Moab Workload Manager on Cray XT3

NATIONAL CENTER
FOR COMPUTATIONAL SCIENCES

presented by

Don Maxwell (ORNL)

Michael Jackson (Cluster Resources, Inc.)

Oak Ridge National Laboratory
U.S. Department of Energy

MOAB Workload Manager on Cray XT3

- Why MOAB?
- Requirements
- Features
- Support/Futures

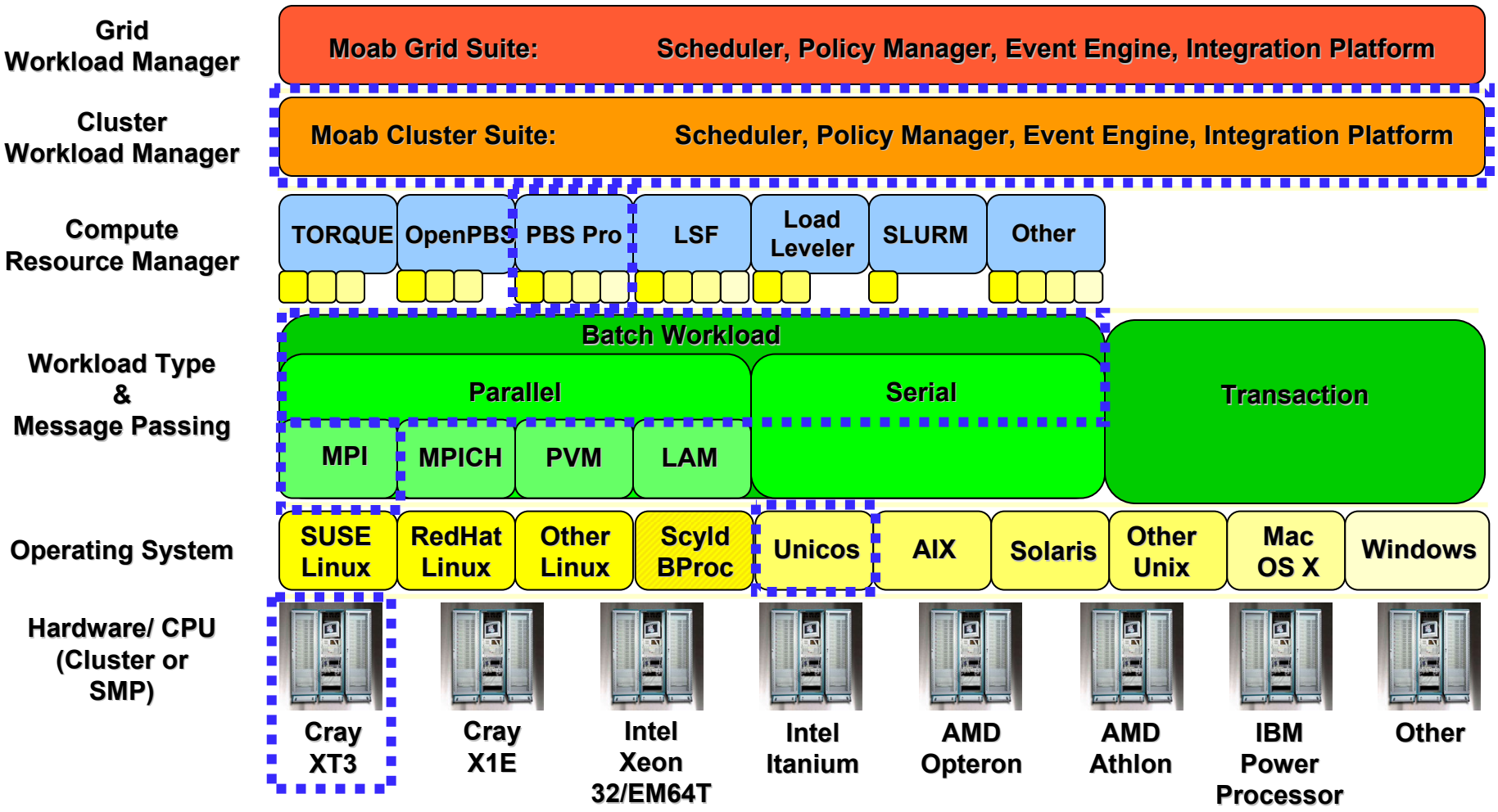
Why Moab?

- Existing platforms
  - **Cray X1E/XT3 and SGI Altix running PBSPro**
  - **Opteron-based visualization cluster running SLURM**
  - **IBM Power 4 running LoadLeveler**
- Traditional resource management features missing
  - **Dynamic Backfill for high resource utilization**
  - **“Run this Job Next” (e.g. llfavorjob)**
  - **Diagnostics too limited (e.g. determine “topdog”)**
- More flexible scheduling for NLCF resources
  - **Integration with resource allocation system (RATS)**
  - **Dynamic prioritization of jobs**
- Reputation of Cluster Resources, Inc.
  - **Developers of leading scheduling systems Maui and Moab**
  - **New ORNL staff with prior experience with Cluster Resources**

