8665
8671 1994,uname=Linux torque-devtest-01 2.6.32-358.el6.x86_64 #1 SMP
Fri Feb 22 00:31:26 UTC 2013 x86_64,opsys=linux
  mom_service_port = 15002
  mom_manager_port = 15003
  total_sockets = 2
  total_numachips = 4
  total_cores = 16
  total_threads = 32
  available_sockets = 1
  available_numachips = 3
  available_cores = 15
  available_threads = 30
```

In looking at the output for this example, you will see that even though only one core was requested the available sockets, numachip, cores and threads were all reduced. This is because the NUMA architecture is hierarchical: socket contains one or more numachips; a numachip contains two or more cores; cores contain one or more threads (one thread in the case of non-threaded cores). In order for a resource to be available, the entire resource must be free. When a job uses one core, the use of that core consumes part of the associated socket, and numa chip

resources. As a result, the affected socket and numachip cannot be used when subsequent jobs request sockets and numachips as resources. Also, because the job asked for one core, the number of threads for that core are consumed. As a result, the number of threads available on the machine is reduced by the number of threads in the core.

As another example, suppose a user makes an job request and they want to use a socket. The pbsnodes output will look like:

```
torque-devtest-01
  state = free
  power_state = Running
  np = 16
  ntype = cluster
  jobs = 113.torque-devtest-01
  status =
rectime=1412732948,macaddr=00:26:6c:f4:66:a0,cpuclock=Fixed,varattr=,jobs=,state=free,
netload=17080856592,gres=,loadave=10.74,ncpus=16,physmem=49416100kb,availmem=50056608k
b,totmem=51480476kb,idletime=29,nusers=2,nsessions=3,sessions=8665
8671 1994,uname=Linux torque-devtest-01 2.6.32-358.el6.x86_64 #1 SMP
Fri Feb 22 00:31:26 UTC 2013 x86_64,opsys=linux
  mom_service_port = 15002
  mom_manager_port = 15003
  total_sockets = 2
  total_numachips = 4
  total_cores = 16
  total_threads = 32
  available_sockets = 1
  available_numa_chips = 2
  available_cores = 8
  available_threads = 16
```

In looking at the output in this example, you will see that not only are the available sockets reduced to one, but all of the numachips, cores, and threads associated with the socket are no longer available. In other words, by requesting a job placement of "socket" all of the resources of the socket are reserved and are no longer available to other jobs.

NUMA-Support Systems

This topic serves as a central information repository for NUMA-support systems. This topic provides basic information and contains links to the various NUMA-aware topics found throughout the documentation.

i Support for NUMA-support systems is available only on large-scale SLES systems using SGI Altix and UV hardware and requires Torque 3.0 or later.

In this chapter:

- [About NUMA-Supported Systems on page 185](#)
- [Torque Installation and Configuration on page 185](#)
- [Moab/Torque NUMA Configuration on page 185](#)

About NUMA-Supported Systems

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.

Torque Installation and Configuration

To enable Torque for NUMA-support, you will need to add the `--enable-numa-support` option during the configure portion of the installation. You will also need create the `mom.layout` file and configure the `server_priv/nodes` file.

See:

[Torque NUMA-Support Configuration on page 53](#)

Moab/Torque NUMA Configuration

Moab *requires* additional configuration to enable NUMA-support.

See:

Moab-NUMA-Support Integration Guide in the *Moab Workload Manager Administrator Guide*.

Chapter 13: 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:

- ["Automatic Queue and Job Recovery" below](#)
- ["Host Resolution" below](#)
- ["Firewall Configuration" on the next page](#)
- ["Torque Log Files" on the next page](#)
- ["Using "tracejob" to Locate Job Failures" on page 190](#)
- ["Using GDB to Locate Job Failures" on page 192](#)
- ["Other Diagnostic Options" on page 193](#)
- ["Stuck Jobs" on page 193](#)
- ["Frequently Asked Questions \(FAQ\)" on page 194](#)
- ["Compute Node Health Check" on page 201](#)
- ["Debugging" on page 203](#)

Automatic Queue and Job Recovery

When `pbs_server` restarts and recovers a job but cannot find that job's queue, it will create a new queue with the original name, but with a `ghost_queue` attribute (as seen in `qmgr`) and then add the job to that queue. This will happen for each queue the server does not recognize. Ghost queues will not accept new jobs, but will allow the jobs in it to run and be in a running state. If users attempt to submit any new jobs to these queues, the user will get an error stating that the queue had an error on recovery, and is in a ghost state. Once the admin reviews and corrects the queue's settings, the admin may remove the ghost setting and then the queue will function normally.


See ["ghost_queue" on page 403](#) for more information.

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.

See [Specifying Compute Nodes on page 22](#) for more information about configuring `pbs_server` to identify compute nodes.

 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 the previous page](#)

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 the previous page](#)

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 loglevel 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.

i You might see some peculiar behavior if you mount the `client_logs` directory for shared access via network-attached storage.

When `trqauthd` first gets access on a particular day, it writes an "open" message to the day's log file. It also writes a "close" message to the last log file it accessed prior to that, which is usually the previous day's log file, but not always. For example, if it is Monday and no client commands were executed over the weekend, `trqauthd` writes the "close" message to Friday's file.

Since the various `trqauthd` binaries on the submit hosts (and potentially, the compute nodes) each write an "open" and "close" message on the first access of a new day, you'll see multiple (seemingly random) accesses when you have a shared log.

The `trqauthd` records the following events along with the date and time of the occurrence:

- When `trqauthd` successfully starts. It logs the event with the IP address and port.
- When a user successfully authenticates with `trqauthd`.
- When a user fails to authenticate with `trqauthd`.
- When `trqauthd` encounters any unexpected errors.

Example 13-1: trqauthd logging sample

```
2012-10-05 15:05:51.8404 Log opened
2012-10-05 15:05:51.8405 Torque authd daemon started and listening on IP:port
101.0.1.0:12345
2012-10-10 14:48:05.5688 User hfrye at IP:port abc:12345 logged in
```

Related Topics

["Troubleshooting" on page 187](#)

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

03/02/2005 17:58:28 S enqueueing 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
(s)
03/02/2005 17:58:41 A user=dev group=dev jobname=STDIN queue=batch ctime=1109811508

qtime=1109811508 etime=1109811508 start=1109811521
exec_host=icluster.org/0 Resource_List.needsnodes=1 Resource_
List.nodelist=1
Resource_List.nodes=1 Resource_List.walltime=00:01:40
03/02/2005 18:02:11 M walltime_210 exceeded limit 100
03/02/2005 18:02:11 M kill_job
03/02/2005 18:02:11 M kill_job found a task to kill
03/02/2005 18:02:11 M sending signal 15 to task
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 18:06:27 M kill_job
03/02/2005 18:06:27 M kill_job done
03/02/2005 18:06:27 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.needsnodes=1 Resource_
List.nodelist=1
Resource_List.nodes=1 Resource_List.walltime=00:01:40
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

...
```

i The `tracejob` command operates by searching the `pbs_server` accounting records and the `pbs_server`, `MOM`, and scheduler logs. To function properly, it must be run on a node and as a user which can access these files. By default, these files are all accessible by the user `root` and only available on the cluster management node. In particular, the files required by `tracejob` are located in the following directories:

```
TORQUE_HOME/server_priv/accounting
```

```
TORQUE_HOME/server_logs
```

```
TORQUE_HOME/mom_logs
```

```
TORQUE_HOME/sched_logs
```

i `tracejob` may only be used on systems where these files are made available. Non-root users may be able to use this command if the permissions on these directories or files are changed appropriately.

i The value of `Resource_List.*` is the amount of resources requested, and the value of `resources_used.*` is the amount of resources actually used.

Related Topics

["Troubleshooting" on page 187](#)

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.

i See the `PBSCOREDUMP` parameter for enabling creation of core files (see ["Debugging" on page 203](#)).

Related Topics

["Troubleshooting" on page 187](#)

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 187](#)

Stuck Jobs

If a job gets stuck in Torque, try these suggestions to resolve the issue:

- Use the **qdel** command to cancel the job.
- 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 187](#)

Frequently Asked Questions (FAQ)

- ["Cannot connect to server: error=15034" on the facing page](#)
- ["Deleting 'stuck' jobs" on the facing page](#)
- ["Which user must run Torque?" on page 196](#)
- ["Scheduler cannot run jobs - rc: 15003" on page 196](#)
- ["PBS_Server: pbsd_init, Unable to read server database" on page 196](#)
- ["qsub will not allow the submission of jobs requesting many processors" on page 198](#)
- ["qsub reports 'Bad UID for job execution'" on page 198](#)
- ["Why does my job keep bouncing from running to queued?" on page 198](#)
- ["How do I use PVM with Torque?" on page 199](#)
- ["My build fails attempting to use the TCL library" on page 199](#)
- ["My job will not start, failing with the message 'cannot send job to mom, state=PRERUN'" on page 199](#)
- ["How do I determine what version of Torque I am using?" on page 199](#)
- ["How do I resolve autogen.sh errors that contain "error: possibly undefined macro: AC_MSG_ERROR"?" on page 200](#)

- ["How do I resolve compile errors with libssl or libcrypto for Torque 4.0 on Ubuntu 10.04?" on page 200](#)
- ["Why are there so many error messages in the client logs \(trqauthd logs\) when I don't notice client commands failing?" on page 200](#)

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 o 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 249](#) `-p` command or can be removed manually using the following steps:

To remove job X

1. Shut down `pbs_server`.

```
> qterm
```

2. Remove job spool files.

```
> rm TORQUE_HOME/server_priv/jobs/X.SC TORQUE_HOME/server_priv/jobs/X.JB
```

3. Restart `pbs_server`

```
> 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 by 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:

```
%> 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 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:

```
> 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:

```
> cat data | qmgr
```

5. Restart pbs_server without the flags:

```
> qterm
> pbs_server
```

This will reinitialize the database to the current version.

 Reinitializing the server database will reset the next jobid to 1

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.

```
> 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 reports 'Bad UID for job execution'

```
[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 38](#).

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`)

i 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 `sshd` 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:

```
libtool: link: gcc -Wall -pthread -g -D_LARGEFILE64_SOURCE -o .libs/trqauthd trq_auth_
daemon.o trq_main.o -ldl -lssl -lcrypto -L/home/adaptive/torques/torque-
4.0.0/src/lib/Libpbs/.libs /home/adaptive/torques/torque-
4.0.0/src/lib/Libpbs/.libs/libtorque.so -lpthread -lrt -pthread
/usr/bin/ld: cannot find -lssl
collect2: ld returned 1 exit status
make[3]: *** [trqauthd] Error 1

libtool: link: gcc -Wall -pthread -g -D_LARGEFILE64_SOURCE -o .libs/trqauthd trq_auth_
daemon.o trq_main.o -ldl -lssl -lcrypto -L/home/adaptive/torques/torque-
4.0.0/src/lib/Libpbs/.libs /home/adaptive/torques/torque-
4.0.0/src/lib/Libpbs/.libs/libtorque.so -lpthread -lrt -pthread
/usr/bin/ld: cannot find -lcrypto
collect2: ld returned 1 exit status
make[3]: *** [trqauthd] Error 1
```

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 187](#)

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 `RMMSGIGNORE` parameter in Moab Parameters in the *Moab Workload Manager Administrator 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" below](#)
- ["Creating the Health Check Script" on the next page](#)
- ["Adjusting Node State Based on the Health Check Output" on page 203](#)
- ["Example Health Check Script" on page 203](#)


Related Topics

["Troubleshooting" on page 187](#)

Configuring MOMs to Launch a Health Check

The health check feature is configured via the `mom_priv/config` file using the parameters described below:

Parameter	Format	Default	Description
<code>\$node_check_script</code>	<STRING>	N/A	(Required) Specifies the fully qualified pathname of the health check script to run

Parameter	Format	Default	Description
\$node_check_interval	<INTEGER>	1	<p>(Optional) Specifies the number of MOM intervals between health checks (by default, each MOM interval is 45 seconds long - this is controlled via the "\$status_update_time" on page 346 node parameter. The integer may be followed by a list of event names (<code>jobstart</code> and <code>jobend</code> are currently supported). See "pbs_mom" on page 222 for more information. xref>.</p> <div>  The node health check may be configured to run before the prologue script by including the "jobstart" option. However, the job environment variables are not in the health check at that point. </div>


Related Topics


["Compute Node Health Check" on the previous page](#)

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.

By default, the script looks for the `EVENT :` keyword to indicate successes. 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 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.

 In order for the node health check script to log a positive run, it is necessary to include the keyword `EVENT :` at the beginning of the message your script returns. Failure to do so may result in unexpected outcomes.

 Both the `ERROR` and `EVENT :` keywords are case insensitive.

Related Topics

["Compute Node Health Check" on the previous page](#)

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. The node health script continues to run at the configured interval (see ["Configuring MOMs to Launch a Health Check" on page 201](#) for more information), and if it does not generate the error message again during one of its later executions, Moab picks that up at the beginning of its next iteration and restores the node to an online state.

Related Topics

["Compute Node Health Check" on page 201](#)

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:

```
#!/bin/sh
/bin/mount | grep global
if [ $? != "0" ]
then
    echo "ERROR cannot locate filesystem global"
fi
```

Related Topics

["Compute Node Health Check" on page 201](#)

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)

PBSCOREDUMP environment variable - If set, it will cause the offending *pbs_mom* or *pbs_server* daemon to create a core file if a *SIGSEGV*, *SIGILL*, *SIGFPE*, *SIGSYS*, or *SIGTRAP* signal is received.

i To enable core dump file creation in RHEL/CentOS 6 or SLES 11 systems, add the following line to the `/etc/init.d/pbs_mom` and `/etc/init.d/pbs_server` scripts (or the `TORQUE_HOME/pbs_environment` file):

```
export DAEMON_COREFILE_LIMIT=unlimited
```

For RHEL/CentOS 7 or SLES 12 systems using `systemd`, add this line to the `trqauthd.service`, `pbs_mom.service`, and `pbs_server.service` unit files (in `/usr/lib/systemd/system/`):

```
LimitCORE=infinity
```

Core dumps will be placed in the daemons' home directories as shown in the table below.

Daemon	Path
trqauthd	/var/spool/torque
pbs_server	/var/spool/torque/server_priv
pbs_mom	/var/spool/torque/mom_priv

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 (See ["Using 'tracejob' to Locate Job Failures" on page 190](#)) for more information.

i *PBSLOGLEVEL* and *PBSCOREDUMP* must be added to the `TORQUE_HOME/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

Error code name	Number	Description
PBSE_FLOOR	15000	No error
PBSE_UNKJOBID	15001	Unknown job identifier

Error code name	Number	Description
PBSE_NOATTR	15002	Undefined attribute
PBSE_ATTRRO	15003	Attempt to set READ ONLY attribute
PBSE_IVALREQ	15004	Invalid request
PBSE_UNKREQ	15005	Unknown batch request
PBSE_TOOMANY	15006	Too many submit retries
PBSE_PERM	15007	No permission
PBSE_IFF_NOT_FOUND	15008	"pbs_iff" not found; unable to authenticate
PBSE_MUNGE_NOT_FOUND	15009	"munge" executable not found; unable to authenticate
PBSE_BADHOST	15010	Access from host not allowed
PBSE_JOBEXIST	15011	Job already exists
PBSE_SYSTEM	15012	System error occurred
PBSE_INTERNAL	15013	Internal server error occurred
PBSE_REGROUTE	15014	Parent job of dependent in rte queue
PBSE_UNKSIG	15015	Unknown signal name
PBSE_BADATVAL	15016	Bad attribute value
PBSE_MODATTRRUN	15017	Cannot modify attribute in run state
PBSE_BADSTATE	15018	Request invalid for job state
PBSE_UNKQUE	15020	Unknown queue name
PBSE_BADCRED	15021	Invalid credential in request
PBSE_EXPIRED	15022	Expired credential in request

Error code name	Number	Description
PBSE_QUNOENB	15023	Queue not enabled
PBSE_QACCESS	15024	No access permission for queue
PBSE_BADUSER	15025	Bad user - no password entry
PBSE_HOPCOUNT	15026	Max hop count exceeded
PBSE_QUEEXIST	15027	Queue already exists
PBSE_ATTRTYPE	15028	Incompatible queue attribute type
PBSE_QUEBUSY	15029	Queue busy (not empty)
PBSE_QUENBIG	15030	Queue name too long
PBSE_NOSUP	15031	Feature/function not supported
PBSE_QUENOEN	15032	Cannot enable queue,needs add def
PBSE_PROTOCOL	15033	Protocol (ASN.1) error
PBSE_BADATLST	15034	Bad attribute list structure
PBSE_NOCONNECTS	15035	No free connections
PBSE_NOSERVER	15036	No server to connect to
PBSE_UNKRESC	15037	Unknown resource
PBSE_EXCQRESC	15038	Job exceeds queue resource limits
PBSE_QUENODFLT	15039	No default queue defined
PBSE_NORERUN	15040	Job not rerunnable
PBSE_ROUTEREJ	15041	Route rejected by all destinations
PBSE_ROUTEEXPD	15042	Time in route queue expired

Error code name	Number	Description
PBSE_MOMREJECT	15043	Request to MOM failed
PBSE_BADSCRIPT	15044	(qsub) Cannot access script file
PBSE_STAGEIN	15045	Stage-In of files failed
PBSE_RESCUNAV	15046	Resources temporarily unavailable
PBSE_BADGRP	15047	Bad group specified
PBSE_MAXQUED	15048	Max number of jobs in queue
PBSE_CKPBSY	15049	Checkpoint busy, may be retries
PBSE_EXLIMIT	15050	Limit exceeds allowable
PBSE_BADACCT	15051	Bad account attribute value
PBSE_ALRDYEXIT	15052	Job already in exit state
PBSE_NOCOPYFILE	15053	Job files not copied
PBSE_CLEANEOUT	15054	Unknown job id after clean init
PBSE_NOSYNCMSTR	15055	No master in sync set
PBSE_BADDEPEND	15056	Invalid dependency
PBSE_DUPLIST	15057	Duplicate entry in list
PBSE_DISPROTO	15058	Bad DIS based request protocol
PBSE_EXECTHERE	15059	Cannot execute there
PBSE_SISREJECT	15060	Sister rejected
PBSE_SISCOMM	15061	Sister could not communicate
PBSE_SVRDOWN	15062	Requirement rejected -server shutting down

Error code name	Number	Description
PBSE_CKPSHORT	15063	Not all tasks could checkpoint
PBSE_UNKNODE	15064	Named node is not in the list
PBSE_UNKNODEATR	15065	Node-attribute not recognized
PBSE_NONODES	15066	Server has no node list
PBSE_NODENBIG	15067	Node name is too big
PBSE_NODEEXIST	15068	Node name already exists
PBSE_BADNDATVAL	15069	Bad node-attribute value
PBSE_MUTUALEX	15070	State values are mutually exclusive
PBSE_GMODERR	15071	Error(s) during global modification of nodes
PBSE_NORELYMOM	15072	Could not contact MOM
PBSE_NOTSNODE	15073	No time-shared nodes
PBSE_JOBTYPE	15074	Wrong job type
PBSE_BADACLHOST	15075	Bad ACL entry in host list
PBSE_MAXUSERQUED	15076	Maximum number of jobs already in queue for user
PBSE_BADDISALLOWTYPE	15077	Bad type in "disallowed_types" list
PBSE_NOINTERACTIVE	15078	Interactive jobs not allowed in queue
PBSE_NOBATCH	15079	Batch jobs not allowed in queue
PBSE_NORERUNABLE	15080	Rerunable jobs not allowed in queue
PBSE_NONONRERUNABLE	15081	Non-rerunable jobs not allowed in queue
PBSE_UNKARRAYID	15082	Unknown array ID

Error code name	Number	Description
PBSE_BAD_ARRAY_REQ	15083	Bad job array request
PBSE_TIMEOUT	15084	Time out
PBSE_JOBNOTFOUND	15085	Job not found
PBSE_NOFAULTTOLERANT	15086	Fault tolerant jobs not allowed in queue
PBSE_NOFAULTINTOLERANT	15087	Only fault tolerant jobs allowed in queue
PBSE_NOJOBARRAYS	15088	Job arrays not allowed in queue
PBSE_RELAYED_TO_MOM	15089	Request was relayed to a MOM
PBSE_MEM_MALLOC	15090	Failed to allocate memory for memmgr
PBSE_MUTEX	15091	Failed to allocate controlling mutex (lock/unlock)
PBSE_TRHEADATTR	15092	Failed to set thread attributes
PBSE_THREAD	15093	Failed to create thread
PBSE_SELECT	15094	Failed to select socket
PBSE_SOCKET_FAULT	15095	Failed to get connection to socket
PBSE_SOCKET_WRITE	15096	Failed to write data to socket
PBSE_SOCKET_READ	15097	Failed to read data from socket
PBSE_SOCKET_CLOSE	15098	Socket closed
PBSE_SOCKET_LISTEN	15099	Failed to listen in on socket
PBSE_AUTH_INVALID	15100	Invalid auth type in request
PBSE_NOT_IMPLEMENTED	15101	Functionality not yet implemented
PBSE_QUENOTAVAILABLE	15102	Queue is not available

Related Topics

["Troubleshooting" on page 187](#)

Appendices

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

- [Commands Overview on page 213](#)
- [Server Parameters on page 299](#)
- [Node Manager \(MOM\) Configuration on page 327](#)
- [Diagnostics and Error Codes on page 351](#)
- [Considerations Before Upgrading on page 359](#)
- [Large Cluster Considerations on page 361](#)
- [Prologue and Epilogue Scripts on page 369](#)
- [Running Multiple Torque Servers and MOMs on the Same Node on page 377](#)
- [Security Overview on page 379](#)
- [Job Submission Filter \("qsub Wrapper"\) on page 381](#)
- ["torque.cfg" Configuration File on page 383](#)
- [Torque Quick Start Guide on page 389](#)
- [BLCR Acceptance Tests on page 393](#)
- [Queue Attributes on page 399](#)

Commands Overview

Torque Services

Command	Description
pbs_iff	Interprocess authentication service. Deprecated.
<code>"pbs_mom"</code> on page 222	PBS batch execution mini-server. Runs on each Torque compute node.
<code>"pbs_server"</code> on page 229	Batch system manager daemon. Runs on the Torque master node.
<code>"pbs_track"</code> on page 232	Process launcher. Starts a new process and tells pbs_mom to track its lifecycle and resource usage.
<code>"trqauthd"</code> on page 296	Torque authorization daemon.

Client Commands

Command	Description
<code>"momctl"</code> on page 215	Manage/diagnose MOM (node execution) daemon
<code>"pbsdsh"</code> on page 234	Launch tasks within a parallel job
<code>"pbsnodes"</code> on page 236	View/modify batch status of compute nodes
<code>"qalter"</code> on page 239	Modify queued batch jobs

Command	Description
"qchkpt" on page 248	Checkpoint batch jobs
"qdel" on page 249	Delete/cancel batch jobs
"qgpumode" on page 251	Specifies new mode for GPU
"qgpureset" on page 253	Reset the GPU
"qhold" on page 253	Hold batch jobs
"qmgr" on page 255	Manage policies and other batch configuration
"qmove" on page 259	Move batch jobs
"qorder" on page 260	Exchange order of two batch jobs in any queue
"qrerun" on page 261	Rerun a batch job
"qrls" on page 263	Release batch job holds
"qrun" on page 264	Start a batch job
"qsig" on page 265	Send a signal to a batch job
"qstat" on page 267	View queues and jobs
"qsub" on page 275	Submit jobs

Command	Description
"qterm" on page 295	Shutdown pbs server daemon
tracejob	Trace job actions and states recorded in Torque logs (see "Using "tracejob" to Locate Job Failures" on page 190)

Related Topics

- [" Node Manager \(MOM\) Configuration" on page 327](#)
- [" Server Parameters" on page 299](#)

momctl

(PBS MOM Control)

Synopsis

