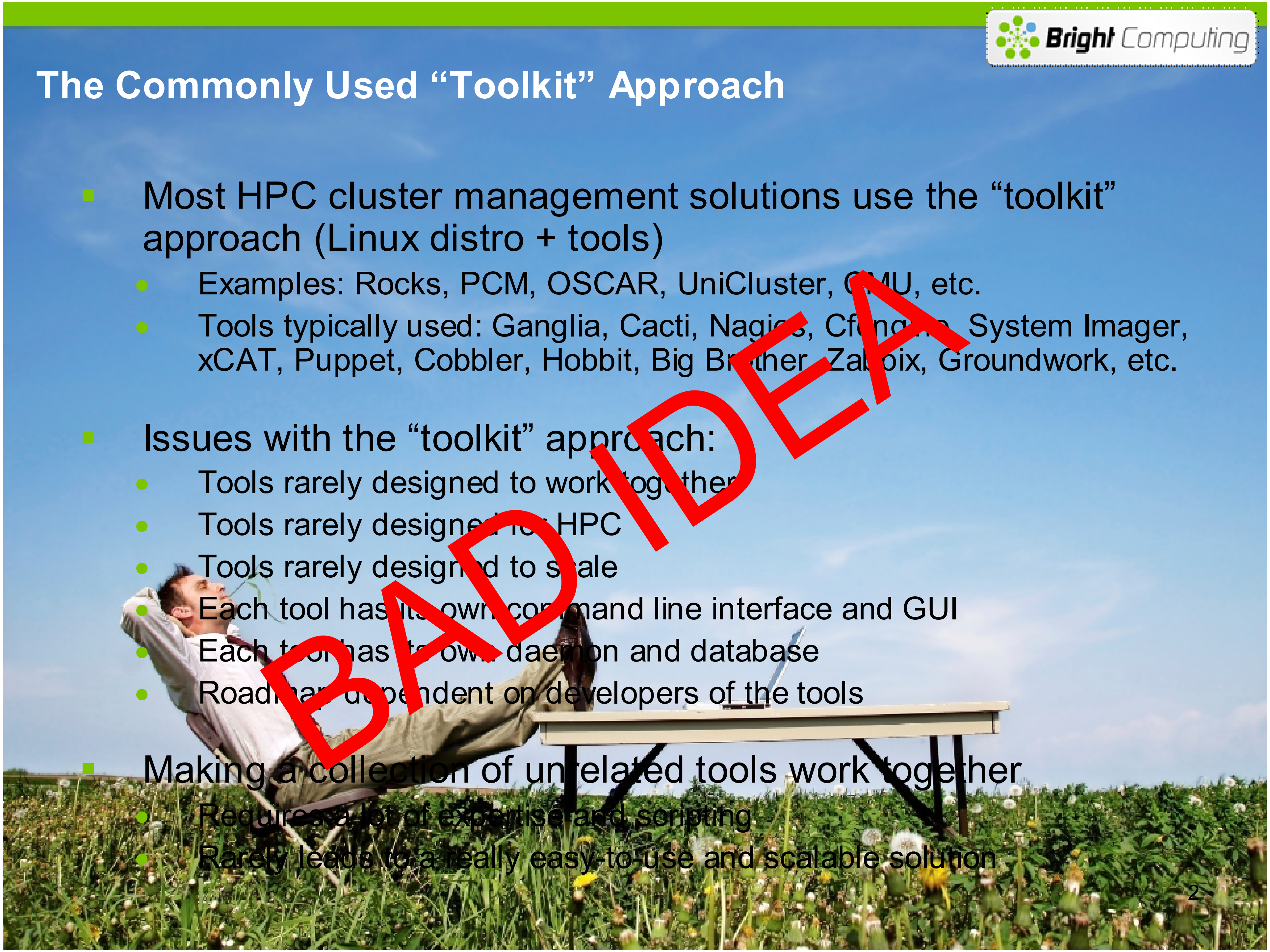
Bright Cluster Manager

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



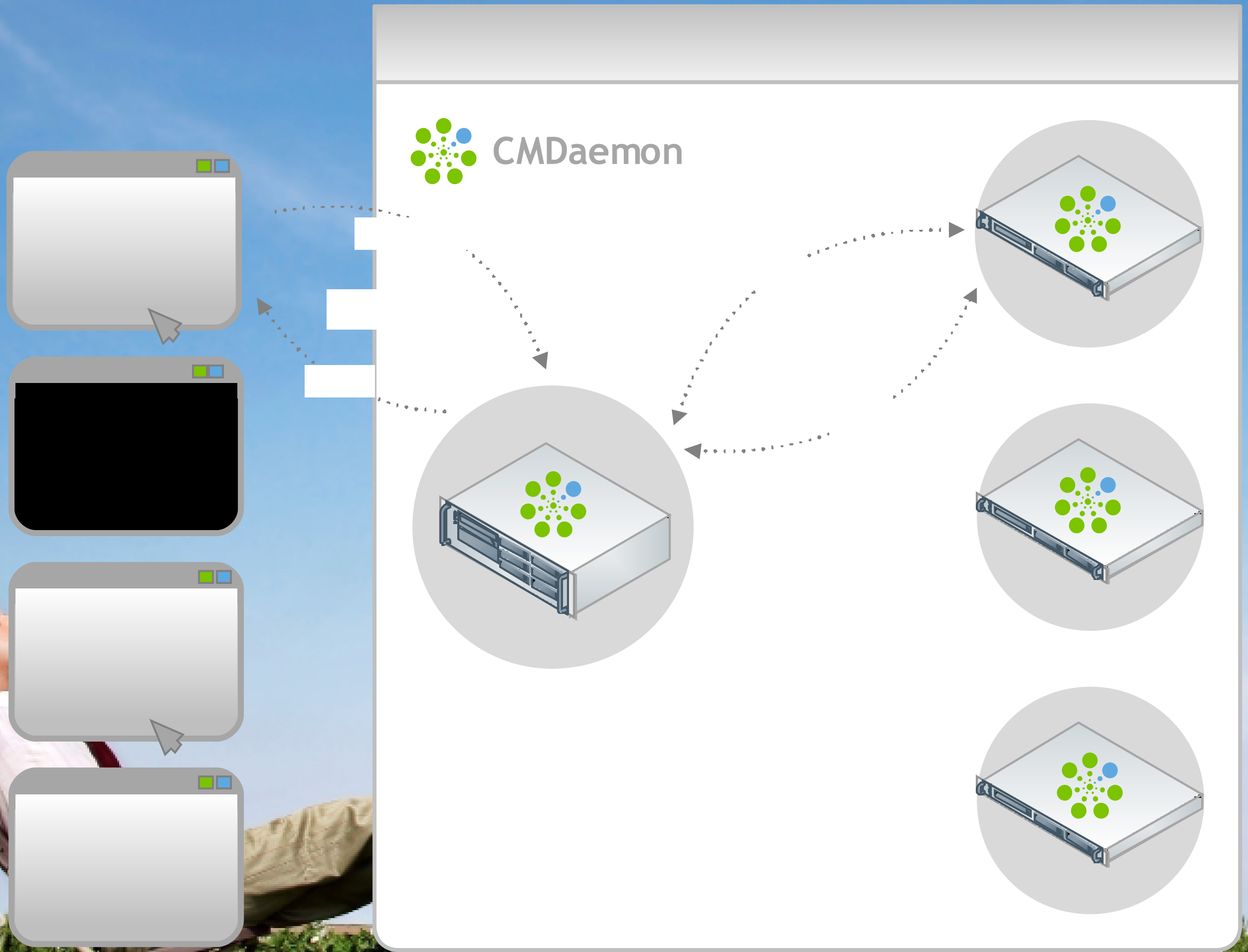
- Most HPC cluster management solutions use the "toolkit" approach (Linux distro + tools)
  - Examples: Rocks, PCM, OSCAR, UniCluster, CMU, etc.
  - Tools typically used: Ganglia, Cacti, Nagios, Cfengine, System Imager, xCAT, Puppet, Cobbler, Hobbit, Big Brother, Zabbix, Groundwork, etc.
- Issues with 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

