Moab and Torque on Cray Architectures

© Cluster Resources, Inc.

Outline

- The Motivation
- The Solution
  - Why Moab?
  - Why Torque?
- Model Comparison
- Conclusion

© Cluster Resources, Inc.

The Motivation

© Cluster Resources, Inc.

The Motivation

- You have invested in a large Cray system
  - XT3/XT4, X1E, XD1…
- You want
  - High Utilization/ROI (happy investors)
  - Enforce Site Objectives (happy managers)
  - Manageability (happy admins)
  - Usability (happy users)

© Cluster Resources, Inc.

Let’s look at some examples

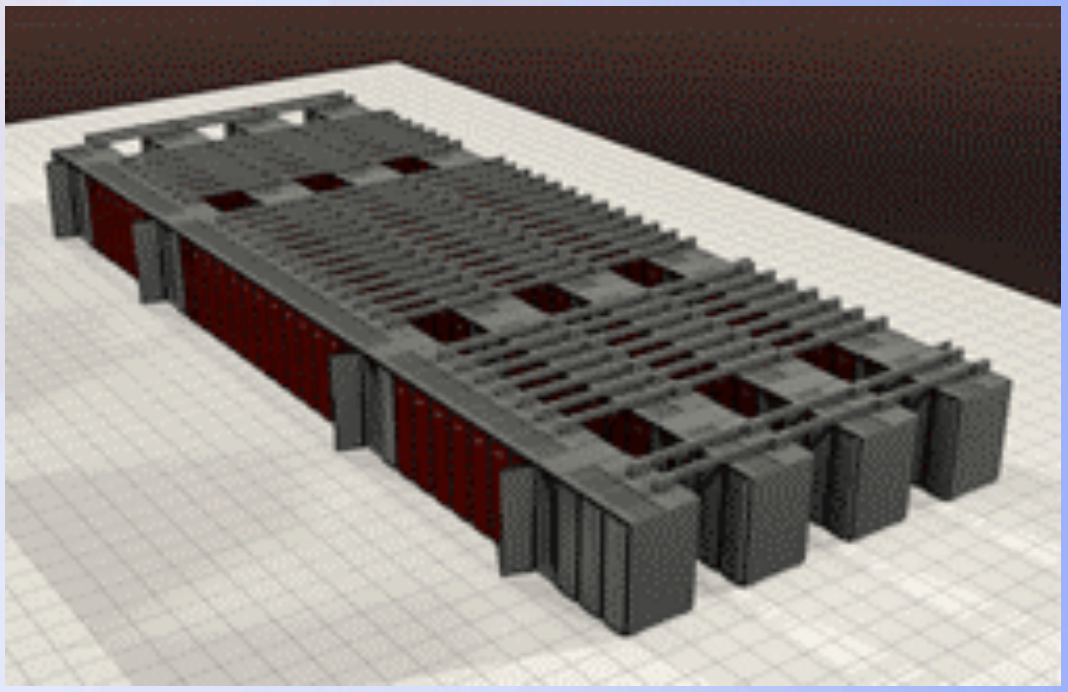
© Cluster Resources, Inc.

Managing Leadership Systems w/ Moab

Sandia – Red Storm

Red Storm: Cray XT3
12,960 CPUs

- **124.42 teraOPS theoretical peak performance**
- **135 racks**
- **AMD Opteron™**
- **40 terabytes of DDR memory**
- **340 terabytes of disk storage**
- **Linux/Catamount OS**
- **<2.5 megawatts power & cooling**



© Cluster Resources, Inc.

Design: Sandia

CLUSTER RESOURCES INC.