```
momctl -c { <JOBID> | all }
momctl -C
momctl -d { <INTEGER> | <JOBID> }
momctl -f <FILE>
momctl -h <HOST>[,<HOST>]...
momctl -l
momctl -p <PORT_NUMBER>
momctl -q <ATTRIBUTE>
momctl -r { <FILE> | LOCAL:<FILE> }
momctl -s
momctl -u
```

Overview

The *momctl* command allows remote shutdown, reconfiguration, diagnostics, and querying of the pbs_mom daemon.

Format

-c — Clear	
Format	{ <JOBID> all }

-c — Clear	
Default	---
Description	Makes the MOM unaware of the job's existence. It does not clean up any processes associated with the job.
Example	<pre>momctl -c 2 -h node1</pre>

-C — Cycle	
Format	---
Default	---
Description	Cycle pbs_mom (force the MOM to send a status update to pbs_server).
Example	<pre>momctl -C -h node1</pre> <p><i>Cycle pbs_mom on node1.</i></p>

-d — Diagnose	
Format	{ <INTEGER> <JOBID> }
Default	0
Description	Diagnose MOM(s). (For more details, see "Diagnose detail" on page 221 below.)
Example	<pre>momctl -d 2 -h node1</pre> <p><i>Print level 2 and lower diagnose information for the MOM on node1.</i></p>

-f — Host File

Format	<FILE>
Default	---
Description	A file containing only comma or whitespace (space, tab, or new line) delimited hostnames.
Example	<pre>momctl -f hosts.txt -d 0</pre> <p><i>Print diagnose information for the MOMs running on the hosts specified in hosts.txt.</i></p>

-h — Host List

Format	<HOST>[,<HOST>]...
Default	localhost
Description	A comma separated list of hosts.
Example	<pre>momctl -h node1,node2,node3 -d 0</pre> <p><i>Print diagnose information for the MOMs running on node1, node2, and node3.</i></p>

-l — Layout

Format	---
Default	---
Description	Display the layout of the machine as it is understood by Torque.

-l — Layout

Example

```
[root@c04a numa]# momctl -l
Machine (50329748KB)
Socket 0 (33552532KB)
  Chip 0 (16775316KB)
    Core 0 (1 threads)
    Core 1 (1 threads)
    Core 2 (1 threads)
    Core 3 (1 threads)
    Core 4 (1 threads)
    Core 5 (1 threads)
    Core 6 (1 threads)
    Core 7 (1 threads)
  Chip 1 (16777216KB)
    Core 8 (1 threads)
    Core 9 (1 threads)
    Core 10 (1 threads)
    Core 11 (1 threads)
    Core 12 (1 threads)
    Core 13 (1 threads)
    Core 14 (1 threads)
    Core 15 (1 threads)
Socket 1 (16777216KB)
  Chip 2 (8388608KB)
    Core 16 (1 threads)
    Core 17 (1 threads)
    Core 18 (1 threads)
    Core 19 (1 threads)
    Core 20 (1 threads)
    Core 21 (1 threads)
    Core 22 (1 threads)
    Core 23 (1 threads)
  Chip 3 (8388608KB)
    Core 24 (1 threads)
    Core 25 (1 threads)
    Core 26 (1 threads)
    Core 27 (1 threads)
    Core 28 (1 threads)
    Core 29 (1 threads)
    Core 30 (1 threads)
    Core 31 (1 threads)
```

-p — Port

Format

<PORT_NUMBER>

Default

Torque's default port number.

Description

The port number for the specified MOM(s).

Example

```
momctl -p 5455 -h node1 -d 0
```

Request diagnose information over port 5455 on node1.

-q — Query	
Format	<ATTRIBUTE>
Default	---
Description	Query or set <ATTRIBUTE>, where <ATTRIBUTE> is a property listed by pbsnodes -a (see "Query attributes" on the next page for a list of attributes).
Example	<pre>momctl -q physmem</pre> <p><i>Print the amount of physmem on localhost.</i></p> <pre>momctl -h node2 -q loglevel=7</pre> <p><i>Change the current MOM logging on node2 to level 7.</i></p>

-r — Reconfigure	
Format	{ <FILE> LOCAL:<FILE> }
Default	---
Description	Reconfigure MOM(s) with remote or local config file, <FILE>. This does not work if \$remote_reconf is not set to true when the MOM is started.
Example	<pre>momctl -r /home/user1/new.config -h node1</pre> <p><i>Reconfigure MOM on node1 with /home/user1/new.cofig on node1.</i></p>

-s — Shutdown	
Format	---
Default	---
Description	Have the MOM shut itself down gracefully (this includes reporting to server so that pbsnodes marks the node down).

-s — Shutdown

Example

```
momctl -s
```

Shut down the pbs_mom process on localhost.

-u — Update

Format

Default

Description

Update the hardware configuration on pbs_server with a layout from the MOM.

Example

```
momctl -u
```

Update pbs_server hardware configuration.

Query attributes

Attribute	Description
arch	node hardware architecture
availmem	available RAM
loadave	1 minute load average
ncpus	number of CPUs available on the system
netload	total number of bytes transferred over all network interfaces
nsessions	number of sessions active
nusers	number of users active
physmem	configured RAM
sessions	list of active sessions

Attribute	Description
totmem	configured RAM plus configured swap

Diagnose detail

Level	Description
0	<p>Display the following information:</p> <ul style="list-style-type: none"> • Local hostname • Expected server hostname • Execution version • MOM home directory • MOM config file version (if specified) • Duration MOM has been executing • Duration since last request from pbs_server daemon • Duration since last request to pbs_server daemon • RM failure messages (if any) • Log verbosity level • Local job list
1	<p>All information for level 0 plus the following:</p> <ul style="list-style-type: none"> • 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	<p>All information from level 1 plus the following:</p> <ul style="list-style-type: none"> • PID • Event alarm status
3	<p>All information from level 2 plus the following:</p> <ul style="list-style-type: none"> • syslog enabled

Example A-1: MOM diagnostics

```
[root@node01]# momctl -dl

Host: node01/node01   Version: 6.1.1.1   PID: 30404
Server[0]: torque-server (10.2.15.70:15001)
  Last Msg From Server: 1275 seconds (StatusJob)
  Last Msg To Server: 42 seconds
HomeDirectory: /var/spool/torque/mom_priv
stdout/stderr spool directory: '/var/spool/torque/spool/' (15518495 blocks available)
MOM active: 260257 seconds
Check Poll Time: 45 seconds
Server Update Interval: 45 seconds
LogLevel: 7 (use SIGUSR1/SIGUSR2 to adjust)
Communication Model: TCP
MemLocked: TRUE (mlock)
TCP Timeout: 300 seconds
Trusted Client List:
10.2.15.3:15003,10.2.15.5:15003,10.2.15.6:15003,10.2.15.70:0,10.2.15.204:15003,127.0.0
.1:0
Copy Command: /bin/scp -rpB
NOTE: no local jobs detected

diagnostics complete
```

Example A-2: System shutdown

```
> momctl -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

```
pbs_mom [-a alarm] [-A alias] [-C chkdiritory] [-c config] [-d
directory] ["-f" on the facing page] [-F] [-h help] [-H hostname] [-L
logfile] [-M MOMport] [-R RMPport] [-p|-r] [-P purge] [-w] [-x]
```

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 ensure 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

Flag	Name	Description
-a	alarm	Specifies 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.
-A	alias	Specifies this multimom's alias name. The alias name needs to be the same name used in the <i>mom.hierarchy</i> file. It is only needed when running multiple MOMs on the same machine. For more information, see "Torque Multi-MOM" on page 57 .
-c	config	Specifies an alternative configuration file, see description below. If this is a relative file name it will be relative to <code>TORQUE_HOME/mom_priv</code> , (see the -d option). If the specified file cannot be opened, <i>pbs_mom</i> will abort. If the -c option is not supplied, <i>pbs_mom</i> will attempt to open the default configuration file "config" in <code>TORQUE_HOME/mom_priv</code> . If this file is not present, <i>pbs_mom</i> will log the fact and continue.
-C	chkdirectory	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.
-d	directory	Specifies the path of the directory which is the home of the server's working files, <code>TORQUE_HOME</code> . This option is typically used along with -M when debugging MOM. The default directory is given by <code>TORQUE_HOME</code> which is typically <code>/var/spool/torque</code> .
-f	force_update	Forces the server to accept an update of the hardware on the node. Should be used the first time <i>pbs_mom</i> is run after a hardware update on a node.

Flag	Name	Description
-F	fork	Do not fork. <div> This option is useful when running under systemd (Red Hat 7-based or SUSE 12-based systems).</div>
-h	help	Displays the help/usage message.
-H	hostname	Sets the MOM's hostname. This can be useful on multi-homed networks.
-L	logfile	Specifies 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 <code>TORQUE_HOME/mom_logs</code> directory (see the -d option).
-M	port	Specifies the port number on which the mini-server (MOM) will listen for batch requests.
-p	n/a	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 -p 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 -p option is mutually exclusive with the -r and -q options.
-P	purge	Specifies the impact on jobs which were in execution when the mini-server shut down. With the -P 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 -p option, no attempt is made to try and preserve or recover running jobs. All jobs are terminated and removed from the queue.
-q	n/a	Specifies the impact on jobs which were in execution when the mini-server shut down. With the -q option, MOM will allow the processes belonging to jobs to continue to run, but will not attempt to monitor them. The -q option is mutually exclusive with the -p and -r options.

Flag	Name	Description
-r	n/a	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.
-R	port	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.
-w	wait_for_server	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.
-x	n/a	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.

Configuration file

The configuration file, located at `mom_priv/config` by default, can 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).

See the ["Parameters" on page 327](#) page for a full list of `pbs_mom` parameters.

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`.

If a qualifier is sent with no matching token in the command line, `escape[zzz=snafu]`, an error is reported.

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.

Request	Description
cycle	Forces an immediate MOM cycle.
status_update_time	Retrieve or set the <code>\$status_update_time</code> parameter.
check_poll_time	Retrieve or set the <code>\$check_poll_time</code> parameter.
configversion	Retrieve the config version.
jobstartblocktime	Retrieve or set the <code>\$jobstartblocktime</code> parameter.
enablemomrestart	Retrieve or set the <code>\$enablemomrestart</code> parameter.
loglevel	Retrieve or set the <code>\$loglevel</code> parameter.
down_on_error	Retrieve or set the <code>\$down_on_error</code> parameter.

Request	Description
diag0 - diag4	Retrieves varied diagnostic information.
rcpcmd	Retrieve or set the \$rcpcmd parameter.
version	Retrieves the pbs_mom version.

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 `ERROR` keyword to `stdout` followed by an error message. The message (up to 1024 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 `$down_on_error` MOM setting is enabled, the MOM will set itself to state down and report to pbs_server. Additionally, the `$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 the MOM's `$down_on_error` and pbs_server's `down_on_error` features enabled. Also see ["down_on_error" on page 307](#) (in [Server Parameters on page 299](#)).

See [Creating the Health Check Script on page 202](#) for more information.

Files

File	Description
TORQUE_HOME/server_name	File containing the pbs_server hostname
TORQUE_HOME/mom_priv	Directory for configuration files (/var/spool/torque/mom_priv by default)
TORQUE_HOME/mom_logs	Directory for log files recorded by the server

File	Description
TORQUE_HOME/mom_priv/-prologue	The administrative script to be run before job execution
TORQUE_HOME/mom_priv/e-pilogue	The administrative script to be run after job execution

Signal handling

`pbs_mom` handles the following signals:

Signal	Description
SIGHUP	Causes <code>pbs_mom</code> to re-read its configuration file, close and reopen the log file, and reinitialize resource structures.
SIGALRM	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.
SIGINT and SIGTERM	Results in <code>pbs_mom</code> exiting without terminating any running jobs. This is the action for the following signals as well: SIGXCPU, SIGXFSZ, SIGCPULIM, and SIGSHUTDN.
SIGUSR1, SIGUSR2	Causes the MOM to increase and decrease logging levels, respectively.
SIGPIPE, SIGINFO	Are ignored.
SIGBUS, SIGFPE, SIGILL, SIGTRAP, and SIGSYS	Cause a core dump if the PBSCOREDUMP environmental variable is defined.

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 on page 229](#)

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

```
pbs_server [-a active] [-A acctfile] [-c] [-d config_path] [-f force
overwrite] [-F fork] [-H hostname] [--ha] [-l location] [-L logfile]
[-n don't send hierarchy] [-p port] [-S scheduler_port] [-t type] [-
v] [--about] [--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

Option	Name	Description
-a	active	Specifies if scheduling is active or not. This sets the <i>scheduling</i> server attribute. If the option argument is "true" ("True", "t", "T", or "1"), the server is active and the PBS job scheduler will be called. If the argument is "false" ("False", "f", "F", or "0"), the server is idle, and the scheduler will not be called. Jobs would then need to be run manually or via an external scheduler. If this option is not specified, the server will retain the prior value of the scheduling attribute.

Option	Name	Description
-A	acctfile	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 <code>TORQUE_HOME/server_priv/ac-counting</code> directory.
-c	wait_for_moms	This directs <code>pbs_server</code> 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.
-d	config_dir- ectory	Specifies the path of the directory which is home to the server's configuration files, <code>TORQUE_HOME</code> . A host may have multiple servers. Each server must have a different configuration directory. The default configuration directory is given by the symbol <code>TORQUE_HOME</code> which is typically <code>/var/spool/torque</code> .
-f	force over- write	Forces an overwrite of the server database. This can be useful to bypass the <code>yes/no</code> prompt when running something like <code>pbs_server -t create</code> and can ease installation and configuration of Torque via scripts.
-F	fork	Do not fork. <div> This option is useful when running under systemd (Red Hat 7-based or SUSE 12-based systems).</div>
-H	hostname	Causes the server to start under a different hostname as obtained from <code>gethostname</code> (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 <code>gethost name</code> (2).
--ha	high_avail- ability	Starts server in high availability mode (for details, see "Server High Availability" on page 119).
-l	location	Specifies where to find the scheduler (for example, Moab) when it does not reside on the same host as Torque. <div><code>pbs_server -l <other_host>:<other_port></code></div>
-L	logfile	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 <code>TORQUE_HOME/server_logs</code> directory.
-n	no send	This directs <code>pbs_server</code> 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. See Setting Up the MOM Hierarchy (Optional) on page 41 .

Option	Name	Description
-p	port	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.
-S	scheduler_ port	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.
-t	type	<p>If the job is rerunnable or restartable, and <i>-t create</i> is specified, the server will discard any existing configuration files, queues, and jobs, and initialize configuration files to the default values. The server is idled.</p> <div>  If -t is not specified, the job states will remain the same. </div>

Files

File	Description
TORQUE_HOME/server_priv	Default directory for configuration files, typically /var/spool/torque/server_priv
TORQUE_HOME/server_logs	Directory for log files recorded by the server

Signal handling

On receipt of the following signals, the server performs the defined action:

Action	Description
SIGHUP	<p>The current server log and accounting log are closed and new log files opened. This allows for the prior logs to be renamed and new logs started from the time of the signal. Usage example:</p> <pre># mv 20170816 20170816.old && kill -HUP \$(pgrep pbs_server)</pre>
SIGINT	Causes an orderly shutdown of pbs_server.
SIGUSR1, SIGUSR2	Causes server to increase and decrease logging levels, respectively. Use <code>qmgr -c "set server loglevel=7"</code> instead to set it directly.

Action	Description
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

```
pbs_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

Option	Description
-b	Instead of having <i>pbs_track</i> send its PID to Torque, it will <code>fork()</code> first, send the child PID to Torque, and then execute from the forked child. This essentially "backgrounds" <i>pbs_track</i> so that it will return after the new process is launched.
-j <JOBID>	Job ID the new process should be associated with.

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

The *pbsdsh* command distributes tasks to nodes under pbs.

- i** Some limitations exist in the way that *pbsdsh* can be used. Please note the following situations are not currently supported:
- Running multiple instances of *pbsdsh* concurrently within a single job.
 - Using the *-o* and *-s* options concurrently; although requesting these options together is permitted, only the output from the first node is displayed rather than output from every node in the chain.

Synopsis

```
pbsdsh [-c copies] [-o] [-s] [-u] [-v] program [args]
pbsdsh [-n node] [-o] [-s] [-u] [-v] program [args]
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.

- i** Note that under particularly high workloads, the *pbsdsh* command may not function properly.

Options

Option	Name	Description
-c	copies	The program is spawned on the first Copies nodes allocated. This option is mutually exclusive with -n .
-h	hostname	The program is spawned on the node specified.
-n	node	The program is spawned on one node which is the n-th node allocated. This option is mutually exclusive with -c .
-o	---	Capture stdout of the spawned program. Normally stdout goes to the job's output.
-s	---	If this option is given, the program is run in turn on each node, one after the other.
-u	---	The program is run once on each node (unique). This ignores the number of allocated processors on a given node.
-v	---	Verbose output about error conditions and task exit status is produced.

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

```
pbsnodes [-{a|x}] [-q] [-s server] [node[:property]]
pbsnodes -l [-q] [-s server] [state] [nodename[:property] ...]
pbsnodes -m <running|standby|suspend|hibernate|shutdown> <host list>
pbsnodes [-{c|d|o|r}] [-q] [-s server] [-n -l] [-N "note"] [-A
"append 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.



NUMA-Awareness

When Torque is configured with NUMA-awareness and configured with `--enable-groups`, the number of total *and* the number of available sockets, numachips (numa nodes), cores, and threads are returned when the status of nodes are queried by Moab (a call is made to *pbsnodes*).

See [pbsnodes with NUMA-Awareness on page 182](#) for additional information and examples.

Options

Option	Description
-a	All attributes of a node or all nodes are listed. This is the default if no flag is given.

Option	Description
-A	Append a note attribute to existing note attributes. The -N note option will overwrite exiting note attributes. -A will append a new note attribute to the existing note attributes delimited by a ',' and a space.
-c	Clear OFFLINE from listed nodes.
-d	Print MOM diagnosis on the listed nodes. Not yet implemented. Use momctl instead.
-l	<p>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 "active", "all", "busy", "down", "free", "job-exclusive", "job-sharing", "offline", "reserve", "state-unknown", "time-shared", and "up".</p> <ul style="list-style-type: none"> • Using <i>all</i> displays all nodes and their attributes. • Using <i>active</i> displays all nodes which are job-exclusive, job-sharing, or busy. • Using <i>up</i> displays all nodes in an "up state". Up states include job-exclusive, job-sharing, reserve, free, busy and time-shared. • All other strings display the nodes which are currently in the state indicated by the string.
-m	<p>Set the hosts in the specified host list to the requested power state. If a compute node does not support the energy-saving power state you request, the command returns an error and leaves the state unchanged.</p> <p>In order for the command to wake a node from a low-power state, Wake-on-LAN (WOL) must be enabled for the node.</p> <div> <p> In order for the command to wake a node from a low-power state, Wake-on-LAN must be enabled for the node and it must support the  WOL packet. For more information, see "Changing Node Power States" on page 106.</p> </div> <p>The allowable power states are:</p> <ul style="list-style-type: none"> • Running: The node is up and running. • Standby: CPU is halted but still powered. Moderate power savings but low latency entering and leaving this state. • Suspend: Also known as Suspend-to-RAM. Machine state is saved to RAM. RAM is put into self-refresh mode. Much more significant power savings with longer latency entering and leaving state. • Hibernate: Also known as Suspend-to-disk. Machine state is saved to disk and then powered down. Significant power savings but very long latency entering and leaving state. • Shutdown: Equivalent to <i>shutdown now</i> command as root. <p>The host list is a space-delimited list of node host names. See Examples on page 238.</p>
-n	Show the "note" attribute for nodes that are DOWN, OFFLINE, or UNKNOWN. This option requires -l .

Option	Description
-N	Specify a "note" attribute. This allows an administrator to add an arbitrary annotation to the listed nodes. To clear a note, use <code>-N ""</code> or <code>-N n</code> .
-o	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 pbs_server; it is purely for the manager or operator.
-p	Purge the node record from pbs_server. Not yet implemented.
-q	Suppress all error messages.
-r	Reset the listed nodes by clearing OFFLINE and adding DOWN state. pbs_server will ping the node and, if they communicate correctly, free the node.
-s	Specify the PBS server's hostname or IP address.
-x	Same as -a , but the output has an XML-like format.

Examples

Example A-3: host list

```
pbsnodes -m shutdown node01 node02 node03 node04
```

With this command, pbs_server tells the pbs_mom associated with nodes01-04 to shut down the node.

