CLUSTER® RESOURCES INC.

Enabling Moab's Adaptive Computing for Cray XT3/XT4

Michael Jackson, President
Cluster Resources
michael@clusterresources.com
+1 (801) 717-3722

Contents

1. Overview
2. Management Evolution
3. Core Concepts
4. Applicable Areas
5. Scenarios

Technology Focus

Technology Areas

- Cluster – Basic
- Cluster – Holistic (Compute + Storage + Network + Licence, Etc.)
- Grid – Local Area Grid (One Administrative Domain)
- Grid – Wide Area Grid (Multiple Administrative, Data and Security Domains)
- Data Center
- Hosting
- Utility/Adaptive Computing

Platforms

- XT3 / XT4 …
- X1E
- XD1
- Other External Platforms

Session Plug

Moab & TORQUE on Cray Architectures

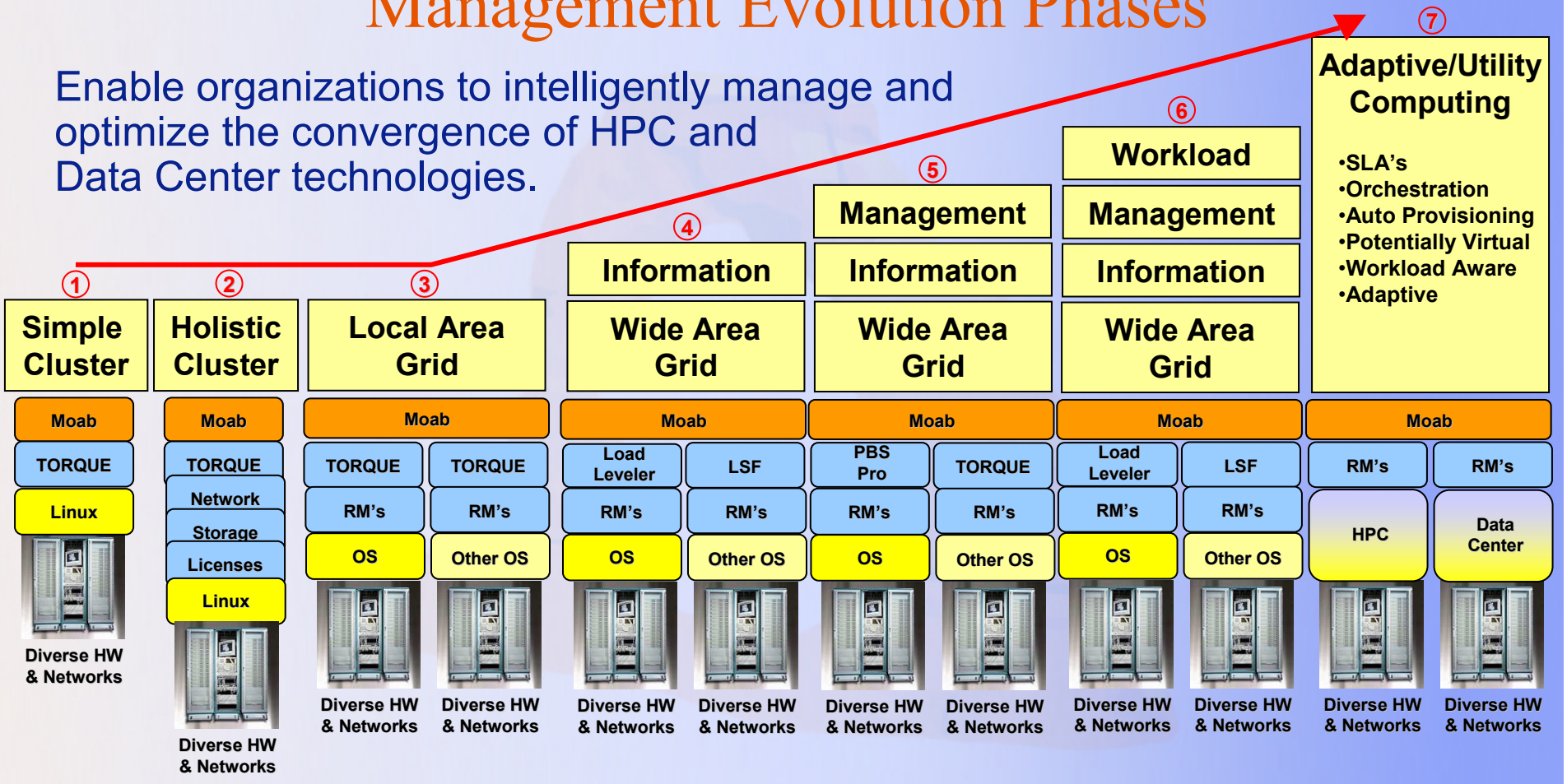
Session 17B

Time: 11:00 Today

Room: San Juan / Whidbey

Management Evolution Phases

Enable organizations to intelligently manage and optimize the convergence of HPC and Data Center technologies.