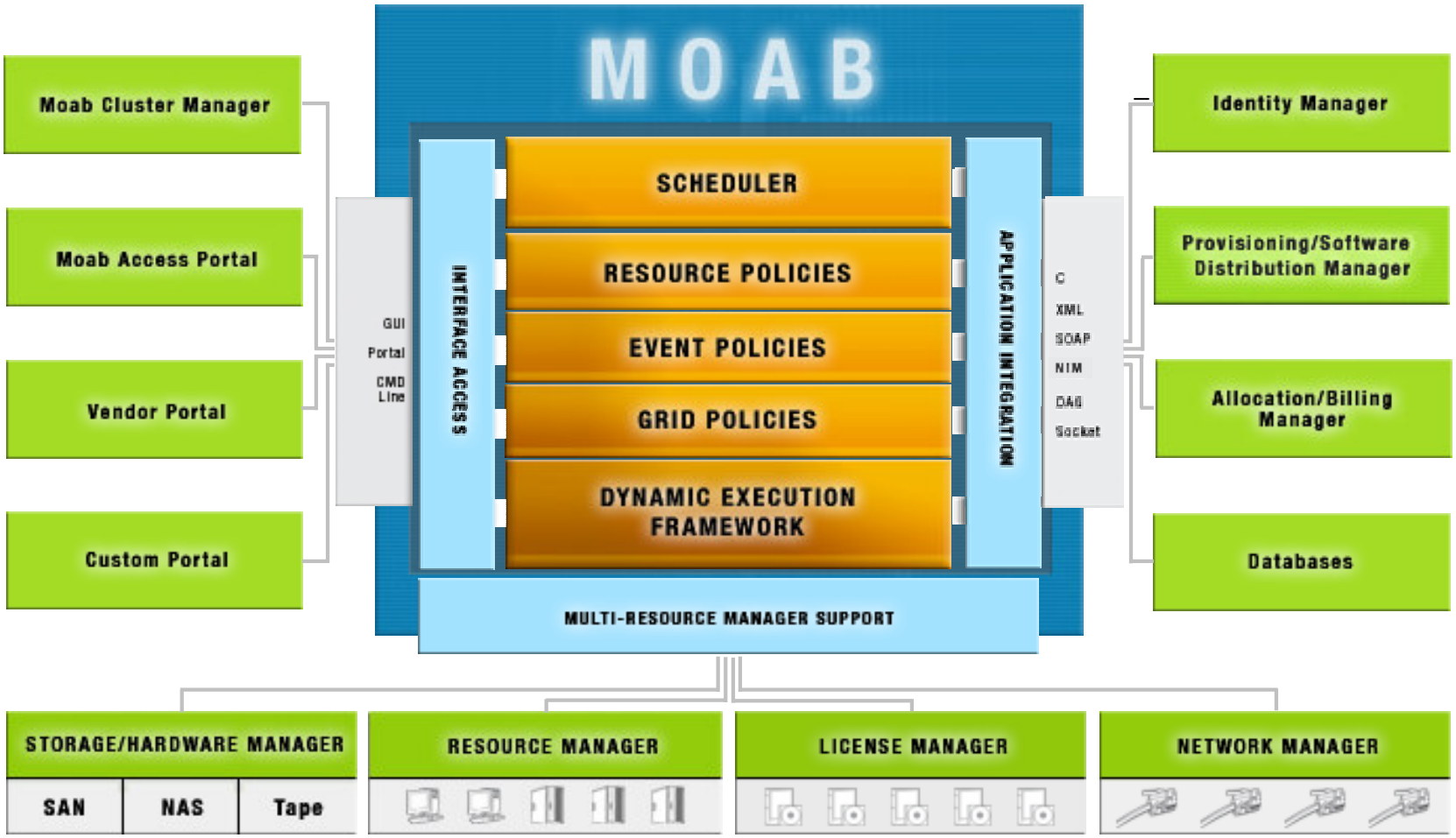
Cluster Resources' Solution Space

- Cluster (Intelligence and Orchestration)
  - **Workload Management**
  - **Resource Management**
  - **Allocation Management**
  - **Event and Condition Management**
- Grid
  - **Grid Workload Management**
  - **Grid Job Submission (Grid Portal)**
  - **Grid Allocation Management**
- Utility/Hosted Computing
- Services
  - **Support**
  - **Consulting (Solution Design, Optimization, etc.)**
  - **Integration**
  - **Development**
  - **Training**