The *pbsnodes* output shows the current power state of nodes. In this example, note that *pbsnodes* returns the MAC addresses of the nodes.

```

pbsnodes
nuc1
  state = free
  power_state = Running
  np = 4
  ntype = cluster
  status = retime=1395765676,macaddr=0b:25:22:92:7b:26
, cpuclock=Fixed, varattr=, jobs=, state=free, netload=1242652020, gres=, loadave=0.16, ncpus=
6, physmem=16435852kb, availmem=24709056kb, totmem=33211016kb, idletime=4636, nusers=3, nses
sions=12, sessions=2758 998 1469 2708 2797 2845 2881 2946 4087 4154 4373
6385, uname=Linux bdaw 3.2.0-60-generic #91-Ubuntu SMP Wed Feb 19 03:54:44 UTC 2014
x86_64, opsys=linux
  note = This is a node note
  mom_service_port = 15002
  mom_manager_port = 15003

nuc2
  state = free
  power_state = Running
  np = 4
  ntype = cluster
  status = retime=1395765678,macaddr=2c:a8:6b:f4:b9:35
, cpuclock=OnDemand:800MHz, varattr=, jobs=, state=free, netload=12082362, gres=, loadave=0.0
0, ncpus=4, physmem=16300576kb, availmem=17561808kb, totmem=17861144kb, idletime=67538, nuse
rs=2, nsessions=7, sessions=2189 2193 2194 2220 2222 2248 2351, uname=Linux nuc2 2.6.32-
431.el6.x86_64 #1 SMP Fri Nov 22 03:15:09 UTC 2013 x86_64, opsys=linux
  mom_service_port = 15002
  mom_manager_port = 15003

```

Related Topics

[pbs_server\(8B\)](#)

Non-Adaptive Computing topics

- [PBS External Reference Specification](#)

qalter

Alter batch job.

Synopsis

```

qalter [-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

Option	Name	Description
-a	date_time	<p>Replaces the time at which the job becomes eligible for execution. The date_time argument syntax is:</p> <pre>[[[CC]YY]MM]DD]hhmm[.SS]</pre> <p>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.</p> <p>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	<p>Replaces the account string associated with the job. This attribute cannot be altered once the job has begun execution.</p>
-c	checkpoint_interval	<p>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.</p> <p>The interval argument is specified as:</p> <ul style="list-style-type: none">• <i>n</i> – No checkpointing is to be performed.• <i>s</i> – Checkpointing is to be performed only when the server executing the job is shutdown.• <i>c</i> – Checkpointing is to be performed at the default minimum cpu time for the queue from which the job is executing.• <i>c=minutes</i> – 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. <div> This value must be greater than zero. If the number is less than the default checkpoint time, the default time will be used.</div> <p>This attribute can be altered once the job has begun execution, but the new value does not take effect unless the job is rerun.</p>

Option	Name	Description
-e	path_name	<p>Replaces the path to be used for the standard error stream of the batch job. The path argument is of the form:</p> <pre>[hostname:]path_name</pre> <p>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 1003.1. The argument will be interpreted as follows:</p> <ul style="list-style-type: none"> • <i>path_name</i> – 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 <code>hostname</code> component. • <i>hostname:path_name</i> – 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 <code>hostname</code>. <p>This attribute can be altered once the job has begun execution, but it will not take effect unless the job is rerun.</p>
-h	hold_list	<p>Updates the types of holds on the job. The <code>hold_list</code> argument is a string of one or more of the following characters:</p> <ul style="list-style-type: none"> • <i>u</i> – Add the USER type hold. • <i>s</i> – Add the SYSTEM type hold if the user has the appropriate level of privilege. (Typically reserved to the batch administrator.) • <i>o</i> – 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.) • <i>n</i> – 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. <p>This attribute can be altered once the job has begun execution, but the hold will not take effect unless the job is rerun.</p>
-j	join	<p>Declares which standard streams of the job will be merged together. The join argument value may be the characters "oe" or "eo", or the single character "n".</p> <p>An argument value of <code>oe</code> 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 <code>eo</code> directs that the standard output and standard error streams of the job will be merged, intermixed, and returned as the standard error.</p> <p>A value of <code>n</code> 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.</p> <div>  If using either the <code>-e</code> or the <code>-o</code> option and the <code>-j eo oe</code> option, the <code>-j</code> option takes precedence and all standard error and output messages go to the chosen output file. </div>

Option	Name	Description
-k	keep	<p>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.</p> <p>The argument is either the single letter "e", "o", or "n", or one or more of the letters "e" and "o" combined in either order.</p> <ul style="list-style-type: none"> • <i>n</i> – No streams are to be retained. • <i>e</i> – 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: <code>job_name.es_{sequence}</code> where <code>job_name</code> is the name specified for the job, and <code>sequence</code> is the sequence number component of the job identifier. • <i>o</i> – 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: <code>job_name.os_{sequence}</code> where <code>job_name</code> is the name specified for the job, and <code>sequence</code> is the sequence number component of the job identifier. • <i>eo</i> – Both the standard output and standard error streams will be retained. • <i>oe</i> – Both the standard output and standard error streams will be retained. <p>This attribute cannot be altered once the job has begun execution.</p>
-l	resource_list	<p>Modifies the list of resources that are required by the job. The <code>resource_list</code> argument is in the following syntax:</p> <pre>resource_name[=[value]] [, resource_name[=[value]] , ...]</pre> <p>For the complete list of resources that can be modified, see "Requesting Resources" on page 65.</p> <p>If a requested modification to a resource would exceed the resource limits for jobs in the current queue, the server will reject the request.</p> <p>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.</p>
-m	mail_options	<p>Replaces the set of conditions under which the execution server will send a mail message about the job. The <code>mail_options</code> argument is a string which consists of the single character "n", or one or more of the characters "a", "b", and "e".</p> <p>If the character "n" is specified, no mail will be sent.</p> <p>For the letters "a", "b", and "e":</p> <ul style="list-style-type: none"> • <i>a</i> – Mail is sent when the job is aborted by the batch system. • <i>b</i> – Mail is sent when the job begins execution. • <i>e</i> – Mail is sent when the job ends.

Option	Name	Description
-M	user_list	<p>Replaces the list of users to whom mail is sent by the execution server when it sends mail about the job.</p> <p>The user_list argument is of the form:</p> <pre>user[@host] [,user[@host],...]</pre>
-n	node-exclusive	<p>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.</p> <pre>> qalter ... -n y #enables exclusive node allocation on a job > qalter ... -n n #disables exclusive node allocation on a job</pre>
-N	name	<p>Renames the job. The name specified may be up to and including 15 characters in length. It must consist of printable, nonwhite space characters with the first character alphabetic.</p>
-o	path	<p>Replaces the path to be used for the standard output stream of the batch job. The path argument is of the form:</p> <pre>[hostname:]path_name</pre> <p>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:</p> <ul style="list-style-type: none"> • <i>path_name</i> – 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 <code>hostname</code> component. • <i>hostname:path_name</i> – 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 <code>hostname</code>. <p>This attribute can be altered once the job has begun execution, but it will not take effect unless the job is rerun.</p>
-p	priority	<p>Replaces the priority of the job. The priority argument must be an integer between -1024 and +1023 inclusive.</p> <p>This attribute can be altered once the job has begun execution, but it will not take effect unless the job is rerun.</p>
-r	[y/n]	<p>Declares whether the job is rerunnable (see the qrerun command). The option argument <code>c</code> is a single character. PBS recognizes the following characters: y and n. If the argument is "y", the job is marked rerunnable.</p> <p>If the argument is "n", the job is marked as not rerunnable.</p>

Option	Name	Description
-S	path	<p>Declares the shell that interprets the job script.</p> <p>The option argument path_list is in the form: path[@host] [,path[@host], ...]</p> <p>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.</p> <p>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.</p> <p>This attribute can be altered once the job has begun execution, but it will not take effect unless the job is rerun.</p>
-t	array_range	<p>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</p> <p>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.</p> <p>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 (%).</p> <pre>qalter 15.napali[] -t %20</pre> <p>Here, the array 15.napali[] is configured to allow a maximum of 20 concurrently running jobs.</p> <p>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.</p>
-u	user_list	<p>Replaces the user name under which the job is to run on the execution system.</p> <p>The user_list argument is of the form: user[@host] [,user[@host], ...]</p> <p>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.</p> <p>This attribute cannot be altered once the job has begun execution.</p>
-W	additional_attributes	<p>The -W option allows for the modification of additional job attributes.</p> <p>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.</p> <p>To see the attributes PBS currently supports within the -W option, see "-W additional_attributes" on the facing page.</p>

-W additional_attributes

The following table lists the attributes PBS currently supports with the -W option.

Attribute	Description
depend=dependency_list	<p>Redefines the dependencies between this and other jobs. The dependency_list is in the form:</p> <pre>type[:argument[:argument...]][,type:argument...]</pre> <p>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: <code>seq_number.server.name</code>, 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).</p> <ul style="list-style-type: none"> • <i>synccount:count</i> – 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. • <i>syncwith:jobid</i> – 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. • <i>after:jobid [:jobid...]</i> – This job may be scheduled for execution at any point after jobs jobid have started execution. • <i>afterok:jobid [:jobid...]</i> – This job may be scheduled for execution only after jobs jobid have terminated with no errors. See the csh warning under "Extended Description". • <i>afternotok:jobid [:jobid...]</i> – This job may be scheduled for execution only after jobs jobid have terminated with errors. See the csh warning under "Extended Description". • <i>afterany:jobid [:jobid...]</i> – This job may be scheduled for execution after jobs jobid have terminated, with or without errors. • <i>on:count</i> – 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. • <i>before:jobid [:jobid...]</i> – When this job has begun execution, then jobs jobid... may begin. • <i>beforeok:jobid [:jobid...]</i> – If this job terminates execution without errors, then jobs jobid... may begin. See the csh warning under "Extended Description". • <i>beforenotok:jobid [:jobid...]</i> – If this job terminates execution with errors, then jobs jobid... may begin. See the csh warning under "Extended Description". • <i>beforeany:jobid [:jobid...]</i> – When this job terminates execution, jobs jobid... may begin. <p>If any of the before forms are used, the job referenced by jobid must have been submitted with a dependency type of on.</p> <p>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.</p> <p>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.</p>

Attribute	Description
group_list=g_list	<p>Alters the group name under which the job is to run on the execution system.</p> <p>The <code>g_list</code> argument is of the form:</p> <pre>group[@host] [,group[@host], ...]</pre> <p>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.</p>
stagein=file_list stageout=file_list	<p>Alters which files are staged (copied) in before job start or staged out after the job completes execution. The <code>file_list</code> is in the form:</p> <pre>local_file@hostname:remote_file[,...]</pre> <p>The name <code>local_file</code> 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 <code>remote_file</code> is the destination name on the host specified by <code>hostname</code>. The name may be absolute or relative to the user's home directory on the destination host.</p>

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.unix.org/online.html>.

Related Topics

- [qdel](#)
- [qhold](#)
- [qrls](#)
- [qsub](#)

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

```
$ # request a checkpoint for job 3233
$ qchkpt 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>|-b <secs>|-m <message>|-p <purge>|-t  
<array_range>|-W <delay>}]  
<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

Option	Name	Description
-a	asynchronous delete	Performs an asynchronous delete. The server responds to the user before contacting the MOM. The option <code>qdel -a all</code> performs <code>qdel all</code> due to restrictions from being single-threaded.
-b	seconds	Defines the maximum number of seconds <code>qdel</code> will block attempting to contact <code>pbs_server</code> . If <code>pbs_server</code> is down, or for a variety of communication failures, <code>qdel</code> will continually retry connecting to <code>pbs_server</code> for job submission. This value overrides the <code>CLIENTRETRY</code> parameter in <code>torque.cfg</code> . This is a non-portable Torque extension. Portability-minded users can use the <code>PBS_CLIENTRETRY</code> environmental variable. A negative value is interpreted as infinity. The default is 0.
-m	message	Specify a comment to be included in the email. The argument <code>message</code> specifies the comment to send. This option is only available to a batch operator or the batch administrator.
-p	purge	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.
-t	array_range	The <code>array_range</code> argument is an integer id or a range of integers. Multiple ids or id ranges can be combined in a comma delimited list (examples: <code>-t 1-100</code> or <code>-t 1,10,50-100</code>). The command acts on the array (or specified range of the array) just as it would on an individual job. <div> When deleting a range of jobs, you must include the subscript notation after the job ID (for example, "<code>qdel -t 1-3 98432[]</code>").</div>
-W	delay	Specifies the wait delay between the sending of the <code>SIGTERM</code> and <code>SIGKILL</code> signals. The argument is the length of time in seconds of the delay.

Operands

The `qdel` command accepts one or more `job_identifier` operands of the form:

```
sequence_number[.server_name][@server]
```

or

```
all
```

Examples

```
# Delete a job array
$ qdel 1234[]

# Delete one job from an array
$ qdel 1234[1]

# Delete all jobs, including job arrays
$ qdel all

# Delete selected jobs from an array
$ qdel -t 2-4,6,8-10 64[]
```



There is not an option that allows you to delete all job arrays without deleting jobs.

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\(1B\)](#)

[qsig\(1B\)](#)

Non-Adaptive Computing topics

- [pbs_deljob\(3B\)](#)

qgpumode



This command is deprecated, use the `nvidia-smi` utility instead. See <https://developer.nvidia.com/nvidia-system-management-interface> and http://developer.download.nvidia.com/compute/cuda/6_0/rel/gdk/nvidia-smi.331.38.pdf for more information.

(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`.

i For additional information about options for configuring GPUs, see NVIDIA GPUs in the *Moab Workload Manager Administrator Guide*.

Options

Option	Description
-g	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.
-H	Specifies the host where the GPU is located.
-m	<p>Specifies the new mode for the GPU:</p> <ul style="list-style-type: none">• 0 (Default/Shared): Default/shared compute mode. Multiple threads can use <code>cudaSetDevice()</code> with this device.• 1 (Exclusive Thread): Compute-exclusive-thread mode. Only one thread in one process is able to use <code>cudaSetDevice()</code> with this device.• 2 (Prohibited): Compute-prohibited mode. No threads can use <code>cudaSetDevice()</code> with this device.• 3 (Exclusive Process): Compute-exclusive-process mode. Many threads in one process are able to use <code>cudaSetDevice()</code> with this device. <div><p>i Support for Exclusive Thread was discontinued with CUDA 8, in favor of Exclusive Process.</p></div> <div><pre>qgpumode -H node01 -g 0 -m 1</pre><p><i>This puts the first GPU on node01 into mode 1 (exclusive)</i></p><div><pre>qgpumode -H node01 -g 0 -m 0</pre><p><i>This puts the first GPU on node01 into mode 0 (shared)</i></p></div></div>

Related Topics

["qgpureset" on the facing page](#)

qgpureset

(reset GPU)

Synopsis

```
qgpureset -H host -g gpuid -p -v
```

Description

The *qgpureset* command resets the GPU.

Options

Option	Description
-g	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.
-H	Specifies the host where the GPU is located.
-p	Specifies to reset the GPU's permanent ECC error count.
-v	Specifies to reset the GPU's volatile ECC error count.

Related Topics

["qgpumode" on page 251](#)

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

Option	Name	Description
-h	hold_list	The hold_list argument is a string consisting of one or more of the letters "u", "o", or "s" in any combination. The hold type associated with each letter is: <ul style="list-style-type: none">• <i>u</i> - USER• <i>o</i> - OTHER• <i>s</i> - 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 *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 299).

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.

i 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 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 backslash "\".

Comments begin with the "#" character and continue to end of the line. Comments and blank lines are ignored by *qmgr*.

Options

Option	Name	Description
-a	---	Abort <i>qmgr</i> on any syntax errors or any requests rejected by a server.
-c	command	Execute a single command and exit <i>qmgr</i> .
-e	---	Echo all commands to standard output.
-n	---	No commands are executed, syntax checking only is performed.
-z	---	No errors are written to standard error.

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:

Command	Description
active	Sets the active objects. If the active objects are specified, and the name is not given in a <i>qmgr</i> cmd the active object names will be used.
create	Is to create a new object, applies to queues and nodes.
delete	Is to destroy an existing object, applies to queues and nodes.
set	Is to define or alter attribute values of the object.
unset	Is to clear the value of attributes of the object. <div> This form does not accept an OP and value, only the attribute name.</div>
list	Is to list the current attributes and associated values of the object.

Command	Description
print	Is to print all the queue and server attributes in a format that will be usable as input to the <i>qmgr</i> 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: <ul style="list-style-type: none"> • "=" – 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 *jobIds* which specify the jobs to be moved to the new destination. The *qmove* command accepts one or more *jobId* operands 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 275](#)

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.

 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 275](#)

["qmove" on page 259](#)

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

Option	Description
-f	Force a rerun on a job

```
qrerun -f 15406
```

i The *qrerun all* command is meant to be run if all of the compute nodes go down. If the machines have actually crashed, then we know that all of the jobs need to be restarted. The behavior if you don't run this would depend on how you bring up the *pbs_mom* daemons, but by default would be to cancel all of the jobs.

Running the command makes it so that all jobs are requeued without attempting to contact the moms on which they should be running.

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.

i If you run *qrls* on an array sub-job, *pbs_server* will correct the slot limit holds for the array to which it belongs.

Options

Command	Name	Description
-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: <ul style="list-style-type: none">• <i>u</i> – USER• <i>o</i> – OTHER• <i>s</i> – 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 [{-a | -H <HOST>}] <JOBID>[ <JOBID>] ...
```

Overview

The *qrun* command runs a job.

Format

-a	
Format	---
Default	---
Description	Run the job(s) asynchronously.
Example	<pre>qrun -a 15406</pre>

-H	
Format	<STRING> Host Identifier
Default	---
Description	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 "worst possible" host on which to execute the job.
Example	<pre>qrun -H hostname 15406</pre>

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

Option	Name	Description
-a	asynchronously	Makes the command run asynchronously.
-s	signal	<p>Declares which signal is sent to the job.</p> <p>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 <i>qsig</i>. If it doesn't recognize the signal name, try issuing the signal number instead.</p> <p>Two special signal names, "suspend" and "resume", are used to suspend and resume jobs. Cray systems use the Cray-specific suspend()/resume() calls.</p> <p>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.</p> <p>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 "S" state. Manager or operator privilege is required to suspend or resume a job.</p> <div> Interactive jobs may not resume properly because the top-level shell will background the suspended child process.</div>

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] [-C] [-f [-1]] [-W site_specific] [job_identifier... |
destination...] [time]
qstat [-a|-i|-r|-e] [-c] [-n [-1]] [-s] [-G|-M] [-R] [-u user_list]
[job_identifier... | destination...]
qstat -Q [-f [-1]] [-c] [-W site_specific] [destination...]
qstat -q [-c] [-G|-M] [destination...]
qstat -B [-c] [-f [-1]] [-W site_specific] [server_name...]
qstat -t [-c] [-C]
```

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.

Options

Option	Description
-a	All jobs are displayed in the alternative format (see "Standard output" on page 270). 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.
-B	Specifies that the request is for batch server status and that the operands are the names of servers.
-c	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.
-C	Specifies that Torque will provide only a condensed output (job name, resources used, queue, state, and job owner) for jobs that have not changed recently. See job_full_report_time on page 309 . Jobs that have recently changed will continue to send a full output.
-e	If the operand is a job id or not specified, only jobs in executable queues are displayed. Setting the <code>PBS_QSTAT_EXEONLY</code> environment variable will also enable this option.
-f	Specifies that a full status display be written to standard out. The <code>[time]</code> value is the amount of walltime, in seconds, remaining for the job. <code>[time]</code> does not account for wall-time multipliers.
-G	Show size information in giga-bytes.

Option	Description
-i	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.
-1	In combination with -n , the -1 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 <i>qstat</i> output.
-M	Show size information, disk or memory in mega-words. A word is considered to be 8 bytes.
-n	In addition to the basic information, nodes allocated to a job are listed.
-q	Specifies that the request is for queue status which should be shown in the alternative format.
-Q	Specifies that the request is for queue status and that the operands are destination identifiers.
-r	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. Note that if there is no walltime given for a job, then elapsed time does not display.
-R	In addition to other information, disk reservation information is shown. Not applicable to all systems.
-s	In addition to the basic information, any comment provided by the batch administrator or scheduler is shown.
-t	<p>Normal <i>qstat</i> output displays a summary of the array instead of the entire array, job for job. <i>qstat -t</i> expands the output to display the entire array. Note that arrays are now named with brackets following the array name; for example:</p> <pre>dbeer@napali:~/dev/torque/array_changes\$ echo sleep 20 qsub -t 0-299 189[].napali</pre> <p>Individual jobs in the array are now also noted using square brackets instead of dashes; for example, here is part of the output of <i>qstat -t</i> for the preceding array:</p> <pre>189[299].napali STDIN[299] dbeer 0 Q batch</pre>

Option	Description
-u	<p>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:</p> <pre>user_name[@host] [, user_name[@host], ...]</pre> <p>Host names may be wild carded on the left end, e.g. "*.nasa.gov". User_name without a "@host" is equivalent to "user_name@*", that is at any host.</p>

Operands

If neither the **-q** nor the **-B** option is given, the operands on the qstat command must be either job identifiers or destinations 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:

Item	Description
C	Job is completed after having run.
E	Job is exiting after having run.
H	Job is held.
Q	Job is queued, eligible to run or routed.
R	Job is running.
S	(Unicos only) Job is suspended.
T	Job is being moved to new location.
W	Job is waiting for its execution time (-a option) to be reached.

- 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 274](#) 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 (not the number of nodes in use)
- 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 job's current state
- the amount of CPU time or wall time used by the job (hh:mm)

When any of the above options or the **-r** option is used to request an alternative display format, a column with the requested memory for the job is displayed. If more than one type of memory is requested for the job, either through server or queue parameters or command line, only one value can be displayed. The value displayed depends on the order the memory types are evaluated with the last type evaluated being the value displayed. The order of evaluation is dmem, mem, pmem, pvmem, vmem.

