Unifying Heterogeneous Cray Resources and Systems into an Intelligent Single-scheduled Environment

Scott Jackson – Engineering
Overview

- Introduction
- Heterogeneous Resources
- Disparate Systems
- Leadership Sites and Moab
- Additional Benefits
- Q&A
Introduction
Introduction

- Manage Life Cycle of Cray Systems
  - **Updated** (New chips, software, OS, etc.)
  - **Enhanced** (Add memory, change network, new RM, etc.)
  - **Extended** (Add resources, add new resource type or family)

- Productive During Transition Period
- Unify User and Admin Experience
- Increase Resource Utilization
Moab Cluster Suite™

What it is:
A workload management solution that provides simple web-based job submission and controls, graphical cluster administration and management reporting tools for high performance computing environments.

What it does:
- **Integrates** and **unifies management** across resources and environments in a cluster
- Controls the **sharing of resource usage** among users, groups and projects
- **Simplifies use, access and control** for both users and administrators
- **Tracks, diagnoses and reports** on cluster workload and status information
- **Automates tasks** to accelerate workload and reduce administration
- Provides a **foundation for future** growth for scalable grid-ready computing

Why you should care:
- Increases **work accomplished by 10-30%** per server, with **90-99% utilization**
- Provides an **integrated workload-management suite** at a **20 to 70% less cost**
- Gives administrators **greater control** over how resources are **shared** among users, projects, and organizations
- **Easy to use**, especially for those who are new to HPC.
- Helps organizations **cut energy costs** as much as **50% on idle nodes** with automated power-management and temperature-balancing policies.

10/23/2008
TORQUE Resource Manager

What it is:
An commercially supported leadership-class open source resource management solution that provides Petascale batch monitoring, submission, queuing and execution management.

Why you should care:
- No cost open source solution
- Dedicated commercial development
- Commercially supported
- Allows Moab to handle partition creation within XT systems
  - Better Failure Recovery
  - Reservations
  - Heterogeneous Resources
  - Node Features
- Used on both of the world’s petaflop systems
- Very large community, with thousands of downloads a month
Scheduling Jobs Across Heterogeneous Nodes
Heterogeneity

- Consumable Resources
  - Processors
  - Memory
  - Disk
- Software/Licenses
- Software Levels (ALPS 2.0, 2.1)
- Architectures (XT3, XT4, XT5)
- Operating Systems
Four Resource Selection Cases

1. **Nodes of Specified Type**
   - Give me nodes with 8 gigabytes of memory

2. **Nodes of Similar Type**
   - Give me all nodes with same amount of memory

3. **Nodes of Different Type**
   - Give me one node with 8 GB memory and 10 nodes with 2 GB memory

4. **Nodes of Any Type**
   - Give me whatever you can find
1. Nodes of Specified Type

A job may request nodes of a specified type
-- i.e. Quad core only, or only nodes with 8 GB memory

- Enabling Technologies
  - Adaptable Resource Manager Interface

- Example Syntax
  - `qsub -l procs=8:quad hello.job`
Confidential and Proprietary

Moab – XT3 Integration

**Node Query**
1. Obtain node class information from Torque
2. Obtain processor information from XTAdmin database
3. Obtain login and yod node information from Torque
4. Obtain *cpa allocation* information from CPA API
5. Return node information to Moab

**Job Query**
1. Obtain job information from Torque
2. Obtain job tasklist information from XTAdmin database
3. Return node information to Moab

**Class Query**
1. Query class info via Torque api

**Job Submit**
1. Submit job via Torque command

**Job Start**
1. Create a *cpa allocation* with cpa api
2. Start job with Torque qrun command
3. Return job status information to Moab

**Job Cancel**
1. Cancel job via Torque api

---

**XTAdmin Database**
- `qstat -q pbsnodes -a`
- `node.query.xt3.pl`
- `processer lustre`
- `node information returned`

**CPA**
- `cpa_lookup_nodes`
- `partition allocation`
- `job.query.xt3.pl`
- `job information returned`

**Torque**
- `pbs_statqueue qsub`
- `job.start.xt3.pl`
- `job start status returned`
- `job.submit.xt3.pl`

---

**Moab – XT3 Integration**
- `pbs_deljob`
2. Nodes of Similar Type

A job may require the nodes to be of the same type, but it does not care which. For example, we may want the job to run entirely across quad core nodes or dual core nodes, but not across both simultaneously.

