TORQUE Tutorial

A Beginner's Guide

Kenneth Nielson

September 16, 2009

TORQUE Resource Manager

What is TORQUE
TORQUE's Role
TORQUE Components
Installation
Configuration
Job Administration
Diagnostics
MPI
Multi-mom and Any mom
Roadmap
Q&A

What is TORQUE?

. Terascale Open-Source Resource and QUEue Manager

. TORQUE is an open source resource manager providing control over batch jobs and distributed compute nodes. It is a community effort based on the original *PBS project and, with more than 1,200 patches, has incorporated significant advances in the areas of scalability, fault tolerance, and feature extensions contributed by NCSA, OSC, USC , the U.S. Dept of Energy, Sandia, PNNL, U of Buffalo, TeraGrid, and many other leading edge HPC organizations.

. *PBS – Portable Batch System*

What is TORQUE

The Portable Batch System, PBS, is a batch job and computer system resource management package. It was developed with the intent to be conformant with the POSIX 1003.2d Batch Environment Standard. As such, it will accept batch jobs, a shell script and control attributes, preserve and protect the job until it is run, run the job, and deliver output back to the submitter. PBS may be installed and configured to support jobs run on a single system, or many systems grouped together. Because of the flexibility of PBS, the systems may be grouped in many fashions.

TORQUE's Role

. Provide job queuing facility
. Monitor resource configuration, utilization, and health
. Provide remote job execution and job management facilities
. Reports information to cluster scheduler
. Receives direction from cluster scheduler
. Handles user client requests

TORQUE Components

Commands

Job Server

Job Executor

Job Scheduler

TORQUE Components

Commands

- Three classes of commands
  - user – any authorized user can execute
  - Operator – special access privileges required
  - Manager – special access privileges required

- User commands
  - qsub, qstat, pbsnodes, qdel

TORQUE Components

Job Server

- pbs_server
  - Central focus of TORQUE
  - All commands and other daemons communicate with pbs_server via TCP/IP and UDP/IP
  - Provides basic batch services
    - Job creation
    - Job modification
    - Job protection
    - Job execution

TORQUE Components

Job Executor

. pbs_mom
- Daemon called MOM – Machine-Oriented Miniserver
- receives copy of jobs from pbs_server
- Places jobs into execution
- Creates new session similar to user login session
- For parallel jobs a Mother Superior manages group of sister nodes
- Returns output to pbs_server or Mother Superior

TORQUE Components

Job Scheduler

- Controls site policy
- TORQUE supports multiple schedulers
  - pbs_sched
    - not supported by Adaptive Computing
  - Maui
    - Open source
    - User Group support only
  - Moab
    - Torque support included
    - For what Moab can do that Maui cannot go to **http://www.clusterresources.com/products/maui/docs/a.kmoabcomp.shtml**

TORQUE Installation

Where to get it.

svn (subversion)

svn://svn.clusterresources.com/torque

/trunk – currently 2.4 beta

/branches/2.3-fixes – snapshot build with latest fixes

/branches/2.3-multimom – allows multiple moms on a single node

www.clusterresources.com

http://www.clusterresources.com/downloads/torque/

torque-2.3.7.tar.gz is the latest released version

TORQUE Installation

Extract and build the distribution to the machine that will act as the TORQUE server.

```
> tar –xzvf torqueXXX.tar.gz
> cd torqueXXX
> ./configure
> make
> make install
```

TORQUE Installation

Torque Install Directory

. Default location /usr/local/

- bin
  . Contains client commands – qstat, pbsnodes, qsub, etc.
  . Needed on server and login/submission hosts

- sbin
  . Contains server and node daemons – pbs_server, pbs_mom, pbs_demux, pbs_sched, momctl

- lib
  . Contains TORQUE libraries – libtorque.so.x

TORQUE Installation

Initial TORQUE Startup

pbs_server

As root type
pbs_server -t create
or
torque.setup < user>

Stop pbs_server before running in production
qterm

TORQUE Installation

```
root@ken-linuxBox:/usr/local/sbin# pbs_server -t create

  Qmgr: p s
  #
  # Set server attributes.
  #
  set server acl_hosts = ken-linuxBox
  set server log_events = 511
  set server mail_from = adm
  set server scheduler_iteration = 600
  set server node_check_rate = 150
  set server tcp_timeout = 6
```

Adaptive COMPUTING

TORQUE Installation

```
ken@ken-linuxBox:~ /dev/torque/2.3-fixes$ sudo ./torque.setup ken

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 = ken-linuxBox
set server default_queue = batch
set server log_events = 511
set server mail_from = adm
set server scheduler_iteration = 600
set server node_check_rate = 150
set server tcp_timeout = 6
set server mom_job_sync = True
set server keep_completed = 300
```