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 job's 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_statque(3)`, and `pbs_statserver(3)`. Each attribute in the second element list whose elements correspond with the `attrl` 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_statque\(3B\)](#)
- [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

```
qsub [-a date_time] [-A account_string] [-b secs]
[-c checkpoint_options] [-C directive_prefix] [-d path] [-D path] [-e
path] [-f] [-F] [-h]
[-I] [-j join] [-k keep] [-K kill_delay] [-l resource_list] [-L NUMA
resource_list]
[-m mail_options] [-M user_list] [-n node_exclusive] [-N name] [-o path]
[-p priority] [-P user[:group]] [-q destination] [-r] [-S path_to_shell
(s)]
[-t array_request] [-v variable_list] [-V] [-w path] [-W additional_
```

```
attributes] [-x] [-X]  
[-z] [script]
```

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 292). 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:



Variable	Description
PBS_ARRAYID	Each member of a job array is assigned a unique identifier (see -t option).
PBS_ENVIRONMENT	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 qsub option).
PBS_GPUFILE	The name of the file containing the list of assigned GPUs. For more information about how to set up Torque with GPUS, see Accelerators in the Moab Workload Manager <i>Administrator Guide</i> .
PBS_JOBID	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-p/\$PBS_JOBID.output).
PBS_JOBNAME	The job name supplied by the user.
PBS_NODEFILE	The name of the file contains the list of nodes assigned to the job (for parallel and cluster systems).
PBS_O_HOST	The name of the host upon which the <i>qsub</i> command is running.

Variable	Description
PBS_O_QUEUE	The name of the original queue to which the job was submitted.
PBS_O_WORKDIR	The absolute path of the current working directory of the <i>qsub</i> command.
PBS_QUEUE	The name of the queue from which the job is executed.
PBS_SERVER	The hostname of the pbs_server which <i>qsub</i> submits the job to.

Options




Option	Name	Description
-a	date_time	<p>Declares the time after which the job is eligible for execution.</p> <p>The date_time argument is in the form:</p> <pre>[[[CC] YY] MM] DD] hhmm [. SS]</pre> <p>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.</p> <p>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.</p> <p>For example, if you submit a job at 11:15 am with a time of <code>-a 1110</code>, the job will be eligible to run at 11:10 am tomorrow.</p>
-A	account_string	<p>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 External Reference Specification (included in the Torque download tarball in docs/v2_2_ers.pdf).</p>
-b	seconds	<p>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, <i>qsub</i> will continually retry connecting to pbs_server for job submission.</p> <p>This value overrides the CLIENTRETRY parameter in <code>torque.cfg</code>. 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>


Option	Name	Description
-c	checkpoint_options	<p>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.</p> <p>Valid checkpoint options are:</p> <ul style="list-style-type: none"> • <i>none</i> – No checkpointing is to be performed. • <i>enabled</i> – Specify that checkpointing is allowed but must be explicitly invoked by either the qhold or qchkpt commands. • <i>shutdown</i> – Specify that checkpointing is to be done on a job at pbs_mom shutdown. • <i>periodic</i> – 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 • <i>interval=minutes</i> – 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. • <i>depth=number</i> – Specify a number (depth) of checkpoint images to be kept in the checkpoint directory. • <i>dir=path</i> – Specify a checkpoint directory (default is /var/spool/torque/checkpoint).
-C	directive_prefix	<p>Defines the prefix that declares a directive to the <i>qsub</i> command within the script file. (See the paragraph on script directives under "Extended description" on page 292.)</p> <p>If the -C option is presented with a directive_prefix argument that is the null string, <i>qsub</i> will not scan the script file for directives.</p>
-d	path	<p>Defines the working directory path to be used for the job. If the -d option is not specified, the default working directory is the home directory. This option sets the environment variable PBS_O_INITDIR.</p>
-D	path	<p>Defines the root directory to be used for the job. This option sets the environment variable PBS_O_ROOTDIR.</p>


Option	Name	Description
-e	path	<p>Defines the path to be used for the standard error stream of the batch job. The path argument is of the form:</p> <pre>[hostname:]path_name</pre> <p>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.</p> <div>  When specifying a directory for the location you need to include a trailing slash. </div> <p>The argument will be interpreted as follows:</p> <ul style="list-style-type: none"> • <i>path_name</i> – where <code>path_name</code> is not an absolute path name, then the <code>qsub</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 <code>hostname</code> component. • <i>hostname:path_name</i> – where <code>path_name</code> is not an absolute path name, then the <code>qsub</code> 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 <code>hostname</code> system. • <i>path_name</i> – where <code>path_name</code> specifies an absolute path name, then the <code>qsub</code> will supply the name of the host on which it is executing for the <code>hostname</code>. • <i>hostname:path_name</i> – where <code>path_name</code> specifies an absolute path name, the path will be used as specified. <p>If the <code>-e</code> option is not specified, the default file name for the standard error stream will be used. The default name has the following form:</p> <ul style="list-style-type: none"> • <i>job_name.esquence_number</i> – where <code>job_name</code> is the name of the job (see the qsub name option) and <code>sequence_number</code> is the job number assigned when the job is submitted.
-f	---	<p>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.</p> <div>  If Torque is compiled with <code>PBS_NO_POSIX_VIOLATION</code> (there is no config option for this), you have to use <code>-W fault_tolerant=true</code> to mark the job as fault tolerant. </div>

Option	Name	Description
-F	---	<p>Specifies the arguments that will be passed to the job script when the script is launched. The accepted syntax is:</p> <pre>qsub -F "myarg1 myarg2 myarg3=myarg3value" myscript2.sh</pre> <div>  Quotation marks are required. <i>qsub</i> will fail with an error message if the argument following <i>-F</i> is not a quoted value. The <i>pbs_mom</i> server will pass the quoted value as arguments to the job script when it launches the script. </div>
-h	---	Specifies that a user hold be applied to the job at submission time.
-I	---	<p>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 <i>qsub</i> to the terminal session in which <i>qsub</i> is running. Interactive jobs are forced to not rerunable. See "Extended description" on page 292 for additional information of interactive jobs.</p>
-j	join	<p>Declares if the standard error stream of the job will be merged with the standard output stream of the job.</p> <p>An option argument value of <i>oe</i> directs that the two streams will be merged, intermixed, as standard output. An option argument value of <i>eo</i> directs that the two streams will be merged, intermixed, as standard error.</p> <p>If the join argument is <i>n</i> or the option is not specified, the two streams will be two separate files.</p> <div>  If using either the <i>-e</i> or the <i>-o</i> option and the <i>-j eo oe</i> option, the <i>-j</i> option takes precedence and all standard error and output messages go to the chosen output file. </div>

Option	Name	Description
-k	keep	<p>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.</p> <p>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".</p> <ul style="list-style-type: none"> • <i>e</i> – 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: <code>job_name.e_{sequence}</code> where <code>job_name</code> is the name specified for the job, and <code>sequence</code> is the sequence number component of the job identifier. • <i>o</i> – 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: <code>job_name.o_{sequence}</code> where <code>job_name</code> is the name specified for the job, and <code>sequence</code> is the sequence number component of the job identifier. • <i>eo</i> – Both the standard output and standard error streams will be retained. • <i>oe</i> – Both the standard output and standard error streams will be retained. • <i>n</i> – Neither stream is retained.
-K	kill_delay	<p>When set on a job, overrides server and queue kill_delay settings. The kill_delay value is a positive integer. The default is 0. See kill_delay on page 312 for more information.</p>

Option	Name	Description
-l	resource_ list	<p>Defines the resources that are required by the job and establishes a limit to the amount of resource that can be consumed. See Requesting Resources on page 65 for more information.</p> <p>If not set for a generally available resource, such as CPU time, the limit is infinite. The resource_list argument is of the form:</p> <pre>resource_name [= [value]] [, resource_name [= [value]] , ...]</pre> <div>  In this situation, you should request the more inclusive resource first. For example, a request for procs should come before a gres request. </div> <p>In Torque 3.0.2 or later, <i>qsub</i> supports the mapping of <code>-l gpus=X</code> to <code>-l gres=gpus:X</code>. This allows users who are using NUMA systems to make requests such as <code>-l ncpus=20:gpus=5</code> 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.</p> <div>  <code>-l</code> supports some Moab-only extensions. See Requesting Resources on page 65 for more information on native Torque resources. <code>qsub -W x=</code> is recommended instead (supports more options). See -W for more information. </div> <p>For information on specifying multiple types of resources for allocation, see Multi-Req Support in the <i>Moab Workload Manager Administrator Guide</i>.</p>
-L	req_ information	<div>  Available with Torque 6.0 and later. This uses a different syntax than the <code>-l resource_list</code> option. </div> <p>Defines the NUMA-aware resource requests for NUMA hardware. This option will work with non-NUMA hardware.</p> <p>See -L NUMA Resource Request on page 173 for the syntax and valid values.</p>
-m	mail_ options	<p>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 "p", or one or more of the characters "a", "b", "e", and "f".</p> <p>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.</p> <p>If the character "p" is specified, mail will never be sent for the job.</p> <p>For the characters "a", "b", "e", and "f":</p> <ul style="list-style-type: none"> <i>a</i> – Mail is sent when the job is aborted by the batch system. <i>b</i> – Mail is sent when the job begins execution. <i>e</i> – Mail is sent when the job terminates. <i>f</i> – Mail is sent when the job terminates with a non-zero exit code. <p>If the <code>-m</code> option is not specified, mail will be sent if the job is aborted.</p>

Option	Name	Description
-M	user_list	<p>Declares the list of users to whom mail is sent by the execution server when it sends mail about the job.</p> <p>The user_list argument is of the form:</p> <pre>user[@host] [,user[@host],...]</pre> <p>If unset, the list defaults to the submitting user at the <i>qsub</i> host, i.e. the job owner.</p>
-n	node_exclusive	<p>Allows a user to specify an exclusive-node access/allocation request for the job. This will set <code>node_exclusive = True</code> in the output of <code>qstat -f <job ID></code>.</p> <div> <p> For Moab, the following options are equivalent to "-n":</p> <pre>> qsub -l naccesspolicy=singlejob jobscript.sh # OR > qsub -W x=naccesspolicy:singlejob jobscript.sh</pre> </div> <p>By default, this only applies for cpusets, and only for compatible schedulers (see Linux Cpuset Support on page 111).</p> <p>For systems that use Moab <i>and</i> have cgroups enabled, the recommended manner for assigning all cores is to use NUMA syntax: <code>"-L tasks=<count>:lprocs=all:place=node"</code>.</p> <p>With cgroups, the ("-l") syntax (lowercase L) will, by default, restrict to the number of cores requested, or to the <code>resources_default.procs</code> value (i.e., 1 core, typically). In order to override this behavior and have Moab assign all the cores on a node while using <code>"-l...singlejob"</code> and/or <code>"-n"</code> (in other words, without <code>"-L ...lprocs=all..."</code>), you must also set <code>RMCFG[<torque>] FLAGS=MigrateAllJobAttributes</code> in <code>moab.cfg</code>.</p>
-N	name	<p>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.</p> <p>If the <code>-N</code> 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>

Option	Name	Description
-o	path	<p>Defines the path to be used for the standard output stream of the batch job. The path argument is of the form:</p> <pre>[hostname:]path_name</pre> <p>where <i>hostname</i> is the name of a host to which the file will be returned, and <i>path_name</i> is the path name on that host in the syntax recognized by POSIX.</p> <div>  When specifying a directory for the location you need to include a trailing slash. </div> <p>The argument will be interpreted as follows:</p> <ul style="list-style-type: none"> • <i>path_name</i> – where <i>path_name</i> is not an absolute path name, then the <i>qsub</i> 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 <i>hostname</i> component. • <i>hostname:path_name</i> – where <i>path_name</i> is not an absolute path name, then the <i>qsub</i> 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 <i>hostname</i> system. • <i>path_name</i> – where <i>path_name</i> specifies an absolute path name, then the <i>qsub</i> will supply the name of the host on which it is executing for the <i>hostname</i>. • <i>hostname:path_name</i> where <i>path_name</i> specifies an absolute path name, the path will be used as specified. <p>If the <i>-o</i> option is not specified, the default file name for the standard output stream will be used. The default name has the following form:</p> <ul style="list-style-type: none"> • <i>job_name.osequence_number</i> – where <i>job_name</i> is the name of the job (see the qsub name option) and <i>sequence_number</i> is the job number assigned when the job is submitted.
-p	priority	<p>Defines the priority of the job. The priority argument must be a integer between -1024 and +1023 inclusive. The default is no priority which is equivalent to a priority of zero.</p>
-P	user [:group]	<p>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 <code>[:group]</code> option; it is available in Torque 2.4.8 and later.</p>

Option	Name	Description
-q	destination	<p>Defines the destination of the job. The destination names a queue, a server, or a queue at a server.</p> <p>The <i>qsub</i> 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.</p> <p>If the <i>-q</i> option is not specified, the <i>qsub</i> command will submit the script to the default server. (See "Environment variables" on page 292 and the PBS ERS section 2.7.4, "Default Server".)</p> <p>If the <i>-q</i> option is specified, it is in one of the following three forms:</p> <ul style="list-style-type: none"> • queue • @server • queue@server <p>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.</p> <p>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.</p> <p>If the destination argument names both a queue and a server, the job will be submitted to the named queue at the named server.</p>
-r	y/n	<p>Declares whether the job is rerunnable (see the qrerun command). The option argument is a single character, either y or n.</p> <p>If the argument is "y", the job is rerunnable. If the argument is "n", the job is not rerunnable. The default value is y, rerunnable.</p>
-S	path_list	<p>Declares the path to the desired shell for this job.</p> <pre>qsub script.sh -S /bin/tcsh</pre> <p>If the shell path is different on different compute nodes, use the following syntax:</p> <pre>path[@host] [,path[@host],...]</pre> <pre>qsub script.sh -S /bin/tcsh@node1,/usr/bin/tcsh@node2</pre> <p>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.</p> <p>If the <i>-S</i> option is not specified, the option argument is the null string, or no entry from the <i>path_list</i> is selected, the execution will use the user's login shell on the execution host.</p>

Option	Name	Description
-t	array_request	<p>Specifies the task ids of a job array. Single task arrays are allowed.</p> <p>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: <code>-t 1-100</code> or <code>-t 1,10,50-100</code></p> <p>An optional <i>slot limit</i> 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 (%).</p> <pre>qsub script.sh -t 0-299%5</pre> <p>This sets the slot limit to 5. Only 5 jobs from this array can run at the same time.</p> <p>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.</p>
-v	variable_list	<p>Expands the list of environment variables that are exported to the job.</p> <p>In addition to the variables described in the "Description" section above, variable_list names environment variables from the <i>qsub</i> 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.</p>
-V	---	<p>Declares that all environment variables in the <i>qsub</i> commands environment are to be exported to the batch job.</p>
-w	path	<p>Defines the working directory path to be used for the job. If the <code>-w</code> option is not specified, the default working directory is the current directory. This option sets the environment variable <code>PBS_O_WORKDIR</code>.</p>

Option	Name	Description
-W	additional_attributes	<div> <p>i Use "-W x=" as pass-through for scheduler-only job extensions. See Resource Manager Extensions in the <i>Moab Workload Manager Administrator Guide</i> for a list of scheduler-only job extensions.</p> <p>For legacy purposes, qsub -l will continue to support some scheduler-only job extensions. However, when in doubt, use "-W x=".</p> </div> <p>The -W option allows for the specification of additional job attributes. The general syntax of -W is in the form:</p> <pre>-W attr_name=attr_value.</pre> <p>You can use multiple -W options with this syntax:</p> <pre>-W attr_name1=attr_value1 -W attr_name2=attr_value2.</pre> <div> <p>i 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.</p> </div> <p>PBS currently supports the following attributes within the -W option:</p> <ul style="list-style-type: none"> • <i>depend=dependency_list</i> – Defines the dependency between this and other jobs. The dependency_list is in the form: <pre>type[:argument[:argument...]][,type:argument...]</pre> <p>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 <code>seq_number.server.name</code>, 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 289.</p> • <i>group_list=g_list</i> – Defines the group name under which the job is to run on the execution system. The g_list argument is of the form: <pre>group[@host][,group[@host],...]</pre> <p>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.</p> • <i>interactive=true</i> – If the interactive attribute is specified, the job is an interactive job. The qsub option is an alternative method of specifying this attribute. • <i>job_radix=<int></i> – 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 64. • <i>stagein=file_list</i> • <i>stageout=file_list</i> – Specifies which files are staged (copied) in before job start


Option	Name	Description
		<p>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 <code>file_list</code> is in the form:</p> <pre>local_file@hostname:remote_file[,...]</pre> <p>regardless of the direction of the copy. The name <code>local_file</code> 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 <code>remote_file</code> is the destination name on the host specified by <code>hostname</code>. 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 (<code>rcp</code>) call on the execution system in the follow manner:</p> <ul style="list-style-type: none"> For stagein: <code>rcp hostname:remote_file local_file</code> For stageout: <code>rcp local_file hostname:remote_file</code> <p>Data staging examples:</p> <pre>-W stagein=/tmp/input.txt@headnode:/home/user/input.txt -W stageout=/tmp/output.txt@headnode:/home/user/output.txt</pre> <p>If Torque has been compiled with <code>wordexp</code> support, then variables can be used in the specified paths. Currently only <code>\$PBS_JOBID</code>, <code>\$HOME</code>, and <code>\$TMPDIR</code> are supported for stagein.</p> <ul style="list-style-type: none"> <code>umask=XXX</code> – Sets <code>umask</code> used to create stdout and stderr spool files in <code>pbs_mom</code> spool directory. Values starting with 0 are treated as octal values, otherwise the value is treated as a decimal <code>umask</code> value.
-x	---	<p>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 <code>-x</code> option allows the script to be executed in the interactive job and then the job completes. For example:</p> <pre>script.sh #!/bin/bash ls ---end script---</pre> <pre>qsub -I script.sh qsub: waiting for job 5.napali to start dbeer@napali:~# <displays the contents of the directory, because of the ls command> qsub: job 5.napali completed</pre>
-X	---	Enables X11 forwarding. The <code>DISPLAY</code> environment variable must be set.
-z	---	Directs that the <code>qsub</code> command is not to write the job identifier assigned to the job to the commands standard output.

depend=dependency_list valid dependencies

i For job dependencies to work correctly, you must set the ["keep_completed" on page 311](#) server parameter.

i Jobs can depend on single job dependencies and array dependencies at the same time.

Dependency	Description
<code>after:jobid[:jobid...]</code>	This job may be scheduled for execution at any point after jobs jobid have started execution.
<code>afterany:jobid[:jobid...]</code>	This job may be scheduled for execution after jobs jobid have terminated, with or without errors.
<code>afternotok:jobid[:jobid...]</code>	This job may be scheduled for execution only after jobs jobid have terminated with errors. See the csh warning under "Extended description" on page 292 .
<code>afterok:jobid[:jobid...]</code>	This job may be scheduled for execution only after jobs jobid have terminated with no errors. See the csh warning under "Extended description" on page 292 .
<code>before:jobid[:jobid...]</code>	When this job has begun execution, then jobs jobid... may begin.
<code>beforeany:jobid[:jobid...]</code>	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.
<code>beforenotok:jobid[:jobid...]</code>	If this job terminates execution with errors, then jobs jobid... may begin. See the csh warning under "Extended description" on page 292 .
<code>beforeok:jobid[:jobid...]</code>	If this job terminates execution without errors, then jobs jobid... may begin. See the csh warning under "Extended description" on page 292 .

Dependency	Description
<code>on:count</code>	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 "before" forms (see below).
<code>synccount:count</code>	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.
<code>syncwith:jobid</code>	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.
<div>  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 the facing page. </div>	
<code>afteranyarray:arrayid[count]</code>	This job may be scheduled for execution after jobs in arrayid have terminated, with or without errors.
<code>afternotokarray:arrayid[count]</code>	This job may be scheduled for execution only after jobs in arrayid have terminated with errors.
<code>afterokarray:arrayid[count]</code>	This job may be scheduled for execution only after jobs in arrayid have terminated with no errors.
<code>afterstartarray:arrayid[count]</code>	After this many jobs have started from arrayid, this job may start.
<code>beforeanyarray:arrayid[count]</code>	<p>When this job terminates execution, jobs in arrayid may begin.</p> <p>If any of the before forms are used, the jobs referenced by arrayid must have been submitted with a dependency type of on.</p> <p>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.</p>
<code>beforenotokarray:arrayid[count]</code>	If this job terminates execution with errors, then jobs in arrayid may begin.
<code>beforeokarray:arrayid[count]</code>	If this job terminates execution without errors, then jobs in arrayid may begin.

Dependency	Description
beforestartarray:arrayid[count]	Before this many jobs have started from arrayid, this job may start.



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.

Dependency examples

```
qsub -W depend=afterok:123.big.iron.com /tmp/script
```

```
qsub -W depend=before:234.hunk1.com:235.hunk1.com
```

```
/tmp/script
```

```
qsub script.sh -W depend=afterokarray:427[]
```

(This assumes every job in array 427 has to finish successfully for the dependency to be satisfied.)

```
qsub script.sh -W depend=afterokarray:427[][5]
```

(This means that 5 of the jobs in array 427 have to successfully finish in order for the dependency to be satisfied.)

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 **qsub** and **qsub** 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 TORQUE_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. See ["torque.cfg" Configuration File on page 383](#) for more information on these parameters.