- Enabling Technologies
  - Node Sets

- Example Syntax
  - qsub –l procs=8,nodeset=oneof:feature:dual:quad hello.job
Default Node Set Policy

moab.cfg:

# By default, jobs will be allocated nodes of a single core size
NODESETPOLICY ONEOF
NODESETATTRIBUTE FEATURE
NODESETLIST DUAL,QUAD

# Try to keep jobs within similar resource types, but have the flexibility
# to run earlier if a preferred resource type is not available
NODESETISOPTIONAL TRUE
3. Nodes of Different Types

A job may specifically request disparate chunks of nodes of multiple varieties. For example, the user may want the job to run a single master task on one quad core node having 8 GB memory, and 20 slave tasks on 10 dual core nodes.

- **Enabling Technologies**
  - CPA partition linking
  - Enhanced yod supporting the BATCH_TUPLE# environment variables

- **Example Syntax**
  - `qsub -l select=1:mem=8gb:quad+20:dual hello.job`
Dynamic Yod Environment Variables

The following pair of environment variables are set by Moab and request a single master task on one quad core node having 8 GB memory, and 20 slave tasks on 10 dual core nodes:

BATCH_TUPLE0=1:8:quad
BATCH_TUPLE1=20:0:dual

yod hello.exe
4. Nodes of Any Type

A job may not care if it allocated across heterogeneous node types. This gives the scheduler the greatest flexibility in maximizing utilization of the resources and avoiding fragmentation. The user’s job is likely to run sooner. For example, a job might request to run on 8 cores.

- **Enabling Technologies**
  - Moab heterogeneous node scheduling
  - Enhanced yod supporting dynamic allocation

- **Example Syntax**
  - `qsub –l procs=8 hello.job`
What about XT4/XT5?

Heterogeneous node support can be extended to the XT4/XT5 system and the ALPS partition manager with the exception of the fourth case just described. The ALPS job launcher (aprun) does not currently support a dynamic form of heterogeneous node chunking. Although aprun does support a colon delimited syntax which allows a command to be launched on chunks of heterogeneous nodes, the aprun command must be explicitly pre-constructed using command-line options in the job script and must anticipate the heterogeneous characteristics of the allocated nodes. This does not allow Moab the freedom to support dynamic heterogeneous node allocation.
Scheduling Jobs Across Disparate Systems

- Ahh, but can you schedule jobs across different ALPS domains?

- Yes! To do this we can use one Moab interfacing with multiple Native Resource Managers.

- Motivation
  - Single point of submission
  - Load balancing
  - Unified Job Accounting
  - Unified Policies (Fairshare, etc)
Multiple Resource Managers

- **Independent Head Node**
  - Moab Server
  - Torque 1 CLI
  - Torque 2 CLI

- **Cluster1 Head Node**
  - Torque Server 1
  - ALPS Domain 1
  - Moab CLI

- **Cluster2 Head Node**
  - Torque Server 2
  - ALPS Domain 2
  - Moab CLI

- **Cluster1 Login Node**
  - Torque Client (Mom)
  - Moab CLI

- **Cluster2 Login Node**
  - Torque Client (Mom)
  - Moab CLI

- **Cluster1 Compute Nodes**

- **Cluster2 Compute Nodes**
**Configuration Files**

**moab.cfg:**

RMCFG[cluster1] TYPE=NATIVE:XT4 SERVER=cluster1-pbs SUBMITCMD=/opt/torque-cluster1/bin/qsub
RMCFG[cluster2] TYPE=NATIVE:XT4 SERVER=cluster2-pbs SUBMITCMD=/opt/torque-cluster2/bin/qsub

**config.xt4.pl:**

$alpsUser = "root";
%alpsHost = ( cluster1 => "cluster1-login", cluster2 => "cluster2-login" );
%torquePath = ( cluster1 => "/opt/torque-cluster1/bin", cluster2 => "/opt/torque-cluster2/bin" );
Multi-RM Scheduling Flow

- Node information is collected for each cluster (combines info from Torque + ALPS – prefixing node ids with cluster name)
- Job information is gathered for each cluster (combines info from Torque + ALPS)
- Once the scheduler decides to start a job, an ALPS partition is created (via ssh) and the partition id recorded in a job variable
- The job is started via the associated resource manager api
- Stale ALPS partitions are cleaned up
- Moab handles user interface requests (job submissions, job cancellations, queries)
- Moab handles pending resource manager events (job finishing, job cancellation, submission via Torque)
Scheduling Jobs Across Completely Different Architectures

- What about scheduling jobs across completely different architectures (like XT3/CPA and XT4/ALPS)?
- But of course, using the Moab Grid Suite!
Managing Leadership Systems w/ Moab

ORNL

Jaguar: Cray XT/XT5
~181,000 cores
1.64 Petaflop

World’s Most Powerful Computer. For Science!