Adaptive COMPUTING

TORQUE Configuration

TORQUE Home Directory

- Default /var/spool/torque -- $TORQUE_HOME, $PBS_HOME, etc.
  - /var/spool/torque
    - server_name – Name of host where pbs_server resides. Can have multiple host names for high availability
  - server_priv
    - jobs
    - nodes
  - server_logs
    - files of the form yyyymmdd (i.e. 20090916)
  - mom_priv
    - jobs
    - config
  - mom_logs
    - files of the form yyyymmdd (i.e. 20090916)

TORQUE Configuration

pbs_server Configuration -- nodes file

- server_priv/nodes
  - contains list of mom host names and attributes
    - attributes
      - np – number of processes
      - note – administrator note
      - properties – administrators choice
- nodes file syntax
  - host np= X note= string property1 property2...propertyn
  - example:
    - host1 np= 4 note= new intel_i7 data
    - host2 np= 4 x86
    - host3 np= 8 amd_64

TORQUE Configuration

pbs_server node configuration

- Restart pbs_server
- Run pbsnodes

```
host1
    state = down
    np = 4
    properties = intel_i7,data
    ntype = cluster
    note = new

host2
    state= down
    np= 4
```

Adaptive COMPUTING

TORQUE Configuration

pbs_server node configuration

- Dynamic node configuration
  - > qmgr -c “create node node003”

Manually edit the nodes file

- *$TORQUEHOME/server_priv/nodes*
- Restart pbs_server daemon after change

TORQUE Configuration

- **pbs_server queue configuration**
  - Attributes
    - queue_type
      - execution, route
    - resources_default
      - default resource requirements for jobs (walltime, nodes)
    - enabled
      - Specifies whether queue accepts new jobs. (Default FALSE)
    - started
      - specifies whether jobs in queue are allowed to execute. (Default Fales)

TORQUE Configuration

- **pbs_server queue configuration**
  - default queue batch
  - create new queue
    - qmgr
      - create queue reg
      - set queue reg queue_type= Execution
      - set queue reg resources_default.node= 1
      - set queue reg resources_default.walltime= 01:00:00
      - set queue reg enabled= True
      - set queue reg started= True
  - setting default queue
    - qmgr -c “set server default_queue= reg”

Note: A queue is called a class in Moab

TORQUE Configuration

pbs_mom Configuration

- As root run pbs_mom
  - No special configuration needed to start
  - use mom_priv/config for options

- mom_priv/config
  - Allows custom configuration of mom node
  - Syntax
    - $<option> value
    - example

```
$loglevel 3
$usecp *.fte.com:/data /usr/local/data
```

TORQUE Configuration

- For shared filesystems use the $usecp parameter in the mom_priv/config file

```
$usecp *.fte.com:/data /usr/local/data
```

- For local, non-shared filesystems, rcp or scp must be configured to allow direct copy without prompting for passwords (key authentication, etc.)

http://www.clusterresources.com/products/torque/docs/6.1scpsetup.shtml

TORQUE Configuration

Scheduler Configuration

- Follow directions for scheduler of choice
- Moab configuration
  - http://www.clusterresources.com/products/mwm/docs/2.0installation.shtml

Advanced Configuration

Customizing the Install

- Most recommended configure options have been selected as default.
- Some often used options
  - --with-debug – for use with gdb
  - --prefix=<DIR> -- change install directory
  - --exec-prefix=<DIR> -- change only executable install directory
  - --disable-gcc-warnings – Use with care.
- ./configure --help will give all options

Advanced Configuration

- Configuring Job Submission Hosts
  - Use acl_hosts
  - Use torque.cfg submithosts,allowcomputehosts
  - /etc/hosts.equiv
- Configuring TORQUE on a Multi-Homed Server
- Specifying Non-Root Administrators

```
> qmgr

Qmgr: set server managers += josh@*.fsc.com
Qmgr: set server operators += josh@*.fsc.com
Qmgr: set server log_level= 3
```

Job Administration

Job Flow

- pbs_server receives new job
- Informs the scheduler
- When nodes available, scheduler sends instructions and nodeslist to pbs_server
- pbs_server sends job to the first node in the nodelist
- The first node, or Mother Superior, launches the job and passes it to the rest of the nodes in the nodelist, or the Sister moms
-

Job Administration

qsub

- Batch and Interactive
- Requesting Resources

Examples

- To ask for 2 processors on each of four nodes:
  - qsub -l nodes=4:ppn=2
- The following job will wait until node01 is free with 200 MB of available memory:
  - qsub -l nodes=node01,mem=200mb /home/user/script.sh

Directives can be embedded into job script

- example on next page

Job Administration

```
# !/bin/sh

# PBS -N ds14FeedbackDefaults
# PBS -q testqueue
# PBS -l nodes=1:ppn=2,walltime=240:00:00
# PBS -M user@mydomain.com


source ~/.bashrc

cat $PBS_NODEFILE
cat $PBS_O_JOBID
```