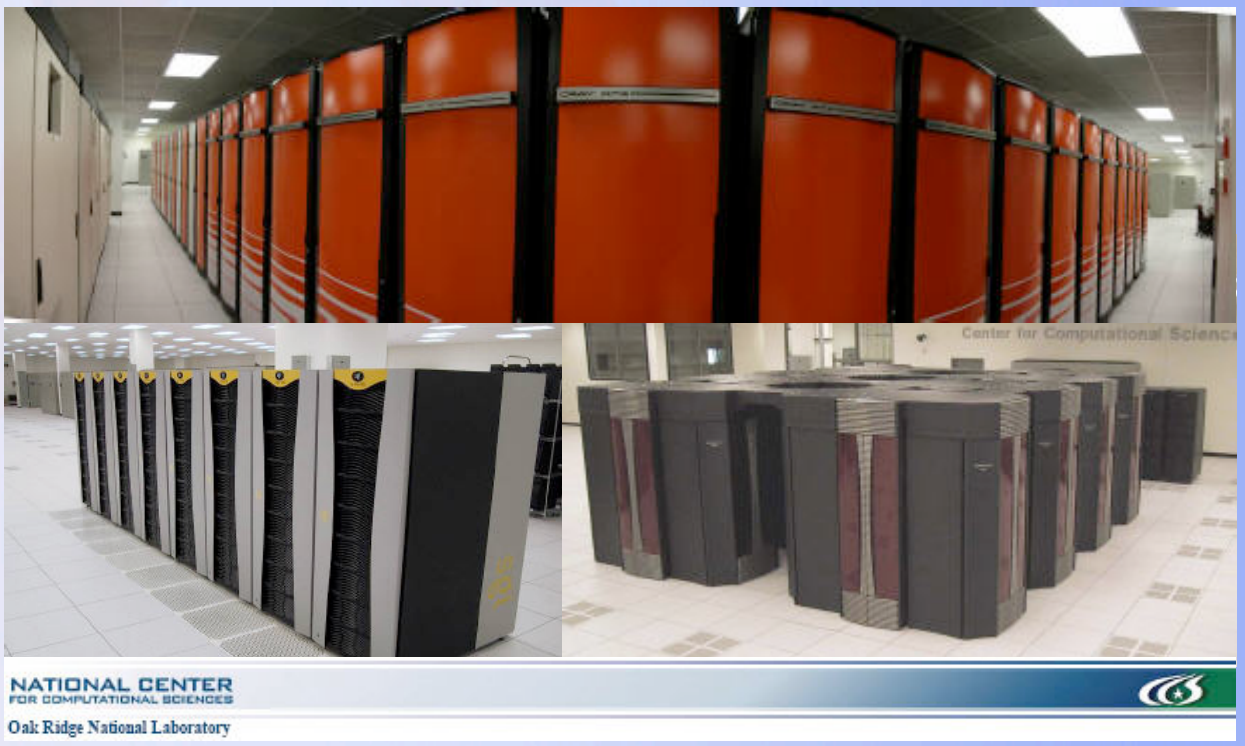
Managing Leadership Systems w/ Moab

ORNL

Jaguar: **Cray XT3**
~18,000 cores
moving to1 Petaflop

Phoenix: **Cray X1E**
1,024 cores

RAM: **SGI Altix**
256 cpus



© Cluster Resources, Inc. Photo: ORNL

Managing Leadership Systems w/ Moab

Other Leading Government Site

Unannounced:
Cray XT3
Over 18,000 cores

- **AMD Opteron™**
- **~100 racks**

© Cluster Resources, Inc. Photo:

The Solution

© Cluster Resources, Inc.

Moab and Torque on Cray

Funding Managers

- High utilization / Return On Investment
- System cycles delivered to specific workload and groups according to commissioned objectives
- Statistics and reports provide evidence of delivered performance and utilization

Site Managers

- Service Level Enforcement/Guarantees
- Flexible policies to meet performance objectives
- Enforce political resource sharing
- Reports and simulations for capacity planning
- Graphical charting tools