“The Jaguar system at ORNL provides immense computing power in a balanced, stable system that is allowing scientists and engineers to tackle some of the world’s most challenging problems.”
—2006, Kelvin Droegemeier, Meteorology Professor, University of Oklahoma.
Managing Leadership Systems w/ Moab

Sandia – Red Storm

**Red Storm**: Cray XT3
- 12,960 nodes
- 38,400 cores

- 284 teraOPS theoretical peak performance
- 135 racks
- AMD Opteron™
- 78 terabytes of memory
- 1.7 petabytes of disk storage
- Linux/Catamount OS
- 2.5 megawatts power & cooling

Design: Sandia
Managing Leadership Systems w/ Moab

Other Leading Government Site

Cray XT4
Over 18,000 cores

- AMD Opteron™
- ~100 racks
Market Usage

• **Billions of Dollars** worth of Hardware run Moab

• **Worlds Largest computer** runs Moab (1 Petaflop – over 100,000 processor cores used)

• **Future Largest Systems** (w/ planned Moab use):
  • Another 1 Petaflop System
  • 2 Petaflop System
  • 5 Petaflop System
  • 25 Petaflop System

• ~25% of the resources of the **Top 100** systems in the world use Moab (Using Top500.org - 2008)

• **98+% Customer Retention** (By Revenue)
Conclusion
Conclusion

- Moab and Torque can be used on Cray systems to:
  - Improve utilization
  - Enforce site policies
- Moab’s Intelligent Integration with ALPS and CPA Allow:
  - Support for heterogeneous resources
  - Unification of disparate XT systems into a grid resource

This means better utilization and easier transitions during the life cycle of the system as you update, enhance and expand your Cray systems.
For more information

Contact: Scott Jackson
Cluster Resources, Inc.
scottmo@clusterresources.com
(801) 717-3708
http://www.clusterresources.com
Appendix
The Moab Product Family Tree

- Moab Cluster Suite
- Moab Grid Suite
- Moab Hybrid Cluster Suite
- Moab Cluster Builder for SUSE Linux
- Moab Adaptive Cluster Suite
- Moab Adaptive Energy Suite
- Provisioning xCAT, HP SA, Virtualization, Etc.

Moab Adaptive Computing Suite

- Adaptive Operating Environment
- Moab Adaptive Energy Suite
- Private cloud
- PaaS
- business-process automation
- project-space creation
- workload-aware green computing
- data center
- automated project-space creation
- full turnkey cluster software (SLES)
- cluster workload manager

Moab

HPC

HPC grid

multi-OS hybrid cluster

Data Center

Cloud

External

Internal

1/2/2009
Moab Grid Suite™

What it is:
A workload management solution that provides simple web-based job submission and controls, graphical grid administration and management reporting tools for a group of high performance computing environments unified into a grid.

What it does:
- Enables rapid unification of multiple clusters into a managed grid environment
- Intelligently applies policies which enforce guidelines provided by owners of the resources
- Optimizes resource usage for timing, best fit resource usage and location
- Tracks usage for billing purposes

Why you should care:
- Improves utilization of resources by 10 to 30% and provides access to unique resources
- Enables collaboration between teams without the complexity of interacting manually with multiple systems and overcoming the politics of sharing
- Aids organizations to share costs of infrastructure investment and to properly apply the investment to projects and needs in a timely and controlled basis
Multi-OS Hybrid Cluster

Example: Holland Computing – 2300 Server Hybrid
Workload-Aware Green Computing™

What it is:
A workload and environment management solution that monitors energy use, workload needs, resources within and environment and then orchestrates optimal placement of workload, state of resource power usage and delivery on mission objectives.

What it does:
- **Intelligent power management** places idle servers in power-saving modes
- **Workload consolidation** uses workload packing and virtualization technologies to consolidate workload
- **Cost- and temperature-based scheduling** routes workload to cost-efficient servers and allows hot servers to cool down
- **Advanced monitoring and reporting** enables reports on power consumption and carbon credits per user, project, or resource

Why you should care:
- Servers with no workload still consume 60% power, Moab can automatically put these idle servers in power savings mode
- Pack workload onto servers more efficiently, **improving utilization by up to 60 to 80%**.
- **Reduce cooling costs by up to 25%** with temperature-based workload placement
- Help organizations **achieve their green computing objectives** with energy tracking, optimization, usage enforcement and carbon credit tracking