Initial Cluster

Managed Cluster

Single Owner or Department
- Consolidating Multiple Clusters

Commercial
- Capacity Planning
- Cost Sharing
- Awareness Focused
- IT Consolidation

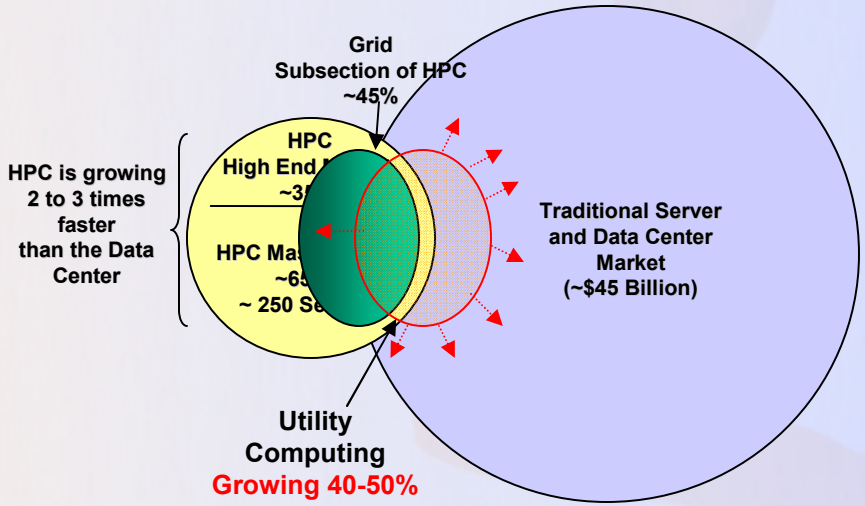
Commercial or Government
- Capacity Planning
- Cost Sharing
- Lower Administration
- Improve Allocation
- Cost Reduction

Government or EDU Research Collab.
- Collaboration
- Improve User Productivity
- Service Delivery
- Centralization
- Optimization

Commercial or EDU
- Re-Inventing IT

Management Evolution Adoption

Market Size and Convergence*



Data Center & Traditional Server Market

- ~$45B/yr – growing 3% to 5% /yr

HPC

- ~$11B/yr – growing 2½ to 3x general server market
  - Cluster / High-End Cluster – ~35%
  - Mass Market Cluster - ~65%
  - Grid – 45% (growing faster than cluster)

Utility Computing (Adaptive HPC & Data Center)

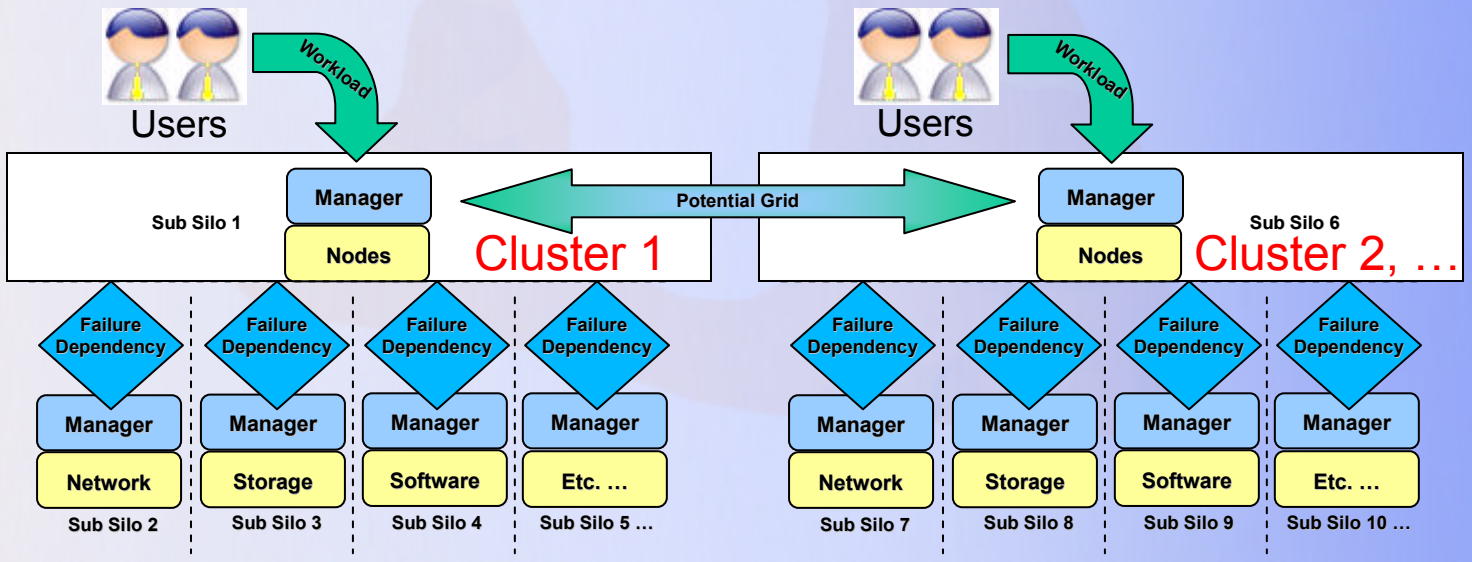
- ~$5B/yr – ***Growing 40% to 50% /yr (faster than Data Center or HPC)***

