Bright Cluster Manager

Advanced HPC cluster management made easy

株式会社ベストシステムズ
代表取締役　西　克也

The Commonly Used "Toolkit" Approach

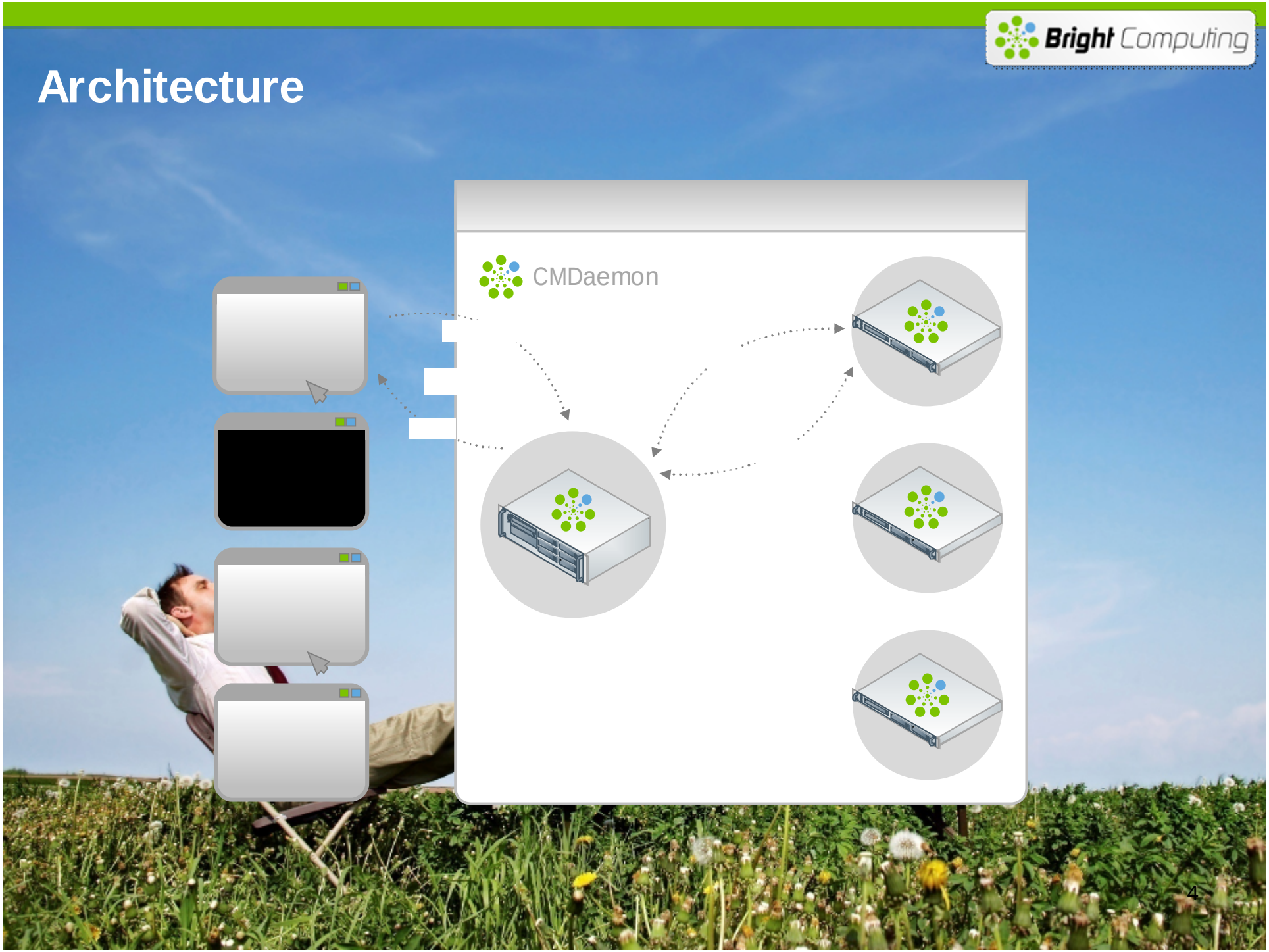
- Most HPC cluster management solutions use the "toolkit" approach (Linux distro + tools)
  - Examples: Rocks, PCM, OSCAR, UniCluster, CMU, etc.
  - Tools typically used: Ganglia, Cacti, Nagios, Cfengine, System Imager, xCAT, Puppet, Cobbler, Hobbit, Big Brother, Zabbix, Groundwork, etc.
- Issues with the "toolkit" approach:
  - Tools rarely designed to work together
  - Tools rarely designed for HPC
  - Tools rarely designed to scale
  - Each tool has its own command line interface and GUI
  - Each tool has its own daemon and database
  - Roadmap dependent on developers of the tools
- Making a collection of unrelated tools work together
  - Requires a lot of expertise and scripting
  - Rarely leads to a really easy-to-use and scalable solution