Moab Workload Manager

TORQUE Resource Manager

System Administrators

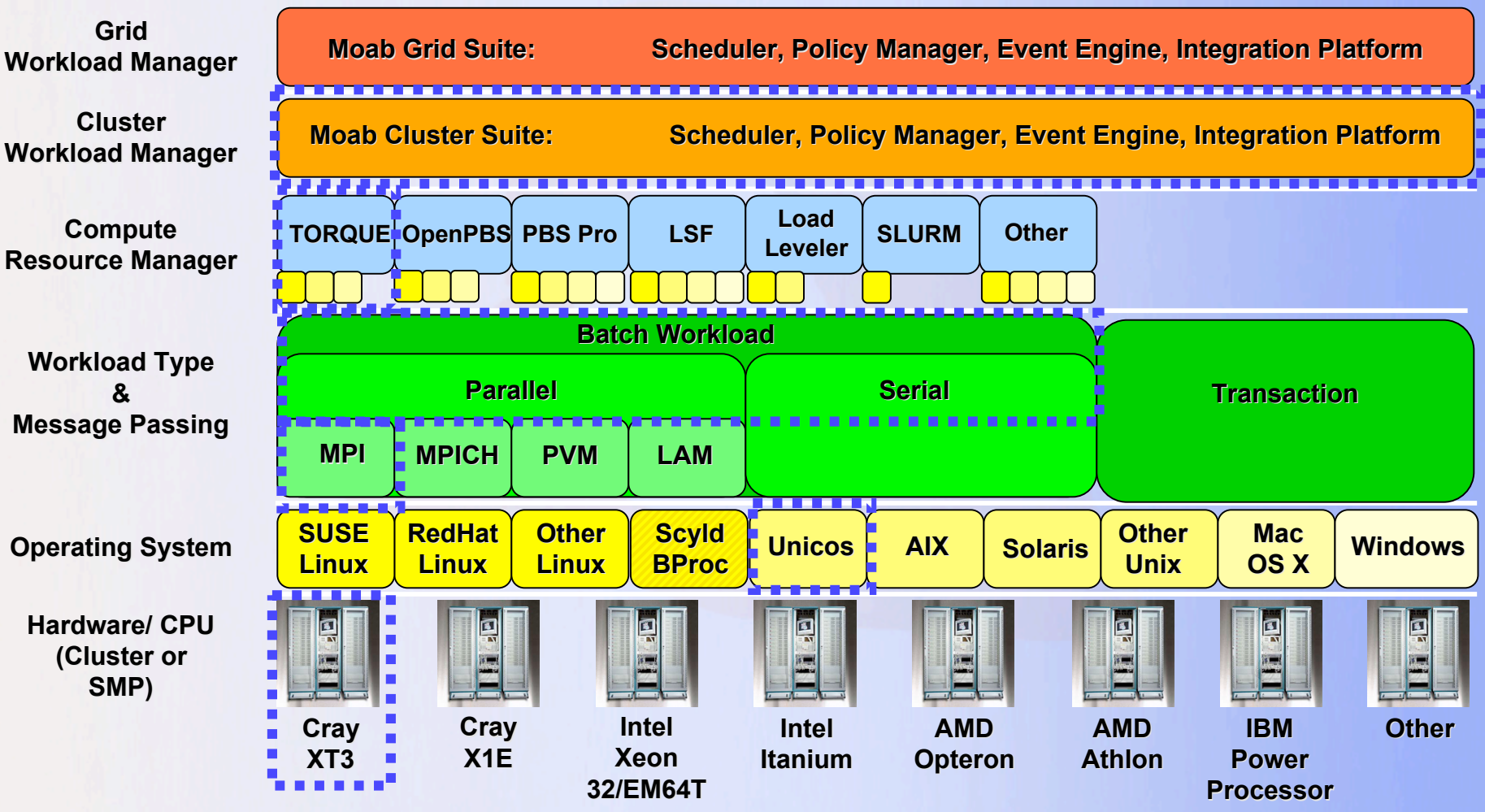
- Unified batch management
- Task automation
- Powerful diagnostics / monitoring
- Evaluate impact of new policies
- Self-help for users
- Graphical Administrative Interface