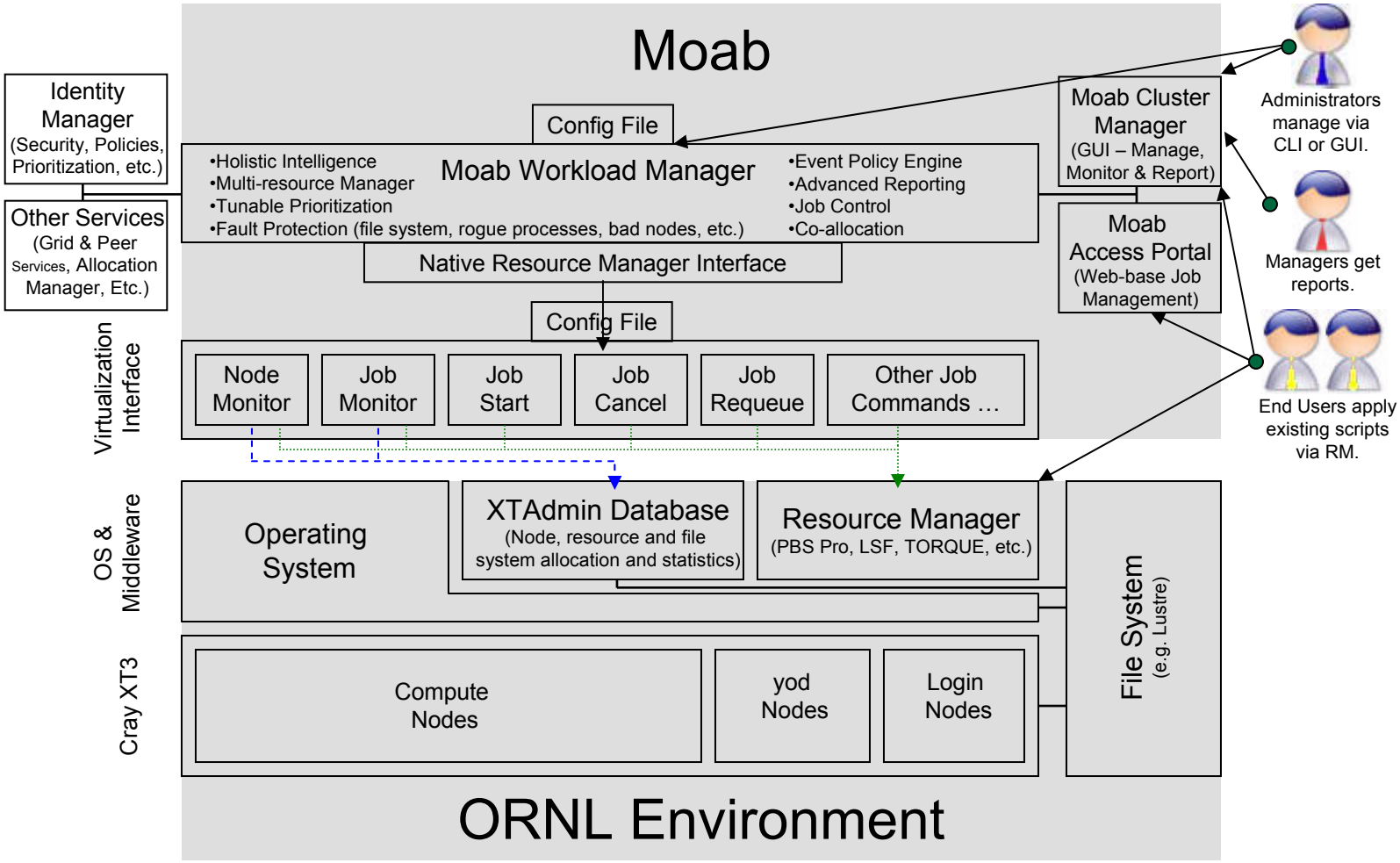
Solution Framework: Where Does it Fit?



Integration Points:



Moab – XT3 Integration at ORNL



Requirements

- Port Moab to XT3, deploy It, integrate it with 3 tools, test it and go production in 3 Weeks!
- No disruption of production work
  - No change in the interface to the batch system on NLCF machines
  - Monitor and Simulation Mode allows evaluation of priority changes without impacting the current job queue
- Dynamic Backfill
  - Static backfill (a.k.a. FIFO backfill) = underutilization
  - Cathy Scheduler
- More control over workload to meet objectives better
  - "run-this-job-next"
- More flexible prioritization supporting fairshare
- Preemption for visualization workload

Moab Features for a System Administrator

- **Monitor and Simulation mode**
  - Installing new versions
  - Changing policies

Moab: Safe Evaluation and Deployment

- **Monitor Mode**
  - Moab monitors all information, processes and policies, decides what it would do, but does not implement it.
- **Simulation Mode**
  - Moab imports historical or scenario based information and allows offline evaluation of changes to resources, workload and policies.
- **Interactive Mode**
  - Moab fully schedules workload, but asks administrator to approve each decision before implementing it.
- **Partition Mode**
  - Moab applies full workload management to a subset of resources, all other resources are unaffected. Also allows for mixed production and test environments.
- **Normal Mode**
  - Normal production mode with full capabilities.

Moab Features for a System Administrator

- **Monitor and Simulation mode**
  - Installing new versions
  - Changing policies
- **Diagnostics**
  - Load the scheduler is placing on the server
  - Accounts and associated attributes (priorities, QOS, etc.)
  - Job priorities
  - Prior failures and reasons
  - Configuration problems

