Cluster Resources, Inc.

Cluster Resources, Inc.

Grid Computing
"Beyond Enablement"

Next Step Capabilities
building on the basic grid infrastructure.

David Jackson
CTO
Cluster Resources, Inc.

Josh Butikofer
Moab Grid Scheduler Developer
Cluster Resources, Inc.

Feb 2005

Cluster Resources, Inc.

Outline

- Who is Cluster Resources, Inc. (CRI)
- Grid Types
  - Local Area Grid (Campus Grid)
  - Wide Area Grid (Collaboration Grid)
- Next Step Grid Supporting Capabilities and Policies
  - Intelligent Data Staging
  - Co-Allocation & Multi-Sourcing
  - Sovereignty (Local vs. Central Management Policies)
  - Virtual Private Cluster and Virtual Private Grid
  - Service Monitoring and Management
- Usage Cases

Cluster Resources, Inc.

Cluster Resources - Background

- Cluster Resources, Inc.™ is a leading provider of workload and resource management software and services for cluster, grid and utility-based computing environments.
- As the developers of the popular Maui Scheduler and the next generation Moab Workload Manager™, Moab Grid Scheduler™, and other associated products, Cluster Resources has come to be recognized as a leader in innovation and return on investment.
- Providing grid and cluster middleware supporting the largest and most complex clusters and grids in the world, as well as small computing sites, Cluster Resources is able to apply best practices to most any industry and environment.
- With well over 1,500 clients worldwide, and drawing upon over a decade of industry experience, Cluster Resources delivers the software products and services that enable an organization to understand, control, and fully optimize their compute resources.
- Recognized Leaders in Technology, Openness, Service Expertise, and Affordability

Cluster Resources, Inc.

Cluster Resources – Background cont.

- Industry Leading Schedulers
  - Most advanced
    - Functionality
    - Manageability and control
  - Scheduler and Policy Engine of choice
    - Selected by President's Information Technology Advisory Committee as Scheduler of choice (PITAC)
    - US Department of Energy, & Department of Defence
    - Global 1000 biotechnology, energy, manufacturing & computing companies
    - Managing the largest clusters in the world
    - World's largest grid
    - World's largest data center
    - Department of Energy Scalable System Software Project
    - Founding contributor to Global Grid Forum
    - Utility-based computing focused
  - Openness: Supports the widest range of resource manager and platform environments
    - Ensures the customer is not locked into any particular environment, and allows collaboration with other parties without requiring similar environments

Cluster Resources, Inc.

Solution Framework for Grid Computing

End Users: Submit jobs via web form Interface

Administrators: Sets policies and manages via Moab Cluster Manager

Globus

Moab Grid Scheduler

Moab Grid Scheduler: Provides Grid-based policy enforcement and workload management, as well as interfaces with GridFTP for data staging and optionally Globus or other secure identity managers for secure management of access rights

Used in: Wide Area Grids

Moab Access Portal

Moab Access Portal: Provides web-based job submission and management for end users, and passes requests to Moab Workload Manager

Moab Cluster Manager

Moab Cluster Manager: Is a graphical tool that uses wizards and forms to improve ease of use and to unify Workload and Resource Manager functions

Moab Workload Manager

Moab Workload Manager: Enforces policies, monitors workload and controls submissions through resource manager

Resource Manager

Resource Manager Submits Jobs to Compute Nodes

Used in: Local Area Grids & Small to Large Clusters

OS and Communication

Compute Nodes

Cluster Resources, Inc.

Local Area Grid (Campus Grid)

Characteristics

- Shared user space (user name and password structure)
- Shared data space (existing network communication)
- Typically geographically close
- Typically used within a fully or mostly trusted environment (e.g. department to department within a business, university or governmental organization)

Benefits of Local Area Grid vs. Independent Clusters

- Scale (larger single cluster image)
- Collaboration (shared resources, data and personnel)
- Reduce costs
  - Higher performance (better utilization due to reduced fragmentation)
  - Sharing of unique resources (applications, hardware attributes, storage, instruments)
  - Administrative (unification of management tools and infrastructure, reduced training)
  - Managerial control (unified reporting and accounting)
- End-User experience (unification of experience)

Cluster Resources, Inc.

Local Area Grid (Campus/Enterprise)
Multi-Resource Management Mode

Administrators:
Sets policies and manages via Moab Cluster Manager

Moab Cluster Manager

Moab Cluster Manager:
Acts as the Interface, using wizards and forms to improve ease of use and to unify interface to Workload and Resource Managers

Moab Access Portal

Web Server & Moab Access Portal:
Accepts job requests and passes requests to Moab Cluster Manager

Single User Space

End Users: (In a Single User Space)
Submit jobs via web form Interface

Moab Workload Manager

Moab Workload Manager:
Enforces policies, monitors workload and controls submissions through resource manager

Local Area Grid
Head Node:

Resource/Queue Manager → Resource/Queue Manager

Cluster Head Node 1:
Manages Submissions to Compute Nodes

OS and Communication

Compute Nodes – Cluster 1

Resource/Queue Manager → Resource/Queue Manager

Cluster Head Node 2.
…

OS and Communication

Compute Nodes – Cluster 2

Resource/Queue Manager → Resource/Queue Manager

Cluster Head Node etc.:
…

OS and Communication

Compute Nodes – Cluster etc.

Unified Data Space

Cluster Resources, Inc.

Wide Area Grid (Collaboration Grid)

Characteristics

- Not based on cost savings other than access to unique resources.
- Disconnected user space (user name and password structure)
- Disconnected data space (no existing network communication)
- Either geographically separated or architecturally separated
- Typically used within a partially trusted public environment (e.g. academic, government or research)
- Have access to a specialized data link (e.g. Internet2 (US), AARNet (Australia), Super Janet (UK), EARN (Europe), IIJ (Japan), etc.)

Benefits

- Collaboration
  - Sharing of unique resources (applications, hardware attributes, storage, instruments)
  - Collaboration (project data, personnel, time reduction)
  - End-User experience (unification of submission and experience)

Cluster Resources, Inc.

Wide Area Grid

Administrators:
Sets policies and manages via Moab Grid Scheduler (Across the Grid) and Moab Workload Manager (For each Cluster)
Admin also must administer multiple user and data spaces

Grid FTP

Moab Grid Scheduler Interacts with Grid FTP to Stage Data to each of the Clusters

Globus

Moab Grid Scheduler Leverages the Security and Access Control provided in Globus

User Space 1 | User Space 2 | User Space etc.