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:

```
:
#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.

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:

```
set EXITVAL = $status
```

and the following line as the last executable line in *.logout*:

```
exit $EXITVAL
```

Interactive Jobs:

If the **qsub** option is specified on the command line or in a script directive, or if the "interactive" job attribute declared true via the **qsub** 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.

Once 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:

Sequence	Description
~.	<i>qsub</i> terminates execution. The batch job is also terminated.
~asusp	Suspend the input half of <i>qsub</i> (terminal to job), but allow output to continue to be displayed. Only works under the C shell. "asusp" is the auxiliary suspend character, usually CNTL-Y.
~susp	Suspend the <i>qsub</i> program if running under the C shell. "susp" is the suspend character (usually CNTL-Z).

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_iris5\(7B\)](#)
- [pbs_resources_sp2\(7B\)](#)
- [pbs_resources_sunos4\(7B\)](#)
- [pbs_resources_unicos8\(7B\)](#)
- [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

In order to execute `qterm`, the user must have PBS Operation or Manager privileges.

Options

Option	Name	Description
-t	type	<p>Specifies the type of shut down. The types are:</p> <ul style="list-style-type: none"> • <i>quick</i> – This is the default action if the <code>-t</code> 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. <div>  The immediate and delay types are deprecated. </div>

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_iris5\(7B\)](#)

[pbs_resources_sp2\(7B\)](#)

[pbs_resources_sunos4\(7B\)](#)

[pbs_resources_unicos8\(7B\)](#)

trqauthd

(Torque authorization daemon)

Synopsis

```
trqauthd -d
```

```
trqauthd -D
```

```
trqauthd -F
```

```
trqauthd -n
```


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.

Beginning in Torque 6.0.3, you have the option when starting `trqauthd` to disable `trqauthd` from logging anything. In addition, the `-F` (don't fork) option is available when running under `systemd`.

 If you run `trqauthd` before starting `pbs_server`, you will receive a warning that no servers are available. To avoid this message, start `pbs_server` before running `trqauthd`.

Options

-d — Terminate	
Format	---
Default	---
Description	Terminate <i>trqauthd</i> .
Example	<code>trqauthd -d</code>

-D — Debug	
Format	---
Default	---
Description	Run <i>trqauthd</i> in debug mode.
Example	<code>trqauthd -D</code>

-F — Fork	
Format	---
Default	---
Description	Prevents the system from forking. Useful when running under <code>systemd</code> (Red Hat 7-based or SUSE 12-based systems).

-F — Fork

Example

```
trqauthd -F
```

-n — No Logging

Format

```
---
```

Default

```
---
```

Description

Disables *trqauthd* from logging anything.

Example

```
trqauthd -n
```

Server Parameters

Torque server parameters are specified using the **qmgr** command. The `set` subcommand is used to modify the `server` object. For example:

```
> qmgr -c 'set server default_queue=batch'
```

Parameters

"acl_group_hosts" on the facing page	"email_batch_seconds" on page 307	"log_events" on page 313	"node_submit_exceptions" on page 320
"acl_host_enable" on the facing page	"exit_code_canceled_job" on page 307	"log_file_max_size" on page 314	"node_submit_exceptions" on page 320
"acl_hosts" on the facing page	"ghost_array_recovery" on page 307	"log_file_roll_depth" on page 314	"np_default" on page 320
"acl_logic_or" on page 302	"gres_modifiers" on page 308	"log_keep_days" on page 315	"operators" on page 321
"acl_user_hosts" on page 302	"interactive_jobs_can_roam" on page 308	"log_level" on page 315	"pass_cpuclock" on page 321
"allow_node_submit" on page 302	"job_exclusive_on_use" on page 308	"mail_body_fmt" on page 315	"poll_jobs" on page 321
"allow_proxy_user" on page 302	"job_force_cancel_time" on page 309	"mail_domain" on page 315	"query_other_jobs" on page 322
"auto_node_np" on page 303	"job_full_report_time" on page 309	"mail_from" on page 316	"record_job_info" on page 322
"automatic_requeue_exit_code" on page 303	"job_log_file_max_size" on page 309	"mail_subject_fmt" on page 316	"record_job_script" on page 322
"cgrouper_per_task" on page 303	"job_log_file_roll_depth" on page 310	"managers" on page 316	"resources_available" on page 322
"checkpoint_defaults" on page 303	"job_log_keep_days" on page 310	"max_job_array_size" on page 317	"scheduling" on page 323
"clone_batch_delay" on page 304	"job_log_nanny" on page 310	"max_slot_limit" on page 317	"submit_hosts" on page 323
"clone_batch_size" on page 304	"job_stat_rate" on page 311	"max_threads" on page 317	"tcp_incoming_timeout" on page 323
"copy_on_rerun" on page 304	"job_stat_rate" on page 311	"max_user_queueable" on page 318	"tcp_timeout" on page 324
"cray_enabled" on page 305	"job_suffix_alias" on page 311	"max_user_run" on page 318	"thread_idle_seconds" on page 324
"default_gpu_mode" on page 305	"job_sync_timeout" on page 311	"min_threads" on page 318	"timeout_for_job_delete" on page 324
"default_queue" on page 305	"keep_completed" on page 311	"moab_array_compatible" on page 318	"timeout_for_job_requeue" on page 324
"disable_automatic_requeue" on page 305	"kill_delay" on page 312	"mom_job_sync" on page 319	"use_jobs_subdirs" on page 325
"disable_server_id_check" on page 306	"legacy_vmem" on page 312	"next_job_number" on page 319	
"display_job_server_suffix" on page 306	"lock_file" on page 312	"node_check_rate" on page 320	
"dont_write_nodes_file" on page 306	"lock_file_check_time" on page 313	"node_check_rate" on page 320	
"down_on_error" on page 307	"lock_file_update_time" on page 313	"node_pack" on page 320	

acl_group_hosts

Format	group@host[group@host]...
Default	---
Description	Users who are members of the specified groups will be able to submit jobs from these otherwise untrusted hosts. Users who aren't members of the specified groups will not be able to submit jobs unless they are specified in acl_user_hosts.

acl_host_enable

Format	<BOOLEAN>
Default	FALSE
Description	When set to TRUE, hosts not in the pbs_server nodes file must be added to the "acl_hosts" below list in order to get access to pbs_server.

acl_hosts

Format	<HOST>[,<HOST>]... or <HOST>[range] or <HOST*> where the asterisk (*) can appear anywhere in the host name
Default	Not set.
Description	<p>Specifies a list of hosts that can have access to pbs_server when acl_host_enable is set to TRUE. This does not enable a node to submit jobs. To enable a node to submit jobs use submit_hosts.</p> <div><p>i Hosts that are in the TORQUE_HOME/server_priv/nodes file do not need to be added to this list.</p><pre>Qmgr: set queue batch acl_hosts="hostA,hostB" Qmgr: set queue batch acl_hosts+=hostC Qmgr: set server acl_hosts="hostA,hostB" Qmgr: set server acl_hosts+=hostC</pre></div> <div><p>i 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.</p><pre>Qmgr: set server acl_hosts = "galaxy*.tom.org" Qmgr: set server acl_hosts += "galaxy[0-50].tom.org"</pre></div>

acl_logic_or

Format	<BOOLEAN>
Default	FALSE
Description	When set to <code>TRUE</code> , the user and group queue ACLs are logically OR'd. When set to <code>FALSE</code> , they are AND'd.

acl_user_hosts

Format	group@host[group@host]...
Default	---
Description	The specified users are allowed to submit jobs from otherwise untrusted hosts. By setting this parameter, other users at these hosts will not be allowed to submit jobs unless they are members of specified groups in <code>acl_group_hosts</code> .

allow_node_submit

Format	<BOOLEAN>
Default	FALSE
Description	When set to <code>TRUE</code> , allows all hosts in the <code>TORQUE_HOME/server_priv/nodes</code> file (MOM nodes) to submit jobs to <code>pbs_server</code> . To only allow <code>qsub</code> from a subset of all MOMs, use " Server Parameters " on page 299.

allow_proxy_user

Format	<BOOLEAN>
Default	FALSE
Description	When set to <code>TRUE</code> , specifies that users can proxy from one user to another. Proxy requests will be either validated by <code>ruserok()</code> or by the scheduler (see " Job Submission " on page 61).

auto_node_np

Format <BOOLEAN>

Default DISABLED

Description 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.

automatic_requeue_exit_code

Format <LONG>

Default ---

Description 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.

cgroup_per_task

Format <BOOLEAN>

Default FALSE

Description When set to `FALSE`, jobs submitted with the `-L` syntax will have *one* cgroup created per host unless they specify otherwise at submission time. This behavior is similar to the pre-6.0 cpuset implementation.

When set to `TRUE`, jobs submitted with the `-L` syntax will have *one* cgroup created per task unless they specify otherwise at submission time.



Some MPI implementations are not compatible with using one cgroup per task.

See "[-L NUMA Resource Request](#)" on page 173 for more information.

checkpoint_defaults

Format <STRING>

Default ---

checkpoint_defaults

Description 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.

```
set queue batch checkpoint_defaults="enabled, periodic, interval=5"
```

clone_batch_delay

Format <INTEGER>

Default 1

Description Specifies the delay (in seconds) between clone batches (see [clone_batch_size](#)).

clone_batch_size

Format <INTEGER>

Default 256

Description Job arrays are created in batches of size *X*. *X* jobs are created, and after the [clone_batch_delay](#), *X* more are created. This repeats until all are created.


copy_on_rerun

Format <BOOLEAN>

Default FALSE

Description When set to `TRUE`, Moab HPC Suite will copy the output and error files over to the user-specified directory when the `grerun` command is executed (i.e. a job preemption). Output and error files are only created when a job is in running state before the preemption occurs.

 *pbs_server* and *pbs_mom* need to be on the same version.

 When you change the value, you must perform a `pbs_server` restart for the change to effect.

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 <i>Moab Workload Manager Administrator Guide</i> .

default_gpu_mode	
Format	<STRING>
Default	exclusive_thread
Description	<p>Determines what GPU mode will be used for jobs that request GPUs but do not request a GPU mode. Valid entries are <code>exclusive_thread</code>, <code>exclusive</code>, <code>exclusive_process</code>, <code>default</code>, and <code>shared</code>.</p> <div> <p>i If you are using CUDA 8 or newer, the default of <code>exclusive_thread</code> is no longer supported. If the server specifies an <code>exclusive_thread</code> setting, the MOM will substitute an <code>exclusive_process</code> mode setting. We recommend that you set the default to <code>exclusive_process</code>.</p> <p>i If you upgrade your CUDA library, you must rebuild Torque.</p> </div>

default_queue	
Format	<STRING>
Default	---
Description	Indicates the queue to assign to a job if no queue is explicitly specified by the submitter.

disable_automatic_requeue	
Format	<BOOLEAN>
Default	FALSE

disable_automatic_requeue

Description	Normally, if a job cannot start due to a transient error, the MOM returns a special exit code to the server so that the job is requeued instead of completed. When this parameter is set, the special exit code is ignored and the job is completed.
--------------------	--

disable_server_id_check

Format	<BOOLEAN>
---------------	-----------

Default	FALSE
----------------	-------

Description	When set to <code>TRUE</code> , 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.
--------------------	---



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 383).

display_job_server_suffix

Format	<BOOLEAN>
---------------	-----------

Default	TRUE
----------------	------

Description	When set to <code>TRUE</code> , Torque will display both the job ID and the host name. When set to <code>FALSE</code> , only the job ID will be displayed.
--------------------	--



If set to `FALSE`, the environment variable `NO_SERVER_SUFFIX` must be set to `TRUE` for `pbs_track` to work as expected.



`display_job_server_suffix` should not be set unless the server has no queued jobs. If it is set while the server has queued jobs, it will cause problems correctly identifying job ids with all existing jobs.

dont_write_nodes_file

Format	<BOOLEAN>
---------------	-----------

Default	FALSE
----------------	-------

dont_write_nodes_file

Description	When set to <code>TRUE</code> , the nodes file cannot be overwritten for any reason; qmgr commands to edit nodes will be rejected.
--------------------	--

down_on_error

Format	<BOOLEAN>
Default	<code>TRUE</code>
Description	When set to <code>TRUE</code> , pbs_server will mark nodes that report an error from their node health check as down and unavailable to run jobs. See Health check on page 227 for more information.

email_batch_seconds

Format	<INTEGER>
Default	0
Description	If set to a number greater than 0, emails will be sent in a batch every specified number of seconds, per addressee. For example, if this is set to 300, then each user will only receive emails every 5 minutes in the most frequent scenario. The addressee would then receive one email that contains all of the information which would've been sent out individually before. If it is unset or set to 0, then emails will be sent for every email event.

exit_code_canceled_job

Format	<INTEGER>
Default	---
Description	When set, the exit code provided by the user is given to any job that is canceled, regardless of the job's state at the time of cancellation.

ghost_array_recovery

Format	<BOOLEAN>
---------------	-----------


ghost_array_recovery

Default	TRUE
Description	When TRUE , array subjobs will be recovered regardless of whether the .AR file was correctly recovered. This prevents the loss of running and queued jobs. However, it may <i>no longer</i> enforce a per-job slot limit or handle array dependencies correctly, as some historical information will be lost. When FALSE , array subjobs will not be recovered if the .AR file is invalid or non-existent.

gres_modifiers

Format	<user>[<user>...]
Default	---
Description	List of users granted permission to modify the gres resource of their own running jobs. Note that users do not need special permission to modify the gres resource of their own queued jobs.

interactive_jobs_can_roam

Format	<BOOLEAN>
Default	FALSE
Description	<p>By default, interactive jobs run from the login node that they submitted from. When TRUE, interactive jobs may run on login nodes other than the one where the jobs were submitted from. See Installation Notes for Moab and Torque for Cray in the <i>Moab Workload Manager Administrator Guide</i>.</p> <div> With <code>interactive_jobs_can_roam</code> enabled, jobs will only go to nodes with the <code>alps_login</code> property set in the nodes file.</div>

job_exclusive_on_use

Format	<BOOLEAN>
Default	FALSE
Description	When <code>job_exclusive_on_use</code> is set to TRUE , pbsnodes will show job-exclusive on a node when there's at least one of its processors running a job. This differs with the default behavior which is to show job-exclusive on a node when all of its processors are running a job.

job_exclusive_on_use

Example

```
set server job_exclusive_on_use=TRUE
```

job_force_cancel_time

Format

<INTEGER>

Default

Disabled

Description

If a job has been deleted and is still in the system after *x* seconds, the job will be purged from the system. This is mostly useful when a job is running on a large number of nodes and one node goes down. The job cannot be deleted because the MOM cannot be contacted. The *qdel* fails and none of the other nodes can be reused. This parameter can be used to remedy such situations.

job_full_report_time

Format

<INTEGER>

Default

300 seconds

Description

Sets the time in seconds that a job should be fully reported after any kind of change to the job, even if condensed output was requested.

job_log_file_max_size

Format

<INTEGER>

Default

Description

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 <filename> to <filename.1> and a new empty log is opened. If the current day file size exceeds the maximum size a second time, the <filename.1> log file is rolled to <filename.2>, the current log is rolled to <filename.1>, 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).

job_log_file_roll_depth

Format	<INTEGER>
Default	---
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>.

job_log_keep_days

Format	<INTEGER>
Default	---
Description	This maintains logs for the number of days designated. If set to 4, any log file older than 4 days old is deleted.


job_nanny

Format	<BOOLEAN>
Default	FALSE
Description	When set to <code>TRUE</code> , enables the experimental "job deletion nanny" feature. All job cancels will create a repeating task that will resend KILL signals if the initial job cancel failed. Further job cancels will be rejected with the message "job cancel in progress." This is useful for temporary failures with a job's execution node during a job delete request.

job_start_timeout

Format	<INTEGER>
Default	---
Description	Specifies the <i>pbs_server</i> to <i>pbs_mom</i> TCP socket timeout in seconds that is used when the <i>pbs_server</i> sends a job start to the <i>pbs_mom</i> . It is useful when the MOM has extra overhead involved in starting jobs. If not specified, then the tcp_timeout parameter is used.

job_stat_rate	
Format	<INTEGER>
Default	300 (30 in Torque 1.2.0p5 and earlier)
Description	If the mother superior has not sent an update by the specified time, at the specified time pbs_server requests an update on job status from the mother superior.

job_suffix_alias	
Format	<STRING>
Default	---
Description	<p>Allows the job suffix to be defined by the user.</p> <div>  job_suffix_alias should not be set unless the server has no queued jobs. If it is set while the server has queued jobs, it will cause problems correctly identifying job ids with all existing jobs. </div>
Example	<pre>qmgr -c 'set server job_suffix_alias = biology'</pre> <p>When a job is submitted after this, its jobid will have .biology on the end: 14.napali.biology. If display_job_server_suffix is set to false, it would be named 14.biology.</p>

job_sync_timeout	
Format	<INTEGER>
Default	60
Description	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.

keep_completed	
Format	<INTEGER>
Default	300

keep_completed

Description	The amount of time (in seconds) a job will be kept in the queue after it has entered the completed state. <code>keep_completed</code> <i>must</i> be set for job dependencies to work. For more information, see "Keeping Completed Jobs" on page 84 .
--------------------	---

kill_delay

Format	<INTEGER>
Default	If using <code>qdel</code> , 2 seconds If using <code>qrerun</code> , 0 (no wait)
Description	<p>Specifies the number of seconds between sending a SIGTERM and a SIGKILL to a job 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.</p> <div><p>i All MOMs must be configured with <code>\$exec_with_exec true</code> in order for <code>kill_delay</code> to work, even when relying on default <code>kill_delay</code> settings.</p><p>i If <code>kill_delay</code> is set for a queue, the queue setting overrides the server setting. See <code>kill_delay</code> in Queue Attributes on page 399.</p></div>
Example	<pre>qmgr -c "set server kill_delay=30"</pre>

legacy_vmem

Format	<BOOLEAN>
Default	FALSE
Description	When set to true, the vmem request will be the amount of memory requested for each node of the job. When it is unset or false, vmem will be the amount of memory for the entire job and will be divided accordingly

lock_file

Format	<STRING>
Default	<code>torque/server_priv/server.lock</code>

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 <code>torque/server_priv</code> .
--------------------	--

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.

lock_file_update_time

Format	<INTEGER>
Default	3
Description	Specifies how often (in seconds) the thread will update the lock file.

log_events

Format	Bitmap
Default	---

log_events

Description

By default, the server logs all events. To customize this, perform Boolean OR operations on the binary representation of each of the following bitmaps (or "enablement bits") to put into effect, then convert the end result to decimal and assign it to `log_events`:

```
#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_CLIENTAUTH 0x0200 /* TRQAUTHD login events */
#define PBSEVENT_SYSLOG 0x0400 /* pass this event to the syslog as well (if
defined) */
#define PBSEVENT_FORCE 0x8000 /* set to force a message */
```

For example, if you want to log only internal error, system/server, job-related, and job-usage events, set `log_events` to 27 (1 (0x01) + 2 (0x02) + 8 (0x08) + 16 (0x10)) in *qmgr*:

```
Qmgr: set server log_events = 27
```

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

If `log_file_max_size` is set, controls how deep the current day log files will be rolled before they are deleted.

log_keep_days	
Format	<INTEGER>
Default	0
Description	Specifies how long (in days) a server or MOM log should be kept.

log_level	
Format	<INTEGER>
Default	0
Description	Specifies the pbs_server logging verbosity. Maximum value is 7.

mail_body_fmt	
Format	A printf-like format string
Default	PBS Job Id: %i Job Name: %j Exec host: %h %m %d
Description	<p>Override the default format for the body of outgoing mail messages. A number of printf-like format specifiers and escape sequences can be used:</p> <ul style="list-style-type: none"> \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	
Format	<STRING>

mail_domain	
Default	---
Description	Override the default domain for outgoing mail messages. If set, emails will be addressed to <user->@<hostdomain>. If unset, the job's Job_Owner attribute will be used. If set to never, Torque will never send emails.

mail_from	
Format	<STRING>
Default	adm
Description	Specify the name of the sender when Torque sends emails.

mail_subject_fmt	
Format	A printf-like format string
Default	PBS JOB %i
Description	<p>Override the default format for the subject of outgoing mail messages. A number of printf-like format specifiers and escape sequences can be used:</p> <ul style="list-style-type: none"> \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 %

managers	
Format	<user>@<host.sub.domain>[, <user>@<host.sub.domain>...]

managers	
Default	root@localhost
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	<p>This is the maximum number of jobs that can run concurrently in any job array. Slot limits can be applied at submission time with <i>qsub</i>, or it can be modified with <i>qalter</i>.</p> <pre>qmgr -c 'set server max_slot_limit=10'</pre> <p>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."</p>

max_threads	
Format	<INTEGER>
Default	min_threads * 20
Description	This is the maximum number of threads that should exist in the thread pool at any time. See "Setting min_threads and max_threads" on page 136 for more information.

max_user_queuable

Format	<INTEGER>
Default	Unlimited
Description	When set, <code>max_user_queuable</code> places a system-wide limit on the amount of jobs that an individual user can queue.
Example	<pre>qmgr -c 'set server max_user_queuable=500'</pre>

max_user_run

Format	<INTEGER>
Default	Unlimited
Description	This limits the maximum number of jobs a user can have running for the given server.
Example	<pre>qmgr -c "set server max_user_run=5"</pre>

min_threads

Format	<INTEGER>
Default	(2 * the number of procs listed in <code>/proc/cpuinfo</code>) + 1. If Torque is unable to read <code>/proc/cpuinfo</code> , the default is 10.
Description	This is the minimum number of threads that should exist in the thread pool at any time. See " Setting min_threads and max_threads " on page 136 for more information.

moab_array_compatible

Format	<BOOLEAN>
Default	TRUE
Description	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.

mom_job_sync

Format <BOOLEAN>

Default TRUE


Description 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).

next_job_number

Format <INTEGER>

Default ---

Description 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.

 If you use Moab Workload Manager (and have configured it to synchronize job IDs with Torque), then Moab will generate the job ID and *next_job_number* will have no effect on the job ID. See Resource Manager Configuration in the *Moab Workload Manager Administrator Guide* for more information.

no_mail_force

Format <BOOLEAN>

Default FALSE

Description When set to TRUE, eliminates all e-mails when *mail_options* (see "[qsub](#)" on page 275) 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.

node_check_rate

Format <INTEGER>

Default 600

Description Specifies the minimum duration (in seconds) that a node can fail to send a status update before being marked down by the *pbs_server* daemon.

node_pack

Description This is deprecated.

node_ping_rate

Format <INTEGER>

Default 300

Description Specifies the maximum interval (in seconds) between successive "pings" sent from the *pbs_server* daemon to the *pbs_mom* daemon to determine node/daemon health.

node_submit_exceptions

Format String

Default ---

Description When set in conjunction with *allow_node_submit*, these nodes will not be allowed to submit jobs.

np_default

Format <INTEGER>

Default ---

np_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.



np_default sets a minimum number of np per node. Nodes with less than the np_default get additional execution slots.

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.

pass_cpuclock

Format

<BOOLEAN>

Default

TRUE

Description

If set to TRUE, the *pbs_server* daemon passes the option and its value to the *pbs_mom* daemons for direct implementation by the daemons, making the CPU frequency adjustable as part of a [resource request](#) by a job submission.

If set to FALSE, the *pbs_server* daemon creates and passes a PBS_CPUCLOCK job environment variable to the *pbs_mom* daemons that contains the value of the *cpuclock* attribute used as part of a resource request by a job submission. The CPU frequencies on the MOMs are not adjusted. The environment variable is for use by prologue and epilogue scripts, enabling administrators to log and research when users are making *cpuclock* requests, as well as researchers and developers to perform CPU clock frequency changes using a method outside of that employed by the Torque *pbs_mom* daemons.

poll_jobs

Format

<BOOLEAN>

Default

TRUE (FALSE in Torque 1.2.0p5 and earlier)

poll_jobs

Description	<p>If set to <code>TRUE</code>, <i>pbs_server</i> will poll job info from MOMs over time and will not block on handling requests which require this job information.</p> <p>If set to <code>FALSE</code>, no polling will occur and if requested job information is stale, <i>pbs_server</i> may block while it attempts to update this information. For large systems, this value should be set to <code>TRUE</code>.</p>
--------------------	--


query_other_jobs

Format	<BOOLEAN>
Default	FALSE
Description	When set to <code>TRUE</code> , specifies whether or not non-admin users may view jobs they do not own.

record_job_info

Format	<BOOLEAN>
Default	FALSE
Description	This must be set to <code>TRUE</code> in order for job logging to be enabled.

record_job_script

Format	<BOOLEAN>
Default	FALSE
Description	<p>If set to <code>TRUE</code>, this adds the contents of the script executed by a job to the log.</p> <div> For <code>record_job_script</code> to take effect, "record_job_info" above must be set to <code>TRUE</code>.</div>

resources_available

Format	<STRING>
Default	---

resources_available

Description	Allows overriding of detected resource quantities (see "Assigning Queue Resource Limits" on page 409). <i>pbs_server</i> must be restarted for changes to take effect. Also, <i>resources_available</i> is constrained by the smallest of <i>queue.resources_available</i> and the <i>server.resources_available</i> .
--------------------	---

scheduling

Format	<BOOLEAN>
Default	---
Description	Allows <i>pbs_server</i> to be scheduled. When <i>FALSE</i> , <i>pbs_server</i> is a resource manager that works on its own. When <i>TRUE</i> , Torque allows a scheduler, such as Moab or Maui, to dictate what <i>pbs_server</i> should do.

submit_hosts

Format	<HOSTNAME>[,<HOSTNAME>]...
Default	Not set.
Description	<p>Hosts in this list are able to submit jobs. This applies to any node whether within the cluster or outside of the cluster.</p> <p>If "acl_host_enable" on page 301 is set to <i>TRUE</i> and the host is not in the <code>TORQUE_HOME/server_priv/nodes</code> file, then the host must also be in the "acl_hosts" on page 301 list.</p> <p>To allow <i>qsub</i> from all compute nodes instead of just a subset of nodes, use "allow_node_submit" on page 302.</p>

tcp_incoming_timeout

Format	<INTEGER>
Default	600
Description	<p>Specifies the timeout for incoming TCP connections to <i>pbs_server</i>. Functions exactly the same as "tcp_timeout" on the next page, but governs incoming connections while <i>tcp_timeout</i> governs only outgoing connections (or connections initiated by <i>pbs_server</i>).</p>



If you use Moab Workload Manager, prevent communication errors by giving **tcp_incoming_timeout** at least twice the value of the Moab `RMPOLLINTERVAL`. See `RMPOLLINTERVAL` for more information.

tcp_timeout

Format <INTEGER>

Default 300

Description Specifies the timeout for idle outbound 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 361 for additional information.

Use "[tcp_incoming_timeout](#)" on the previous page to specify the timeout for idle inbound TCP connections.

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.

timeout_for_job_delete

Format <INTEGER> (seconds)

Default 120

Description The specific timeout used when deleting jobs because the node they are executing on is being deleted.

timeout_for_job_queue

Format <INTEGER> (seconds)

Default 120

timeout_for_job_requeue

Description	The specific timeout used when requeuing jobs because the node they are executing on is being deleted.
--------------------	--

use_jobs_subdirs

Format	<BOOLEAN>
---------------	-----------

Default	Not set (<code>FALSE</code>).
----------------	---------------------------------

Description	<p>Lets an administrator direct the way pbs_server will store its job-related files.</p> <ul style="list-style-type: none">• When <code>use_jobs_subdirs</code> is unset (or set to <code>FALSE</code>), job and job array files will be stored directly under <code>\$PBS_HOME/server_priv/jobs</code> and <code>\$PBS_HOME/server_priv/arrays</code>.• When <code>use_jobs_subdirs</code> is set to <code>TRUE</code>, job and job array files will be distributed over 10 subdirectories under their respective parent directories. This method helps to keep a smaller number of files in a given directory.
--------------------	---



This setting does not automatically move existing job and job array files into the respective subdirectories. If you choose to use this setting (`TRUE`), you must first

- set `use_jobs_subdirs` to `TRUE`,
- shutdown the Torque server daemon,
- in the `contrib` directory, run the "use_jobs_subdirs_setup" python script with `-m` option,
- start the Torque server daemon.

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" below](#)
- ["Node Features and Generic Consumable Resource Specification" on page 349](#)
- ["Command-line Arguments" on page 350](#)

Related Topics

[" Commands Overview" on page 213](#)


[" Prologue and Epilogue Scripts" on page 369](#)


Parameters

These parameters go in the `mom_priv/config` file. They control various behaviors for the MOMs.

"arch" below	"\$signwalltime" on page 333	"\$max_load " on page 338	"\$reject_job_submission" on page 344
"\$attempt_to_make_dir" on the facing page	"\$jobdirectory_sticky" on page 334	"\$memory_pressure_duration" on page 339	"\$resend_join_job_wait_time" on page 344
"\$clienthost" on the facing page	"\$job_exit_wait_time" on page 334	"\$memory_pressure_threshold" on page 339	"\$restricted" on page 345
"\$check_poll_time" on the facing page	"\$job_output_file_umask" on page 334	"\$mom_hierarchy_retry_time" on page 340	"size[fs=<FS>]" on page 345
"\$configversion" on the facing page	"\$job_starter" on page 334	"\$mom_host" on page 340	"\$source_login_batch" on page 345
"\$cputmult" on page 330	"\$job_starter_run_privileged" on page 335	"\$node_check_script" on page 340	"\$source_login_interactive" on page 345
"\$cray_check_rur" on page 330	"\$log_directory" on page 335	"\$node_check_interval" on page 340	"\$spool_as_final_name" on page 346
"\$cuda_visible_devices" on page 330	"\$log_file_suffix" on page 335	"\$nodefile_suffix" on page 341	"\$status_update_time" on page 346
"\$down_on_error" on page 331	"\$logevent" on page 336	"\$nospool_dir_list" on page 341	"\$thread_unlink_calls" on page 346
"\$enablemomrestart" on page 331	"\$loglevel" on page 336	"\$opsys" on page 342	"\$timeout" on page 347
"\$exec_with_exec" on page 331	"\$log_file_max_size" on page 337	"\$pbsclient" on page 342	"\$tmpdir" on page 347
"\$ext_pwd_retry" on page 332	"\$log_file_roll_depth" on page 337	"\$pbsserver" on page 342	"\$usecp" on page 347
"\$force_overwrite" on page 332	"\$log_keep_days" on page 337	"\$prologalarm" on page 343	"\$use_smt" on page 347
"\$ideal_load" on page 332	"\$max_conn_timeout_micro_sec" on page 337	"\$rcpcmd" on page 343	"\$varattr" on page 348
"\$igncput" on page 333	"\$max_join_job_wait_time " on page 338	"\$remote_reconfig" on page 343	"\$wallmult" on page 348
"\$ignmem" on page 333	"\$max_physical_memory" on page 338	"\$remote_checkpoint_dirs" on page 344	"\$xauthpath" on page 349
"\$ignvmem" on page 333	"\$max_swap_memory" on page 339	"\$reduce_prolog_checks" on page 344	

arch	
Format	<STRING>
Description	Specifies the architecture of the local machine. This information is used by the scheduler only.
Example	arch ia64

\$attempt_to_make_dir	
Format	<BOOLEAN>
Default	FALSE
Description	<p>When set to <i>TRUE</i>, specifies that you want Torque to attempt to create the output directories for jobs if they do not already exist.</p> <div>  Torque uses this parameter to make the directory as the <i>user</i> and not as <i>root</i>. Torque will create the directory (or directories) ONLY if the user has permissions to do so. </div>
Example	\$attempt_to_make_dir true

\$clienthost	
Format	<STRING>
Description	<p>Specifies the machine running pbs_server.</p> <div>  This parameter is deprecated. Use \$pbsserver. </div>
Example	\$clienthost node01.teracluster.org

\$check_poll_time	
Format	<STRING>
Default	45
Description	Amount of time (in seconds) between checking running jobs, polling jobs, and trying to resend obituaries for jobs that haven't sent successfully.
Example	\$check_poll_time 90

\$configversion	
Format	<STRING>

\$configversion

Description	Specifies the version of the config file data.
--------------------	--

Example	<code>\$configversion 113</code>
----------------	----------------------------------

\$cputmult

Format	<FLOAT>
---------------	---------

Description	CPU time multiplier.
--------------------	----------------------

 If set to 0.0, MOM level cputime enforcement is disabled.

Example	<code>\$cputmult 2.2</code>
----------------	-----------------------------

\$cray_check_rur

Format	<BOOLEAN>
---------------	-----------

Default	TRUE
----------------	-------------


Description	When set to FALSE , login MOMs (Cray only) will not look at the energy resource information used for each job. Bypassing Resource Utilization Reporting (RUR) checking may improve performance.
--------------------	--

Example	<code>\$cray_check_rur false</code>
----------------	-------------------------------------

\$cuda_visible_devices

Format	<BOOLEAN>
---------------	-----------

Default	TRUE
----------------	-------------

 This is disabled by default when cgroups are enabled, because it becomes repetitive at that time. If you still wish to have the environment variable with cgroups enabled, then you need to set this parameter to **TRUE**.

\$cuda_visible_devices

Description

When set to `TRUE`, the MOM will set the `CUDA_VISIBLE_DEVICES` environment variable for jobs using NVIDIA GPUs. If set to `FALSE`, the MOM will not set `CUDA_VISIBLE_DEVICES` for any jobs.



For `CUDA < 7`, `$CUDA_VISIBLE_DEVICES` is set to the absolute indices of the GPUs your job will use, so if you are using GPUs 2 and 3, then the variable will be set to 2,3. If you are using `CUDA >= 7.0`, then it will be set to the relative index, starting from 0, so if you are using GPUs 2 and 3, the variable will be set to 0,1. This is necessary because of a change in the CUDA implementation that came out in version 7.

Example

```
$cuda_visible_devices false
```

\$down_on_error

Format

<BOOLEAN>

Default

TRUE

Description

Causes the MOM to report itself as state "down" to *pbs_server* in the event of a failed health check. See [Health check on page 227](#) for more information.

Example

```
$down_on_error true
```

\$enablemomrestart

Format

<BOOLEAN>

Description

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 226](#) for more information.

Example

```
$enablemomrestart true
```

\$exec_with_exec

Format

<BOOLEAN>

`$exec_with_exec`

Default	FALSE
Description	<code>pbs_mom</code> uses the <code>exec</code> 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.
Example	<code>\$exec_with_exec true</code>

`$ext_pwd_retry`

Format	<INTEGER>
Default	3
Description	(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.
Example	<code>\$ext_pwd_retry = 5</code>

`$force_overwrite`

Format	<BOOLEAN>
Description	(Available in Torque 6.0.3 and later.) When set to true, forces the output files to be overwritten each time a job is started.
Example	<code>\$force_overwrite true</code>

`$ideal_load`

Format	<FLOAT>
Description	Ideal processor load.
Example	<code>\$ideal_load 4.0</code>

\$igncput	
Format	<BOOLEAN>
Default	FALSE
Description	Ignores limit violation pertaining to CPU time.
Example	<code>\$igncput true</code>

\$ignmem	
Format	<BOOLEAN>
Default	FALSE
Description	Ignores limit violations pertaining to physical memory.
Example	<code>\$ignmem true</code>

\$ignvmem	
Format	<BOOLEAN>
Default	FALSE
Description	Ignores limit violations pertaining to virtual memory.
Example	<code>\$ignvmem true</code>

\$ignwalltime	
Format	<BOOLEAN>
Default	FALSE
Description	Ignore walltime (do not enable MOM based walltime limit enforcement).
Example	<code>\$ignwalltime true</code>

\$jobdirectory_sticky

Format	<BOOLEAN>
Default	FALSE
Description	When this option is set (TRUE), the job directory on the MOM can have a sticky bit set.
Example	<code>\$jobdirectory_sticky true</code>

\$job_exit_wait_time

Format	<INTEGER>
Default	600
Description	This is the timeout (in seconds) 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 <code>JOB_SUBSTATE_EXIT_WAIT</code> . 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 600 seconds (10 minutes).
Example	<code>\$job_exit_wait_time 300</code>

\$job_output_file_umask

Format	<STRING>
Description	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 "userdefault" will use the user's default umask. This parameter is in version 2.3.0 and later.
Example	<code>\$job_output_file_umask 027</code>

\$job_starter

Format	<STRING>
---------------	----------

\$job_starter

Description	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.
--------------------	--

Example	<pre>\$job_starter /var/torque/mom_priv/job_starter.sh > cat /var/torque/mom_priv/job_starter.sh #!/bin/bash export FOOHOME=/home/foo ulimit -n 314 \$*</pre>
----------------	--

\$job_starter_run_privileged

Format	<BOOLEAN>
---------------	-----------

Default	FALSE
----------------	--------------

Description	When set to TRUE, specifies that you want Torque to execute the \$job_starter script with elevated privileges.
--------------------	--

Example	<code>\$job_starter_run_privileged true</code>
----------------	--

\$log_directory

Format	<STRING>
---------------	----------

Default	TORQUE_HOME/mom_logs/
----------------	------------------------------

Description	Changes the log directory. 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	<code>\$log_directory /opt/torque/mom_logs/</code>
----------------	--


\$log_file_suffix

Format	<STRING>
---------------	----------

\$log_file_suffix

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	<pre>\$log_file_suffix %h = 20100223.mybox \$log_file_suffix foo = 20100223.foo</pre>

\$logevent

Format	<INTEGER>																																	
Description	<p>Creates an event mask enumerating which log events will be recorded in the MOM logs. By default all events are logged.</p> <p>These are the events which can be chosen:</p> <div><table><tr><td>ERROR</td><td>0x0001</td><td>internal errors</td></tr><tr><td>SYSTEM</td><td>0x0002</td><td>system (server) & (trqauthd) events</td></tr><tr><td>ADMIN</td><td>0x0004</td><td>admin events</td></tr><tr><td>JOB</td><td>0x0008</td><td>job related events</td></tr><tr><td>JOB_USAGE</td><td>0x0010</td><td>End of Job accounting</td></tr><tr><td>SECURITY</td><td>0x0020</td><td>security violation events</td></tr><tr><td>SCHED</td><td>0x0040</td><td>scheduler events</td></tr><tr><td>DEBUG</td><td>0x0080</td><td>common debug messages</td></tr><tr><td>DEBUG2</td><td>0x0100</td><td>less needed debug messages</td></tr><tr><td>CLIENTAUTH</td><td>0x0200</td><td>TRQAUTHD login events</td></tr><tr><td>SYSLOG</td><td>0x0400</td><td>pass this event to the syslog as well</td></tr></table></div> <div><p> The listed events are shown here with hexadecimal values; however, a decimal value must be used when setting \$logevent.</p></div>	ERROR	0x0001	internal errors	SYSTEM	0x0002	system (server) & (trqauthd) events	ADMIN	0x0004	admin events	JOB	0x0008	job related events	JOB_USAGE	0x0010	End of Job accounting	SECURITY	0x0020	security violation events	SCHED	0x0040	scheduler events	DEBUG	0x0080	common debug messages	DEBUG2	0x0100	less needed debug messages	CLIENTAUTH	0x0200	TRQAUTHD login events	SYSLOG	0x0400	pass this event to the syslog as well
ERROR	0x0001	internal errors																																
SYSTEM	0x0002	system (server) & (trqauthd) events																																
ADMIN	0x0004	admin events																																
JOB	0x0008	job related events																																
JOB_USAGE	0x0010	End of Job accounting																																
SECURITY	0x0020	security violation events																																
SCHED	0x0040	scheduler events																																
DEBUG	0x0080	common debug messages																																
DEBUG2	0x0100	less needed debug messages																																
CLIENTAUTH	0x0200	TRQAUTHD login events																																
SYSLOG	0x0400	pass this event to the syslog as well																																
Example	<p>\$logevent 1039 will log ERROR, SYSTEM, ADMIN, JOB and SYSLOG events. This has a hexadecimal value of 0x40F.</p>																																	

\$loglevel

Format	<INTEGER>
Description	Specifies the verbosity of logging with higher numbers specifying more verbose logging. Values may range between 0 and 7.
Example	<pre>\$loglevel 4</pre>

\$log_file_max_size

Format	<INTEGER>
Description	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.
Example	<code>\$log_file_max_size = 100</code>

\$log_file_roll_depth

Format	<INTEGER>
Description	Specifies how many times a log file will be rolled before it is deleted.
Example	<code>\$log_file_roll_depth = 7</code>


\$log_keep_days

Format	<INTEGER>
Description	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.
Example	<code>\$log_keep_days 10</code>

\$max_conn_timeout_micro_sec

Format	<INTEGER>
Default	10000
Description	Specifies how long (in microseconds) pbs_mom should wait for a connection to be made. Default value is 10,000 (.01 sec).
Example	<code>\$max_conn_timeout_micro_sec 30000</code> This sets the connection timeout on the MOM to .03 seconds.



\$max_join_job_wait_time

Format	<INTEGER>
Default	600
Description	<p>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 600 seconds (10 minutes).</p> <div> If a MOM is completely idle, it can take as long as the next MOM-to-server update time to requeue a failed job.</div>
Example	<code>\$max_join_job_wait_time 300</code>


\$max_load

Format	<FLOAT>
Description	Maximum processor load.
Example	<code>\$max_load 4.0</code>

\$max_physical_memory

Format	<INTEGER> <unit>
Description	<p>Restrict the amount of memory available to jobs on this node to the specified amount, which may not exceed the amount of memory on the machine and must be greater than 0. Default is to use all available memory on the host.</p> <div> When cgroups are enabled, this limits the whole of the machine and doesn't specifically limit each socket or NUMA node. If you have 2 NUMA nodes and 32 GB of memory, but you limit the machine to 30, it won't force a job requesting 16 GB to span NUMA nodes, but once that jobs starts, there would only be 14 GB remaining in use for jobs.</div> <div> If you are using this setting, availmem (as reported in pbsnodes) is no longer accurate, as we do not know what portion of used memory and swap are by jobs and what portion are from the operating system. Since availmem is no longer accurate, you need to set NODEAVAILABILITYPOLICY to DEDICATED if you are using Moab or Maui.</div>
Example	<code>\$max_physical_memory 30gb</code>

\$max_swap_memory

Format	<INTEGER> <unit>
Description	<p>Restrict the amount of swap available to jobs on this node to the specified amount, which may not exceed the amount of swap on the machine and must be greater than 0. If you wish to disallow swap, this must be set to a very low value instead of 0. Default is to use all available memory on the host.</p> <div> If you are using this setting, availmem (as reported in pbsnodes) is no longer accurate, as we do not know what portion of used memory and swap are by jobs and what portion are from the operating system. Since availmem is no longer accurate, you need to set NODEAVAILABILITYPOLICY to DEDICATED if you are using Moab or Maui.</div>
Example	<code>\$max_swap_memory 5gb</code>

\$memory_pressure_duration

Format	<INTEGER>
Description	<p><i>(Applicable in version 3.0 and later.)</i> 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.</p>
Example	<code>\$memory_pressure_duration 5</code>

\$memory_pressure_threshold

Format	<INTEGER>
Description	<p><i>(Applicable in version 3.0 and later.)</i> 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.</p> <p>To enable, log in as the super user and execute the command <code>echo 1 >> /dev/cpuset/memory_pressure_enabled</code>. See the cpuset man page for more information concerning memory pressure.</p>
Example	<code>\$memory_pressure_threshold 1000</code>

\$mom_hierarchy_retry_time

Format	<SECONDS>
Description	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.
Example	<code>\$mom_hierarchy_retry_time 30</code>

\$mom_host

Format	<STRING>
Description	Sets the local hostname as used by <i>pbs_mom</i> .
Example	<code>\$mom_host node42</code>


\$node_check_script

Format	<STRING>
Description	Specifies the fully qualified pathname of the health check script to run (see " Compute Node Health Check " on page 201 for more information).
Example	<code>\$node_check_script /opt/batch_tools/nodecheck.pl</code>

\$node_check_interval

Format	<STRING>
---------------	----------

`$node_check_interval`

Description	<p>Specifies the number of MOM intervals between subsequent executions of the specified health check. This value defaults to 1 indicating the check is run every MOM interval (see "Compute Node Health Check" on page 201 for more information).</p> <p><code>\$node_check_interval</code> has two special strings that can be set:</p> <ul style="list-style-type: none">• <i>jobstart</i> – makes the node health script run when a job is started (before the prologue script).• <i>jobend</i> – makes the node health script run after each job has completed on a node (after the epilogue script). <div> The node health check may be configured to run before or after the job with the "jobstart" and/or "jobend" options. However, the job environment variables do not get passed to node health check script, so it has no access to those variables at any time.</div>
Example	<code>\$node_check_interval 5</code>

`$nodefile_suffix`

Format	<STRING>
Description	Specifies the suffix to append to a host names to denote the data channel network adapter in a multi-homed compute node.
Example	<code>\$nodefile_suffix i</code> with the suffix of "i" and the control channel adapter with the name <i>node01</i> , the data channel would have a hostname of <i>node01i</i> .

`$nospool_dir_list`

Format	<STRING>
---------------	----------

\$nospool_dir_list

Description

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.



This should generally be used only when the job can run on the same machine as where the output file goes, or if there is a shared filesystem. If not, this parameter can slow down the system or fail to create the output file.

Example