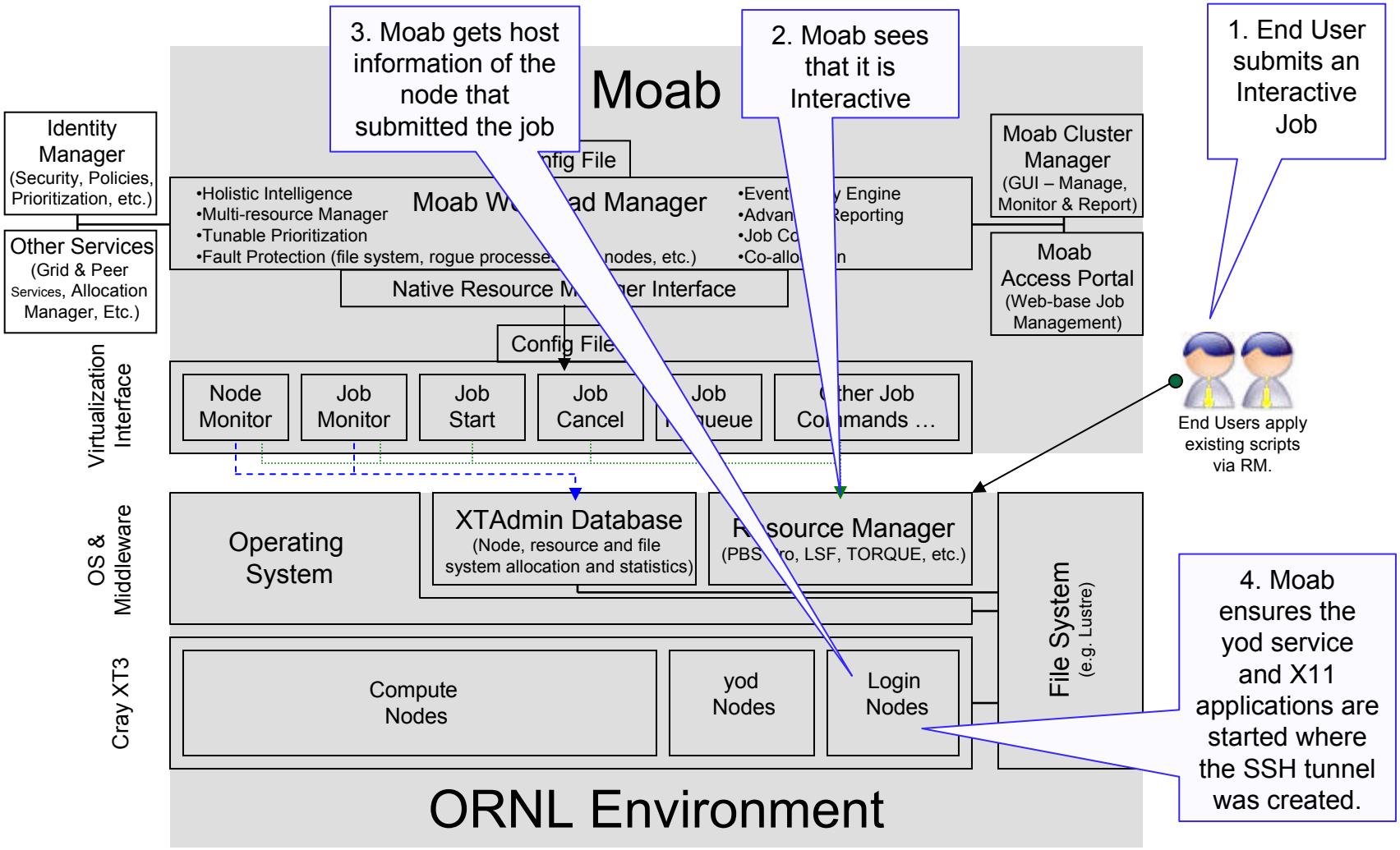
Moab Features for a System Administrator

- **Monitor and Simulation mode**
  - Installing new versions
  - Changing policies
- **Diagnostics**
  - Load the scheduler is placing on the server
  - Accounts and associated attributes (priorities, QOS, etc.)
  - Job priorities
  - Prior failures and reasons
  - Configuration problems
- **Admin levels**
  - User support could see all and influence jobs
  - Operators could see all but not change anything

Moab Features for a System Administrator

- **Monitor and Simulation mode**
  - Installing new versions
  - Changing policies
- **Diagnostics**
  - Load the scheduler is placing on the server
  - Accounts and associated attributes (priorities, QOS, etc.)
  - Job priorities
  - Prior failures and reasons
  - Configuration problems
- **Admin levels**
  - User support could see all and influence jobs
  - Operators could see all but not change anything
- **SSH X11 forwarding**
  - Multiple login and yod nodes complicated with batch
  - Moab perl interface (Virtualization Layer)
    - Easy placement of jobs
    - Knowledge of submitting host and interactive state retrieved from PBS Pro