Job Administration

Manually Administrating Jobs

```
> qsub scatter
4807.ken-linuxbox


> qstat

Job id              Name             User             Time Use S Queue
----------------- ---------------- ---------------- -------- - -----
4807                scatter          user01           12:56:34 Q batch
```

Adaptive COMPUTING

Job Administraton

Manually Administrating Jobs

```
> qrun 4807

> qstat

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

>qstat

Job id              Name             User             Time Use S Queue
----------------- ---------------- ---------------- -------- - -----
4807                scatter          user01           12:56:34 C batch
```

Adaptive COMPUTING

Job Administration

Canceling Jobs

qdel

-w *delay*
Specify the delay between the sending of the SIGTERM and SIGKILL signals.

-p *purge*
Forcibly purge the job from the server. This option is only available to a batch operator or the batch administrator.

-m *message*
Specify a comment to be included in the email. The argument message specifies the comment to send. This option is only available to a batch operator or the batch administrator.

[all|ALL]

Delete all jobs in the queue

Job Administration

Automating Job Administration

Integrate with an external scheduler

Moab Workload Manager

Job Arrays

submit multiple jobs at once

Submit Filters

Job Preemption

Job Administration

- Job Arrays
  - TORQUE 2.3 and later
  - Allows single line submission of multiple jobs for a single script
  - Job can be monitored as a group

Example

```
> qsub -t 0-3 scatter
  33.hostname
> qstat
Job id            Name             User             Time Use S Queue
----------------- ---------------- ---------------- -------- - -----
33-0              scatter-0        user01           12:56:34 R batch
33-1              scatter-1        user01           12:56:34 R batch
33-2              scatter-2        user01           12:56:34 R batch
```

Job Administration

Submit Filters

When submit filters exist TORQUE sends command file to the script/executable which modifies the request based on site policies.

Submit filter designated in torque.cfg.
- Found in /var/spool/torque
- Keyword SUBMITFILTER

Example torque.cfg
- SUBMITFILTER /home/user/submit_filter

Job Administration

Submit Filter Examples

/home/user/submit_filter

```
# !/bin/sh

#  add default memory constraints and add a e-mail notification address to all
requests
#  that did not specify it in user's script or command line

echo “# PBS -l mem=16MB”
echo “# PBS -M ken@adaptivecomputing.com”

while read I
  do
    echo $i
  done
```

Job Administration

Submit Filter Examples

listtest.sh

```
# !/bin/sh
ls -alR /
```

```
qsub listtest.sh
10.kmn.cridomain
```

```
cat /var/spool/torque/server_priv/jobs/10.kmn.cridomain.SC
```

```
# PBS -l mem=16MB
# PBS -M ken@adaptivecomputing.com
ls -alR /
```

TORQUE Administration

Job Preemption

Torque has three basic tools

- Cancel – qdel
- re-que – qrerun
- checkpoint

The scheduler uses the basic tools to enable job preemption.
See Moab for more information

http://www.clusterresources.com/products/mwm/docs/8.4preemption.shtml

TORQUE Administration

Monitoring Resources

TORQUE reports a number of attributes broken into 3 major categories:

Configuration

- Includes both detected hardware configuration, and specified batch attributes
- Can report static 'generic resources' via specification in the mom config file

Utilization

- Includes information regarding the amount of node resources currently available (in use) as well as information about who or what is consuming it
- Can report dynamic 'generic resources' via specification of a 'monitor script' in the mom config file

State

- Includes administrative status, general node *health* information, and general usage status

TORQUE Adminstration

Monitoring Resources

```
> pbsnodes

ken-linuxBox
     state = free
     np = 2
     properties = bldg1,intel_i7
     ntype = cluster
     status = opsys=linux,uname=Linux ken-linuxBox 2.6.24-23-
    generic #1 SMP Wed Apr 1 21:47:28 UTC 2009
    i686,sessions=4983 5873 6220 6331 6335 6360 6369 6402
    6456 6460 6489 6582, nsessions=12, nusers=2, idletime=1,
    totmem=8123824kb, availmem=7584648kb,
    physmem=2067360kb, ncpus=2,loadave=0.05,
    netload=36957532, state=free, jobs=,varattr=,
    rectime=1252467787
       note = backed_up
```

TORQUE Administration

Node States

States
- down (down)
- offline (drained)
- job-exclusive (busy)
- free (idle/running)
- reserve
- job-sharing
- busy
- time-shared
- state-unknown

Changing node state
- Offline
  - pbsnodes -o < nodename>
- Online
  - pbsnodes -c < nodename>

Viewing nodes of a particular state
- pbsnodes -l

TORQUE Administration

Node Properties

- Node Property Attributes
  - Can apply multiple properties per node
  - Properties are 'opaque'
  - Each property can be applied to multiple nodes
  - Properties can not be consumed