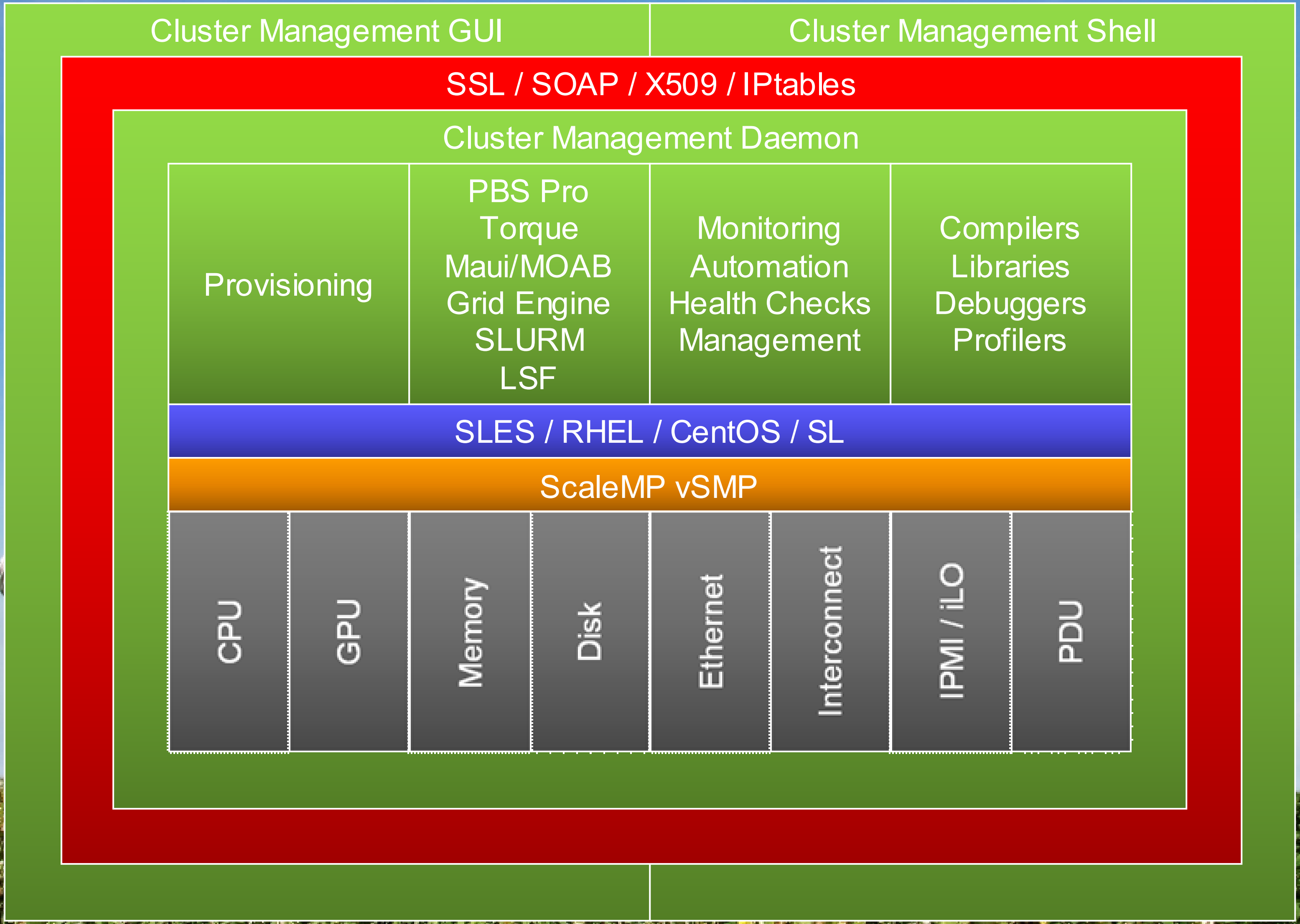
About Bright Cluster Manager

- 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



Cluster Management GUI

Cluster Management Shell

SSL / SOAP / X509 / IPtables

Cluster Management Daemon

Provisioning