End Users: (In Multiple User Spaces)
Submit jobs via command Line Interface

Moab Grid Scheduler

Grid Head Node

Moab Workload Manager

Moab Workload Manager: Enforces policies, monitors workload and controls submissions through resource manager

Resource/Queue Manager

Cluster Head Node 1: Manages Submissions to Compute Nodes

Resource/Queue Manager

OS and Communication

Data Space 1

Compute Nodes – Cluster 1

Moab Workload Manager

…

Resource/Queue Manager

Cluster Head Node 2.
…

Resource/Queue Manager

OS and Communication

Data Space 2

Compute Nodes – Cluster 2

Moab Workload Manager

…

Resource/Queue Manager

Cluster Head Node etc.:
…

Resource/Queue Manager

OS and Communication

Data Space etc.

Compute Nodes – Cluster etc.

Cluster Resources, Inc.

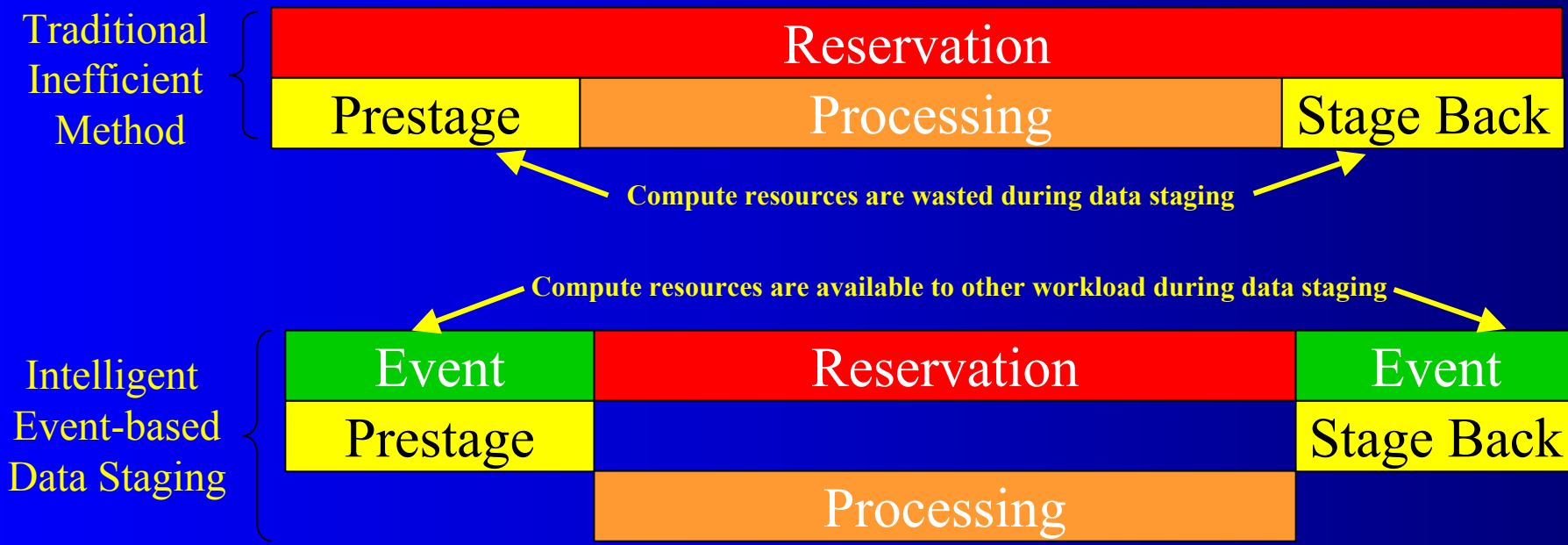
Once your grid can finally communicate…

…what "next step capabilities" are important to make the grid work effectively to accomplish its real goals?

Cluster Resources, Inc.

Intelligent Data Staging

- Automatically pre-stages input data and stages back output data with event policies
- Coordinate data stage time with compute resource allocation
- Use GASS, gridftp, and scp for data management
- Optionally reserve network resources to guarantee data staging and inter-process communication



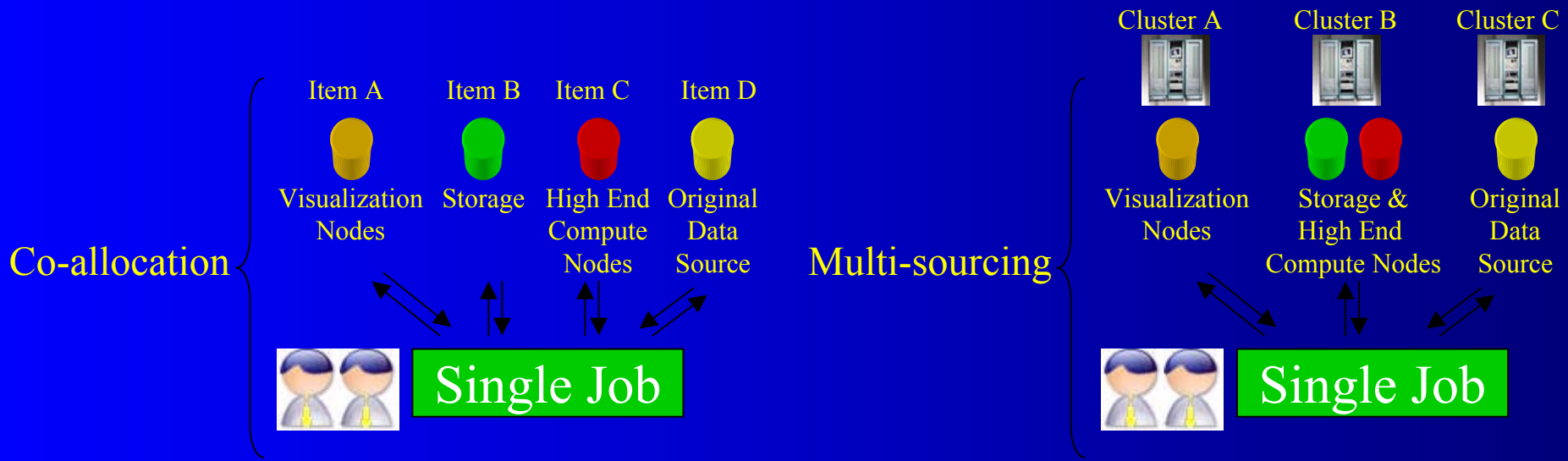
Cluster Resources, Inc.