```
$nospool_dir_list /home/mike/jobs/, /var/tmp/spool/
```

opsys

Format

<STRING>

Description

Specifies the operating system of the local machine. This information is used by the scheduler only.

Example

```
opsys RHEL3
```

\$pbsclient

Format

<STRING>

Description

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.


Example

```
$pbsclient node01.teracluster.org
```

\$pbsserver

Format

<STRING>

\$pbsserver	
Description	Specifies the machine running pbs_server. <div>  This parameter replaces the deprecated parameter \$clienthost. </div>
Example	<code>\$pbsserver node01.teracluster.org</code>

\$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). When running parallel jobs, this is also the maximum time a sister node will wait for a job to start.
Example	<code>\$prologalarm 60</code>

\$rcpcmd	
Format	<STRING>
Description	Specifies the full path and optional additional command line args to use to perform remote copies.
Example	<code>mom_priv/config:</code> <code>\$rcpcmd /usr/local/bin/scp -i /etc/sshauth.dat</code>

\$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	<code>\$remote_reconfig true</code>

\$remote_checkpoint_dirs

Format	<STRING>
Description	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.
Example	<pre>\$remote_checkpoint_dirs /checkpointFiles /bigStorage /fast</pre> <p><i>This informs the MOM that the /checkpointFiles, /bigStorage, and /fast directories are remotely mounted checkpoint directories.</i></p>

\$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	<pre>\$reduce_prolog_checks true</pre>



\$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	<pre>\$reject_job_submission job01</pre>

\$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	<pre>\$resend_join_job_wait_time 120</pre>

\$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	<code>\$restricted *.teracluster.org</code>

size[fs=<FS>]	
Format	N/A
Description	<p>Specifies that the available and configured disk space in the <FS> filesystem is to be reported to the pbs_server and scheduler.</p> <div>  To request disk space on a per job basis, specify the file resource as in <code>qsub -l nodes=1, file=1000kb</code>. </div> <div>  Unlike most MOM config options, the <i>size</i> parameter is not preceded by a "\$" character. </div>
Example	<code>size[fs=/localscratch]</code> The available and configured disk space in the <code>/localscratch</code> filesystem will be reported.

\$source_login_batch	
Format	<STRING>
Description	Specifies whether or not MOM will source the <code>/etc/profile</code> , etc. type files for <i>batch</i> 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.
Example	<code>\$source_login_batch False</code> MOM will bypass the sourcing of <code>/etc/profile</code> , etc. type files.

\$source_login_interactive	
Format	<STRING>

`$source_login_interactive`

Description	Specifies whether or not MOM will source the <code>/etc/profile</code> , etc. type files for <i>interactive</i> 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.
Example	<code>\$source_login_interactive False</code> MOM will bypass the sourcing of <code>/etc/profile</code> , etc. type files.

`$spool_as_final_name`

Format	<BOOLEAN>
Description	This makes the job write directly to its output destination instead of a spool directory. This allows users easier access to the file if they want to watch the jobs output as it runs.
Example	<code>\$spool_as_final_name true</code>

`$status_update_time`

Format	<INTEGER>
Description	Specifies the number of seconds between subsequent MOM-to-server update reports. Default is 45 seconds.
Example	<code>status_update_time:</code> <code>\$status_update_time 120</code> MOM will send server update reports every 120 seconds.

`$thread_unlink_calls`

Format	<BOOLEAN>
Description	This directs the MOM to create a thread to handle job deletion. Default is true. If it is set to TRUE, <code>pbs_mom</code> will use a thread to delete the job's files which increases job performance but also increases the base memory footprint of <code>pbs_mom</code> .
Example	<code>thread_unlink_calls:</code> <code>\$thread_unlink_calls true</code>

\$timeout	
Format	<INTEGER>
Description	<p>Specifies the number of seconds before a TCP connection on the MOM will timeout. Default is 300 seconds.</p> <p>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.</p>
Example	<pre>\$timeout 120</pre> <p>A TCP connection will wait up to 120 seconds before timing out.</p> <p>For 3.x and earlier, MOM-to-MOM communication will allow up to 120 seconds before timing out.</p>

\$tmpdir	
Format	<STRING>
Description	Specifies a directory to create job-specific scratch space.
Example	<pre>\$tmpdir /localscratch</pre>

\$usecp	
Format	<HOST>:<SRCDIR> <DSTDIR>
Description	Specifies which directories should be staged (see "NFS and Other Networked Filesystems" on page 142)
Example	<pre>\$usecp *.fte.com:/data /usr/local/data</pre>

\$use_smt	
Format	<BOOLEAN>
Default	TRUE

\$use_smt

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.



\$use_smt is deprecated. Please use the "[-L NUMA Resource Request](#)" on page 173 syntax to control whether or not threads or cores are used.

If SMT is used, you will need to set the *np* attribute so that each logical processor is counted.

Example

```
$use_smt false
```

\$varattr

Format

<INTEGER> <STRING>

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:

```
name=value
```

Include any arguments after the script's full path. These features are visible in the output of [pbsnodes](#)-a

```
varattr=Matlab=7.1;Octave=1.0.
```

For information about using \$varattr to request dynamic features in Moab, see REQATTR in the *Moab Workload Manager Administrator Guide*.

Example

```
$varattr 25 /usr/local/scripts/nodeProperties.pl arg1 arg2 arg3
```

\$wallmult

Format

<FLOAT>

Description

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 cputmult is used for cpu time.



If set to 0.0, MOM level walltime enforcement is disabled.

Example

```
$wallmult 2.2
```

\$xauthpath	
Format	<STRING>
Description	Specifies the path to the xauth binary to enable X11 forwarding.
Example	<code>\$xauthpath /opt/bin/xauth</code>

Related Topics

[" Node Manager \(MOM\) Configuration" on page 327](#)

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 **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 327](#)

Command-line Arguments

Below is a table of *pbs_mom* command-line startup flags.

Flag	Description
a <integer>	Alarm time in seconds.
c <file>	Config file path.
C <dir-ectory>	Checkpoint path.
d <dir-ectory>	Home directory.
L <file>	Log file.
M <integer>	MOM port to listen on.
p	Perform 'poll' based job recovery on restart (jobs persist until associated processes terminate).
P	On restart, deletes all jobs that were running on MOM (Available in 2.4.X and later).
q	On restart, requeues all jobs that were running on MOM (Available in 2.4.X and later).
r	On restart, kills all processes associated with jobs that were running on MOM, and then requeues the jobs.
R <integer>	MOM 'RM' port to listen on.
S <integer>	pbs_server port to connect to.
v	Display version information and exit.
x	Disable use of privileged port.
?	Show usage information and exit.

For more details on these command-line options, see ["pbs_mom" on page 222](#).

Related Topics

["Node Manager \(MOM\) Configuration" on page 327](#)

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 'p s'`. 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 D-1: Torque error codes

Error code name	Number	Description
PBSE_FLOOR	15000	No error
PBSE_UNKJOBID	15001	Unknown job ID error
PBSE_NOATTR	15002	Undefined attribute
PBSE_ATTRRO	15003	Cannot set attribute, read only or insufficient permission
PBSE_IVALREQ	15004	Invalid request
PBSE_UNKREQ	15005	Unknown request
PBSE_TOOMANY	15006	Too many submit retries
PBSE_PERM	15007	Unauthorized Request
PBSE_IFF_NOT_FOUND	15008	trqauthd unable to authenticate

Error code name	Number	Description
PBSE_MUNGE_NOT_FOUND	15009	Munge executable not found, unable to authenticate
PBSE_BADHOST	15010	Access from host not allowed, or unknown host
PBSE_JOBEXIST	15011	Job with requested ID already exists
PBSE_SYSTEM	15012	System error
PBSE_INTERNAL	15013	PBS server internal error
PBSE_REGROUTE	15014	Dependent parent job currently in routing queue
PBSE_UNKSIG	15015	Unknown/illegal signal name
PBSE_BADATVAL	15016	Illegal attribute or resource value for
PBSE_MODATTRUN	15017	Cannot modify attribute while job running
PBSE_BADSTATE	15018	Request invalid for state of job
PBSE_UNKQUE	15020	Unknown queue
PBSE_BADCRED	15021	Invalid credential
PBSE_EXPIRED	15022	Expired credential
PBSE_QUNOENB	15023	Queue is not enabled
PBSE_QACCESS	15024	Access to queue is denied
PBSE_BADUSER	15025	Bad UID for job execution
PBSE_HOPCOUNT	15026	Job routing over too many hops
PBSE_QUEEXIST	15027	Queue already exists
PBSE_ATTRTYPE	15028	Incompatible type
PBSE_QUEEBUSY	15029	Cannot delete busy queue

Error code name	Number	Description
PBSE_QUENBIG	15030	Queue name too long
PBSE_NOSUP	15031	No support for requested service
PBSE_QUENOEN	15032	Cannot enable queue, incomplete definition
PBSE_PROTOCOL	15033	Batch protocol error
PBSE_BADATLST	15034	Bad attribute list structure
PBSE_NOCONNECTS	15035	No free connections
PBSE_NOSERVER	15036	No server specified
PBSE_UNKRESC	15037	Unknown resource type
PBSE_EXCQRESC	15038	Job exceeds queue resource limits
PBSE_QUENODFLT	15039	No default queue specified
PBSE_NORERUN	15040	Job is not rerunnable
PBSE_ROUTEREJ	15041	Job rejected by all possible destinations (check syntax, queue resources, ...)
PBSE_ROUTEEXPD	15042	Time in Route Queue Expired
PBSE_MOMREJECT	15043	Execution server rejected request
PBSE_BADSCRIPT	15044	(qsub) cannot access script file
PBSE_STAGEIN	15045	Stage-in of files failed
PBSE_RESCUNAV	15046	Resource temporarily unavailable
PBSE_BADGRP	15047	Bad GID for job execution
PBSE_MAXQUED	15048	Maximum number of jobs already in queue

Error code name	Number	Description
PBSE_CKPBSY	15049	Checkpoint busy, may retry
PBSE_EXLIMIT	15050	Resource limit exceeds allowable
PBSE_BADACCT	15051	Invalid Account
PBSE_ALRDYEXIT	15052	Job already in exit state
PBSE_NOCOPYFILE	15053	Job files not copied
PBSE_CLEANEOUT	15054	Unknown job id after clean init
PBSE_NOSYNCMSTR	15055	No master found for sync job set
PBSE_BADDEPEND	15056	Invalid Job Dependency
PBSE_DUPLIST	15057	Duplicate entry in list
PBSE_DISPROTO	15058	Bad DIS based Request Protocol
PBSE_EXECTHERE	15059	Cannot execute at specified host because of checkpoint or stagein files
PBSE_SISREJECT	15060	Sister rejected
PBSE_SISCOMM	15061	Sister could not communicate
PBSE_SVRDOWN	15062	Request not allowed: Server shutting down
PBSE_CKPSHORT	15063	Not all tasks could checkpoint
PBSE_UNKNODE	15064	Unknown node
PBSE_UNKNODEATR	15065	Unknown node-attribute
PBSE_NONODES	15066	Server has no node list
PBSE_NODENBIG	15067	Node name is too big

Error code name	Number	Description
PBSE_NODEEXIST	15068	Node name already exists
PBSE_BADNDATVAL	15069	Illegal value for
PBSE_MUTUALEX	15070	Mutually exclusive values for
PBSE_GMODERR	15071	Modification failed for
PBSE_NORELYMOM	15072	Server could not connect to MOM
PBSE_NOTSNODE	15073	No time-share node available
PBSE_JOBTYPE	15074	Wrong job type
PBSE_BADACLHOST	15075	Bad ACL entry in host list
PBSE_MAXUSERQUED	15076	Maximum number of jobs already in queue for user
PBSE_BADDISALLOWTYPE	15077	Bad type in disallowed_types list
PBSE_NOINTERACTIVE	15078	Queue does not allow interactive jobs
PBSE_NOBATCH	15079	Queue does not allow batch jobs
PBSE_NORERUNABLE	15080	Queue does not allow rerunable jobs
PBSE_NONONRERUNABLE	15081	Queue does not allow nonrerunable jobs
PBSE_UNKARRAYID	15082	Unknown Array ID
PBSE_BAD_ARRAY_REQ	15083	Bad Job Array Request
PBSE_BAD_ARRAY_DATA	15084	Bad data reading job array from file
PBSE_TIMEOUT	15085	Time out
PBSE_JOBNOTFOUND	15086	Job not found
PBSE_NOFAULTTOLERANT	15087	Queue does not allow fault tolerant jobs

Error code name	Number	Description
PBSE_NOFAULTINTOLERANT	15088	Queue does not allow fault intolerant jobs
PBSE_NOJOBARRAYS	15089	Queue does not allow job arrays
PBSE_RELAYED_TO_MOM	15090	Request was relayed to a MOM
PBSE_MEM_MALLOC	15091	Error allocating memory - out of memory
PBSE_MUTEX	15092	Error allocating controlling mutex (lock/unlock)
PBSE_THREADATTR	15093	Error setting thread attributes
PBSE_THREAD	15094	Error creating thread
PBSE_SELECT	15095	Error in socket select
PBSE_SOCKET_FAULT	15096	Unable to get connection to socket
PBSE_SOCKET_WRITE	15097	Error writing data to socket
PBSE_SOCKET_READ	15098	Error reading data from socket
PBSE_SOCKET_CLOSE	15099	Socket close detected
PBSE_SOCKET_LISTEN	15100	Error listening on socket
PBSE_AUTH_INVALID	15101	Invalid auth type in request
PBSE_NOT_IMPLEMENTED	15102	This functionality is not yet implemented
PBSE_QUENOTAVAILABLE	15103	Queue is currently not available
PBSE_TMPDIFFOWNER	15104	tmpdir owned by another user
PBSE_TMPNOTDIR	15105	tmpdir exists but is not a directory
PBSE_TMPNONAME	15106	tmpdir cannot be named for job
PBSE_CANTOPENSOCKET	15107	Cannot open demux sockets

Error code name	Number	Description
PBSE_CANTCONTACTSISTERS	15108	Cannot send join job to all sisters
PBSE_CANTCREATETMPDIR	15109	Cannot create tmpdir for job
PBSE_BADMOMSTATE	15110	Mom is down, cannot run job
PBSE_SOCKET_INFORMATION	15111	Socket information is not accessible
PBSE_SOCKET_DATA	15112	Data on socket does not process correctly
PBSE_CLIENT_INVALID	15113	Client is not allowed/trusted
PBSE_PREMATURE_EOF	15114	Premature End of File
PBSE_CAN_NOT_SAVE_FILE	15115	Error saving file
PBSE_CAN_NOT_OPEN_FILE	15116	Error opening file
PBSE_CAN_NOT_WRITE_FILE	15117	Error writing file
PBSE_JOB_FILE_CORRUPT	15118	Job file corrupt
PBSE_JOB_RERUN	15119	Job can not be rerun
PBSE_CONNECT	15120	Can not establish connection
PBSE_JOBWORKDELAY	15121	Job function must be temporarily delayed
PBSE_BAD_PARAMETER	15122	Parameter of function was invalid
PBSE_CONTINUE	15123	Continue processing on job. (Not an error)
PBSE_JOBSTATE	15124	Current sub state does not allow trasaction.
PBSE_CAN_NOT_MOVE_FILE	15125	Error moving file
PBSE_JOB_RECYCLED	15126	Job is being recycled
PBSE_JOB_ALREADY_IN_QUEUE	15127	Job is already in destination queue.

Error code name	Number	Description
PBSE_INVALID_MUTEX	15128	Mutex is NULL or otherwise invalid
PBSE_MUTEX_ALREADY_LOCKED	15129	The mutex is already locked by this object
PBSE_MUTEX_ALREADY_UNLOCKED	15130	The mutex has already been unlocked by this object
PBSE_INVALID_SYNTAX	15131	Command syntax invalid
PBSE_NODE_DOWN	15132	A node is down. Check the MOM and host
PBSE_SERVER_NOT_FOUND	15133	Could not connect to batch server
PBSE_SERVER_BUSY	15134	Server busy. Currently no available threads

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 Resource Manager" on page 10](#)). 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 350](#)), 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.
- 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.

i 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 226](#) 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.

```
> 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.

```
> 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.

i If you have 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" below](#)
- ["End-User Command Caching" on the next page](#)
- ["Moab and Torque Configuration for Large Clusters" on page 364](#)
- ["Starting Torque in Large Environments" on page 365](#)
- ["Other Considerations" on page 366](#)

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 327](#)). 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.

i 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:  
fs.file-max = 65536
```

Related Topics

[" Large Cluster Considerations" on the previous page](#)

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** -f' 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`

            AGE=$(( $NOW - $TMPFILEMTIME ))
        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"

```

```
cat $TMPFILE
exit 0
```

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 361](#)

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 $(2 * \text{number of cores} + 1) * 10$.
6. Versions 5.1.3, 6.0.2, and later: if you have the server send emails, set `email_batch_seconds` appropriately. Setting this parameter will prevent `pbs_server` from forking too frequently and increase the server's performance. See "[email_batch_seconds](#)" on page 307 for more information on this server parameter.

Related Topics

[" Large Cluster Considerations" on page 361](#)

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 `pbs_mom` 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_mom` without the `-w` option. Starting it with `-w` causes the MOMs to appear to be down.

Related Topics

[" Large Cluster Considerations" on page 361](#)

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.

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 Moab Parameters in 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-filesystem` argument with *configure*. While having filesystem 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. Filesync 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

# Don't allow the arp table to become bigger than this
net.ipv4.neigh.default.gc_thresh3 = 4096
# Tell the gc when to become aggressive with arp table cleaning.
# Adjust this based on size of the LAN.
net.ipv4.neigh.default.gc_thresh2 = 2048
# Adjust where the gc will leave arp table alone
net.ipv4.neigh.default.gc_thresh1 = 1024
# Adjust to arp table gc to clean-up more often
net.ipv4.neigh.default.gc_interval = 3600
# ARP cache entry timeout
net.ipv4.neigh.default.gc_stale_time = 3600
```

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 361](#)

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. The "Mother Superior" is the pbs_mom on the first node allocated for a job. While it is technically a sister node, it is not a "Sister" for the purposes of the following table.

 The initial working directory for each script is `TORQUE_HOME/mom_priv/`.

Script	Execution location	Script Location	Execute as	File per- missions
prologue	Mother Superior	8th argu- ment	root	Readable and execut- able by root and NOT writable by any- one but root (e.g., - r-x-----)
epilogue		11th argu- ment	root	
prologue.user		---	user	Readable and execut- able by root and other (e.g., -r-x--- r-x)
epilogue.user			user	
prologue.parallel	Sister	---	root	Readable and execut- able by root and NOT writable by any- one but root (e.g., -
epilogue.parallel			root	
prologue.user.parallel			user	Readable and execut- able by root and other (e.g., -r-x--- r-x)
epilogue.user.parallel			user	
epilogue.precancel	Mother Superior This script runs after a job cancel request is received from pbs_server and before a kill signal is sent to the job process.		user	

i `epilogue.parallel` is available in version 2.1 and later.

This section contains these topics:

- ["Script Order of Execution" below](#)
- ["Script Environment" below](#)
- ["Per Job Prologue and Epilogue Scripts" on page 372](#)
- ["Prologue and Epilogue Scripts Time Out" on page 373](#)
- ["Prologue Error Processing" on page 373](#)

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.

i 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 372](#) for more information.)

i The node health check may be configured to run before or after the job with the "jobstart" and/or "jobend" options. However, the job environment variables do not get passed to node health check script, so it has no access to those variables at any time.

i Root squashing is now supported for `epilogue` and `prologue` scripts.

Related Topics

[" Prologue and Epilogue Scripts" on the previous page](#)

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:

Argument	Description
<code>argv[1]</code>	job id
<code>argv[2]</code>	job execution user name
<code>argv[3]</code>	job execution group name
<code>argv[4]</code>	job name (Torque 1.2.0p4 and higher only)
<code>argv[5]</code>	list of requested resource limits (Torque 1.2.0p4 and higher only)
<code>argv[6]</code>	job execution queue (Torque 1.2.0p4 and higher only)
<code>argv[7]</code>	job account (Torque 1.2.0p4 and higher only)
<code>argv[8]</code>	job script location

Epilogue Environment

Torque supplies the following arguments to the `epilogue`, `epilogue.user`, `epilogue.precancel`, and `epilogue.parallel` scripts:

Argument	Description
<code>argv[1]</code>	job id
<code>argv[2]</code>	job execution user name
<code>argv[3]</code>	job execution group name
<code>argv[4]</code>	job name
<code>argv[5]</code>	session id
<code>argv[6]</code>	list of requested resource limits

Argument	Description
<code>argv[7]</code>	list of resources used by job
<code>argv[8]</code>	job execution queue
<code>argv[9]</code>	job account
<code>argv[10]</code>	job exit code
<code>argv[11]</code>	job script location

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.

i The cancel job command (**qdel**) will take as long to return as the `epilogue.precancel` script takes to run. For example, if the script runs for 5 minutes, it takes 5 minutes for **qdel** to return.

For all scripts, the environment passed to the script is empty. However, if you submit the job using `msub` rather than `qsub`, some Moab environment variables are available in the Torque prologue and epilogue script environment: `MOAB_CLASS`, `MOAB_GROUP`, `MOAB_JOBARRAYINDEX`, `MOAB_JOBARRAYRANGE`, `MOAB_JOBID`, `MOAB_JOBNAME`, `MOAB_MACHINE`, `MOAB_NODECOUNT`, `MOAB_NODELIST`, `MOAB_PARTITION`, `MOAB_PROCCOUNT`, `MOAB_QOS`, `MOAB_TASKMAP`, and `MOAB_USER`. See **msub** on page 1 in the *Moab Workload Manager Administrator Guide* for more information.