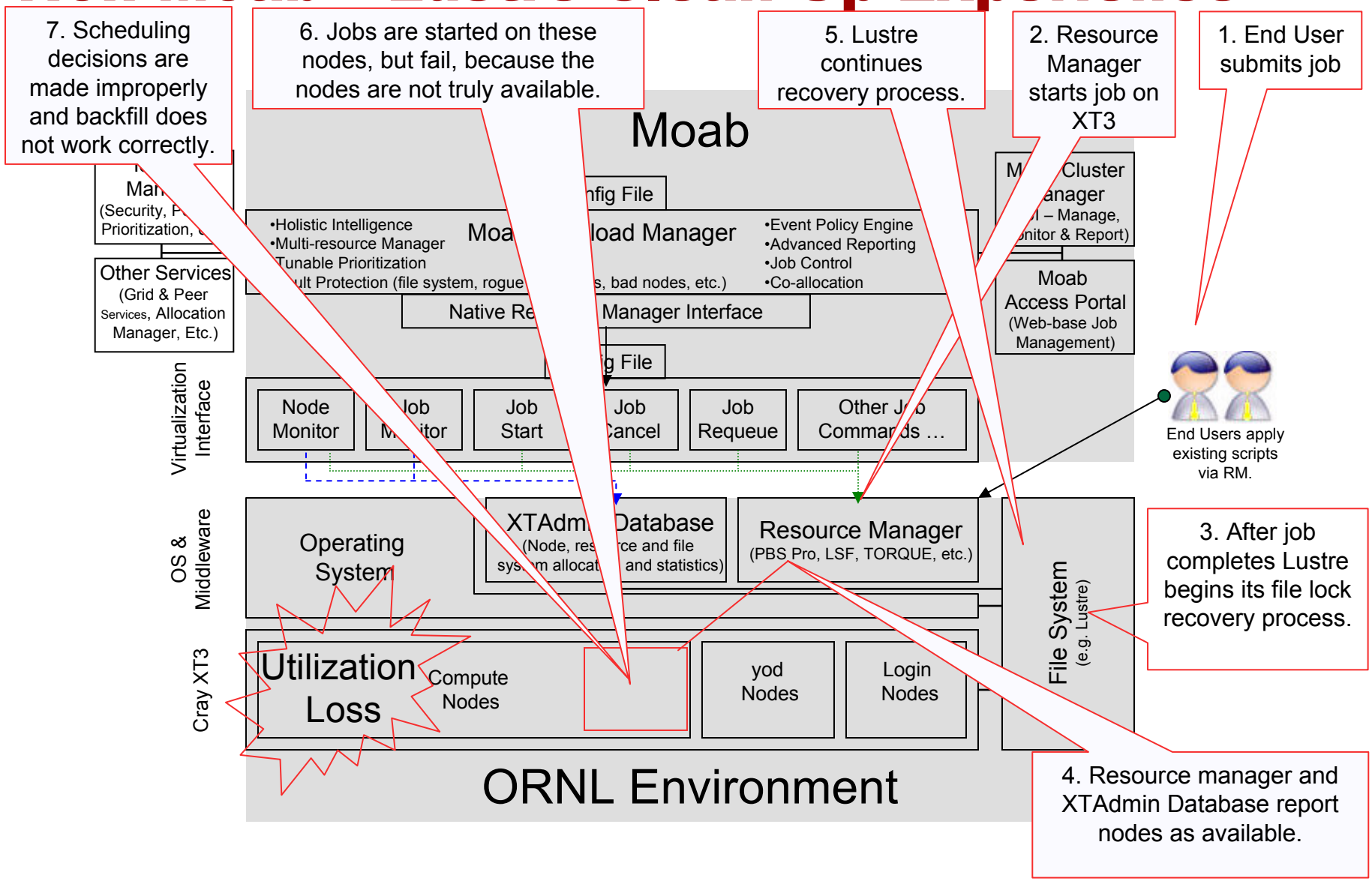
Moab – SSH X11 Interactive Sessions



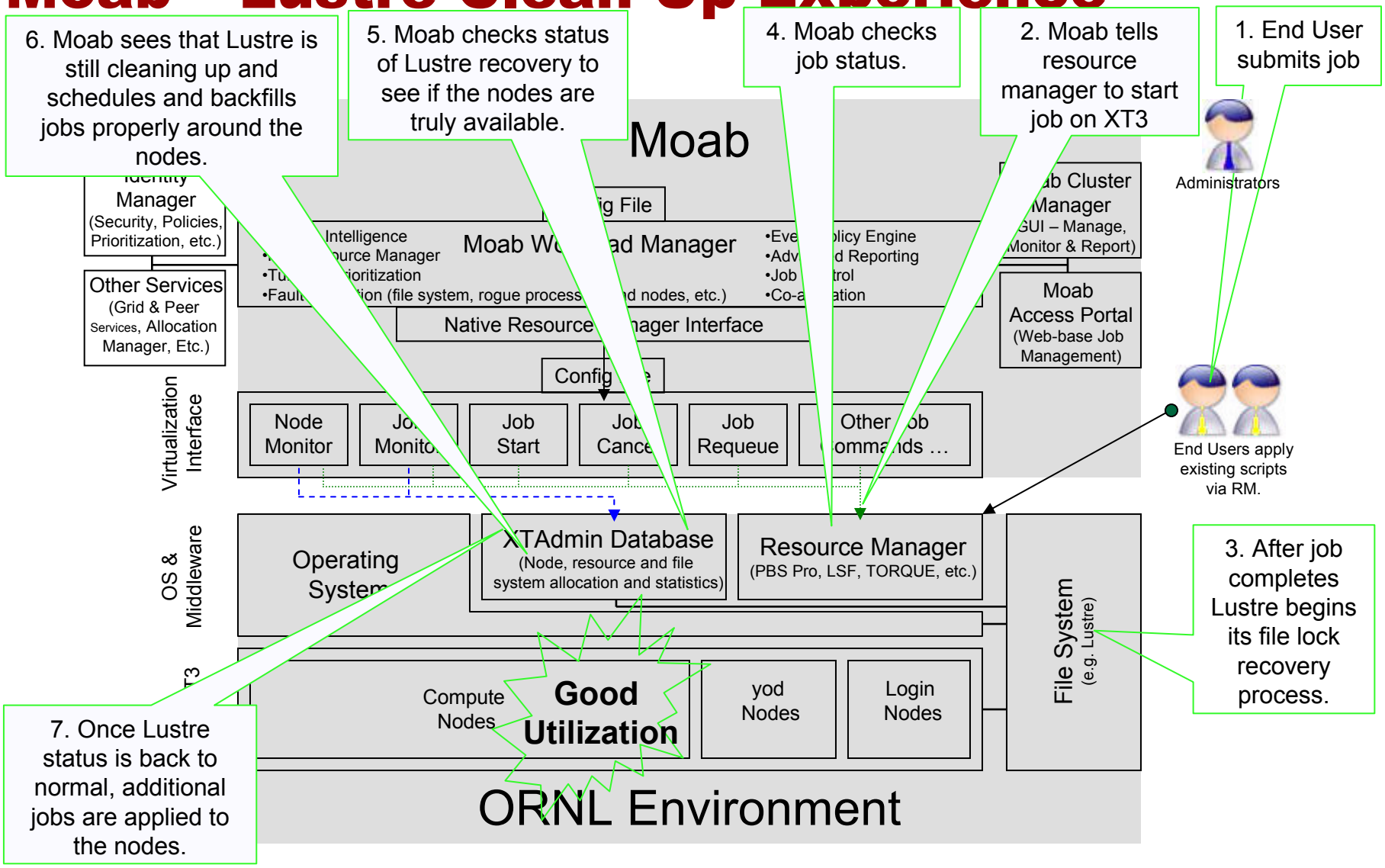
Moab Features for a System Administrator