Optimization from Intelligent Data Staging

Cluster Resources, Inc.

Co-Allocation and Multi-Sourcing

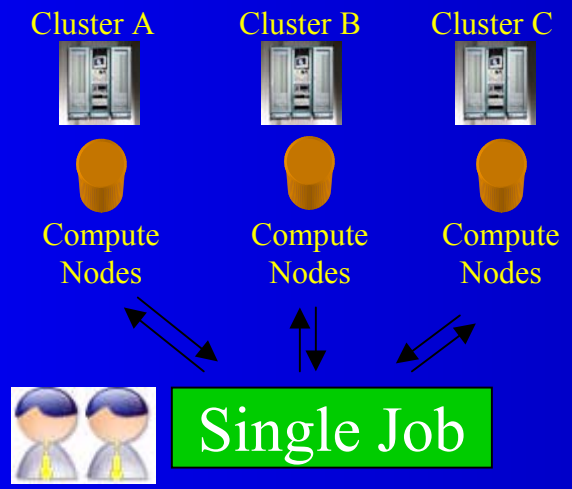
- Workload obtains resource, job, policy, & user information from multiple sources
  - Computational hardware, storage, software licenses, network (bandwidth), and other resources
- Uses and drives multiple services
  - Data managers, job staging services, resource monitors, identity managers, allocation managers and other services
- Similar resources from multiple clusters can be allocated to a job
- Distinct resource types from multiple clusters can be allocated to a job



Cluster Resources, Inc.

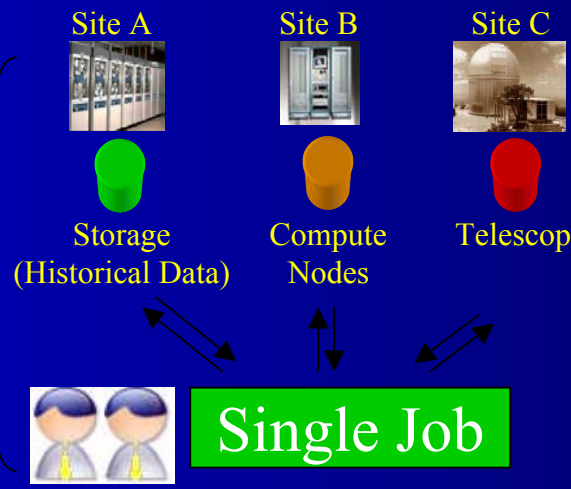
Co-Allocation and Multi-Sourcing Examples

Massively Scalable



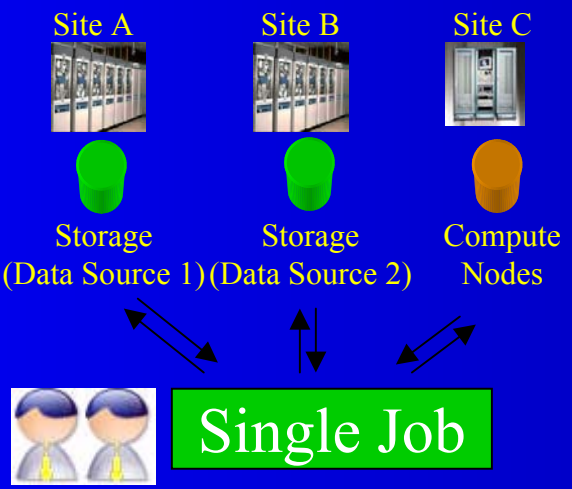
Collaborative Research
(Share Resources)

e.g. Astro Physics



Collaborative Research
(Share Data)

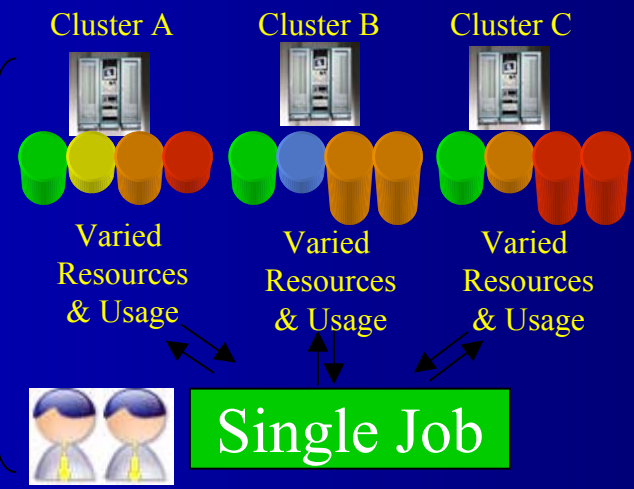
e.g. Genome Project



Cost Optimized Computing
(Unify Departments)

Share licenses
Share capacity
Share instruments
Share costs

Unify administration
Unify experience



Cluster Resources, Inc.

Local vs. Central Management Policies

Local Admin can apply policies to manage:

Local user access to local cluster resources

Local user access to grid resources

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.)

Grid Allocated Resources

Local Admin

Each Admin can manage their own cluster

Portion Allocated to Grid

Local Cluster A Resources

Local Users

Submit to either:

- Local cluster
- Specified cluster(s) in the grid
- Generically to the grid

Outside Grid Users

Cluster Resources, Inc.

Example: Cluster and Grid Policies

Local Cluster

- Local user access to local cluster resources
  - Usage limit of 16 processors at a time for any particular user
  - A department may have up to 30% of the nodes reserved at any one time
- Local user access to grid resources
  - A user may use up to 50,000 credits using grid resources
  - Jobs requiring access to a specific dataset (e.g. bio database) may not submit to the grid
- Outside grid user access to local cluster resources
  - Only 64 processors are made available to Grid jobs during business hours
  - Outside users may only submit jobs with durations of less than 8 hours
  - All outside jobs that run during 8 AM to 6 PM Monday through Friday must be preemptible

Grid

- No site may use more than 50% of the grid resources at any one time, without receiving a reduced priority (Grid Level Fairshare)
- Local users always have affinity for running on grid resources provided by their own site (Grid Level Resource Allocation Policies)
- Jobs with 10 gigabytes of data may only be submitted to sites A, B & C, and no job may be submitted which has data files larger than 100 gigabytes (Grid Level Resource Access Policies)
- For the next two weeks project A has guaranteed access to 2,000 processors spread across three clusters (Grid Level Reservations)

Cluster Resources, Inc.