Also, standard input for both scripts is connected to a system dependent file. Currently, for all systems this is `/dev/null`. Except for epilogue scripts of an interactive job, `prologue.parallel`, `epilogue.precancel`, and `epilogue.parallel`, the standard output and error are connected to output and error files associated with the job. For an interactive job, since the pseudo terminal connection is released after the job completes, the standard input and error point to `/dev/null`. For `prologue.parallel` and `epilogue.parallel`, the user will need to redirect `stdout` and `stderr` manually.

Related Topics

[" Prologue and Epilogue Scripts" on page 369](#)

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 24 2009-11-09 16:11 prologue_script.sh
-r-x----- 1 usertom usertom 24 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.

Related Topics

[" Prologue and Epilogue Scripts" on page 369](#)

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 **\$prologalarm** parameter in the `mom_priv/config` file.

i While Torque is executing the epilogue, `epilogue.user`, or `epilogue.precancel` scripts, the job will be in the *E* (exiting) state.

If an `epilogue.parallel` script cannot open the `.OU` or `.ER` files, an error is logged but the script is continued.

Related Topics

[" Prologue and Epilogue Scripts" on page 369](#)

Prologue Error Processing

If the 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 `pbs_mom` will report the script's exit status to `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.

Error	Description	Action
-4	The script timed out	Job will be requeued
-3	The wait(2) call returned an error	Job will be requeued
-2	Input file could not be opened	Job will be requeued
-1	Permission error (script is not owned by root, or is writable by others)	Job will be requeued
0	Successful completion	Job will run
1	Abort exit code	Job will be aborted
>1	other	Job will be requeued

Example G-2:

Following are example prologue and epilogue scripts that write the arguments passed to them in the job's standard out file:

prologue	
Script	<pre>#!/bin/sh echo "Prologue Args:" echo "Job ID: \$1" echo "User ID: \$2" echo "Group ID: \$3" echo "" exit 0</pre>
stdout	<pre>Prologue Args: Job ID: 13724.node01 User ID: user1 Group ID: user1</pre>

epilogue	
Script	<pre>#!/bin/sh echo "Epilogue Args:" echo "Job ID: \$1" echo "User ID: \$2" echo "Group ID: \$3" echo "Job Name: \$4" echo "Session ID: \$5" echo "Resource List: \$6" echo "Resources Used: \$7" echo "Queue Name: \$8" echo "Account String: \$9" echo "" exit 0</pre>
stdout	<pre>Epilogue Args: Job ID: 13724.node01 User ID: user1 Group ID: user1 Job Name: script.sh Session ID: 28244 Resource List: neednodes=node01,nodes=1,walltime=00:01:00 Resources Used: cput=00:00:00,mem=0kb,vmem=0kb,walltime=00:00:07 Queue Name: batch Account String:</pre>

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.

i Having a separate temporary directory on each node is probably not as good as having a good, high performance parallel filesystem.

```

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 -- grpid
#
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 -- grpid
# 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

```

i 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 369](#)

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" below](#)
- ["Configuring the Second Torque" below](#)
- ["Bringing the First Torque Server online" below](#)
- ["Bringing the Second Torque Server Online" on the next page](#)

Configuring the First Torque

```
./configure --with-server-home=/var/spool/torque1 --bindir=/var/spool/torque1/bin --  
sbindir=/var/spool/torque1/sbin
```

Then make and make install will place the first Torque into /var/spool/torque1 with the executables in their corresponding directories.

Configuring the Second Torque

```
./configure --with-server-home=/var/spool/torque2 --bindir=/var/spool/torque2/bin --  
sbindir=/var/spool/torque2/sbin
```

Then make and make install will place the second Torque into /var/spool/torque2 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 and communicate with the MOMs on port 35001. The first Torque MOM will try to connect to the server on port 35000 and will listen for requests from the server on port 35001. (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 -t create  
> pbs_mom -S 35000 -M 35001
```

Afterwards, when using a client command to make a batch request it is necessary to specify the server name and server port (35000):

```
> pbsnodes -a -s node01:35000
```

Submitting jobs can be accomplished using the `-q` option ([queue][@host[:port]]):

```
> 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 36001, and communicate via TCP 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 TCP on port 36002.

```
> 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):

```
> 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.

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. 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 and must be available on each of the nodes where users may submit jobs. Because the submit filter is likely to run multiple times for each job submission, all operations in the submit filter must be idempotent, i.e., they must produce the same results if called more than once with the same input parameters.

By default, the submit filter must be located at `/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 383), as in the following example:

```
torque.cfg:
```

```
SUBMITFILTER /opt/torque/submit.pl  
...
```

 Initial development courtesy of Oak Ridge National Laboratories.

"torque.cfg" Configuration File

Administrators can configure the `torque.cfg` file (located in `PBS_SERVER_HOME` (`/var/spool/torque` by default)) to alter the behavior of the `qsub` command on specific host machines where the file resides. This file contains a list of parameters and values separated by whitespace. This only affects `qsub`, and only on each specific host with the file.

Configuration Parameters

"CLIENTRETRY" below
"DEFAULTCKPT" below
"FAULT_TOLERANT_BY_DEFAULT" on the next page
"HOST_NAME_SUFFIX" on the next page

"INTERACTIVE_PORT_RANGE" on page 385
"QSUBHOST" on page 385
"QSUBSENDUID" on page 385
"QSUBSLEEP" on page 386

"RERUNNABLEBYDEFAULT" on page 386
"SERVERHOST" on page 386
"SUBMITFILTER" on page 387

"TRQ_IFNAME" on page 387
"VALIDATEGROUP" on page 387
"VALIDATEPATH" on page 388

CLIENTRETRY

Format <INT>

Default 0

Description Seconds between retry attempts to talk to `pbs_server`.

Example

```
CLIENTRETRY 10
```

Torque waits 10 seconds after a failed attempt before it attempts to talk to `pbs_server` again.

DEFAULTCKPT

For mat One of None, Enabled, Shutdown, Periodic, Interval=minutes, depth=number, or dir=r=path

Default None

DEFAULTCKPT

Description Default value for job's checkpoint attribute. For a description of all possible values, see ["qsub" on page 275](#)

 This default setting can be overridden at job submission with the `qsub -c` option.

Example

```
DEFAULTCKPT Shutdown
```

By default, Torque checkpoints at pbs_mom shutdown.

FAULT_TOLERANT_BY_DEFAULT

Format <BOOLEAN>

Default FALSE

Description Sets all jobs to fault tolerant by default. (See `qsub -f` for more information on fault tolerance.)

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

Format <STRING>

Default ---

Description Specifies a hostname suffix. When `qsub` 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.

Example

```
HOST_NAME_SUFFIX -ib
```

When a job is submitted, the -ib suffix is appended to the host name.

INTERACTIVE_PORT_RANGE	
Format	<INTEGER>-<INTEGER>
Default	---
Description	Sets a range of ports for interactive jobs. The minimum port must be greater than 1024, and the maximum port must be greater than the minimum port, or else the setting will be ignored.
Example	<pre>INTERACTIVE_PORT_RANGE 20000-20100</pre> <p><i>Force all interactive listening ports on this host to be between 20000 and 20100, inclusive.</i></p>

QSUBHOST	
Format	<HOSTNAME>
Default	---
Description	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.
Example	<pre>QSUBHOST host1</pre> <p><i>The default hostname associated with a job is host1.</i></p>

QSUBSENDUID	
Format	N/A
Default	---
Description	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.
Example	<pre>QSUBSENDUID</pre> <p><i>Torque assigns a unique ID to a job when it is submitted by qsub.</i></p>

QSUBSLEEP

Format	<INT>
Default	0
Description	Specifies time, in seconds, to sleep between a user's submitting and Torque's starting a <i>qsub</i> command. Used to prevent users from overwhelming the scheduler.
Example	<pre>QSUBSLEEP 2</pre> <p><i>When a job is submitted with <i>qsub</i>, it will sleep for 2 seconds.</i></p>

RERUNNABLEBYDEFAULT

Format	<BOOLEAN>
Default	TRUE
Description	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 qsub -r option. (New in Torque 2.4.)
Example	<pre>RERUNNABLEBYDEFAULT FALSE</pre> <p><i>By default, <i>qsub</i> jobs cannot be rerun.</i></p>

SERVERHOST

Format	<STRING>
Default	localhost
Description	If set, the " qsub " on page 275 command will open a connection to the host specified by the SERVERHOST string.
Example	<pre>SERVERHOST orion15</pre> <p><i>The server will open socket connections and and communicate using serverhost orion15.</i></p>

SUBMITFILTER	
Format	<STRING>
Default	/usr/local/sbin/torque_submitfilter
Description	Specifies the location of the submit filter (see "Job Submission Filter ("qsub Wrapper")" on page 381 used to pre-process job submission.
Example	<pre>SUBMITFILTER /usr/local/sbin/torque_submitfilter</pre> <p><i>The location of the submit filter is specified as /usr/local/sbin/torque_submitfilter.</i></p>

TRQ_IFNAME	
Format	<STRING>
Default	null
Description	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 <i>eth0</i> or <i>eth1</i> , depending on which interface you want to use.
Example	<pre>TRQ_IFNAME eth1</pre> <p><i>Outbound Torque requests are handled by eth1.</i></p>

VALIDATEGROUP	
Format	<BOOLEAN>
Default	FALSE
Description	Validate submit user's group on qsub commands. For Torque builds released after 2/8/2011, <i>VALIDATEGROUP</i> also checks any groups requested in <i>group_list</i> at the submit host. Set <i>VALIDATEGROUP</i> to "TRUE" if you set disable_server_id_check to TRUE.
Example	<pre>VALIDATEGROUP TRUE</pre> <p><i>qsub verifies the submitter's group ID.</i></p>

VALIDATEPATH	
Format	<BOOLEAN>
Default	TRUE
Description	Validate local existence of <code>-d</code> and/or <code>-w</code> working directories.
Example	<div>VALIDATEPATH FALSE</div> <div><i>qsub does not validate the path.</i></div>

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 6.0.4 6.0.4
```

To download a different version, replace each 6.0.4 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.

i 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
```

i OSX 10.4 users need to change the `#define __TDARWIN` in `src/include/pbs_config.h` to `#define __TDARWIN_8`.


i 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`.

i 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-x86_64.sh main:/tmp/; psh main /tmp/torque-package-linux-x86_64.sh --install`). Optionally, distribute and install the clients package.

Configure Torque on the Compute Nodes

- For each compute host, you must configure the MOM daemon to trust the `pbs_server` host. The recommended method for doing this is to create the `TORQUE_HOME/server_name` file with the server hostname in it. Alternatively, you may add a `$pbsserver` line to the `TORQUE_HOME/mom_priv/config` file. (Versions older than 2.0.op5 require this method).
- Additional config parameters may be added to `TORQUE_HOME/mom_priv/config` (see [Node Manager \(MOM\) Configuration on page 327](#) for details).
- See [Specifying Compute Nodes on page 22](#) for more information about configuring `pbs_server` to identify compute nodes.

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 `TORQUE_HOME/server_priv/nodes` file:

```
server_priv/nodes
computenode001.cluster.org
computenode002.cluster.org
computenode003.cluster.org
```

Start the pbs_mom Daemons on Compute Nodes

- Next start the `pbs_mom` daemon on each compute node by running the `pbs_mom` executable.

Run the `trqauthd` daemon to run client commands (see ["Configuring trqauthd for Client Commands" on page 27](#)). This enables running client commands.

Verify Correct Torque Installation

The `pbs_server` daemon was started on the Torque server when the `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.

Enable the Scheduler

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 using Maui/Moab, see **Moab-Torque Integration Guide** on page 1 in the *Moab Workload Manager Administrator Guide*. If using `pbs_sched`, start this daemon now.

i 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 28](#)

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" below](#)
- ["Test 1 - Basic Operation" below](#)
- ["Test 2 - Persistence of Checkpoint Images" on page 396](#)
- ["Test 3 - Restart After Checkpoint" on page 396](#)
- ["Test 4 - Multiple Checkpoint/Restart" on page 397](#)
- ["Test 5 - Periodic Checkpoint" on page 397](#)
- ["Test 6 - Restart from Previous Image" on page 398](#)

Test Environment

All these tests assume the following test program and shell script, `test.sh`.

```
#include
int main( int argc, char *argv[] )
{
    int i;

    for (i=0; i<100; i++)
    {
        printf("i = %d\n", i);
        fflush(stdout);
        sleep(1);
    }
}
#!/bin/bash

/home/test/test
```

Related Topics

[" BLCR Acceptance Tests" above](#)

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.needsnodes = 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/s
bin:/usr/local/bin:/usr/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 393](#)

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 393](#)), 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 393](#)

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 393](#)), 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 393](#)

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.

```
> 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 393](#)

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.

```
> 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 393](#)

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 393](#)

Queue Attributes

This appendix provides information on the different queue attributes.

This appendix also lists some queue resource limits. See [Assigning Queue Resource Limits on page 409](#).

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

Attributes

"acl_groups" below
 "acl_group_enable" on the next page
 "acl_group_sloppy" on the next page
 "acl_hosts" on the next page
 "acl_host_enable" on page 401
 "acl_logic_or" on page 401
 "acl_users" on page 401
 "acl_user_enable" on page 402

"disallowed_types" on page 402
 "enabled" on page 402
 "features_required" on page 402
 "ghost_queue" on page 403
 "keep_completed" on page 403
 "kill_delay" on page 403
 "max_queueable" on page 404
 "max_running" on page 404

"max_user_queueable" on page 404
 "max_user_run" on page 405
 "priority" on page 405
 "queue_type" on page 405
 "req_information_max" on page 406
 "req_information_min" on page 406
 "required_login_property" on page 406
 "resources_available" on page 407

"resources_default" on page 407
 "resources_max" on page 408
 "resources_min" on page 408
 "route_destinations" on page 408
 "started" on page 409

acl_groups

Format <GROUP>[@<HOST>][+<USER>[@<HOST>]]...

Default ---

Description 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.

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

acl_groups

Example

```
> qmgr -c "set queue batch acl_groups=staff"  
> qmgr -c "set queue batch acl_groups+=ops@h1"  
> qmgr -c "set queue batch acl_groups+=staff@h1"
```



Used in conjunction with [acl_group_enable](#).

acl_group_enable

Format

<BOOLEAN>

Default

FALSE

Description

If TRUE, constrains Torque to only allow jobs submitted from groups specified by the [acl_groups](#) parameter.

Example

```
qmgr -c "set queue batch acl_group_enable=true"
```

acl_group_sloppy

Format

<BOOLEAN>

Default

FALSE

Description

If TRUE, [acl_groups](#) will be checked against all groups of which the job users is a member.

Example

acl_hosts

Format

<HOST>[+<HOST>]...

Default

Description

Specifies the list of hosts that may submit jobs to the queue.

Example

```
qmgr -c "set queue batch acl_hosts=h1+h1+h1"
```



Used in conjunction with [acl_host_enable](#).

acl_host_enable

Format <BOOLEAN>

Default FALSE

Description If TRUE, constrains Torque to only allow jobs submitted from hosts specified by the [acl_hosts](#) parameter.

Example

```
qmgr -c "set queue batch acl_host_enable=true"
```

acl_logic_or

Format <BOOLEAN>

Default FALSE

Description 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.

Example

```
qmgr -c "set queue batch acl_logic_or=true"
```

acl_users

Format <USER>[@<HOST>][+<USER>[@<HOST>]]...

Default ---

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@h1"
> qmgr -c "set queue batch acl_users+=stevek@h1"
```

 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	<pre>qmgr -c "set queue batch acl_user_enable=true"</pre>

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, rerunable, nonrerunable, 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	<pre>qmgr -c "set queue batch disallowed_types = interactive" qmgr -c "set queue batch disallowed_types += job_array"</pre>

enabled

Format	<BOOLEAN>
Default	FALSE
Description	Specifies whether the queue accepts new job submissions.
Example	<pre>qmgr -c "set queue batch enabled=true"</pre>

features_required

Format	feature1[,feature2[,feature3...]]
---------------	-----------------------------------

features_required	
Default	---
Description	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.
Example	<pre>qmgr -c 's q batch features_required=fast'</pre>

ghost_queue	
Format	<BOOLEAN>
Default	FALSE
Description	Intended for automatic, internal recovery (by the server) only. If set to TRUE, the queue rejects new jobs, but permits the server to recognize the ones currently queued and/or running. Unset this attribute in order to approve a queue and restore it to normal operation. See "Automatic Queue and Job Recovery" on page 187 for more information regarding this process.
Example	<pre>qmgr -c "unset queue batch ghost_queue"</pre>

keep_completed	
Format	<INTEGER>
Default	0
Description	Specifies the number of seconds jobs should be held in the Completed state after exiting. For more information, see "Keeping Completed Jobs" on page 84 .
Example	<pre>qmgr -c "set queue batch keep_completed=120"</pre>

kill_delay	
Format	<INTEGER>
Default	2

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.



This setting overrides the server setting. See `kill_delay` in [Server Parameters on page 299](#).

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

Description	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.
--------------------	---

Example	<pre>qmgr -c "set queue batch max_user_queuable=20"</pre>
----------------	---

max_user_run

Format	<INTEGER>
---------------	-----------

Default	unlimited
----------------	-----------

Description	This limits the maximum number of jobs a user can have running from the given queue.
--------------------	--

Example	<pre>qmgr -c "set queue batch max_user_run=5"</pre>
----------------	---

priority

Format	<INTEGER>
---------------	-----------

Default	0
----------------	---

Description	Specifies the priority value associated with the queue.
--------------------	---

Example	<pre>qmgr -c "set queue batch priority=20"</pre>
----------------	--

queue_type

Format	One of <i>e</i> , <i>execution</i> , <i>r</i> , or <i>route</i> (see "Creating a Routing Queue" on page 118)
---------------	---



Default	---
----------------	-----



Description	Specifies the queue type.
--------------------	---------------------------



This value must be explicitly set for all queues.

Example	<pre>qmgr -c "set queue batch queue_type=execution"</pre>
----------------	---

req_information_max	
Format	<STRING>
Default	---
Description	<p>Specifies the maximum resource limits allowed for jobs submitted to a queue.</p> <div>  These limits apply only to the qsub -L job submission option. </div> <p>Valid values are lprocs, node, socket, numachip, core, thread, memory, swap, and disk.</p> <div>  If a maximum core count is specified, jobs with usecores must have lprocs<= the maximum core count; jobs without usecores are rejected. If a maximum thread count is specified, lprocs must be <= the maximum thread count. </div>
Example	<pre>qmgr -c "set queue batch req_information_max.lprocs=8"</pre>

req_information_min	
Format	<STRING>
Default	---
Description	<p>Specifies the minimum resource limits allowed for jobs submitted to a queue.</p> <div>  These limits apply only to the qsub -L job submission option. </div> <p>Valid values are lprocs, node, socket, numachip, core, thread, memory, swap, and disk.</p> <div>  If a minimum core count is specified, jobs with usecores must have lprocs>= the minimum core count; jobs without usecores are rejected. If a minimum thread count is specified, lprocs must be >= the minimum thread count. </div>
Example	<pre>qmgr -c "set queue batch req_information_min.lprocs=2"</pre>

required_login_property	
Format	<STRING>


required_login_property	
Default	---
Description	Adds the specified login property as a requirement for all jobs in this queue.
Example	<pre>qmgr -c 's q <queue name> required_login_property=INDUSTRIAL'</pre>

resources_available	
Format	<STRING>
Default	---
Description	Specifies to cumulative resources available to all jobs running in the queue. See "qsub will not allow the submission of jobs requesting many processors" on page 198 for more information.
Example	<pre>qmgr -c "set queue batch resources_available.nodect=20"</pre> <div> <p>i You must restart pbs_server for changes to take effect.</p> <p>Also, resources_available is constrained by the smallest of queue.resources_available and server.resources_available.</p> </div>

resources_default	
Format	<STRING>
Default	---
Description	Specifies default resource requirements for jobs submitted to the queue.
Example	<pre>qmgr -c "set queue batch resources_default.walltime=3600"</pre> <div> <p>i See Setting Queue Resource Controls with Resource Request Syntax 2.0 on page 116 for more information about setting queue resource requirements and the use of <code>-l</code> and <code>-L</code> job submission syntaxes.</p> </div>

resources_max	
Format	<STRING>
Default	---
Description	Specifies the maximum resource limits for jobs submitted to the queue.
Example	<pre>qmgr -c "set queue batch resources_max.nodect=16"</pre>


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

route_destinations	
Format	<queue>[@<host>]
Default	---
Description	<p>Specifies the potential destination queues for jobs submitted to the associated routing queue.</p> <div>  This attribute is only valid for routing queues (see "Creating a Routing Queue" on page 118). </div>
Example	<pre>> 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"</pre> <p>To set multiple queue specifications, use multiple commands:</p> <pre>> qmgr -c 's s route_destinations=batch' > qmgr -c 's s route_destinations+=long' > qmgr -c 's s route_destinations+=short'</pre>

started	
Format	<BOOLEAN>
Default	FALSE
Description	Specifies whether jobs in the queue are allowed to execute.
Example	<pre>qmgr -c "set queue batch started=true"</pre>

Assigning Queue Resource Limits

Administrators can use resources 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*, *ncpus*, *nodect*, *nodes*, *pmem*, *procct*, *pvmem*, *vmem*, and *walltime*.

Resource	Format	Description
arch	string	Specifies the administrator defined system architecture required.
mem	size	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.)
ncpus	integer	Sets the number of processors in one task where a task cannot span nodes. <div> You cannot request both ncpus and nodes in the same queue.</div>
nodect	integer	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. <i>nodect</i> 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.
nodes	integer	Specifies the number of nodes.
pmem	size	Specifies the maximum amount of physical memory to be used by any single process of the job. (Ignored on Fujitsu. Not implemented on Digital Unix and HPUX.)

Resource	Format	Description
procct	integer	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 procct of 16. $2*2$ (2:ppn=2) + $3*4$ (3:ppn=4).
pvmem	size	Amount of virtual memory used by any single process in a job.
vmem	size	Amount of virtual memory used by all concurrent processes in the job.
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.

Suffix		Multiplier
b	w	1
kb	kw	1024
mb	mw	1,048,576
gb	gw	1,073,741,824
tb	tw	1,099,511,627,776

Related Topics

[Queue Configuration on page 115](#)