- **Monitor and Simulation mode**
  - Installing new versions
  - Changing policies
- **Diagnostics**
  - Load the scheduler is placing on the server
  - Accounts and associated attributes (priorities, QOS, etc.)
  - Job priorities
  - Prior failures and reasons
  - Configuration problems
- **Admin levels**
  - User support could see all and influence jobs
  - Operators could see all but not change anything
- **SSH X11 forwarding**
  - Multiple login and yod nodes complicated with batch
  - Moab perl interface (Virtualization Layer)
    - Easy placement of jobs
    - Knowledge of submitting host and interactive state retrieved from PBS Pro
- **Discovered Lustre node recovery state missing in CPA and PBS Pro**
  - Diagnostics revealed jobs were failing
  - Moab modified to consult Lustre recovery table in SDB (database)

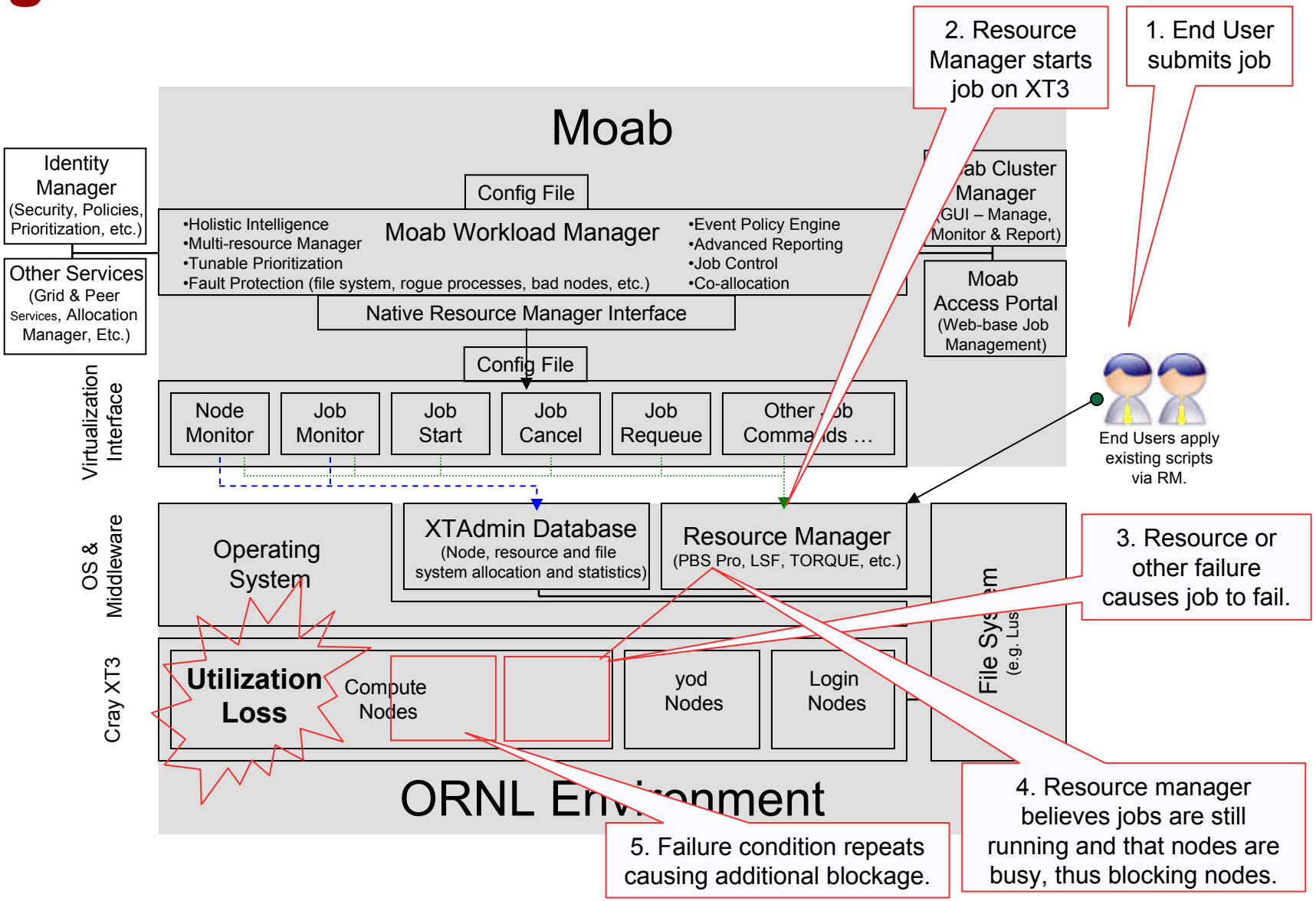
Non Moab – Lustre Clean Up Experience



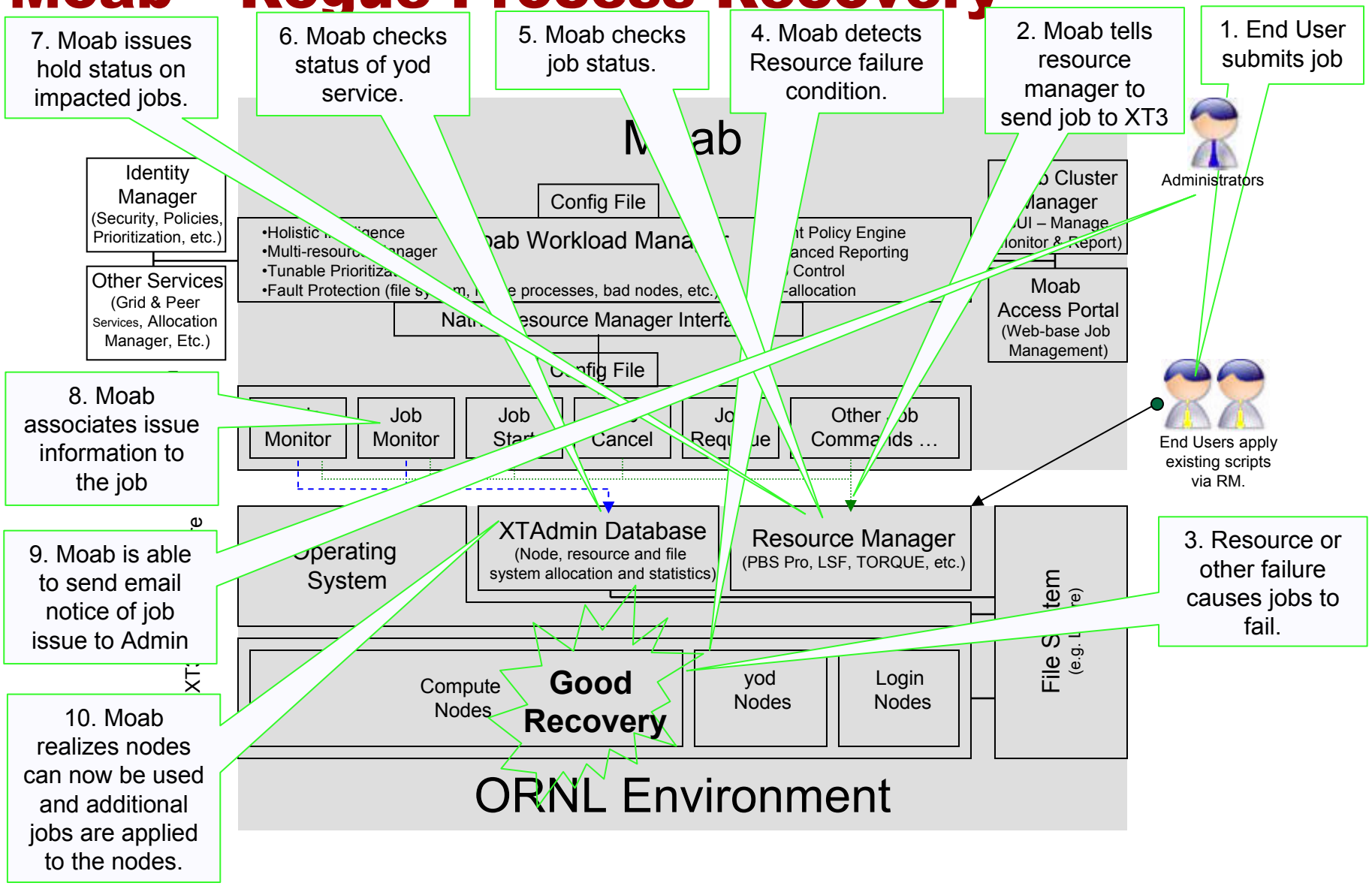
Moab – Lustre Clean Up Experience



Rogue Process Issues



Moab – Rogue Process Recovery



ORNL's Moab Priority Implementation

Factor	Unit of Weight	Actual Weight (Minutes)	Value
Quality of Service	# of days	1440	High (7)
Account Priority	# of days	1440	Allocated Hours (0)
			No Hours (-365)
Queue	# of days	1440	Debug (5)
			Batch (0)
Job Size	1 day / 1000cpu	1	Provided by Moab
Queue Time	1 minute	1	Provided by Moab