***Source: IDC & Modeling**

Traditional HPC

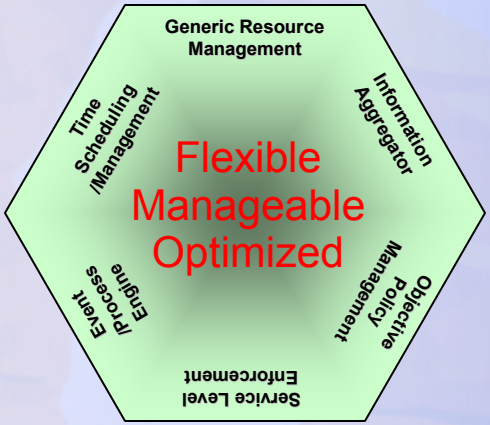
Silo-ed Resources + Silo-ed Management + Silo-ed Workload =

Inefficiency, Disjointed Management and Broad Sets of Dependency Based Failure Points



Adaptive HPC

Unified Policies, Resources & Workload =



Users

Workload

Adaptive Computing Engine: Policies, Events/Workflow\BP, Service Level, Scheduling, Workload, UI, Reporting

Manager	Manager	Manager	Manager	Manager
Network	Storage	Nodes	Software	Etc. …

Cluster 1

Manager	Manager	Manager	Manager	Manager
Network	Storage	Nodes	Software	Etc. …

Cluster 2, Etc. …

Adaptive Computing Scenarios

Current Uses of Moab's Adaptive Computing Technology:

- **Intelligent/Adaptive HPC Center** – Provide centralized intelligent engine that translates high-level business objectives and SLAs into optimized orchestration of adaptive computing and effectively responds to changing environmental conditions.
- **Data Center** – Provide centralized intelligent engine that translates high-level business objectives and SLAs into optimized orchestration of adaptive computing
- **Import Hosted Resources** – Allocate tailor-provisioned hosted resources instantly and transparently to overcome workload spikes or resource failures
- **Host for Internal Use** – Host excess/unused resources or apps to other departments or groups
- **Host for External Use** – Host company resources (data mining, custom apps, specialized hardware, etc.)

What does Cluster Resources do?

Moab Cluster Suite

Consists of a policy-based workload management and scheduling engine, a graphical cluster admin interface, and a Web-based end user job submission and management portal.

Moab Grid Suite

A grid management solution that provides scheduling, job and data migration, credential mapping, and reporting across multiple independent clusters while maintaining full cluster sovereignty.

Moab Utility/Hosting Suite

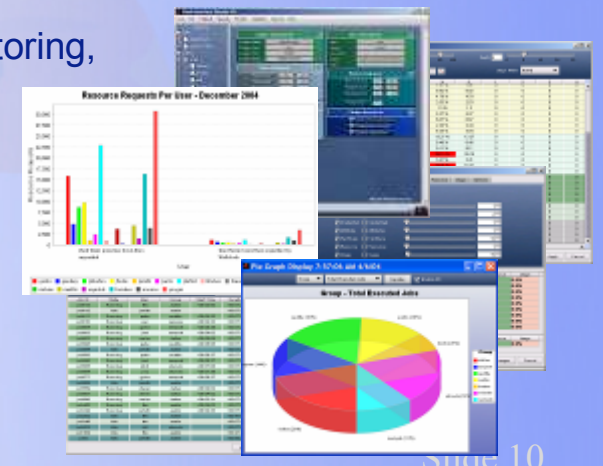
Allows HPC sites to host out or immediately access resources and compute environments; Moab orchestrates underlying services to meet mission objectives and SLA/QoS guarantees.

TORQUE Resource Manager

An open source resource manager providing job queuing, node monitoring, and parallel job execution.

Other

Maui Scheduler and Gold Allocation Manager (open source).



Awarded Largest HPC Management Contract in History

US Department of Energy

250,000+ processors, 300+ clusters

"Partnerships such as this one are a key element of the ASC Program's success in pushing the frontiers of high performance scientific computing. Only by working with leading innovators in HPC can we develop and maintain the large scale systems and increasingly complex simulation environments vital to our national security missions."