Virtual Private Cluster

Local Admin:
Centrally manages all resources

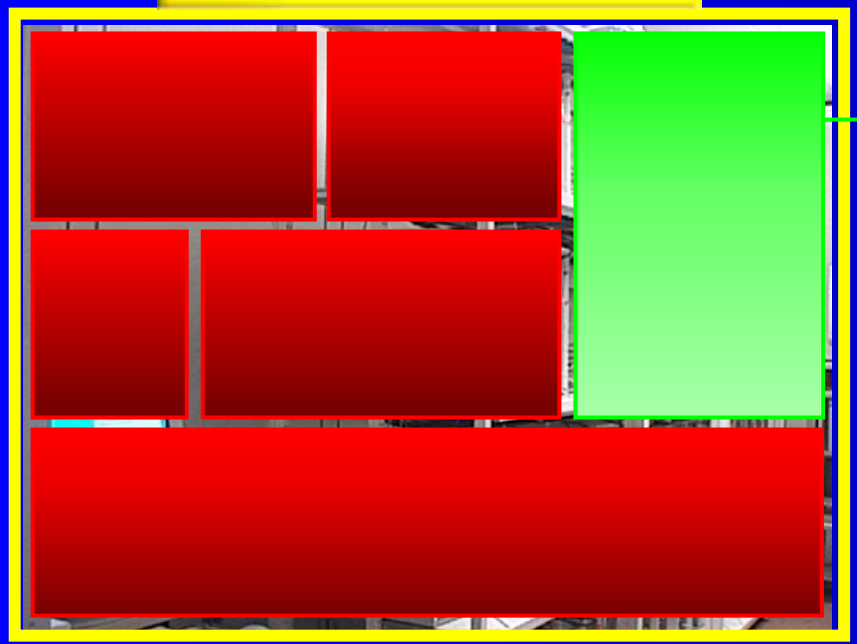
Sub Administrator:
Only sees and manages resources partitioned to his/her group – policies and queues can apply to these virtual clusters

Users:
Only see and access a virtual partition of resources

Virtual Cluster B:

Virtual Cluster etc.:

Compute Resources



Virtual Cluster A:
Virtual Partition of all resources types
(fixed or floating)
and then allocated to a group

Storage

Software Licenses

Network Bandwidth

Etc.

Cluster Resources, Inc.

Virtual Private Grid (Q3 - 2005)

Similar to a Virtual Private Cluster, a Virtual Private Grid is a collection of cluster & grid resources presented in such a way that users only see and interact with those resources for which the user or organization has rights to.

Example:

- Users only see 50% of the available grid resources due to a grid policy which sets a maximum usage limit at 50%
  - This would help users from even trying to submit jobs that would violate grid policies (less frustration, more effective decision making)
- Users do not see resources that have not been allocated to the grid

Cluster Resources, Inc.

Service Monitoring and Management

Service Levels

- Ownership centric
  - Owners can have instant access by preempting others
  - Owners can have the next available resource
  - Owners can have preferred priority levels
- Time dependency centric
  - Deadline Scheduling
  - High Priority, normal priority, low priority (a full spectrum of levels)
- Fairness centric
  - Percentage based Fairshare
  - Target (Quota) based Fairshare
- Cost centric
  - Resource value based access policies (protect high value resources)

Allocation Management

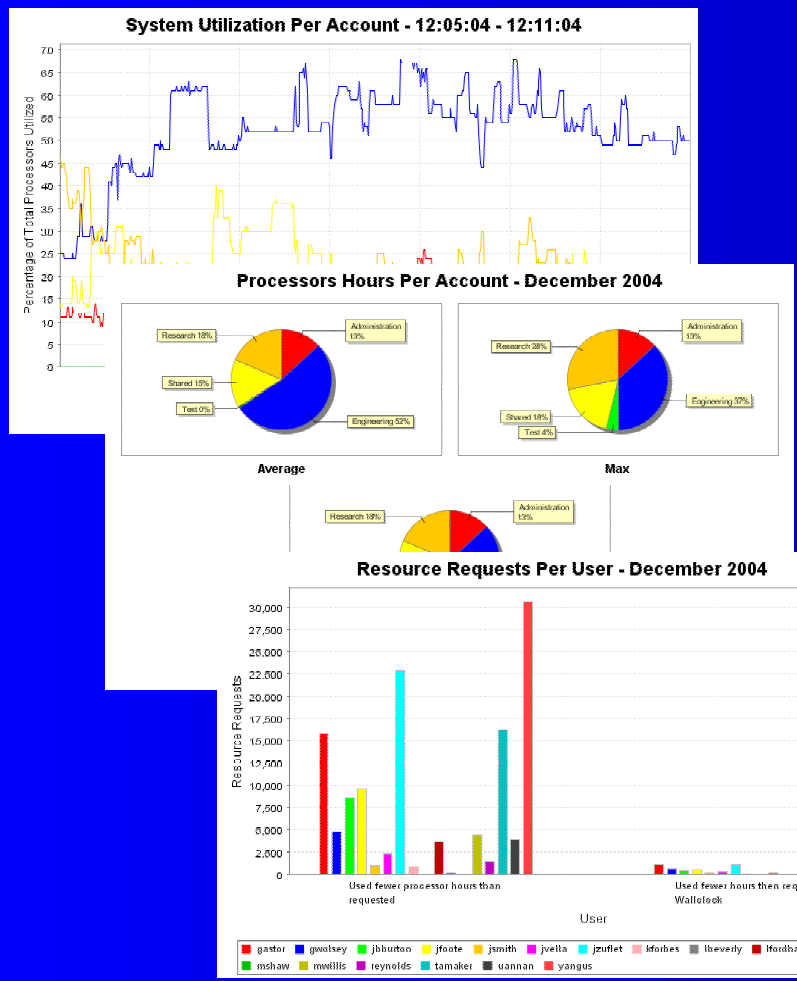
- Credit systems (Gold, QBank)
- CPU hour-based usage limits
- Other resource usage limits (e.g. data, network, license, etc.)