End Users

- Information and control of jobs
- Simple and flexible job submission
- Translations to familiar batch environments
- Prediction of job start
- Reliable cycle delivery
- Web-based Job Submission Portal

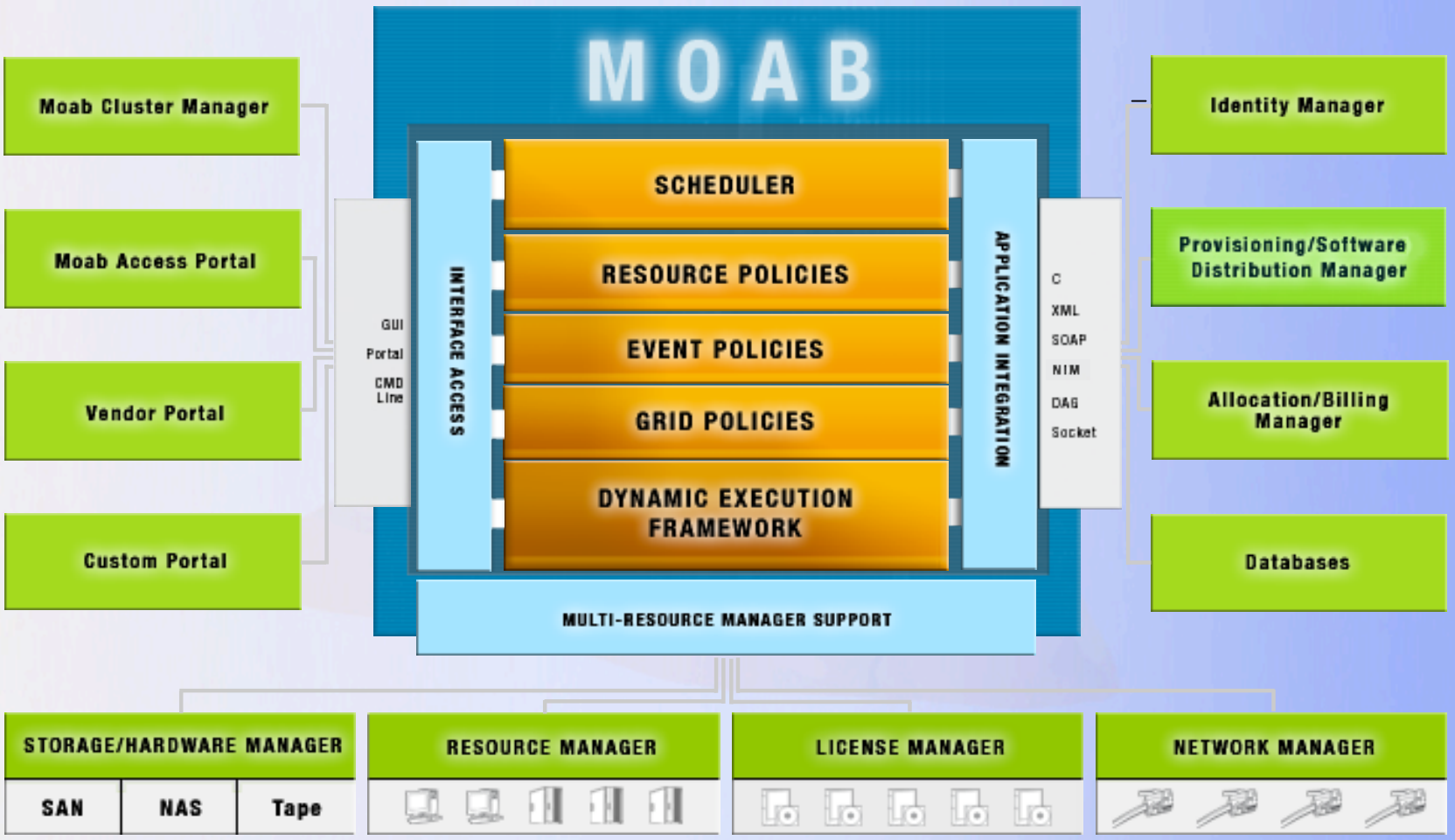
© Cluster Resources, Inc.