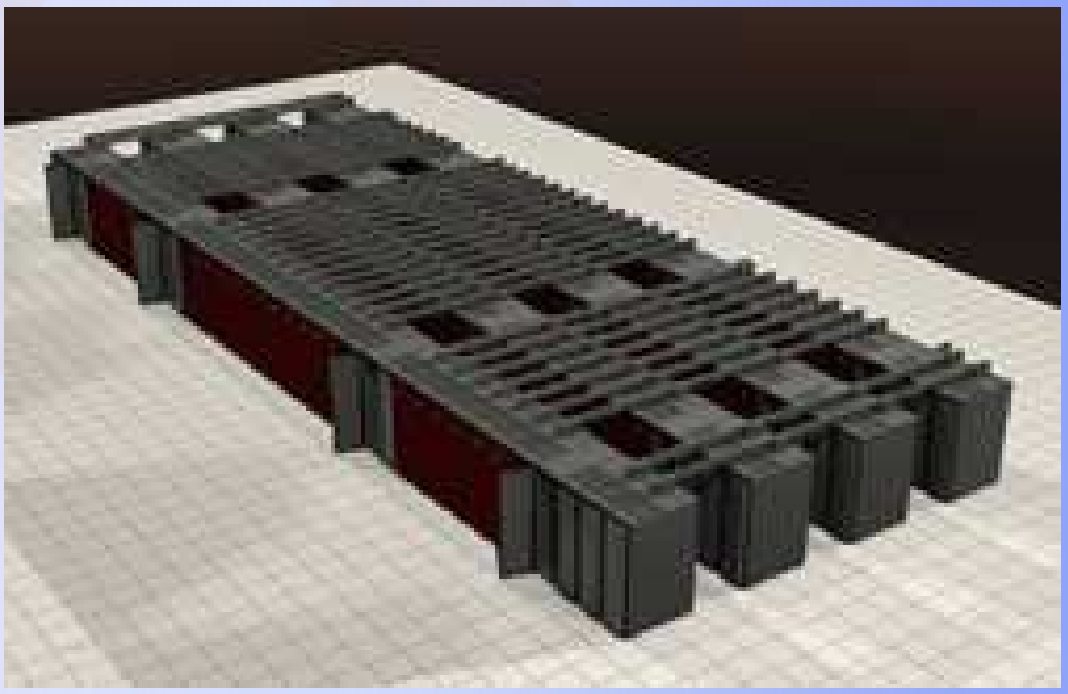
—Brian Carnes, Service and Development Division leader at Lawrence Livermore National Laboratory

Managing Leadership Systems w/ Moab

Sandia – Red Storm

Red Storm:
12,960 CPUs
Cray XT3

- **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**



Design: Sandia

Managing Leadership Systems w/ Moab

ORNL

Jaguar: ~18,000 core **Cray XT3** moving to 1 Petaflop

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

RAM: 256 CPU **SGI Altix**

Photo: ORNL

Other Leading Government Site

Over 18,000 cores
Cray XT3

- **AMD Opteron™**
- **~100 racks**

Photo:

Unified Management & Optimization

Barcelona Supercomputing Center

Europe's Largest Supercomputer/Cluster

5Th Largest HPC System in the World

Unified Management with Moab

Boeing

Using Moab to manage workloads across 4 clusters with hundreds of nodes and more than 1,000 CPUs