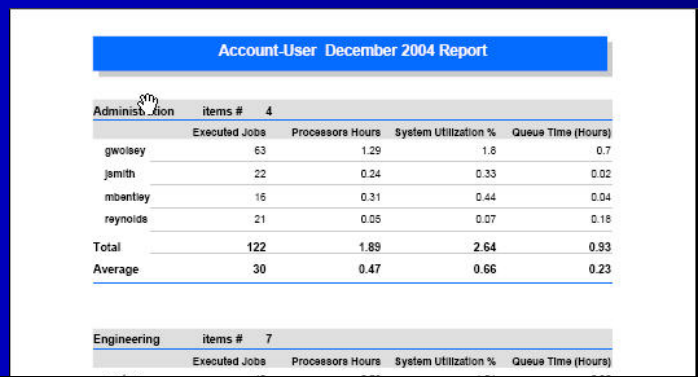
Reporting

- Global, cluster, site, project, user, etc.

Cluster Resources, Inc.

Service Monitoring and Management

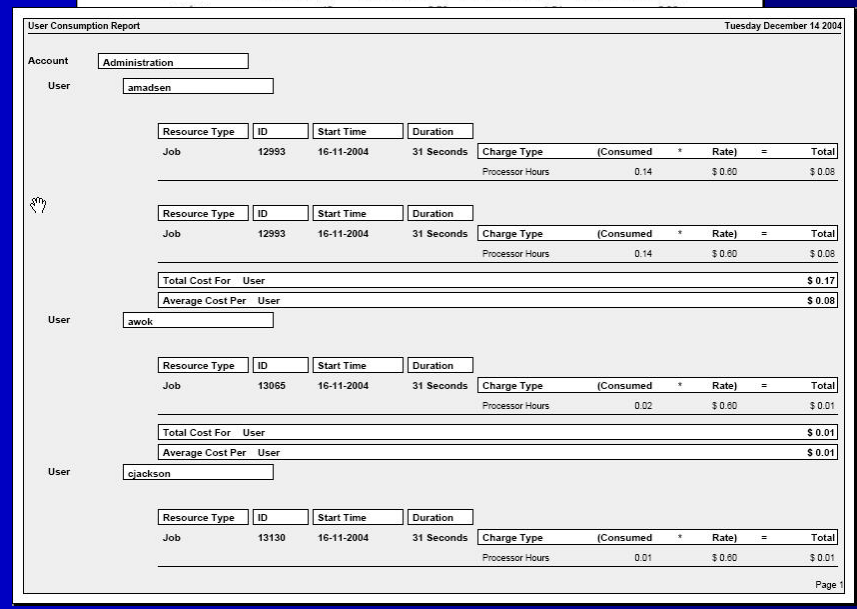




Account-User December 2004 Report

Administration	items #	4			
	Executed Jobs	Processors Hours	System Utilization %	Queue Time (Hours)	
gwolsey	63	1.29	1.8	0.7	
jsmith	22	0.24	0.33	0.02	
mbentley	16	0.31	0.44	0.04	
reynolds	21	0.05	0.07	0.18	
Total	122	1.89	2.64	0.93	
Average	30	0.47	0.66	0.23	

Engineering	items #	7			
	Executed Jobs	Processors Hours	System Utilization %	Queue Time (Hours)	



User Consumption Report — Tuesday December 14 2004

Account: Administration

User: amadsen

Resource Type	ID	Start Time	Duration	Charge Type	(Consumed	*	Rate)	=	Total
Job	12993	16-11-2004	31 Seconds	Processor Hours	0.14		$ 0.60		$ 0.08
Job	12993	16-11-2004	31 Seconds	Processor Hours	0.14		$ 0.60		$ 0.08

Total Cost For User: $ 0.17

Average Cost Per User: $ 0.08

User: awok

Resource Type	ID	Start Time	Duration	Charge Type	(Consumed	*	Rate)	=	Total
Job	13065	16-11-2004	31 Seconds	Processor Hours	0.02		$ 0.60		$ 0.01

Total Cost For User: $ 0.01

Average Cost Per User: $ 0.01

User: cjackson

Resource Type	ID	Start Time	Duration	Charge Type	(Consumed	*	Rate)	=	Total
Job	13130	16-11-2004	31 Seconds	Processor Hours	0.01		$ 0.60		$ 0.01

Page 1

Cluster Resources, Inc.

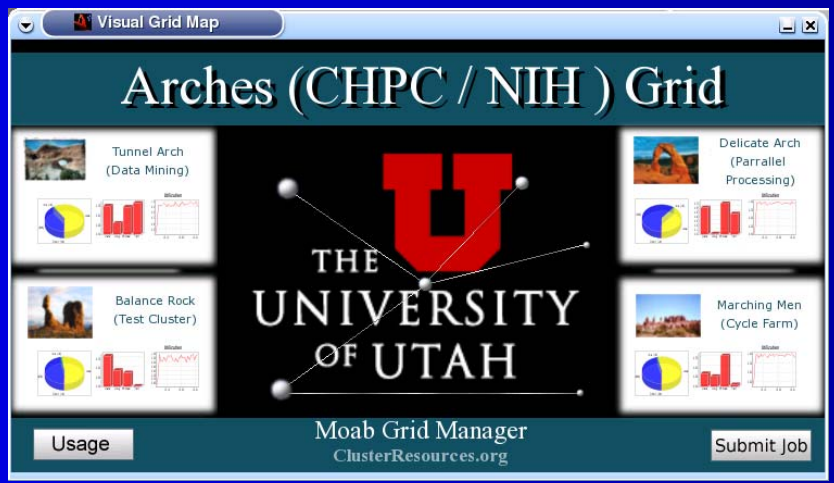
Usage Cases

Real world applications.

Cluster Resources, Inc.

Center for High Performance Computing (CHPC)

- 7 Clusters (Local Area Grid – Common Data and Admin Domain)
  - Specialty Clusters (Clusters optimized for specific job types)
    - Serial jobs, large data, fast interconnect
  - Shared Clusters (Multiple owners)
    - Federation/Condominium-style resources
- Using Moab Grid Scheduler
  - Integrate multiple clusters (single cluster image)
  - Cluster independence (failure isolation)
  - Resource optimization (utilization and responsiveness)
  - Enforce global allocation policies
  - Enabling multi-grid access



Cluster Resources, Inc.

Ohio Supercomputing Center (OSC) & Cluster Ohio Project

- 12 Clusters - moving to 22 (Wide Area Grid – multiple data, user & admin domains)
  - Hardware specific clusters
    - Cray, Altix, Linux clusters, etc.
  - Geographically distributed clusters (multiple university owners)
- Using Moab Grid Scheduler
  - Integrate multiple clusters (resource access & collaboration)
  - Automated data staging
  - Common user interface (mask grid complexities)
  - Resource optimization (utilization and responsiveness)
  - Enforce global allocation policies
  - Enabling multi-grid access

Cluster Resources, Inc.