Job	PRIORITY*	Cred(Accnt:	QOS:Class)	Serv(QTime)	Res(Proc)
Weights		1(1440:	1440: 1440)	1(1)	1(1)
69099	**7298**	98.7(0.0:	0.0: 5.0)	0.0(2.1)	1.3(96.0)

The resulting priority is a simple calculation.

1440 * 5 + 2 + 96 = **7298**

Moab's Flexibility w/ Multi-Factor Prioritization

- Fairshare
  - **User**
  - **Group**
  - **Account**
  - **Class**
  - **Quality of Service**
  - **Jobs per User**
  - **Processor Seconds Per User**
  - **Processors Per User**
- Resources
  - **Node**
  - **Disk**
  - **Processor**
  - **Memory**
  - **Swap**
  - **Processor Seconds**
  - **Processor Equivalent**
  - **Walltime**
- Usage
  - **Consumed**
  - **Remaining**
  - **Percentage Consumed**
- Service Levels
  - **Queue Time**
  - **XFactor**
  - **Policy Violation**
  - **ByPass**
- Target Levels
  - **Queue Time**
  - **XFactor**
- Credential Based
  - **User**
  - **Group**
  - **Account**
  - **Class**
  - **Quality of Service**
- Attribute
- State
- And More….

Moab Features for an End User

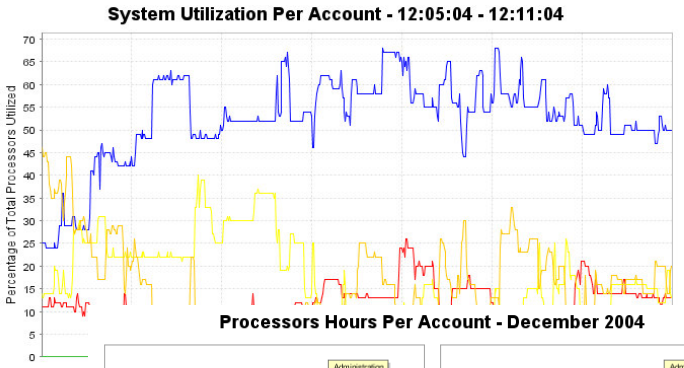
- showbf
  - **Determine the size and length of a job that will backfill at any instance in time**
- checkjob
  - **"Why is my job not starting?"**
  - **Has it tried to run and failed due to a system problem?**
- showstart
  - **An estimated start and end time for jobs currently sitting in the queue**