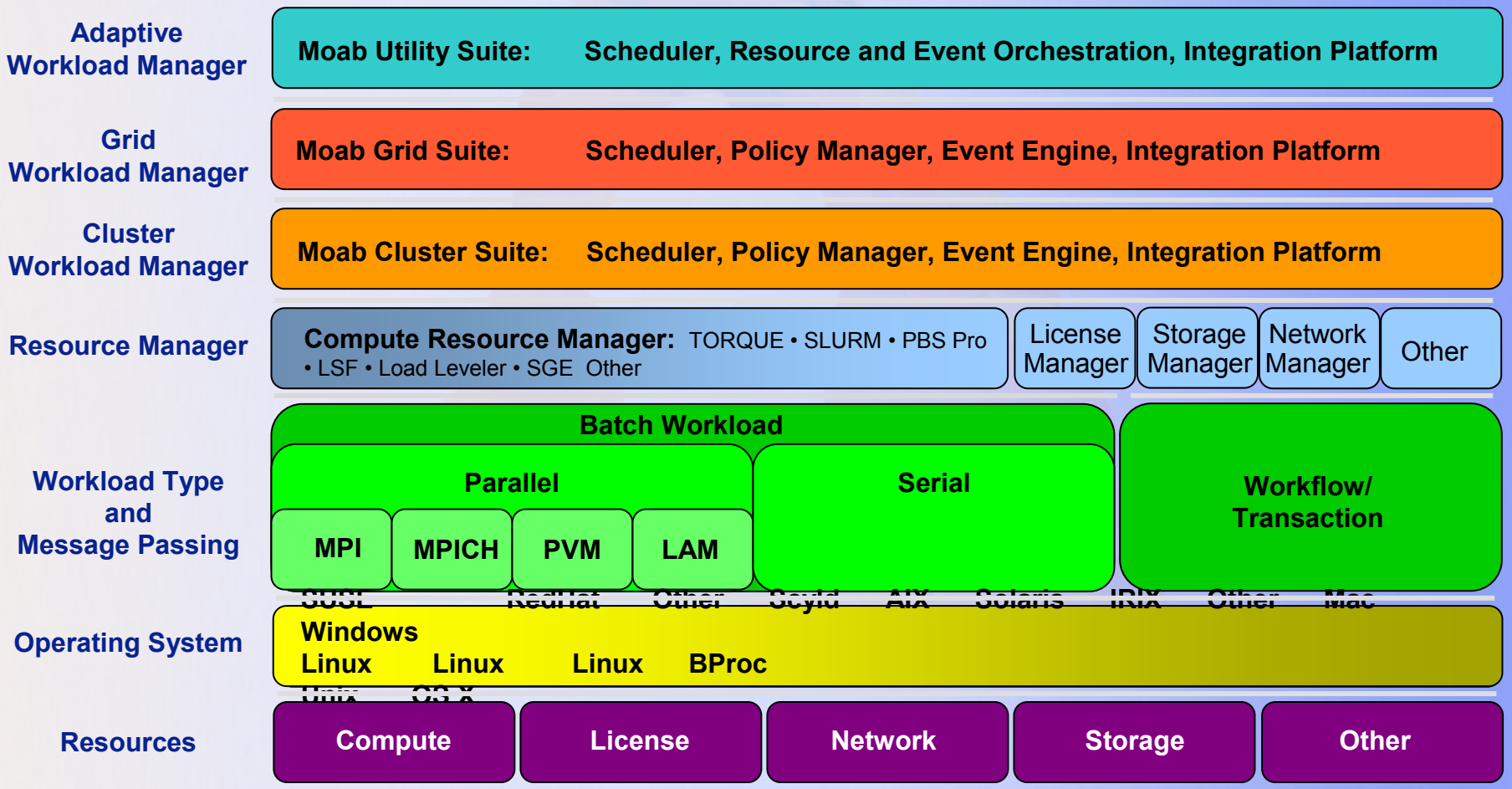
Photo: The Boeing Company

Top500 Systems licensed with Moab

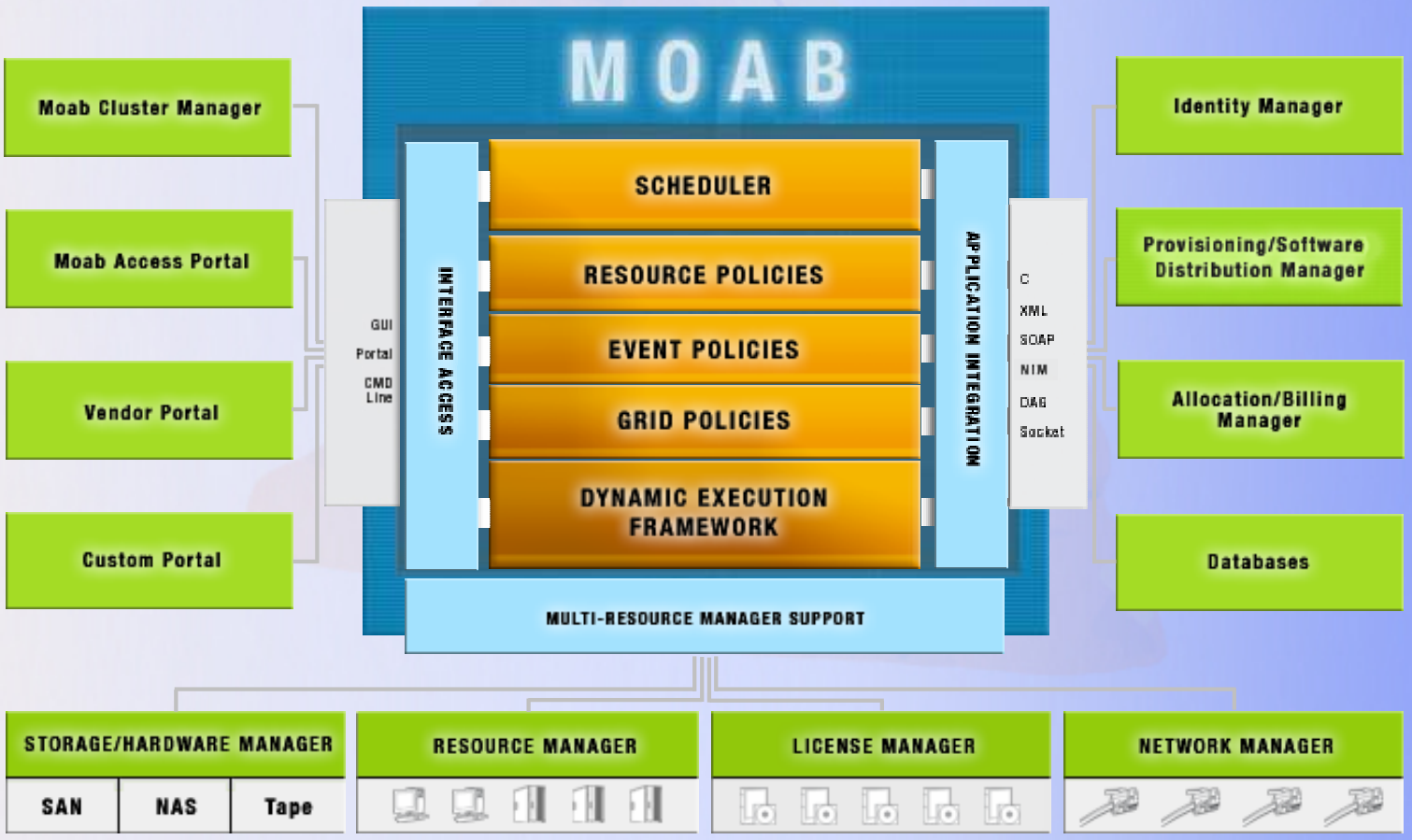
Leadership Class Systems

#1, 2, 4, 5, 6, 10, 11, 19, 20, 23, 24, 25, 28, 35, 37, 44, 51, 54, 58, 60, 62, 69, 71, 75, 78, 79, 84, 87, 96, 114, 140, 142, 143, 146, 176, 202, 203, 230, 296, 297, 356, 366, 412, 445, 450, 451, 454, 488

Company: Moab in the HPC Stack



Moab's Architecture



Moab: Technology Differentiators

Moab is a strategic Adaptive Infrastructure Solution that enables business process management to meet critical objectives.