BAD IDEA

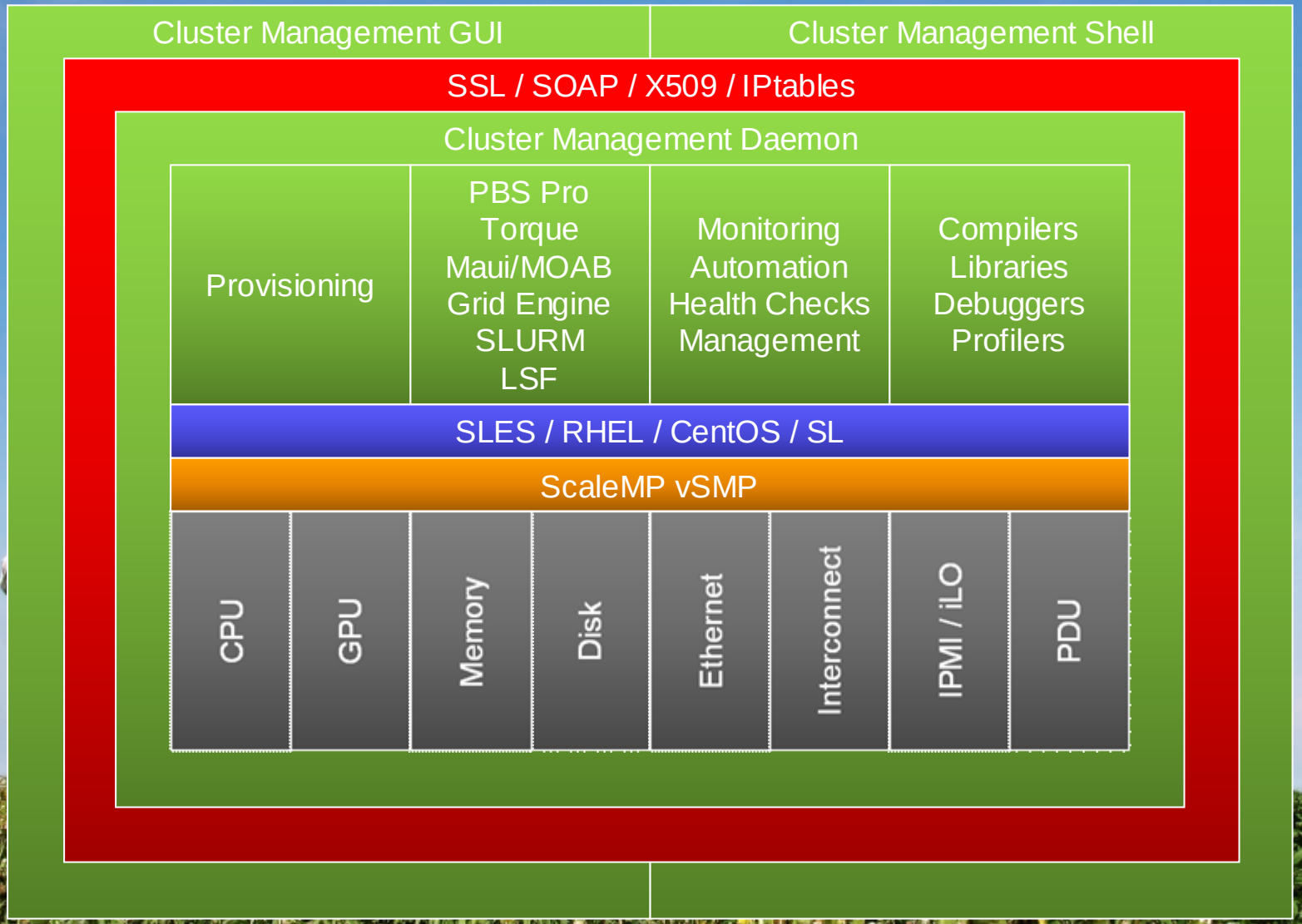
About Bright Cluster Manager

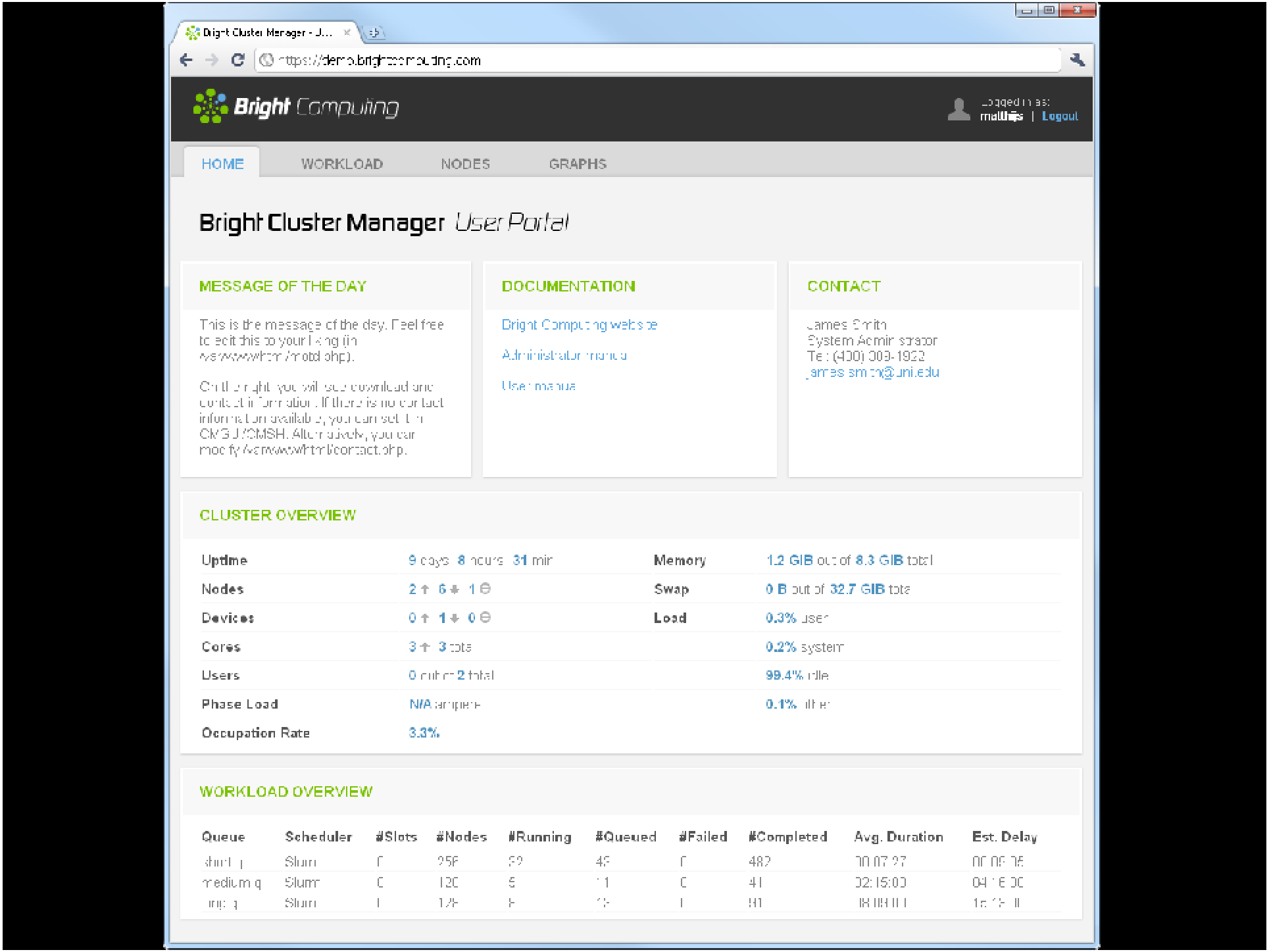
- Bright Cluster Manager takes a much more fundamental & integrated approach
  - Designed and written from the ground up
  - Single cluster management daemon provides all functionality
  - Single, central database for configuration and monitoring data
  - Single CLI and GUI for ALL cluster management functionality
- Which makes Bright Cluster Manager …
  - Extremely easy to use
  - Extremely scalable
  - Secure & reliable
  - Complete
  - Flexible
  - Maintainable

Architecture



Bright Cluster Manager — Elements



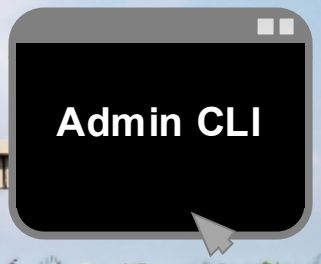

Bright Cluster Manager - U...
https://demo.brightcomputing.com
Bright Computing
Logged in as:
mailto:s | Logout
HOME
WORKLOAD
NODES
GRAPHS
Bright Cluster Manager User Portal
MESSAGE OF THE DAY
This is the message of the day. Feel free to edit this to your liking (in /var/www/html/motd.php).
On the right you will see download and contact information. If there is no contact information available, you can set it in CMGUI/CMSH. Alternatively, you can modify /var/www/html/contact.php.
DOCUMENTATION
Bright Computing website
Administrator manual
User manual
CONTACT
James Smith
System Administrator
Tel: (400) 309-1922
james.smith@uni.edu
CLUSTER OVERVIEW
Uptime
9 days 8 hours 31 min
Memory
1.2 GiB out of 8.3 GiB total
Nodes
2 6 1
Swap
0 B out of 32.7 GiB total
Devices
0 1 0
Load
0.3% user
Cores
3 3 total
0.2% system
Users
0 out of 2 total
99.4% idle
Phase Load
N/A samples
0.1% other
Occupation Rate
3.3%
WORKLOAD OVERVIEW
Queue
Scheduler
#Slots
#Nodes
#Running
#Queued
#Failed
#Completed
Avg. Duration
Est. Delay
short.q
Slurm
medium.q
Slurm
long.q
Slurm

Management Interface

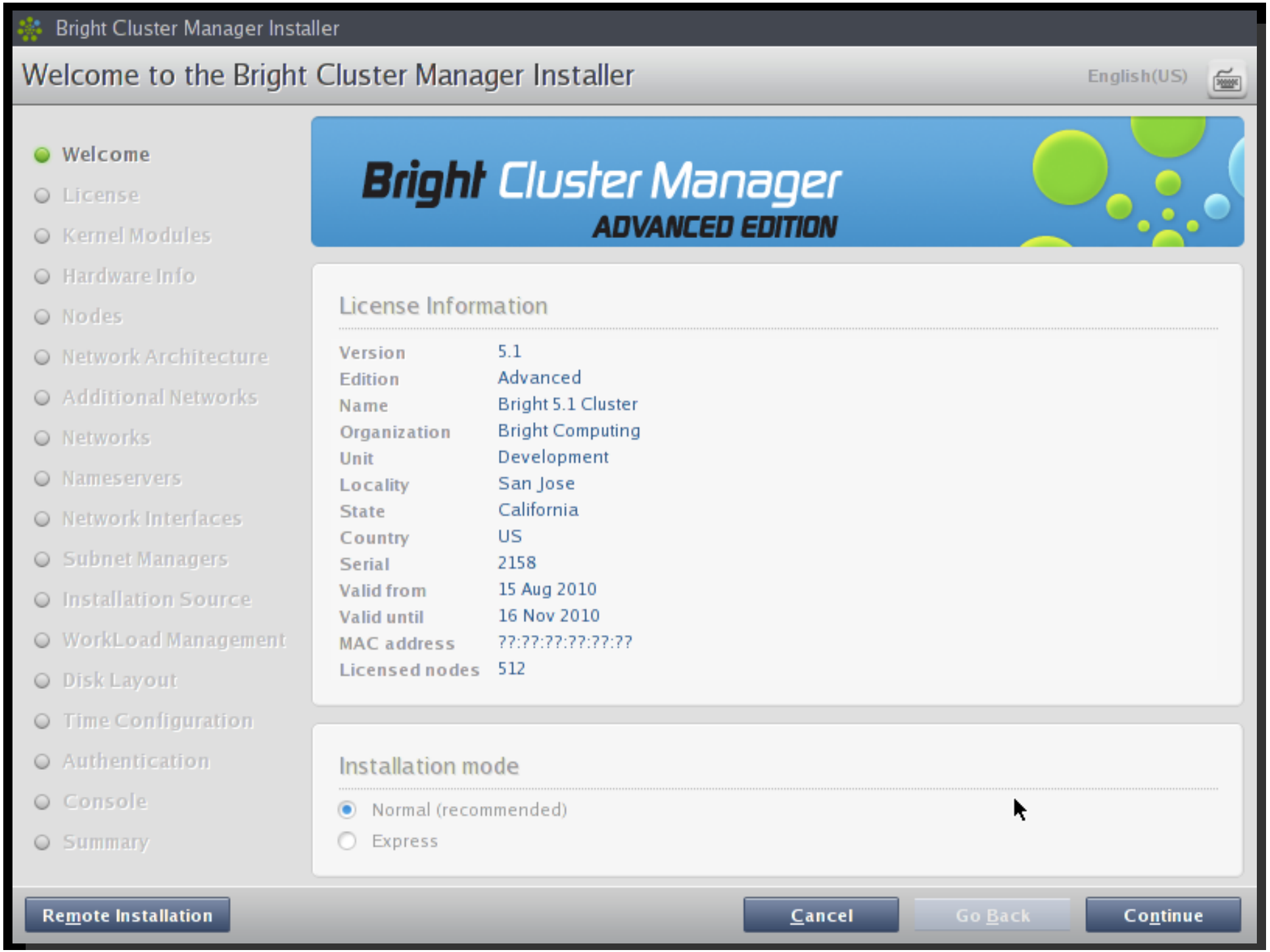
Graphical User Interface (GUI)

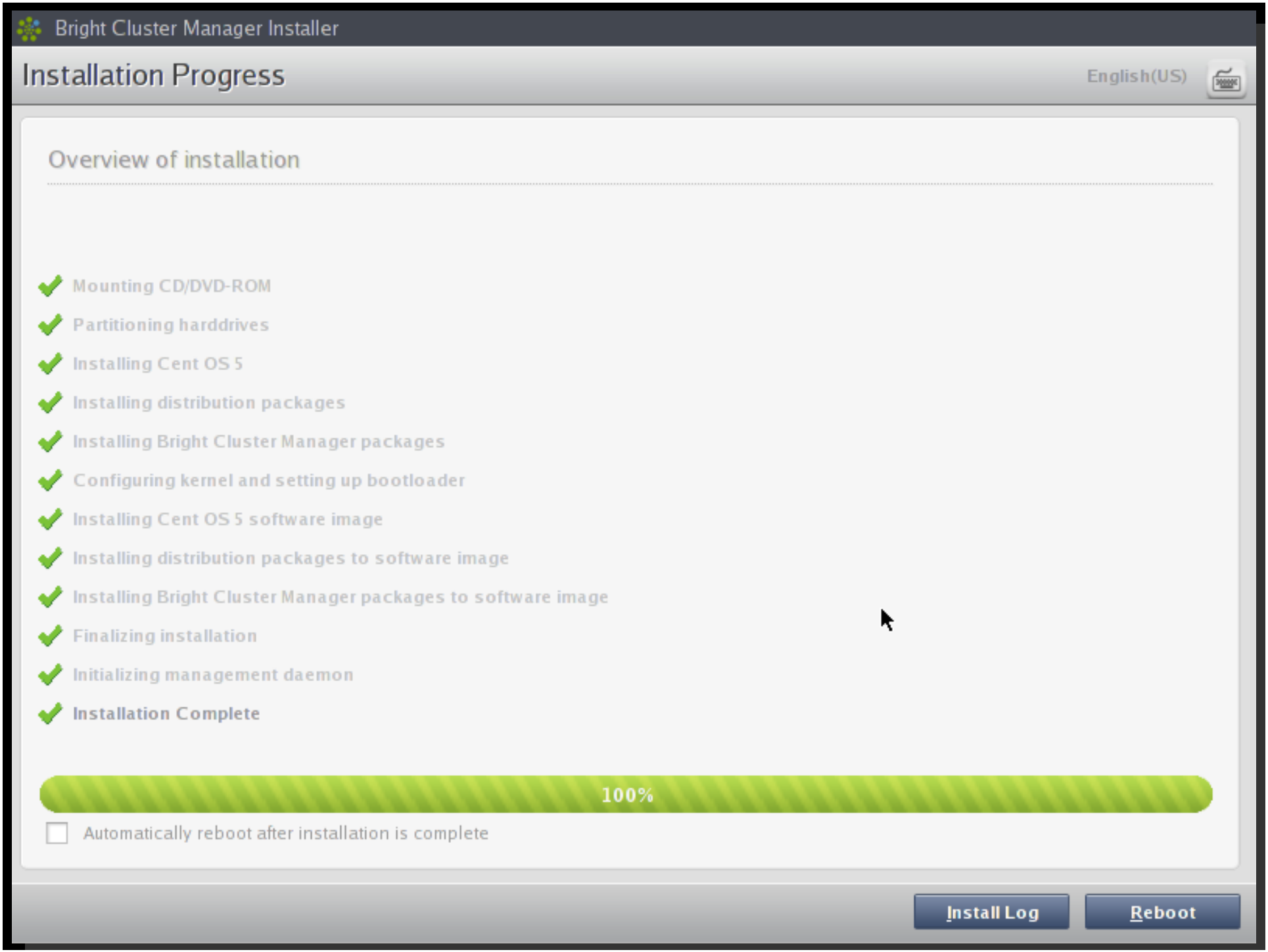
- Offers administrator full cluster control
- Standalone desktop application
- Manages multiple clusters simultaneously
- Runs on Linux, Windows, *MacOS X**
- Built on top of Mozilla XUL engine