Example Participating Grid Sites

- TeraGrid
  - NCSA, SDSC, and other leading edge US-based government and academic sites enabling cluster spanning and co-allocation centric jobs
- China Meteorological Association (CMA – Chinese Weather Grid)
  - Clusters unified as a Local Area Grid for scaling purposes for 3,200 processor system
- University of Tromsa Computing Center (Norway – Part of Nordagrid)
  - Multiple Scandinavian Universities joining resource for collaboration
- WestGrid (Canada)
  - 7 Member sites of Western Canada integrating geographically distributed and specialized resources
- National Oceanic and Atmospheric Association (NOAA – US Weather Grid)
  - 3 Principle sites merging resources for Weather Forecasting

Summary of Globus Integration

- Moab Grid Scheduler integrates with Globus
  - 2.2.x
  - 2.4.x
  - 3.x.x
- Data management
  - Input and output data staging with GASS and GridFTP services
- Job management
  - Job staging with GRAM/Gatekeeper services
- User management
  - Support of Globus user mapping files
- Security
  - X509-based client authentication

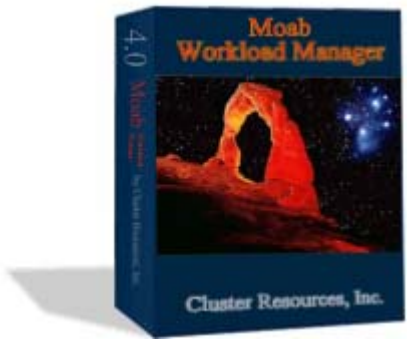
Cluster Resources, Inc.

Q & A

Appendix

Cluster Resources, Inc.

Moab Workload Manager™

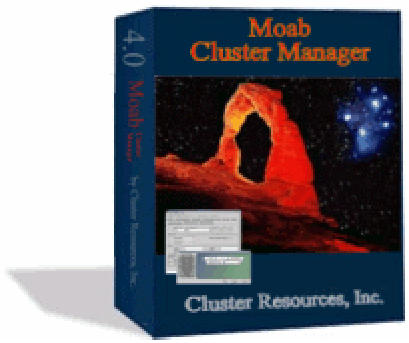


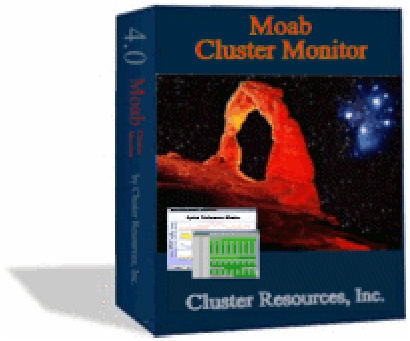
Moab: Our next Generation Product Family

- All Maui scheduling and policy management capabilities
- Supports additional resource types (licenses, filespaces, etc.)
- Event Policies (maintenance tasks, added automation)
- High availability fallback
- Local Area Grid support (campus/enterprise grids)

Cluster Resources, Inc.

Moab Cluster Manager™ & Monitor™



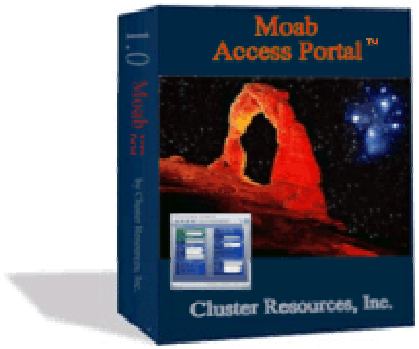


Moab Cluster Manager & Monitor:

- Save significant time in both learning and operating Moab
- Reduce user and admin errors with intelligent wizards & forms
- Diagnose and resolve issues in moments with visual interface
- Report on usage and QoS delivery with customizable reporting
- Use even policies to build automated processes

Cluster Resources, Inc.

Moab Access Portal™

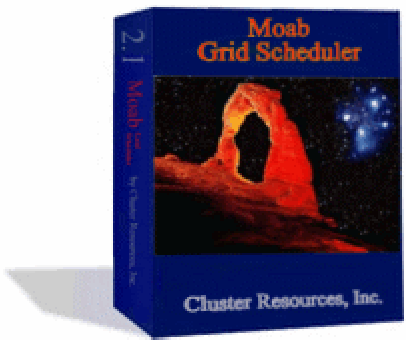


Moab Access Portal:

- Submit workloads/jobs from any location by use of a browser
- Scales to environments with thousands of users
- Reduces administrative work
- End-users are able to review and manage the status of their own current workloads/jobs

Cluster Resources, Inc.

Moab Grid Scheduler (SILVER)

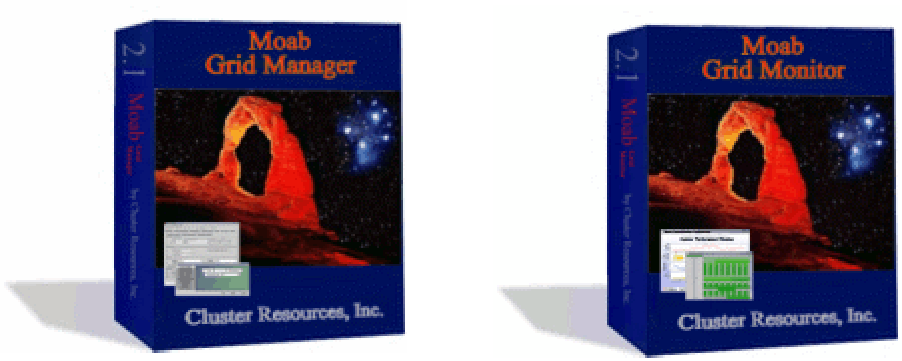


Moab Grid Scheduler:

- Optimize resources across multiple clusters (100+ Clusters)
- Ensure access policies match political/organizational rules
- Allow both global and local workload/usage policies
- Resolve grid usage issues with detailed accounting and diagnostic tools

Cluster Resources, Inc.

Moab Grid Manager & Monitor

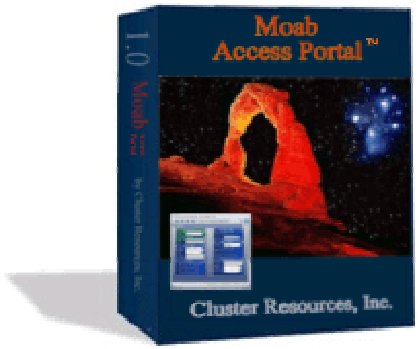


Moab Grid Manager & Monitor:

- Instant visual overview of resource sharing
- Automate reporting of grid usage
- Support allocation management for accounting purposes
- Under development (beta)

Cluster Resources, Inc.