- Dynamically with qmgr

  > qmgr -c "set node node001 properties= bigmem"

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

  - Manually edit server_priv/nodes file
    - always restart pbs_server after modifying nodes file

TORQUE Administration

Accounting Records

- Torque maintains accounting records of jobs in server_priv/accounting
- file of the form yyyymmdd
-

Record Marker	Record Type	Description
D	delete	Job was deleted
E	exit	Job has exited (successfully or unsuccessfully)
Q	queue	Job has been submitted/queued
S	start	an attempt to start the job has been made (if the job fails to properly start, it may have multiple job start records)

- 09/08/2009 22:15:58;Q;9.ken-linuxbox;queue= batch

Diagnostics

Log Files

pbs_server log files
/var/spool/torque/server_logs
qmgr: set server log_level= x

pbs_mom log files
/var/spool/torque/mom_logs
/var/spool/torque/mom_priv/config
$loglevel x

Diagnotics

MOM Diagnostics

momctl

- Diagnoses mom configuration and communication with server
- -d3 option
- Output on next slide

Diagnostics

```
Host: ken-linuxBox/ken-linuxbox   Version: 2.3.8   PID: 12792
Server[0]: ken-linuxBox (127.0.1.1:15001)
  Init Msgs Received:     0 hellos/1 cluster-addrs
  Init Msgs Sent:         1 hellos
  Last Msg From Server:   8 seconds (StatusJob)
  Last Msg To Server:     15 seconds
HomeDirectory:          /var/spool/torque/mom_priv
stdout/stderr spool directory: '/var/spool/torque/spool/' (110542371 blocks available)
NOTE:  syslog enabled
MOM active:             153 seconds
Check Poll Time:        45 seconds
Server Update Interval: 45 seconds
LogLevel:               0 (use SIGUSR1/SIGUSR2 to adjust)
Communication Model:    RPP
MemLocked:              TRUE  (mlock)
TCP Timeout:            20 seconds
Prolog:                 /var/spool/torque/mom_priv/prologue (disabled)
Alarm Time:             0 of 10 seconds
Trusted Client List:    127.0.1.1,127.0.0.1
Copy Command:           /usr/bin/scp -rpB
job[12.ken-linuxbox]  state= RUNNING  sidlist= 12830
Assigned CPU Count:     1

diagnostics complete
```

MPI

MPI (Message Passing Interface)

- Used for parallel jobs
- Augments communication between tasks distributed across cluster
- TORQUE can run with any MPI library
- TORQUE provides limited integration with some MPI libraries
- MPI packages
  - MPICH – Argonne National Lab
  - MPICH-VMI – NCSA
  - Open MPI

MPI

MPIExec Overview

- Replacement for mpirun script
- Initializes a parallel job with a PBS batch or interactive environment
- Uses task manager library of PBS to spawn copies of executable on nodes
- TM interface faster than invoking separate rsh (mpirun)
- Resources used by spawned process accounted correctly with mpiexec
- Tasks that exceed assigned limits (walltime, memory, disk space) are killed
- mpiexec can enforce a security policy. Obviates use of rsh or ssh

See mpiexec home page for more information.
http://www.osc.edu/~djohnson/mpiexec/index.php

Multi-Mom

- Multiple pbs_mom daemons on a single node
- Intended to enhance testing but possible to use in production
- Moms uniquely identified by name and ports
- Default pbs_mom ports
  - 15002
  - 15003
- Use alias in /etc/hosts
  - 192.168.0.10 myhost myhost1 myhost2
  - max alias names?

Multi-Mom

Invoking multi-mom

- syntax – pbs_mom -m -M 30002 -R 30003
- modify nodes file
  - node1 np= 2
  - node2 np= 2 mom_service_port= 30002 mom_manager_port= 30003

- stopping multi-mom
  - momctl -s -p 30003

Any-mom

- Enables any mom node to join a cluster without having an entry in the server_priv/nodes file.
-
- Syntax
- pbs_server -e
-
- Can dynamically add moms to cluster without restarting pbs_server
-
- Creates security issues
- cannot control who joins the cluster
- need outside security policy

TORQUE Roadmap

TORQUE 2.3.8
- Bug fixes only

TORQUE 2.4
- Complete 2.3-fixes merge
- CPU affinity (very basic implementation)
- Multi-mom
- Any mom

TORQUE 2.5
- TORQUE testing framework
- Eliminate need for privileged ports
- CPUsets improvements
- Improve TORQUE HA

TORQUE 3.0
- Alternate communcation model between pbs_server, MOMs and sisters
- scaleabilty for super large systems with large MPI jobs (10,000+ nodes)

Adaptive COMPUTING

TORQUE Q&A