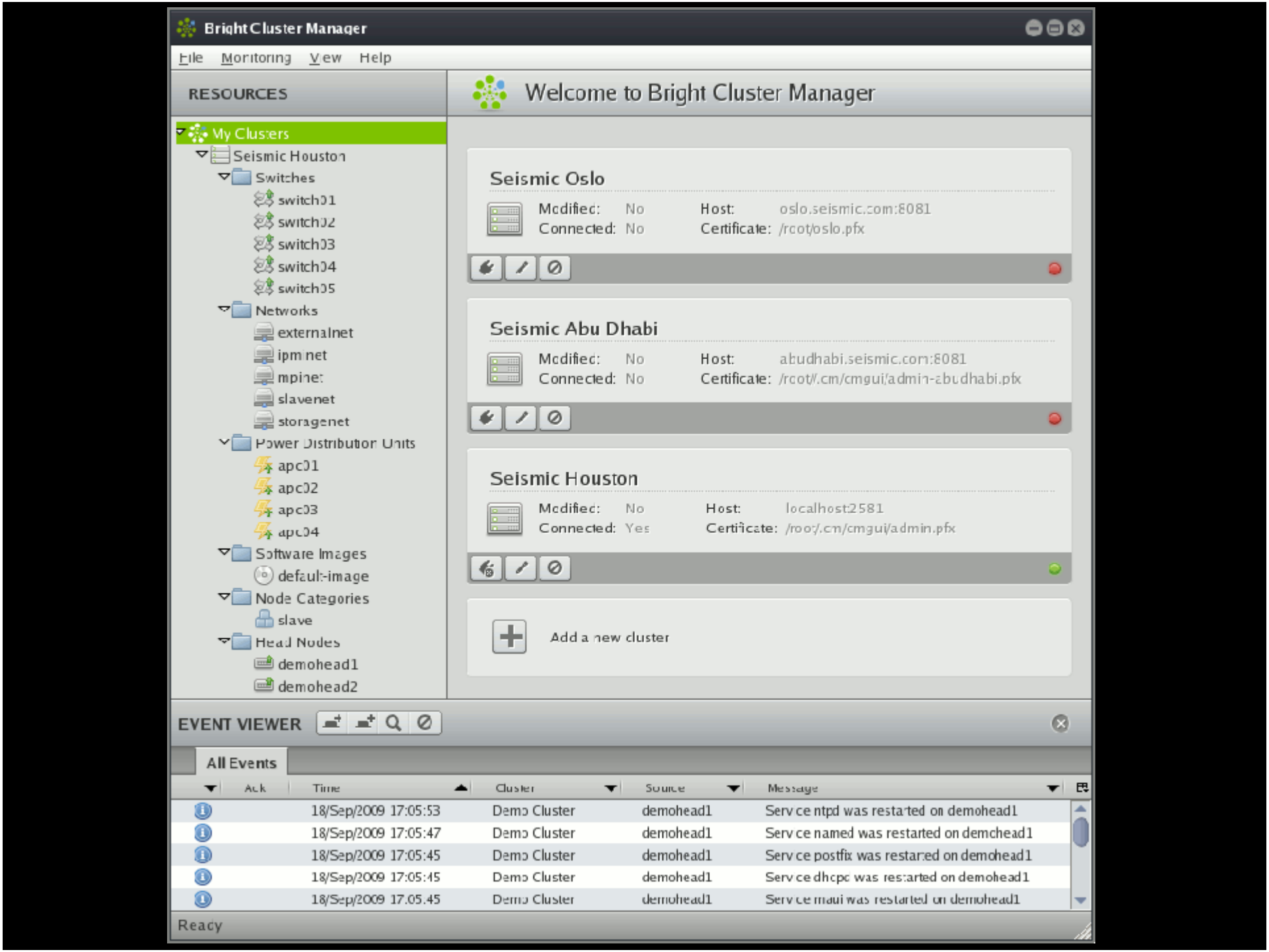
Cluster Management Shell (CMSH)

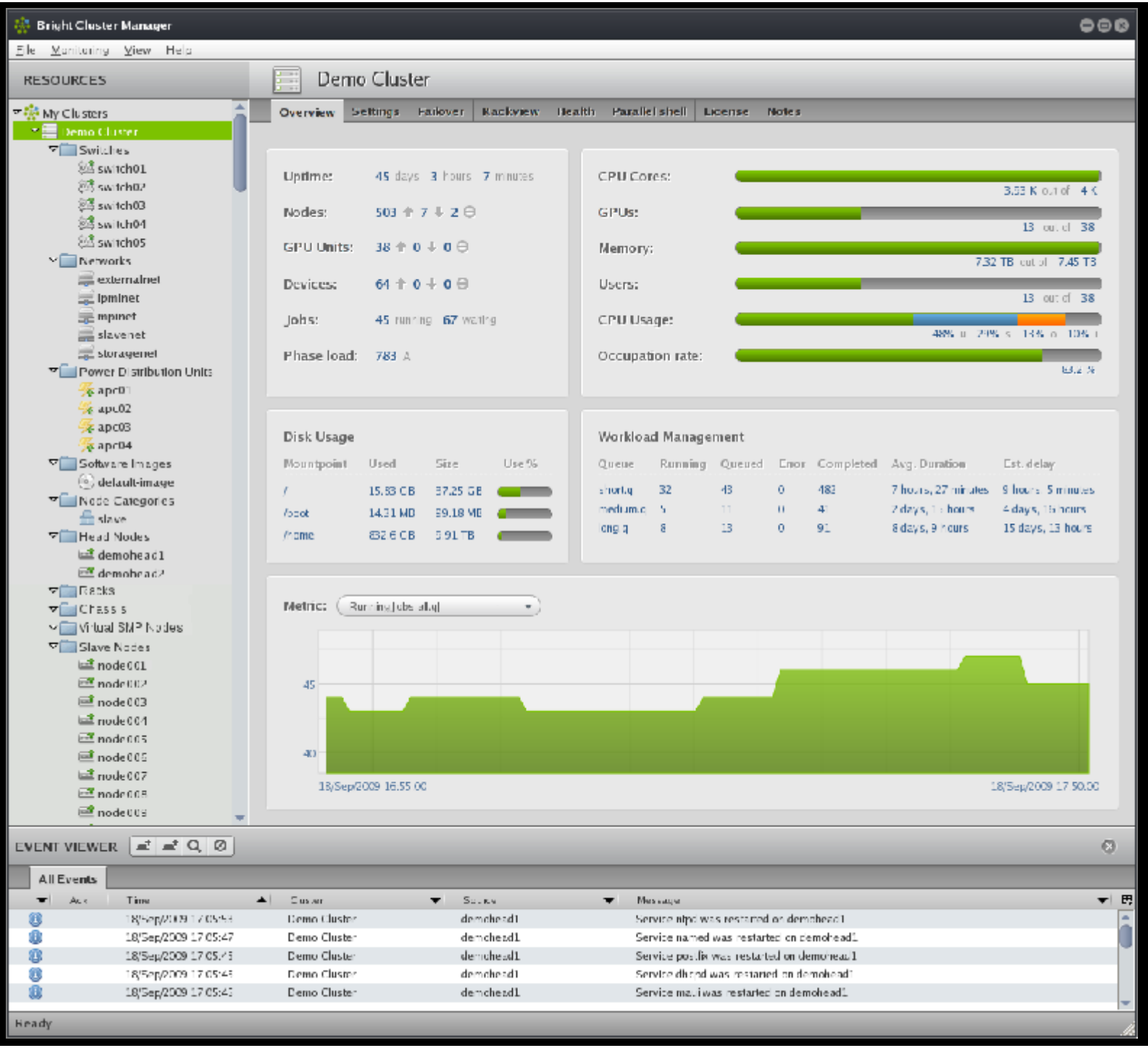


- All GUI functionality also available through Cluster Management Shell
- Interactive and scriptable in batch mode


Bright Cluster Manager Installer
Welcome to the Bright Cluster Manager Installer
English(US)
Welcome
License
Kernel Modules
Hardware Info
Nodes
Network Architecture
Additional Networks
Networks
Nameservers
Network Interfaces
Subnet Managers
Installation Source
WorkLoad Management
Disk Layout
Time Configuration
Authentication
Console
Summary
Bright Cluster Manager
ADVANCED EDITION
License Information
Version 5.1
Edition Advanced
Name Bright 5.1 Cluster
Organization Bright Computing
Unit Development
Locality San Jose
State California
Country US
Serial 2158
Valid from 15 Aug 2010
Valid until 16 Nov 2010
MAC address ??:??:??:??:??:??
Licensed nodes 512
Installation mode
Normal (recommended)
Express
Remote Installation
Cancel
Go Back
Continue