Moab Access Portal for Grids™

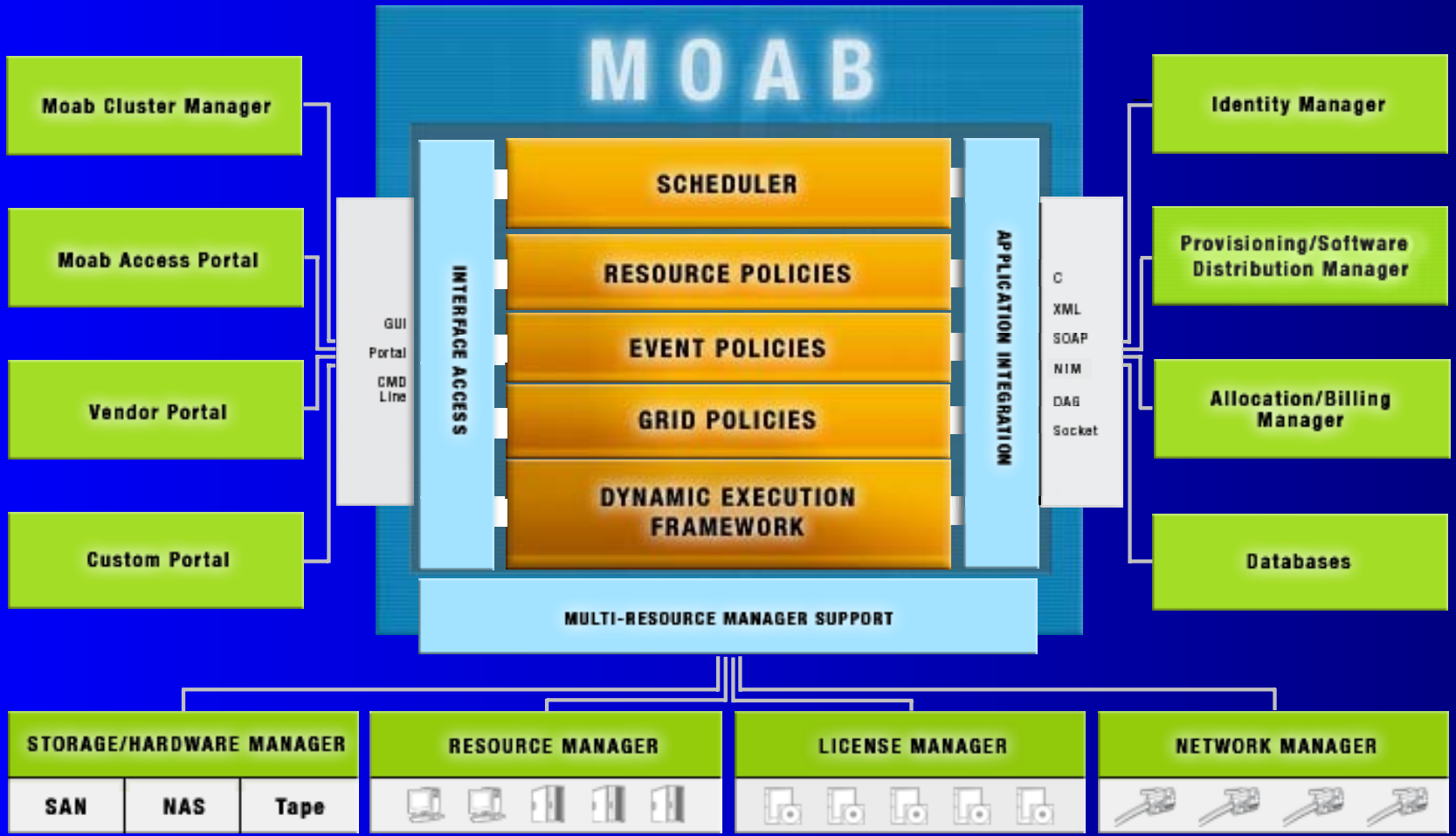


Moab Access Portal for Grids:

- Submit workloads/jobs from any location by use of a browser
- Scales to environments with thousands of users
- Reduces administrative work
- Under development (alpha)

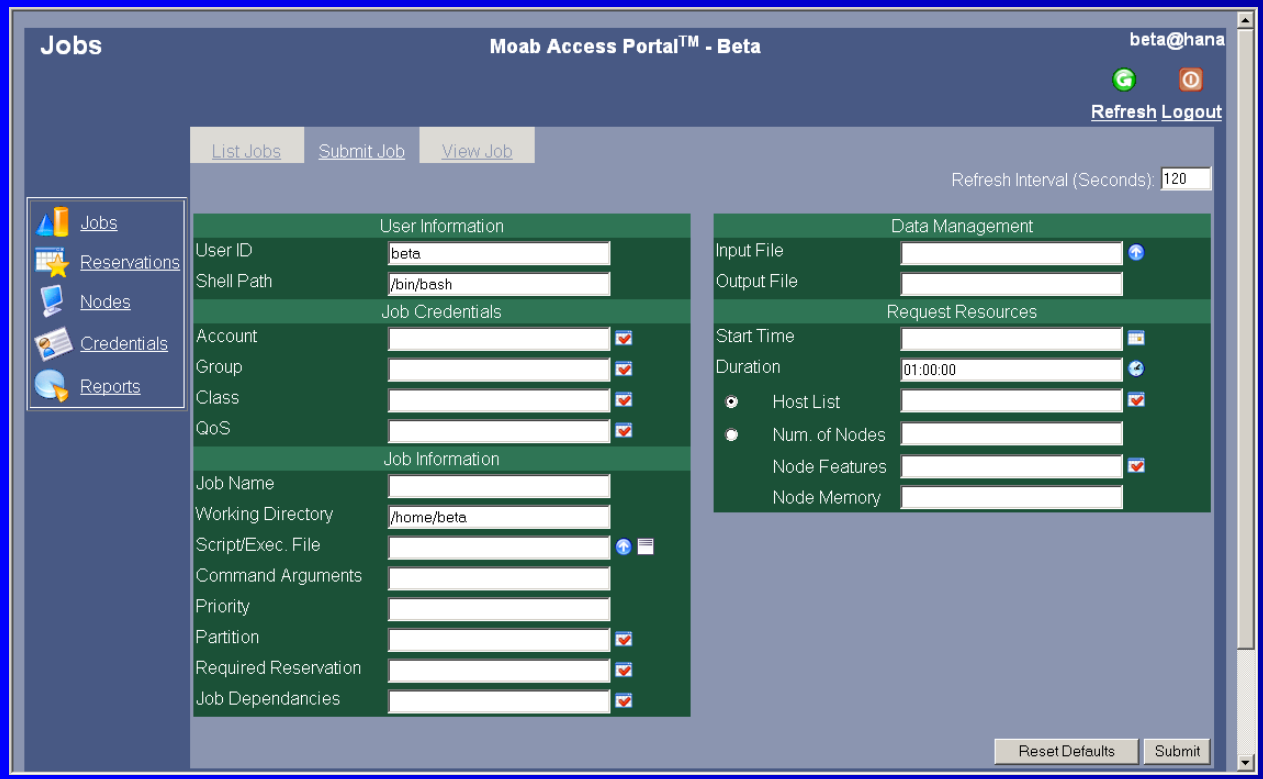
Cluster Resources, Inc.