Solution Framework: Where Does It Fit?



© Cluster Resources, Inc.

CLUSTER RESOURCES INC.



© Cluster Resources, Inc.

Why Moab?

- Improved Utilization
- Reservations (Administrative, Standing)
- Service Level Guarantees (QOS, Priority, Fairshare, Usage Throttling)
- Resource Manager Translation
- Moab can create a Grid across all of your clusters – independent of RM/OS/Arch

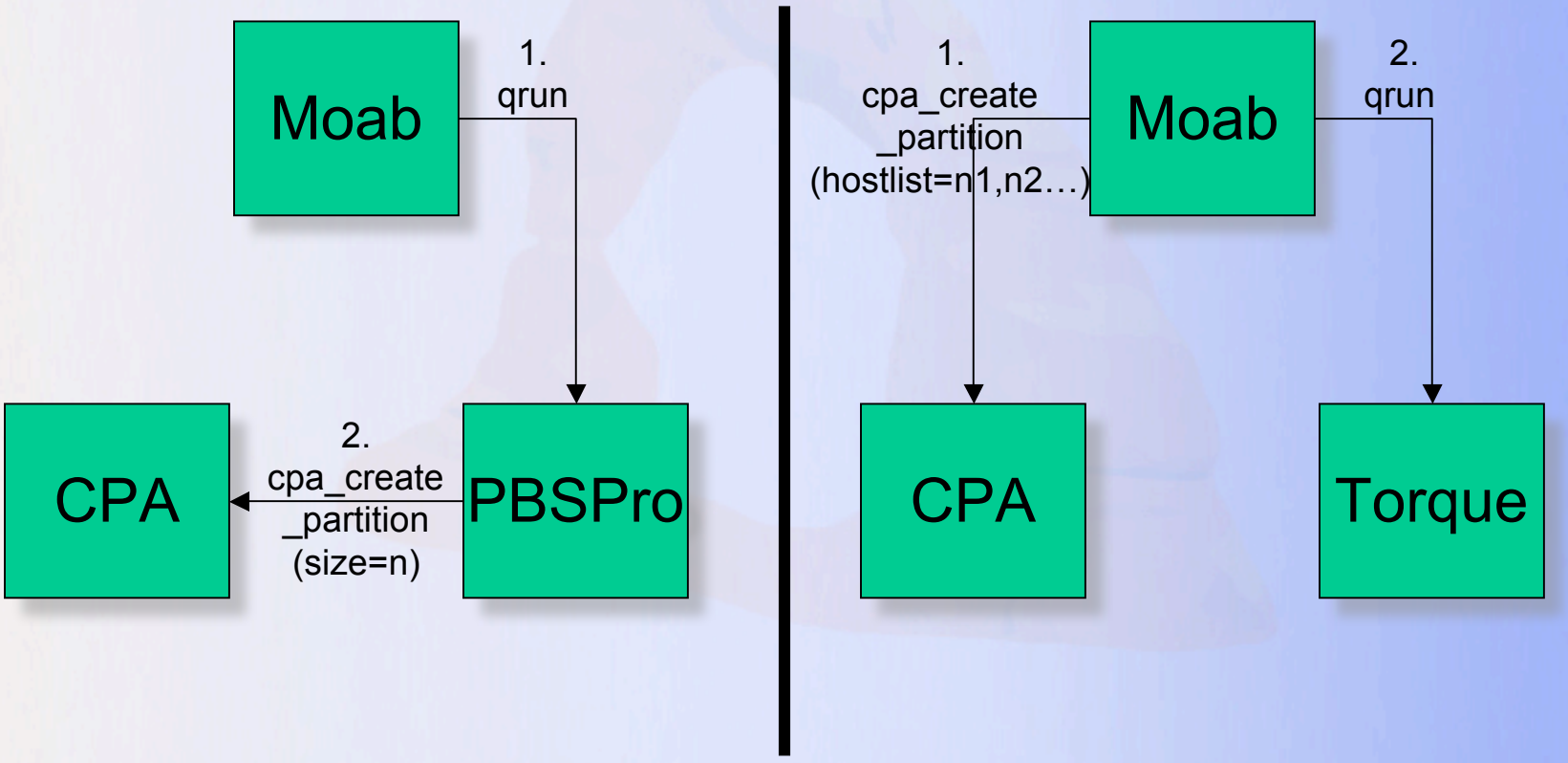
© Cluster Resources, Inc.