PBS Pro
Torque
Maui/MOAB
Grid Engine
SLURM
LSF

Monitoring
Automation
Health Checks
Management

Compilers
Libraries
Debuggers
Profilers

SLES / RHEL / CentOS / SL

ScaleMP vSMP

CPU

GPU

Memory

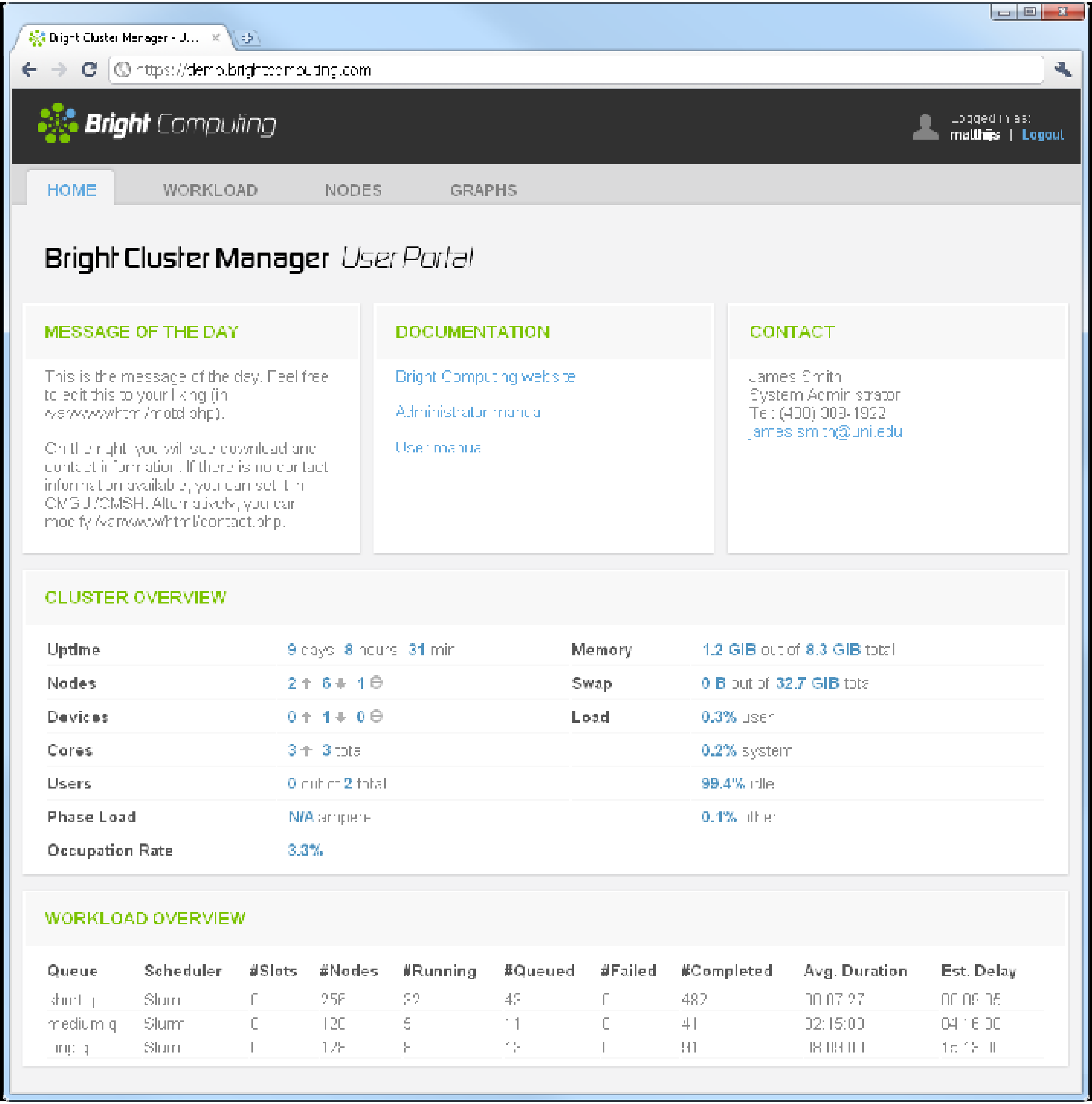
Disk