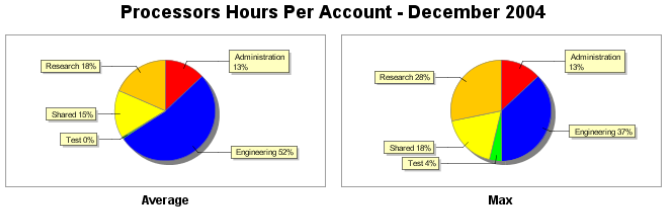
Moab Features for Management

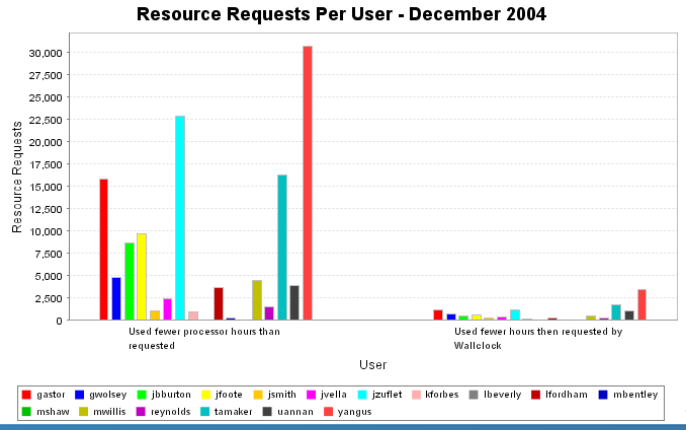
- **Difficult to Visualize Cluster**
  - Usage
  - Performance
  - Historical Data (Big Picture and Capacity Planning)

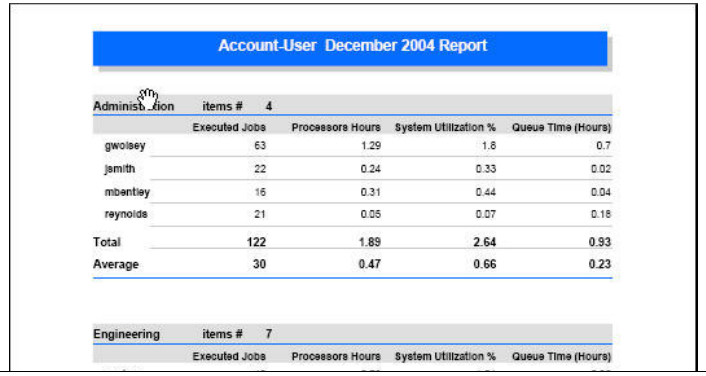
Moab Reports

- Service Monitoring and Management



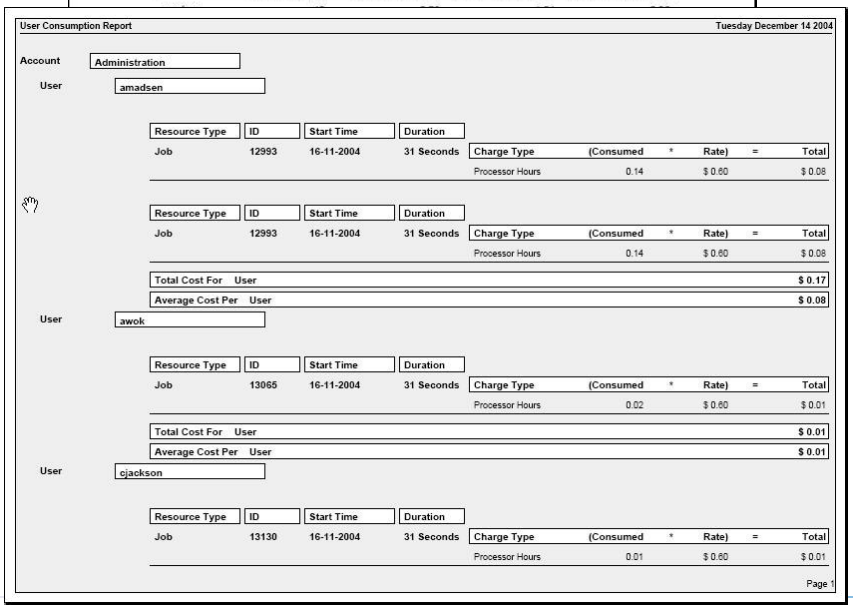






Account-User December 2004 Report

		Executed Jobs	Processors Hours	System Utilization %	Queue Time (Hours)
Administration	items # 4				
	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				



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

Support/Future

- **Port for XT3 completed in about three weeks**
  - X1E ported and running monitor mode currently
- **Quick turnaround for bugs and features**
- **Next step issue resolution or feature requests**
  - Job with top priority sometimes fails
    - Feels like CPA timing problem (CPA_NO_NODES)
  - New feature for multi-dimensional MAXIJOB

Conclusion

- Rollout has gone very well on multiple architectures (XT3, X1E, Altix, Opteron), OSs, Resource Managers (PBS Pro, SLURM), and Interconnects (XT3, Crossbar, Quadrics, Numalink)
- ORNL policies are now properly represented and enforced
- Admin staff time is reduced
- Utilization is increased
- Progress on future projects has accelerated
- Users are happier

- **More science is being delivered!**

Questions?

Or Contact Us Afterwards

•Don Maxwell:
www.ornl.gov
maxwellde@ornl.gov

•Michael Jackson:
www.clusterresources.com
michael@clusterresources.com
+1 (801) 873-3400