Bright Cluster Manager Installer
Installation Progress
English(US)
Overview of installation
Mounting CD/DVD-ROM
Partitioning harddrives
Installing Cent OS 5
Installing distribution packages
Installing Bright Cluster Manager packages
Configuring kernel and setting up bootloader
Installing Cent OS 5 software image
Installing distribution packages to software image
Installing Bright Cluster Manager packages to software image
Finalizing installation
Initializing management daemon
Installation Complete
100%
Automatically reboot after installation is complete
Install Log
Reboot


Bright Cluster Manager
File Monitoring View Help
RESOURCES
Welcome to Bright Cluster Manager
My Clusters
Seismic Houston
Switches
switch01
switch02
switch03
switch04
switch05
Networks
externalnet
ipmi net
mpine:
slavenet
storagenet
Power Distribution Units
apc01
apc02
apc03
apc04
Software Images
default-image
Node Categories
slave
Head Nodes
demohead1
demohead2
Seismic Oslo
Modified: No
Connected: No
Host: oslo.seismic.com:8081
Certificate: /root/oslo.pfx
Seismic Abu Dhabi
Modified: No
Connected: No
Host: abudhabi.seismic.com:8081
Certificate: /root/.cm/cmgui/admin-abudhabi.pfx
Seismic Houston
Modified: No
Connected: Yes
Host: localhost:2581
Certificate: /root/.cm/cmgui/admin.pfx
Add a new cluster
EVENT VIEWER
All Events
Ack
Time
Cluster
Source
Message
18/Sep/2009 17:05:53 Demo Cluster demohead1 Service ntpd was restarted on demohead1
18/Sep/2009 17:05:47 Demo Cluster demohead1 Service named was restarted on demohead1
18/Sep/2009 17:05:45 Demo Cluster demohead1 Service postfix was restarted on demohead1
18/Sep/2009 17:05:45 Demo Cluster demohead1 Service dhcpd was restarted on demohead1
18/Sep/2009 17:05:45 Demo Cluster demohead1 Service cmd was restarted on demohead1
Ready


Bright Cluster Manager
File Monitoring View Help
RESOURCES
Demo Cluster
My Clusters
Demo Cluster
Switches
switch01
switch02
switch03
switch04
switch05
Networks
externalnet
ipminet
mpinet
slavenet
storagenet
Power Distribution Units
apc01
apc02
apc03
apc04
Software Images
default-image
Node Categories
slave
Head Nodes
demohead1
demohead2
Racks
Chassis
Virtual SMP Nodes
Slave Nodes
node001
node002
node003
node004
node005
node006
node007
node008
node009
Overview
Settings
Failover
Rackview
Health
Parallel shell
License
Notes
Uptime: 45 days 3 hours 7 minutes
Nodes: 503 7 2
GPU Units: 38 0 0
Devices: 64 0 0
Jobs: 45 running 67 waiting
Phase load: 783 A
CPU Cores:
GPUs:
Memory: 7.32 TB out of 7.45 TB
Users:
CPU Usage:
Occupation rate:
Disk Usage
Mountpoint Used Size Use%
/ 15.53 GB 37.25 GB
/boot 14.31 MB 89.18 MB
/home 682.6 GB 3.91 TB
Workload Management
Queue Running Queued Error Completed Avg. Duration Est. delay
short.q 32 43 0 483 7 hours, 27 minutes 9 hours, 5 minutes
medium.q 5 11 0 47 2 days, 1 hours 4 days, 15 hours
long.q 8 13 0 91 8 days, 9 hours 15 days, 13 hours
Metric: Running Jobs all.q
45
40
18/Sep/2009 16:55:00
18/Sep/2009 17:50:00
EVENT VIEWER
All Events
Time
Cluster
Source
Message
Demo Cluster demohead1
Demo Cluster demohead1 Service named was restarted on demohead1
Demo Cluster demohead1
Demo Cluster demohead1
Demo Cluster demohead1
Ready

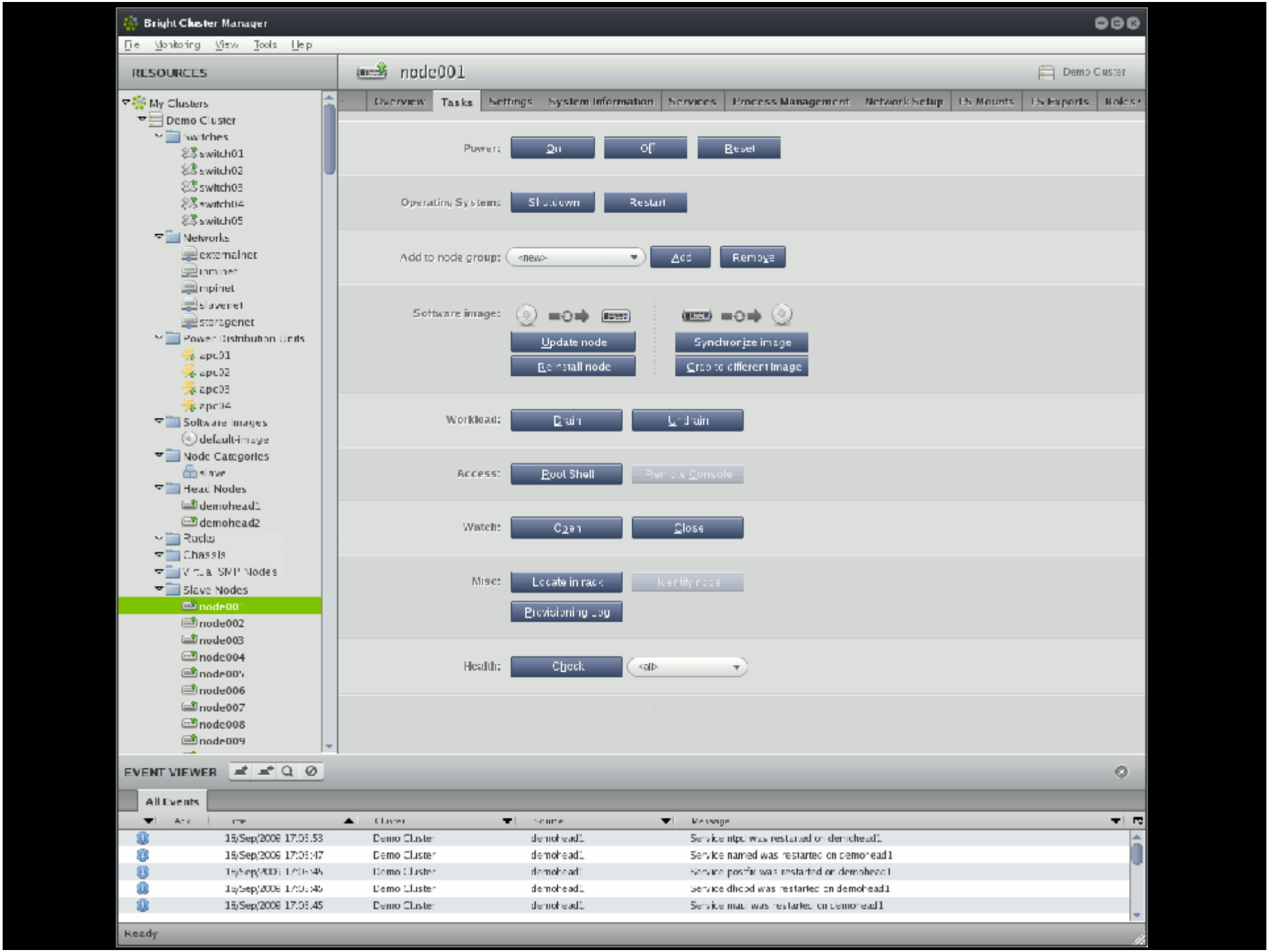
Node Provisioning

Image based

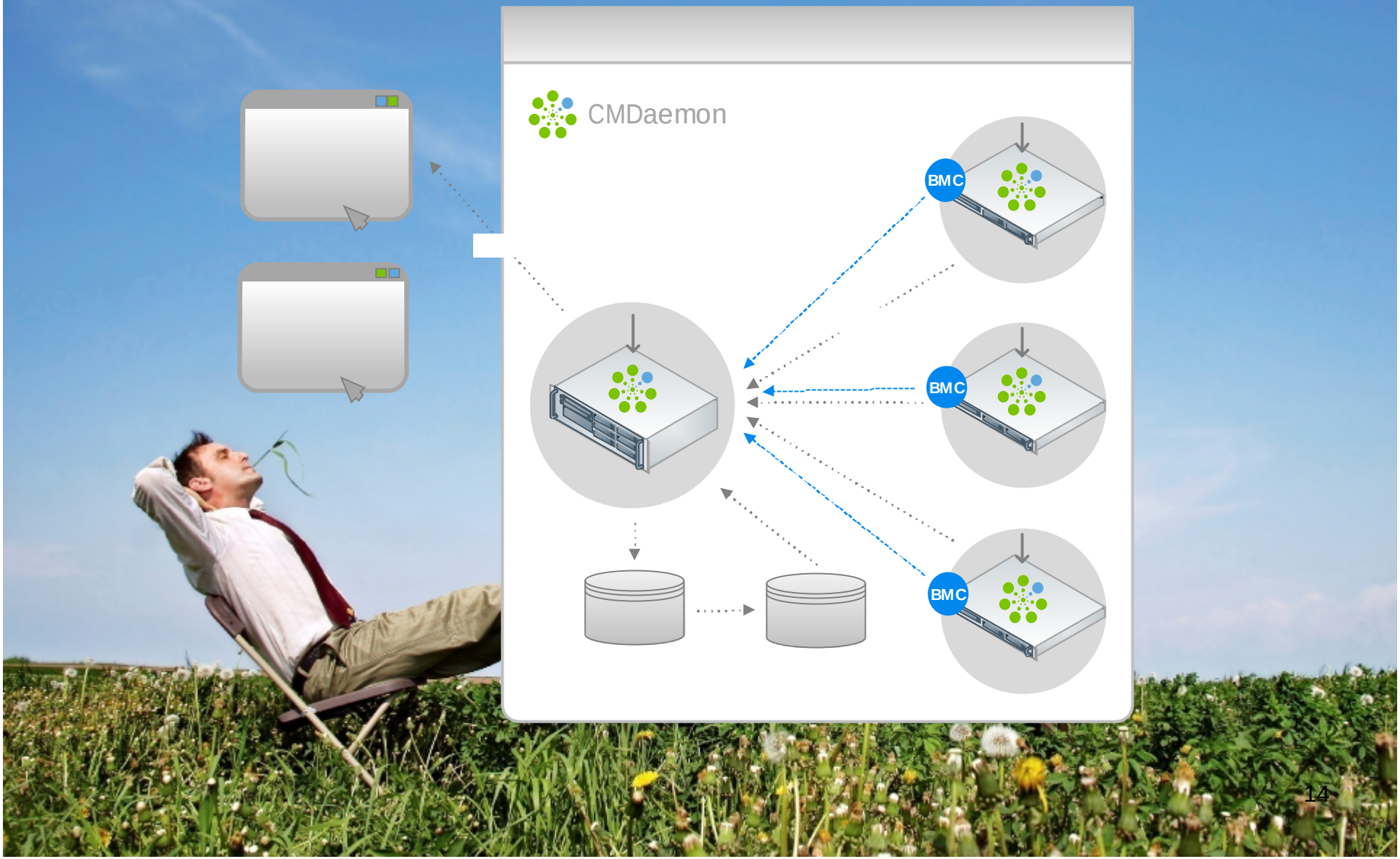
- Regular node image is a directory on the head node
- Unlimited number of images can be created
- Software changes for the regular nodes are made inside the image(s) on the head node
- Provisioning system ensures that changes are propagated to the regular nodes