Ethernet

Interconnect

IPMI / iLO

PDU



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

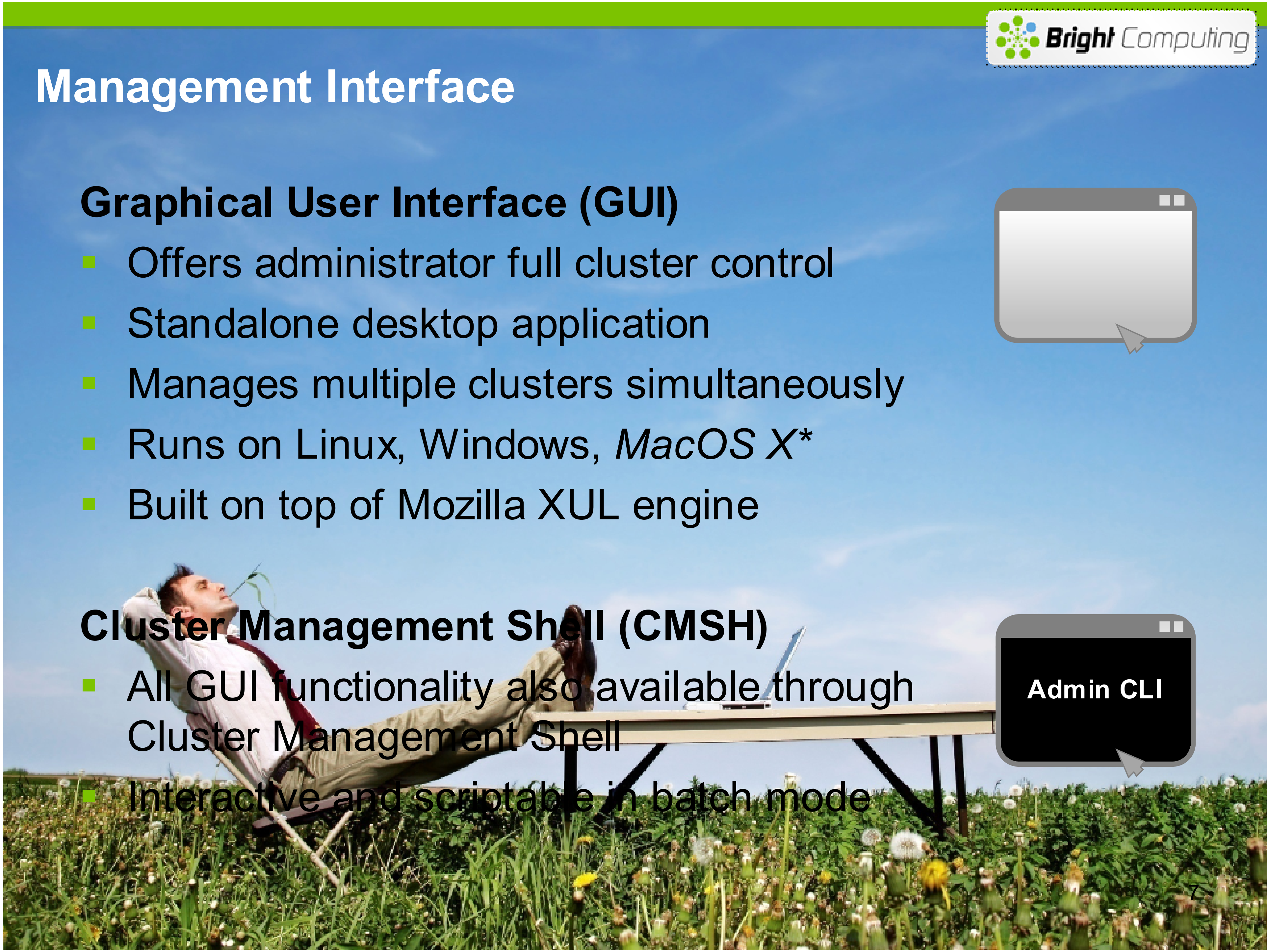
CLUSTER OVERVIEW

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 ampere		0.1% other
Occupation Rate	3.3%		