Why Torque?

- Industry Standard Batch System
- Provides underlying support for Moab's advanced features
- Permits Moab to handle the CPA partition creation which permits
  - Better Failure Recovery
  - Reservations (Admin, Standing, …)
  - Heterogeneous Resources
  - Node Features
- It is free, open source and commercially supported

© Cluster Resources, Inc.

CPA Allocation Model Comparison



© Cluster Resources, Inc.

Prior Model

- PBSPro pbs_mom handles cpa partition creation and management
- Processors could not be selected – PBS asks cpa to allocate a partition of size=n processors
- Very difficult to enforce reservations
- Could not select nodes (processors) for placement so there was no support for heterogeneous resources or node policies of any type
- Poor failure handling. When Moab told PBSPro to start a job – qrun would succeed, but if the cpa allocation failed (such as a lustre recovery issue), the job would just drop back into idle and you would need to examine mom logs to find the cause.

© Cluster Resources, Inc.

Improved Model

- Moab handles cpa partition creation and management before telling Torque to start the job
- Moab can allocate a list of processors of its choice – i.e. asks cpa to allocate a partition across selected processors (3, 5, 7-10, …)
- Moab's scheduling optimizations can now be honored (node availability policies such as minResource, lastAvailable).
- Node (processor) sharing is possible
- Reservations work (admin reservations, job reservations, standing reservations, user reservations)

© Cluster Resources, Inc.

Improved Model continued

- Support for heterogeneous resources
  - Nodes configured with different disk, memory, swap, architecture, opsys
  - Node features (assigning different labels to nodes to be requested in jobs)
  - Generic resources
  - You can steer jobs to specific nodes or sets of nodes (hostlist, nodesets)
  - Per node limits, different policies, rules and constraints per node
  - Track node and avert node issues (load, failures, blocked resources)

© Cluster Resources, Inc.

Improved Model continued

- Allocation (e.g. lustre) failures are detected by Moab ***before*** the job starts
  - This allows intelligent handling and rerouting of jobs to available nodes
  - Failure causes can be caught by Moab, intelligently responded to, and reported to admins