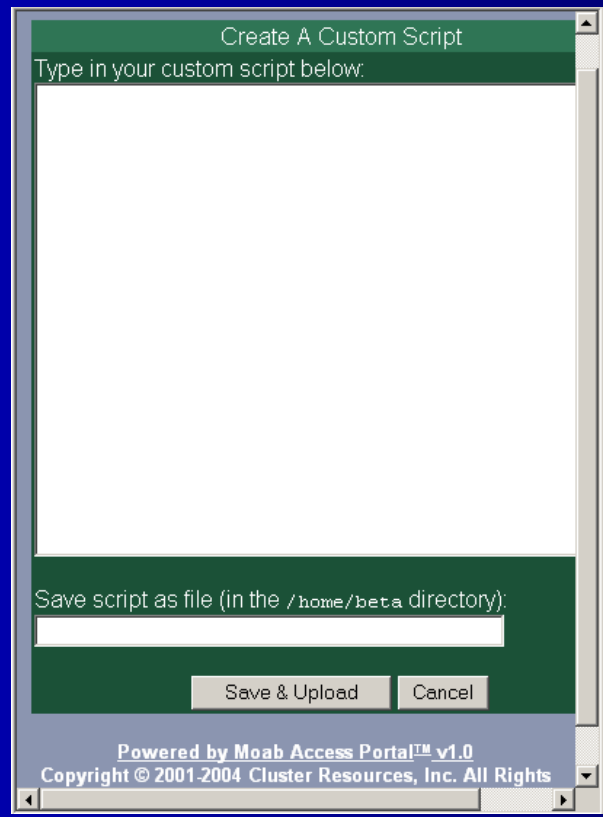
UTILITY - BASED COMPUTING



Cluster Resources, Inc.

Simplified End-User Submission





Cluster Resources, Inc.

Administer through an easy to use GUI

Cluster Resources, Inc.

Moab Grid Scheduler (Silver):

What is it:

- An advanced reservation based job scheduler/policy manager that empowers organizations to optimize distributed workloads to be run across independent clusters.

Benefits:

- Optimize resources across multiple independent clusters (scales to more than 100 clusters)
- Access global resources from a single point with ease (from a few processors to multi-teraflop supercomputers)
- Manage the global resource complexities of the unified system while maintaining local autonomy
- Resolve grid environment issues using detailed accounting of consumed resources and diagnostic tools
- Enforce global usage policies that manage consumption across departments or with outside organizations

Cluster Resources, Inc.

Moab Grid Scheduler (Silver):

Where does it fit:

- Silver fits at the Grid layer above the local or cluster scheduler, in order to empower the access to and optimization of all clusters as a combined whole.

Supported Environment:

Requirements:	Supported Platforms:

• OpenPBS
• TORQUE Resource Manager
• PBSPro
• LoadLeveler
• LSF
  ○ Limited Support
• Scalable System Software (SSS-RM)
  ○ Under Development
	• Linux
	• AIX
	• OSF/Tru-64
	• Solaris
	• HP-UX
	• IRIX
	• FreeBSD
	• Other UNIX platforms

Silver – Security/Privacy

- Account mapping
  - Uses Globus map file
- Job staging
  - Uses Globus Gatekeeper
- Resource availability Query
  - Can only see aggregate resource availability
  - Cannot see nodes
  - Cannot see policies
- Job Management
  - Can only see jobs/reservations it owns
  - Can only manage jobs/reservations it owns

Silver – Fault Tolerance

- Supports object messages
- Reports low level failures via diagnostic commands
  - provides low level failure message originating in globus, moab, torque, or data manager
- Evaluates grid wide resource availability and reports low level reason for blockage

Grid – Admin Commands

- sjobctl (Manage Jobs)
  - Query global and cluster level job state, history, and statistics
  - Diagnose job failures and resource availability
  - Modify job attributes and constraints
  - Cancel hold and force execute job

Grid – Admin Commands

- sresctl (Manage Resources)
  - Modify Resource State
  - List grid jobs and reservations currently utilizing resource
  - View Resource Statistics and History
  - Diagnose Resource Failures

Grid – Admin Commands

- srsvctl (Manage Reservations)
  - Create Single and Multi-Cluster Grid Reservations
  - List Grid Reservations
  - Remove Grid Reservations

Grid – Admin Commands

- suserctl (Manage Users)
  - Query statistics and current usage for grid users
  - Manage resource access and policies for grid users
  - Diagnose user issues

Grid – Admin Commands

- sgridctl (Manage Grid)
  - Query statistics and current usage for grid
  - Set global grid policies and configuration
  - Diagnose grid issues

Grid – Admin Commands

- squery (Query Grid)
  - Query resource availability subject to specific constraints

Grid – Admin Commands

- sqsub (Submit Grid Jobs)
  - Submit Jobs to the Grid

Cluster Resources, Inc.

Grid Level Scheduling Polices

- Management of grid workload prioritization
- Grid level fairshare support
- Per user grid throttling policies

Cluster Resources, Inc.

Examples

- sqsub –l nodes=1@osc,walltime=1:00:00
- srsvctl –c 4:sp2+2@anl –s Feb10 –d 24:00:00
- sjobctl –v 234.gridmaster.edu
- sqsub –l nodes=1600,tpc=200