WORKLOAD OVERVIEW

Queue	Scheduler	#Slots	#Nodes	#Running	#Queued	#Failed	#Completed	Avg. Duration	Est. Delay
short.q	Slurm	0	256	22	45	0	482	00:07:27	00:09:15
medium.q	Slurm	0	120	5	11	0	41	02:15:00	04:16:00
long.q	Slurm	0	128	8	15	0	81	18:00:00	[illegible]

Management Interface



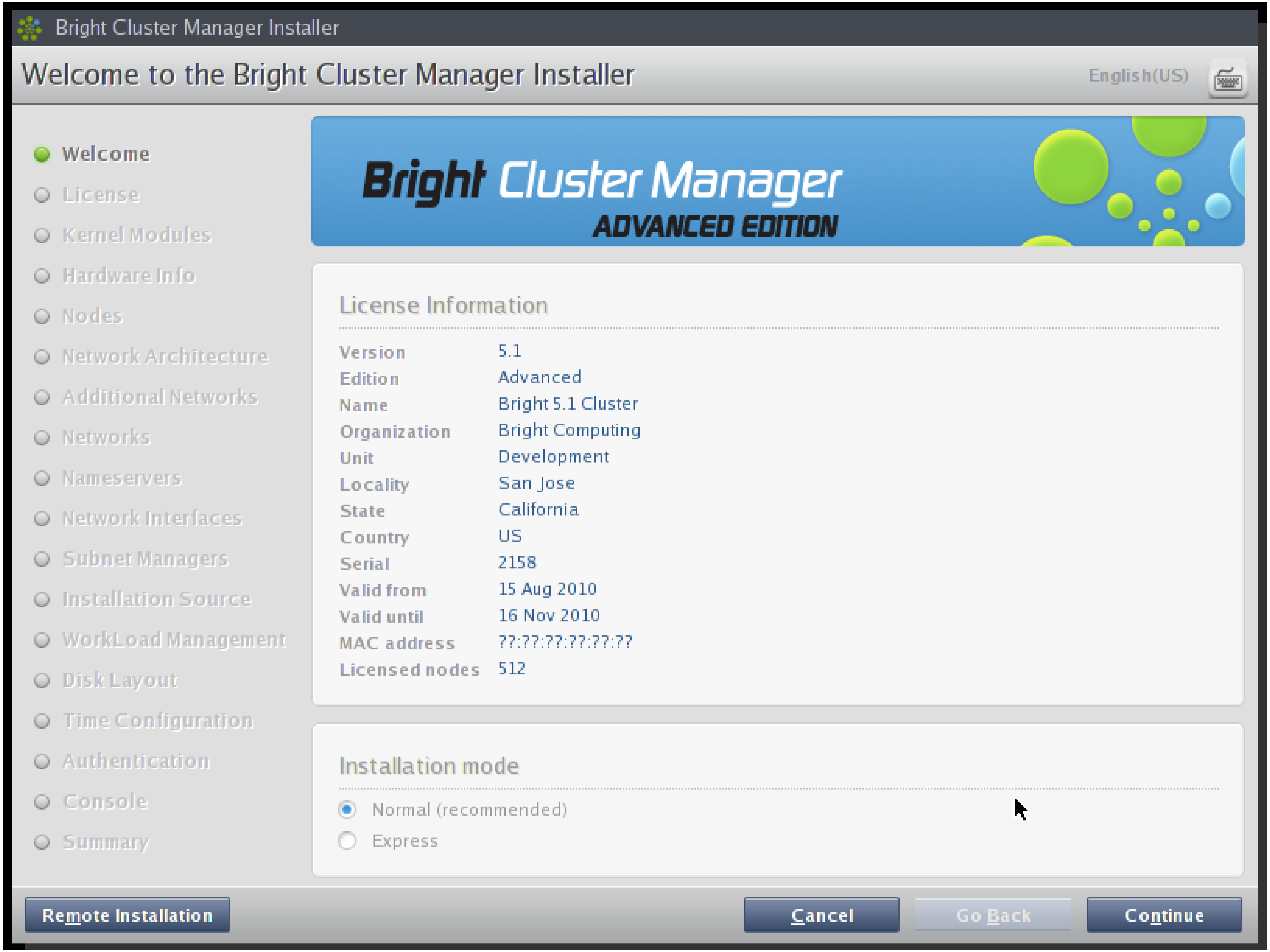
Graphical User Interface (GUI)