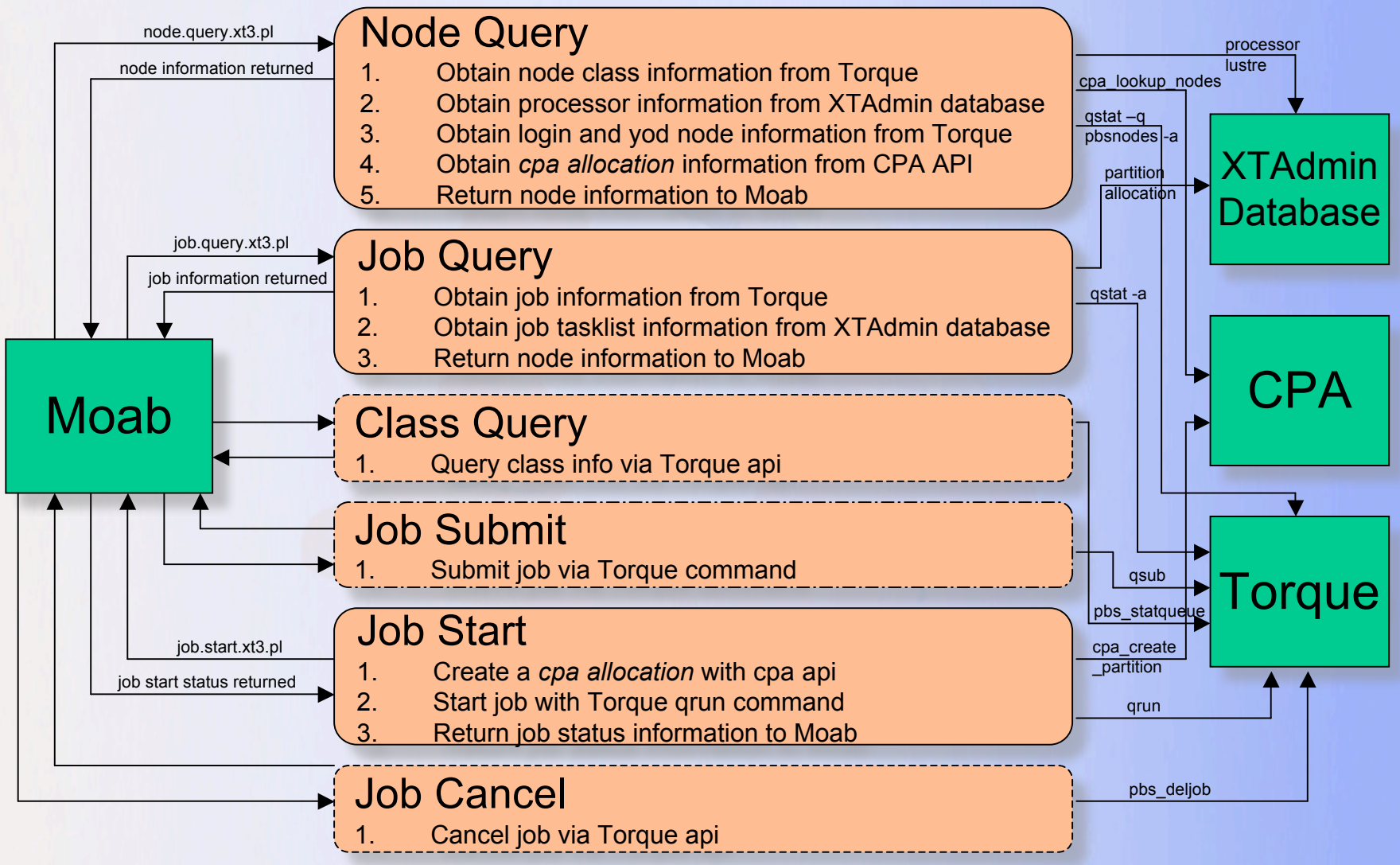
© Cluster Resources, Inc.

Implementation Details

- Hybrid Model (Torque, Native RM Interface)
- Some actions are passed through to torque directly (e.g. Queue Query, Job Submit, Job Cancel)
- Other actions use scripts to aggregate information from multiple sources – Torque, CPA, XTAdmin database (e.g. Node Query, Job Query, Job Start)

© Cluster Resources, Inc.

Moab – XT3 Integration



© Cluster Resources, Inc.

Conclusion

- Moab and Torque can be used on Cray systems to:
  - Improve utilization
  - Enforce site policies
- Managing XT3 CPA allocation via Moab allows better support for:
  - Optimized scheduling
  - Advance reservations
  - Heterogeneous resources
  - Node sharing
  - Improved node allocation failure handling

© Cluster Resources, Inc.

For more information

Contact: Scott Jackson
Cluster Resources, Inc.
scott@clusterresources.com
(801) 717-3708
http://www.clusterresources.com

© Cluster Resources, Inc.