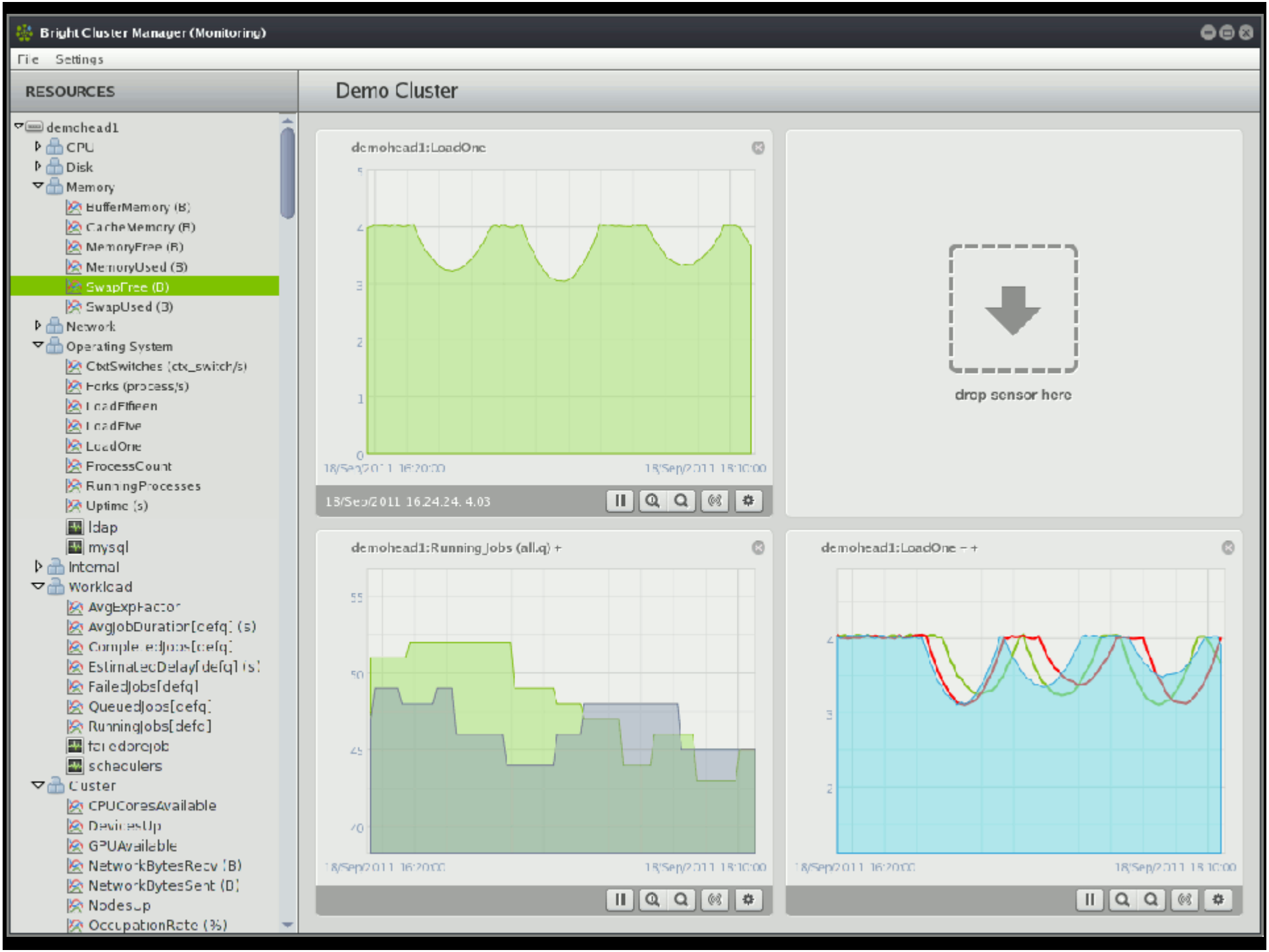
Nodes always boot over the network

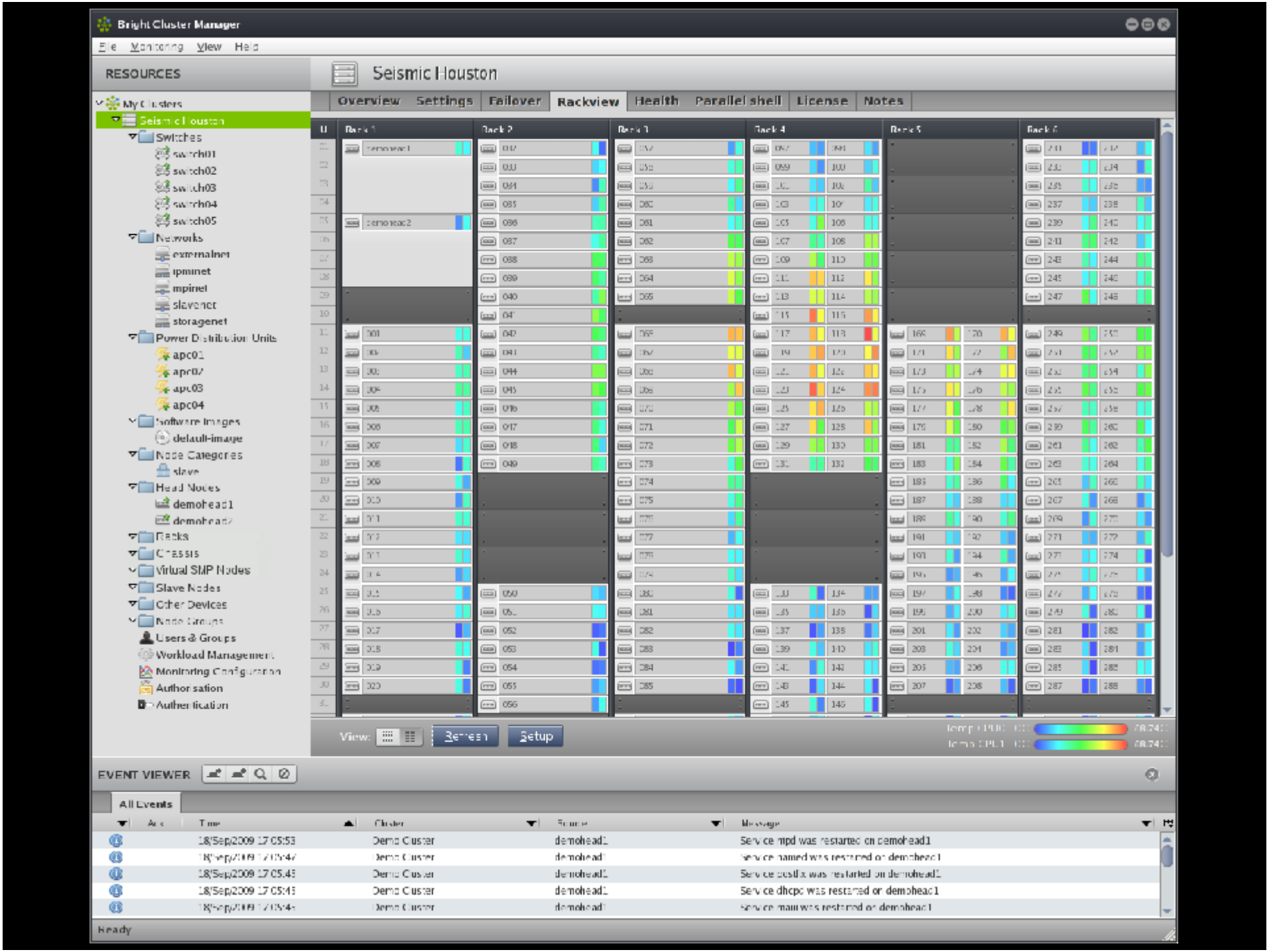
- Regular nodes PXE boot into Node Installer, which
- Identifies node (switch port or MAC based)
- Configures BMC
- Partition disks (if any) and creates file systems (if needed)
- Installs or updates software image (if needed)
- Pivot the root from NFS to the local file system