- **Flexible and Configurable Resource Requirement Controls**
  - Set and dynamically modify workflow requirements, dependencies and associations to meet processes and objectives
  - Enable application-specific actions and conditions that differ per business process to allow for extensive customization

- **Integrated Policy Engine**
  - Incorporate **abstracted** policies & concepts — resource requirements, ownership, SLA guarantees, prioritization, scheduling, allocation, workload, resources, time, etc. — to make **optimal** decisions

- **Modular and Extensible Monitoring and Management Capabilities**
  - Apply unified monitoring and management to virtually any resource or service
  - Harness distributed resources with grid-enabled management

Moab: Technology Differentiators

- **Workflow-Capable Adaptive Event Engine**
  - Adapt resources, rules and environments in real-time to meet SLA guarantees and pro-actively or reactively respond to business surges, failures and changes
  - Automate delivery on business process requirements
  - Create custom responses to key changes in current environments or future needs

- **Solution-Centric Platform Design**
  - Enable customers to meet objectives by offering a seamless migration path from data center or HPC to an enterprise utility environment
  - Apply a federated architecture model to allow for sovereignty, phased extension and scalability requirements
  - Leverage broad interoperability with emerging and legacy technologies and environments
  - Use Moab as a building block to unify and translate monitored information and management interfaces into a composable and consumable service

. . . EQUALS Mature Adaptive Computing Architecture

Core Adaptive Capability Concepts:

Required or Value Added

- Multi-sourcing (Information, Management, Policies)
  - Broad Scope (Resource Managers, Agents, XML, Flat file, DB, SOAP, Socket, C, Java, CLI, etc.)
- Abstracted / Generic Resource Monitoring
- Abstracted / Generic Event Definitions
- Abstracted / Generic Metric Tracking
- Event Engine
  - Internal Management
  - External Management (Outside manager, tools, services, etc.)
  - Workflow Centric (Cascading, Multi-path, Business Process)
- Internal and External Dependency Management
- Extensive Time Based Mechanisms / Scheduling
- Dynamic Workload and Process Management (Jobs, Transactions, System Requests)
- Workload and Policy Hierarchies and relationships (Job Groups, Hierarchical sharing, etc.)
- Virtual Private Resources (Virtual Private Cluster, etc.)
- Reporting, Billing, etc.

Areas of Applicability:

Adaptation for the purposes of:

- Optimizing Resources/Investment/Work Accomplished
  - Mixed hardware types with affinity or requirement based workload placement
  - Mixed network conditions or types with optimization, affinity or requirement based workload placement, network re-configuration or condition resolution
  - Manual Action Replacement
  - Automated Learning
- Quickly Meeting Changing Needs
  - Service Level Delivery / Policy Adaptation
    - Adapt to Changing Organizational & Project Objectives
  - Adaptive Security
  - Resource or Policy Impact Evaluation through Simulation
  - Application Environment Adaptation (Auto provisioning and configuration)
- Automatically Responding to Failure Conditions, Surges or Other Environment Changes
  - File System Failures
  - Machine Room Chiller Failures (Automate safe shutdown)
  - Network Failures
  - CPA Failure Feedback
  - Workload Surges

Usage Scenarios:

- Mixed Vector, MPP, MTA & Scalar Adaptive Supercomputer
- Failure Recovery and Ease of Use Automation
- Automated Application Learning / Optimization
- Other Example Scenarios
  - Dynamically adjusts workload or environment to enforce policies and SLA/QoS
  - Security Adaptation

Usage Scenario: Mixed Vector, MPP, MTA & Scalar

Applying Moab's Adaptive Infrastructure: Adaptive Workflow & Workload Mgmt

- **Configuration Work:**
  - Apply "Node Attributes" so Moab can differentiate architectures
  - Apply information sourcing, co-allocation logic and event interaction to allow Moab to interact w/ each architecture in the appropriate custom way
  - Create "Application Templates" which identify resource ranges, default settings, dependencies and associated workflow activities
  - Create "Affinities" and "Requirements" between application templates and node attributes.
  - Enable "Automated Learning" for the appropriate standard and generic metrics
- **User Transparently Submits to one Location:**
  - Moab evaluates optimal placement of workload considering requirements and service level objectives
  - Moab applies the workload to the resources that have the greatest affinity and match requirements and then applies one or multiple required workflows if necessary to adapt the environment or the workload to achieve the best response time

Usage Scenario: Failure Recovery & Ease of Use Automation

Applying Moab's Adaptive Infrastructure: Active or Reactive Automation

- **Configuration Work:**
  - Identify Failure Conditions (Not a configuration – discovery work)
  - Identify manual method of identifying issue via scripts, tools, monitors, etc.
  - Apply Moab's Native Resource Manager to import failure/state information found in script, flat file, CLI, XML, web service, etc.
  - Identify manual steps, commands to run, people to notify, workflows to apply to resolve issue
  - Apply Moab event mechanisms to interact with remote systems, tools, commands, scripts, notifications, etc. via a workflow capable logic tree

