Production Experiences with the Cray-Enabled TORQUE Resource Manager

Matt Ezell and Don Maxwell
HPC Systems Administrator
Oak Ridge National Laboratory

David Beer
Senior Software Engineer
Adaptive Computing

CUG 2013
May 8, 2013
Napa Valley, CA
Resource Managers on Cray Systems

• The largest systems in the world constantly face issues only seen at extreme scale

• Cray has a local resource manager called ALPS that batch systems must interface with
Cray ALPS

• Stands for “Application Layer Placement Scheduler”
• Maintains System Inventory
  – CPUs
  – Memory
  – Accelerators
• Tracks node state, mode, and reservations
• “Scheduler”, daemons, and client tools
• XML API called BASIL
  – Versioned to allow new features without breaking old software
ALPS High-Level Design

SMW Node
- erd

Boot Node
- apwatch
- apbridge

Moab Node
- Moab
- pbs_server

SDB Node
- apsched
- Shared Files

Login/Batch
- pbs_mom
- apbasil
- apsys
- apstat
- aprun

User Shell

Compute Node
- apinit
- apsheperd
- PEs

Compute Node
- apinit
- apsheperd
- PEs
Previous Moab/ALPS integration

- Moab would talk directly to ALPS
  - Had to run Moab on the Cray
  - Cray crashed, TORQUE/Moab went away
  - Moab used a “native” perl interface

- TORQUE had to talk to ALPS also
  - When confirming reservations

- What if they got out of sync?
New Model Overview

- Now pbs_moms are the only nodes inside of the Cray
- Moab and pbs_server can be outside the Cray (but don't have to be)
  - This allows for HA and/or using larger, faster nodes if desired/needed
- From Moab's perspective, the Cray is just a normal cluster
New Model

Moab

pbs_server

Status of Login Node

ALPS Inven

Normal Status

reporter mom (on the scb)

login mom

login mom

Actual ALPS nodes

login mom

login mom

login mom

login mom
Getting Resource Information

Resource Query

- Moab
- pbs_server
- pbs_mom (reporter mom)
- apbash

NOTE: pbs_server creates a compute node for each node reported by the reporter mom.

: MRMClusterQuery
    - nodelist

: create reservation

Parses XML to create a status update for each node in the inventory.

- inventory request
  - inventory XML
Job Start

Moab

MJobStart

: pbs_runjob
  compute node list

pbs_server

: start job

pbs_server selects a login mom for the job. If it is interactive it will choose the submit host by default. Otherwise, it chooses the login according to a round robin.

pbs_mom (login)

: create reservation nodelist

Store reservation id

: confirm reservation

execute script

apbasil
Job Termination

ALPS

pbs_mom (login)

pbs_server

Moab

Receives a SIGCHLD on the termination of the job’s script.

releases ALPS reservation

NOTE: retries until permanent failure to ensure release.

receives permanent failure

obit to pbs_server

Cluster Query

status of jobs
Release Orphaned Reservation

Checks reservation ids to ensure that there is an active job for each reservation.
Early Work

- Adaptive visited ORNL in June of 2012 for an early beta
- Minor issues discovered
- Beta version left running on 2 test/development systems
Previous NCRC Moab/TORQUE Setup

- Moab01
  - ES TORQUE
  - C1MS Moab
    - C1MS TORQUE
  - T1MS Moab
    - T1MS TORQUE

- Moab02
  - ES TORQUE
  - C2 Moab
    - C2 TORQUE
  - T1 Moab
    - T1 TORQUE
New NCRC Moab/TORQUE Setup
Early Experiences on Gaea c1

• Moved to new version in July 2012
• Hit some fairly major problems that impacted acceptance
• Most difficulties stemmed from bug in features that had nothing to do with Cray
  – Missing PBS_O_\* environment variables
  – Broken environment parsing
  – Multi-threading improvements would sometimes deadlock
  – X11 forwarding didn’t work correctly
• But some Cray-specific bugs also
  – Restarting pbs_server would dump running jobs
  – Unable to delete jobs
INTRODUCING TITAN
Advancing the Era of Accelerated Computing
System Layout

moab1
  moab  pbs_server

dtn-sch1  pbs_mom

dtn-sch2  pbs_mom

dtn-sch3  pbs_mom

PBS Server

sys0

batch1

batch2

batch8

Titan

login1

login2

login8

...
Early Experiences on Titan

• Moved to new architecture in September 2012
• Primary issues has been deadlocks
  – Scripts developed to detect, analyze, and mitigate
  – Many improvements; architectural changes to help
• Problem with submitting jobs when the Cray was down
  – Problem found and fixed
• Two security vulnerabilities discovered
  – Problems fixed and patched
Externalizing TORQUE and Moab

More powerful server hardware

Submit jobs while system is down

Decreased Complexity

Better User Experience
Recent Issues

• ‘Non-digit found where digit expected’ message
  – Patch developed and landed, not running yet

• ‘Invalid Credential’ message
  – Fix upstream, running on Gaea

• Re-used resIDs marked as orphaned
  – Fix upstream, running on Gaea

• Poor interaction with NHC leading to failed jobs
  – Fix upstream, running on Gaea

• ALPS Reservation failures cause jobs to abort
  – Now they requeue, running on Gaea
Recent Changes

• TORQUE 4.2 moved to a C++ compiler
  – Stronger type checking
  – New language constructs
  – Ability to leverage STL

• Emphasis on unit tests and code coverage
  – Should improve quality and avoid bugs over time

• Code moved to GitHub
  – More transparency
  – Improved community involvement
Future Work

• Improvements on large job launch
  – Lots of time spent on internal job ⇔ node bookkeeping and generating the hostlists

• Hostlist compression

• BASIL 1.3 support
  – Adds additional thread placement granularity (especially helpful on XC30 hardware)

• Evaluating event-based ALPS updates
Conclusions

- New TORQUE/ALPS interaction is more straightforward
- Externalizing TORQUE/Moab has improved the user experience
- TORQUE and Moab are now working well on Gaea and Titan
- Overall TORQUE codebase is improving
Questions?

Lunch BOF Tomorrow

ezellma@ornl.gov  mii@ornl.gov  dbeer@adaptivecomputing.com