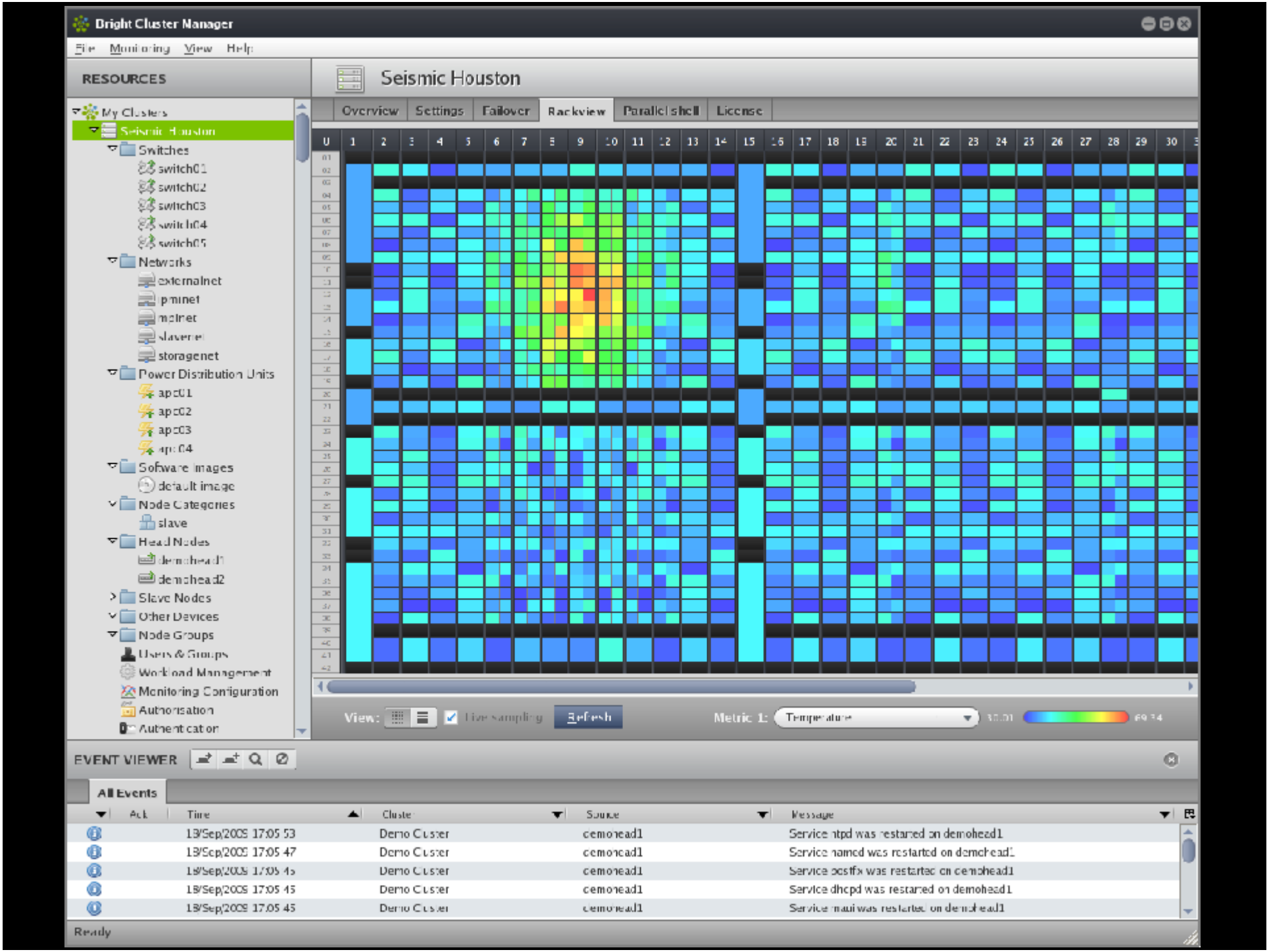

Bright Cluster Manager
File Monitoring View Tools Help
RESOURCES
node001
Demo Cluster
Overview
Tasks
Settings
System Information
Services
Process Management
Network Setup
I/O Mounts
I/O Exports
Roles
My Clusters
Demo Cluster
Switches
switch01
switch02
switch03
switch04
switch05
Networks
externalnet
ipminet
slavenet
storagenet
Power Distribution Units
pdu01
pdu02
pdu03
pdu04
Software Images
default-image
Node Categories
slave
Head Nodes
demohead1
demohead2
Racks
Chassis
Virtual SMP Nodes
Slave Nodes
node001
node002
node003
node004
node005
node006
node007
node008
node009
Power:
On
Off
Reset
Operating System:
Shutdown
Restart
Add to node group:
<new>
Add
Remove
Software image:
Update node
Reinstall node
Synchronize image
Grab to different image
Workload:
Drain
Undrain
Access:
Root Shell
Remote Console
Watch:
Open
Close
Misc:
Locate in rack
Identify node
Provisioning Log
Health:
Check
<all>
EVENT VIEWER
All Events
Age
Time
Cluster
Source
Message
15/Sep/2008 17:03:53
Demo Cluster
demohead1
Service ntpd was restarted on demohead1
15/Sep/2008 17:03:47
Demo Cluster
demohead1
Service named was restarted on demohead1
15/Sep/2008 17:03:45
Demo Cluster
demohead1
Service postfix was restarted on demohead1
15/Sep/2008 17:03:45
Demo Cluster
demohead1
Service dhcpd was restarted on demohead1
15/Sep/2008 17:03:45
Demo Cluster
demohead1
Service mail was restarted on demohead1
Ready

Architecture — Monitoring



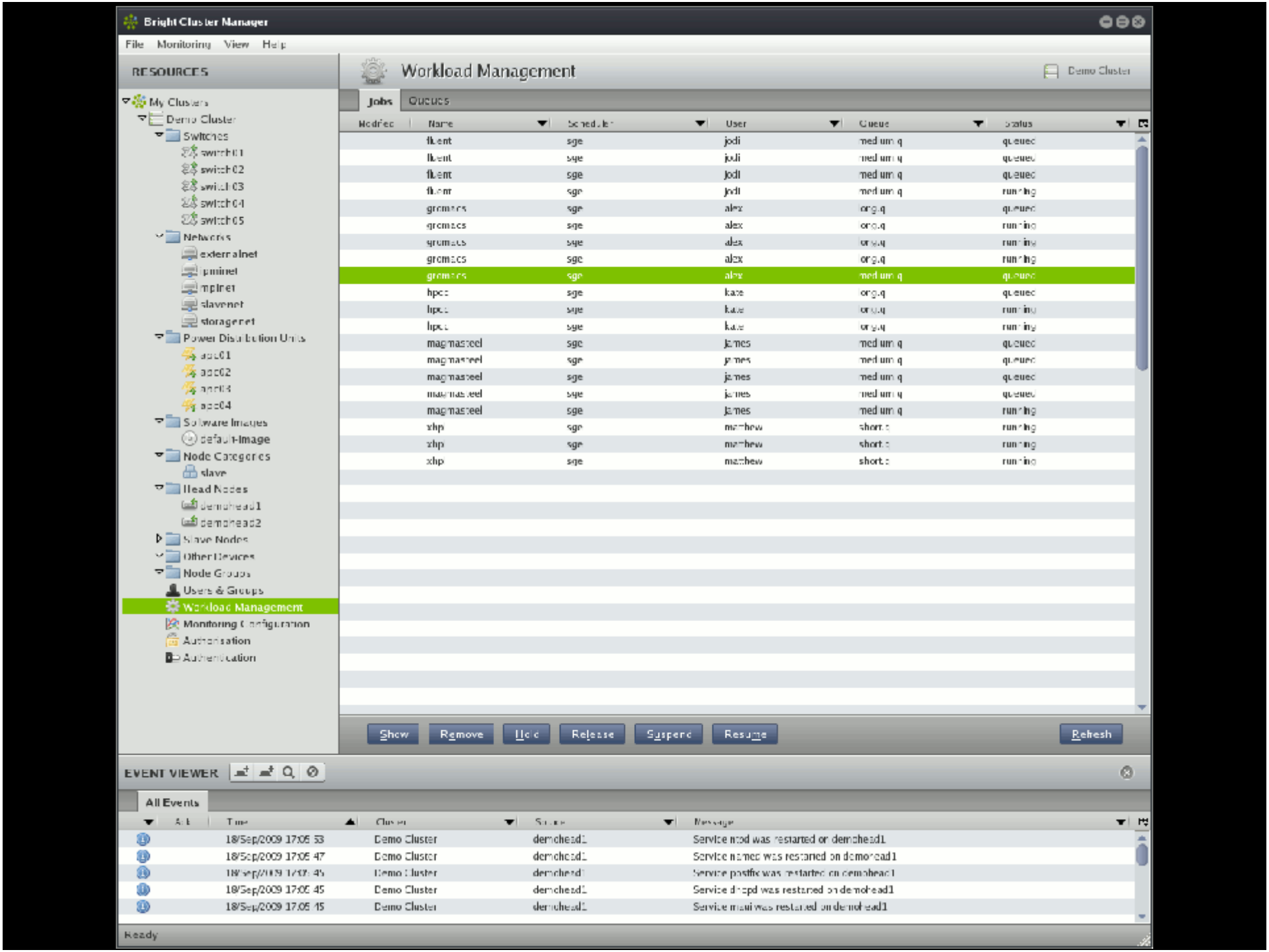

Bright Cluster Manager (Monitoring)
File Settings
RESOURCES
Demo Cluster
demohead1
CPU
Disk
Memory
BufferMemory (B)
CacheMemory (B)
MemoryFree (B)
MemoryUsed (B)
SwapFree (B)
SwapUsed (B)
Network
Operating System
CtxtSwitches (ctx_switch/s)
Forks (process/s)
LoadFifteen
LoadFive
LoadOne
ProcessCount
RunningProcesses
Uptime (s)
ldap
mysql
Internal
Workload
Cluster
CPUCoresAvailable
GPUAvailable
NodesUp
demohead1:LoadOne
drop sensor here
demohead1:Running Jobs (all,q) +
demohead1:LoadOne – +


Bright Cluster Manager
File Monitoring View Help
RESOURCES
Seismic Houston
Overview Settings Failover Rackview Health Parallel shell License Notes
My Clusters
Seismic Houston
Switches
switch01
switch02
switch03
switch04
switch05
Networks
externalnet
ipminet
mpinet
slavenet
storagenet
Power Distribution Units
apc01
apc02
apc03
apc04
Software Images
default-image
Node Categories
slave
Head Nodes
demohead1
demohead2
Racks
Chassis
Virtual SMP Nodes
Slave Nodes
Other Devices
Node Groups
Users & Groups
Workload Management
Monitoring Configuration
Authorisation
Authentication
Rack 1
Rack 2
Rack 3
Rack 4
Rack 5
Rack 6
demohead1
demohead2
View:
Screen
Setup
EVENT VIEWER
All Events
Time
Cluster
Source
Message
18/Sep/2009 17:05:53 Demo Cluster demohead1 Service mpd was restarted on demohead1
18/Sep/2009 17:05:47 Demo Cluster demohead1 Service named was restarted on demohead1
18/Sep/2009 17:05:43 Demo Cluster demohead1 Service postfix was restarted on demohead1
18/Sep/2009 17:05:43 Demo Cluster demohead1 Service dhcpd was restarted on demohead1
Ready