Usage Scenario: Failure Recovery & Ease of Use Automation

- **Example Results:**
  - **Lustre Multi-state Phases Confuse Resource Manager**
    - Interface with Lustre State information tracker to avoid submitting workload when it is not ready
  - **Machine Room Chiller Fails**
    - Notify admins, checkpoint where possible, preempt all jobs, power off nodes
  - **External Power Failure, UPS Triggers**
    - Notify admins, preempt low priority jobs, power off unused nodes
  - **Compute Node Temperature Exceeds Desired Threshold**
    - Notify admins, modify scheduling policies to minimize node usage or apply low processor centric workload
  - **Storage Manager Reports Warnings**
    - Notify admins, block jobs requiring storage manager resources until warnings cease
  - **Compute Node Local Disk Fills Up**
    - Launch script to purge unneeded files on compute node
  - **Effective Node Throughput Drops Below Desired Threshold**
    - Notify admins, launch script to investigate, correct, and recycle node
  - **Major Network Failure**
    - Notify admins, dynamically establish peer relationship w/ alternate organizational resources or connect to remote hosting center

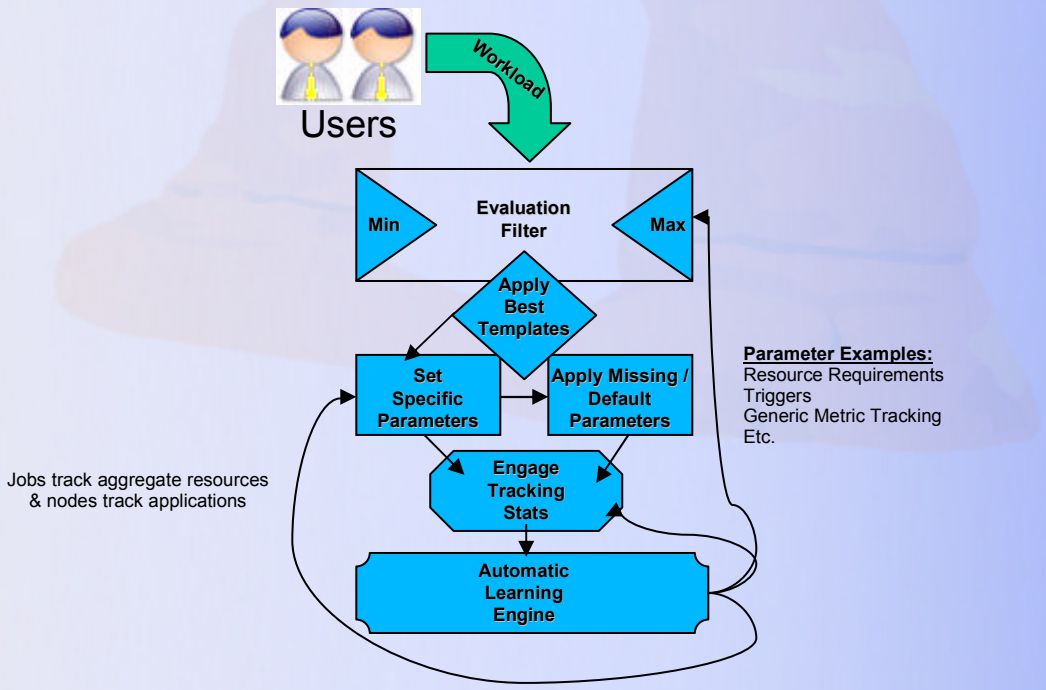
Usage Scenario: Application & Resource Optimization

Applying Moab's Adaptive Infrastructure: Automated Learning

- **Configuration Work:**
  - Apply known template defaults and requirements to applications
  - Teach Moab which attributes to track per application
  - Allow Moab to automatically apply learned optimization or require manual validation
  - If automatic optimization is allowed Moab would fine tune templates and set optimal resource affinities
- **User Transparently Submits Workload with minimal detail**
  - Moab identifies applicable application template(s) as appropriate
  - Moab applies the templates settings (fills in the blanks), workflows, requirements, etc. to optimize workload

Automatic Learning Mechanism

Automatic learning can improve optimization of resources and speed project turn around time.



Usage Scenario: Other Example Scenarios

- **Security Adaptation**
  - Security Zone Enforcement based on workload demand and surges
  - Automation of Interactive Node Enablement (SSH Auto Forwarding)
- **Dynamically adjusts workload or environment to enforce policies and SLA/QoS**
  - Jobs that adapt their resource definitions to fit within available resources
  - Environments that adapt (software provisioning, auto-application compiling and provisioning to match available resources, auto service configuration, etc.)
- **Manages global view of cluster for reporting, billing and charging**

Conclusion

Cluster Resources via Moab provides:

An intelligent Adaptive Computing Foundation which is able to adapt the environment to meet application and workload requirements in an automatic and optimized user-transparent way as well as reduce failure conditions, increase manageability and broaden usability.

Gain “REAL Adaptive Computing Today”!
Start simple,
and then build on a foundation that will maintain
the leadership of your adaptive computing potential.