Bright Cluster Manager
File Monitoring View Help
RESOURCES
Seismic Houston
Overview Settings Failover Rackview ParallelShell License
My Clusters
Seismic Houston
Switches
switch01
switch02
switch03
switch04
switch05
Networks
externalnet
ipminet
mgmtnet
slavenet
storagenet
Power Distribution Units
apc01
apc02
apc03
apc04
Software Images
default image
Node Categories
slave
Head Nodes
demohead1
demohead2
Slave Nodes
Other Devices
Node Groups
Users & Groups
Workload Management
Monitoring Configuration
Authorisation
Authentication
View:
Live sampling
Refresh
Metric 1:
Temperature
EVENT VIEWER
All Events
Ack
Time
Cluster
Source
Message
13/Sep/2005 17:05:53 Demo Cluster demohead1 Service ntpd was restarted on demohead1
13/Sep/2005 17:05:47 Demo Cluster demohead1 Service named was restarted on demohead1
13/Sep/2005 17:05:45 Demo Cluster demohead1 Service postfix was restarted on demohead1
13/Sep/2005 17:05:45 Demo Cluster demohead1 Service dhcpd was restarted on demohead1
13/Sep/2005 17:05:45 Demo Cluster demohead1 Service mail was restarted on demohead1
Ready

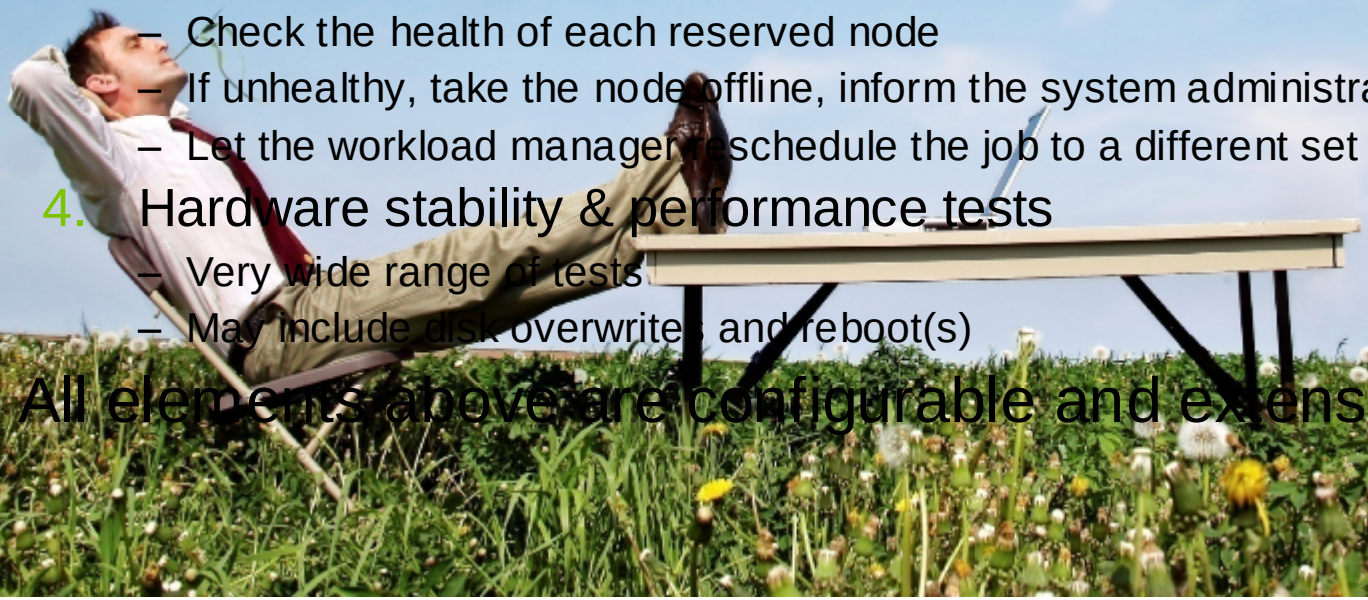
Workload Manager Integration

- Automatic installation
- Automatic configuration
- Sampling, analysis and visualization of workload manager statistics
- Consistent GUI, User Portal and CLI front-end to workload manager
- Bright cluster SOAP API provides consistent access to whole cluster, including workload manager
- Failover of workload manager
- Health checking


Bright Cluster Manager
File Monitoring View Help
RESOURCES
Workload Management
Demo Cluster
Jobs Queues
Modified Name Scheduler User Queue Status
fluent sge jodi medium.q queued
fluent sge jodi medium.q queued
fluent sge jodi medium.q queued
fluent sge jodi medium.q running
gromacs sge alex long.q queued
gromacs sge alex long.q running
gromacs sge alex long.q running
gromacs sge alex long.q running
gromacs sge alex medium.q queued
hpcc sge kate long.q queued
hpcc sge kate long.q running
hpcc sge kate long.q running
magmasteel sge james medium.q queued
magmasteel sge james medium.q queued
magmasteel sge james medium.q queued
magmasteel sge james medium.q queued
magmasteel sge james medium.q running
xhp sge matthew short.q running
xhp sge matthew short.q running
xhp sge matthew short.q running
My Clusters
Demo Cluster
Switches
switch01
switch02
switch03
switch04
switch05
Networks
externalnet
ipminet
mpinet
slavenet
storagenet
Power Distribution Units
pdu01
pdu02
pdu03
pdu04
Software Images
default-image
Node Categories
slave
Head Nodes
demohead1
demohead2
Slave Nodes
Other Devices
Node Groups
Users & Groups
Workload Management
Monitoring Configuration
Authorisation
Authentication
Show
Remove
Hold
Release
Suspend
Resume
Refresh
EVENT VIEWER
All Events
Time Cluster Source Message
18/Sep/2009 17:05:53 Demo Cluster demohead1 Service nscd was restarted on demohead1
18/Sep/2009 17:05:47 Demo Cluster demohead1 Service named was restarted on demohead1
18/Sep/2009 17:05:45 Demo Cluster demohead1 Service postfix was restarted on demohead1
18/Sep/2009 17:05:45 Demo Cluster demohead1 Service dhcpd was restarted on demohead1
18/Sep/2009 17:05:15 Demo Cluster demohead1 Service mailwas restarted on demohead1
Ready

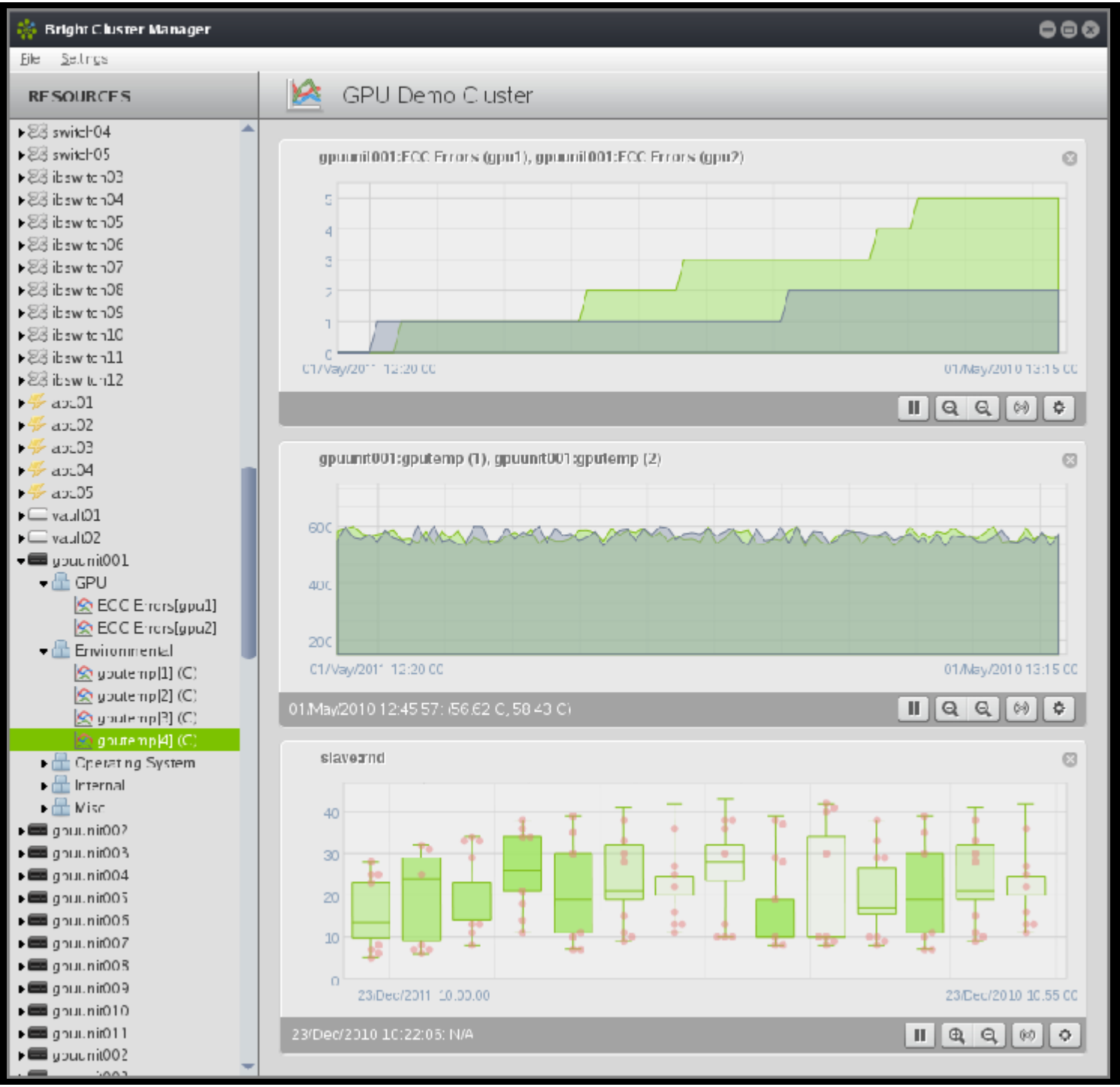
Cluster Health Management

- Goal: provide problem free environment for running jobs
- Four elements
  1. Cluster management automation
  2. Regular health checks
     - Actions that return PASS, FAIL or UNKOWN
     - Can be associated with a settable severity and a message
     - Can launch an action based on any response value
  3. Prejob health checks
     - Let the workload manager hold the job very briefly
     - Check the health of each reserved node
     - If unhealthy, take the node offline, inform the system administrator
     - Let the workload manager reschedule the job to a different set of nodes
  4. Hardware stability & performance tests
     - Very wide range of tests
     - May include disk overwrites and reboot(s)
- All elements above are configurable and extensible



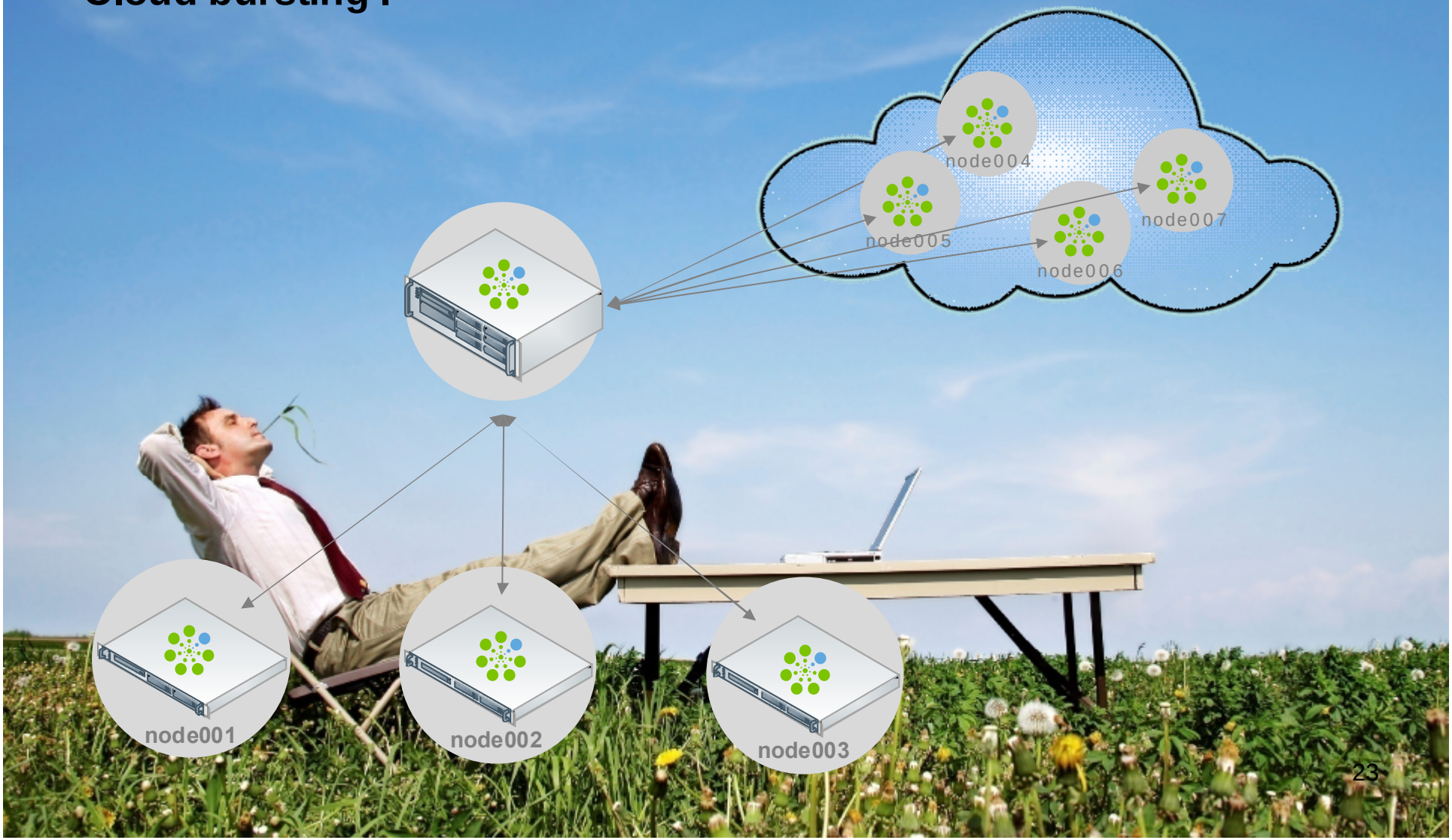
Bright Cluster Manager for GPGPU

- CUDA & OpenCL redistribution rights
- Current and previous versions of CUDA & OpenCL
- Easy switching between CUDA & OpenCL versions
- CUDA driver automatically compiled at boot time
- Support for all NVIDIA GPUs


Bright Cluster Manager
File
Settings
RESOURCES
GPU Demo Cluster
switch04
switch05
ibswitch03
ibswitch04
ibswitch05
ibswitch06
ibswitch07
ibswitch08
ibswitch09
ibswitch10
ibswitch11
ibswitch12
apc01
apc02
apc03
apc04
apc05
vault01
vault02
gpuunit001
GPU
ECC Errors[gpu1]
ECC Errors[gpu2]
Environmental
gputemp[1] (C)
gputemp[2] (C)
gputemp[3] (C)
gputemp[4] (C)
Operating System
Internal
Misc
gpuunit007
gpuunit003
gpuunit004
gpuunit005
gpuunit006
gpuunit007
gpuunit008
gpuunit009
gpuunit010
gpuunit011
gpuunit002
gpuunit001:ECC Errors (gpu1), gpuunit001:ECC Errors (gpu2)
gpuunit001:gputemp (1), gpuunit001:gputemp (2)
slavemd

The Future

Cloud bursting I



The Future

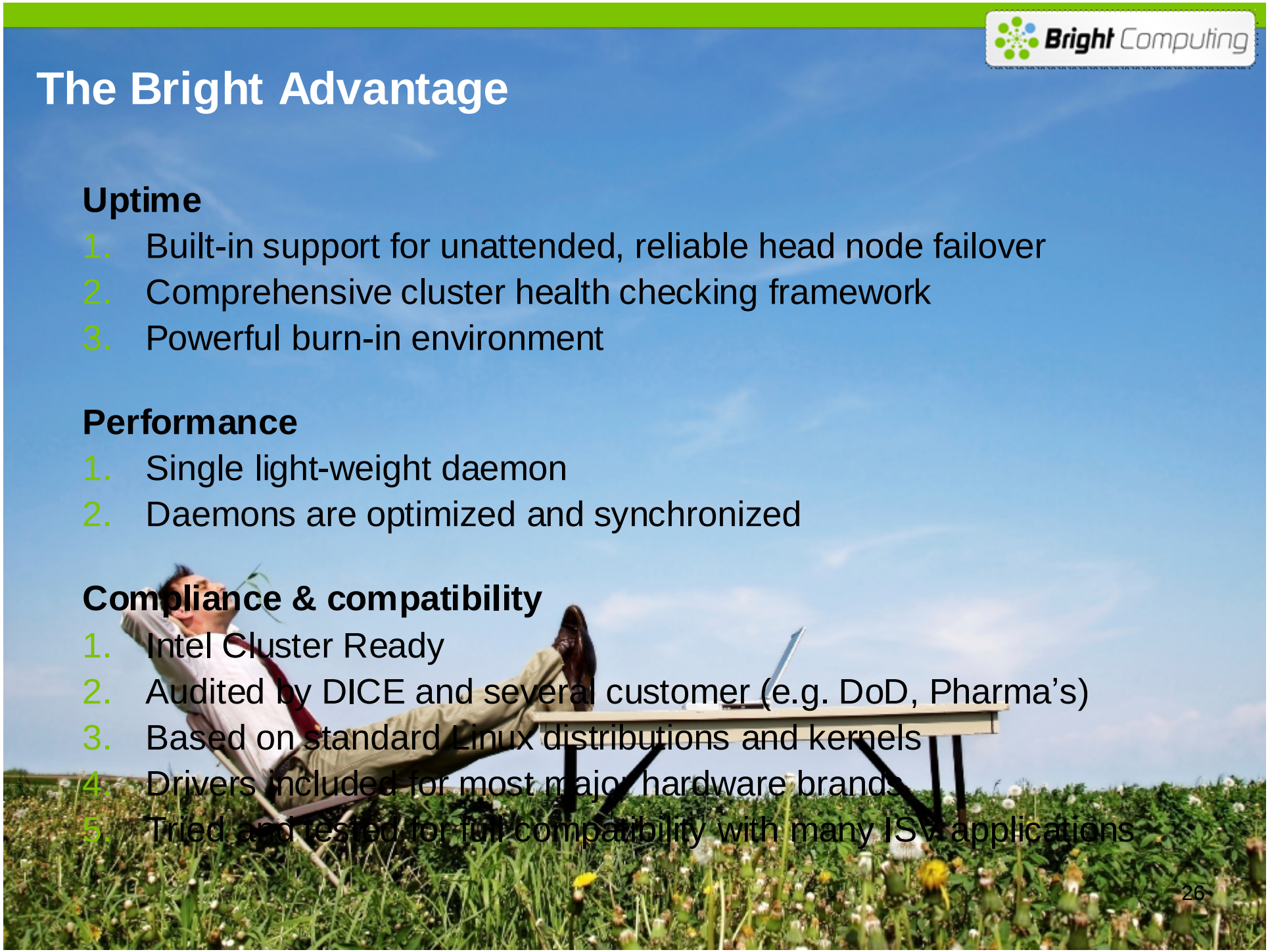
Cloud bursting II

The Bright Advantage

Productivity & Efficiency

1. Easy to learn and use
2. Installation in less than 30 minutes
3. Full insight in and control over the cluster
4. All elements of the cluster are managed (servers, switches, networks, etc.)
5. Flexible provisioning (incremental, live, diskfull, diskless, IB-only, node discovery)
6. Comprehensive monitoring (graphs & rackview)
7. Powerful automation (thresholds, alerts, actions)
8. Vendor-independent workload manager integration
9. Integrated application development environment
10. Multi-cluster functionality
11. Easy, automatic updating from Linux & Bright repositories
12. Comprehensive GPU support
13. Rapid SMP deployment

The Bright Advantage



Uptime

1. Built-in support for unattended, reliable head node failover
2. Comprehensive cluster health checking framework
3. Powerful burn-in environment

Performance

1. Single light-weight daemon
2. Daemons are optimized and synchronized

Compliance & compatibility

1. Intel Cluster Ready
2. Audited by DICE and several customer (e.g. DoD, Pharma's)
3. Based on standard Linux distributions and kernels
4. Drivers included for most major hardware brands
5. Tried and tested for full compatibility with many ISV applications

The Bright Advantage

Scalability

1. Off-loadable provisioning
2. Efficient collection and processing of monitoring metrics
3. Tried & tested on largest clusters in the world

Security

1. Automated security and other updates from PGP signed repositories
2. All internal + external communication encrypted using public/private key encryption through SSH/SSL
3. Authentication based on X509 certificates
4. Role-based access control
5. Auditing of all administrator write actions
6. Firewalls
7. Secure LDAP