Appendix

Cluster: Technology Description

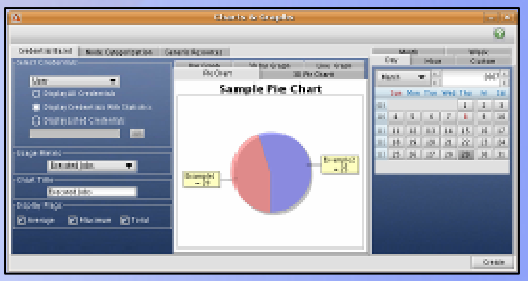
Moab Cluster Suite

- Integrates scheduling, management and reporting
- Integrates hardware, storage, middleware, databases, networks, etc.
- Drives cluster with event and policy-based triggers
- Manages global view of cluster for reporting, billing and charging
- Dynamically adjusts workload to enforce policies and SLA/QoS
- Automates Diagnosis and Failure Response

Cluster: Differentiating Value Propositions

Moab Cluster Suite:

- Integrates with and enhances full HPC stack (i.e. network, storage, middleware, etc.) for holistic management
- Moab translation facilities allow a single solution which can address customers from any background, i.e., TORQUE, SLURM, LSF, PBS Pro, SGE, etc.
- Commonly ½ to ¼ the price of commercial resource managers (LSF, PBS Pro, etc.)
- 'Self-healing' cluster to address common cluster issues
- Provides management GUI showing health and status of standard services
- Advanced infrastructure allows instant upgrade path as needs evolve
- Includes a fully customizable Web-based end user job management and submission portal



Grid: Technology Description

Moab Grid Suite

- Grid workload manager & meta-scheduler
- Maintains individual cluster and group sovereignty
- Works across heterogeneous resources
- Orchestrates scheduling, managing, & monitoring
- Automates job & data migration
- Enforces QoS/SLA policies

Grid: Sovereign Management & Flexible Control

Local Admin can apply policies to manage:

A. Local user access to local cluster resources
B. Local user access to grid resources
C. Outside grid user access to local cluster resources (general or specific policies)

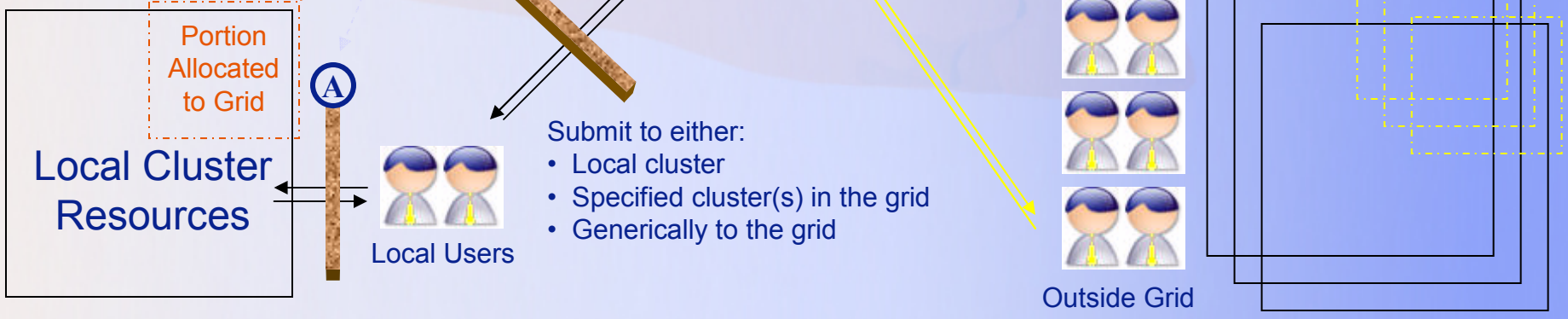
Grid Administration Body

Grid Administration Body can apply policies to manage:

1. General grid policies (Sharing, Priority, Limits, etc.)

Local Admin

Each Admin can manage their own cluster



Grid: Differentiating Value Propositions

Moab Grid Suite:

- Real-world solution
  - Overcomes common grid adoption barriers
- Heterogeneous management
  - Unifies heterogeneous clusters/grids (different RMs, OSs, portals, Globus/non-Globus, architectures, etc.)
- Transparency
  - Requires no change in submission methods due to script and command translation capabilities
- Sovereignty
  - Lets groups/orgs maintain and control their own local resources
- Cost
  - Costs less than competition, leaving more margin for hardware and services

Adaptive: Technology Description

Moab Utility/Hosting Suite:

- Applies mission objectives to nearly any resource (satellites, tape drive robots, networks, licenses, etc.) with its policy engine
- Integrates with existing architecture/framework (DBs, CRM's, custom or vendor specific mgt tools, etc.)
- Monitors and adaptively "grows-and-shrinks" services (Web farm, grid spaces, Databases, etc.)
- Automates manipulation of environments (XEN, VMware, "diskless," diskfull, etc.) with its event engine
- Instant access to additional resources custom built on-the-fly

Adaptive: Differentiating Value Propositions

Moab Utility/Hosting Suite

- Share local resources with internal departments and organizations
- Integrates with HPC and data center monitoring tools
- Gain additional revenues by hosting out & billing access to applications, services or resources
- Guarantee improved service levels and reliability
- Provide resource control to administrators and ease-of-use to end users
- Control information sharing and privacy