- 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

Admin CLI



Welcome to the Bright Cluster Manager Installer

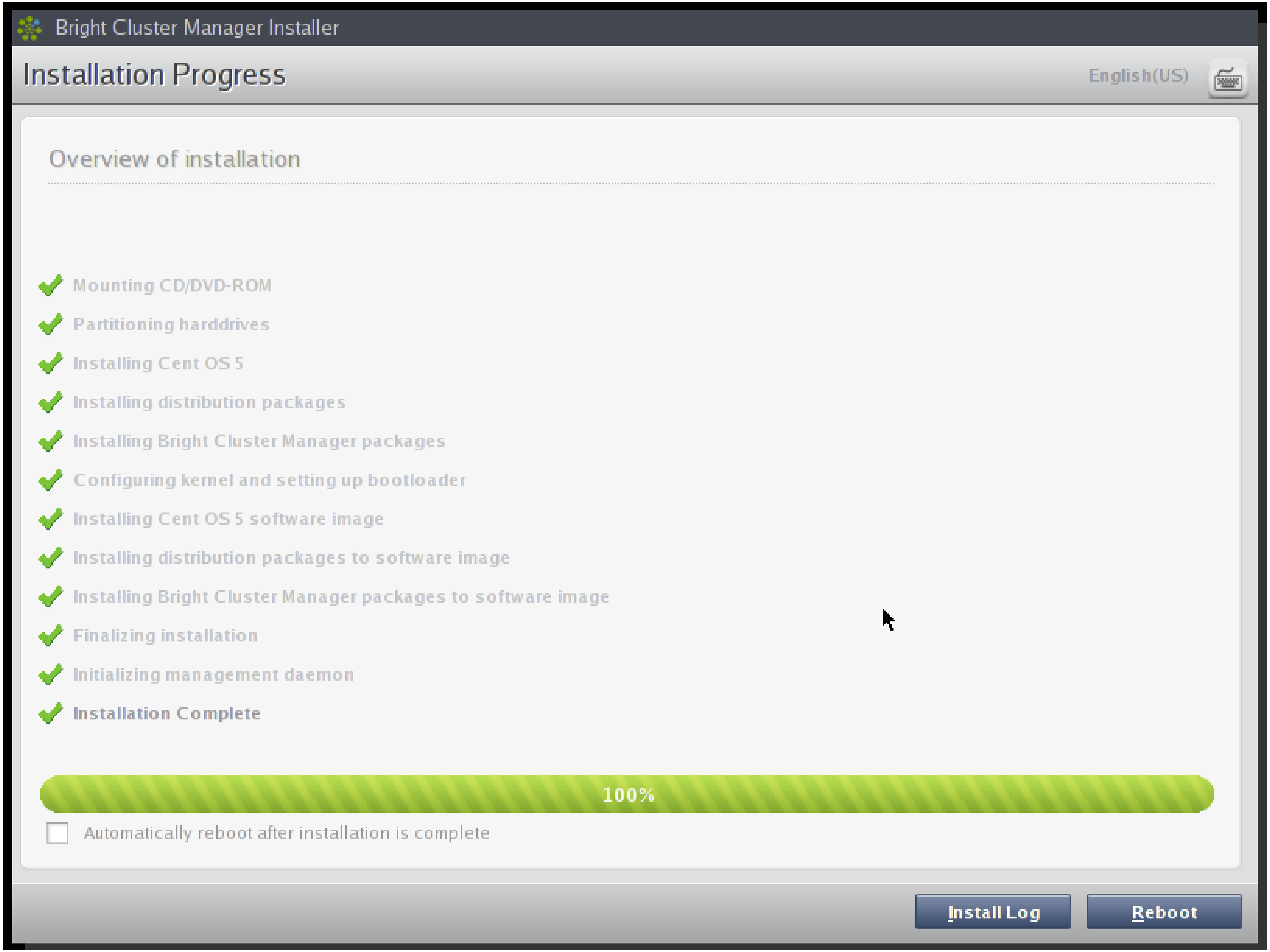
License Information

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



Bright Cluster Manager Installer

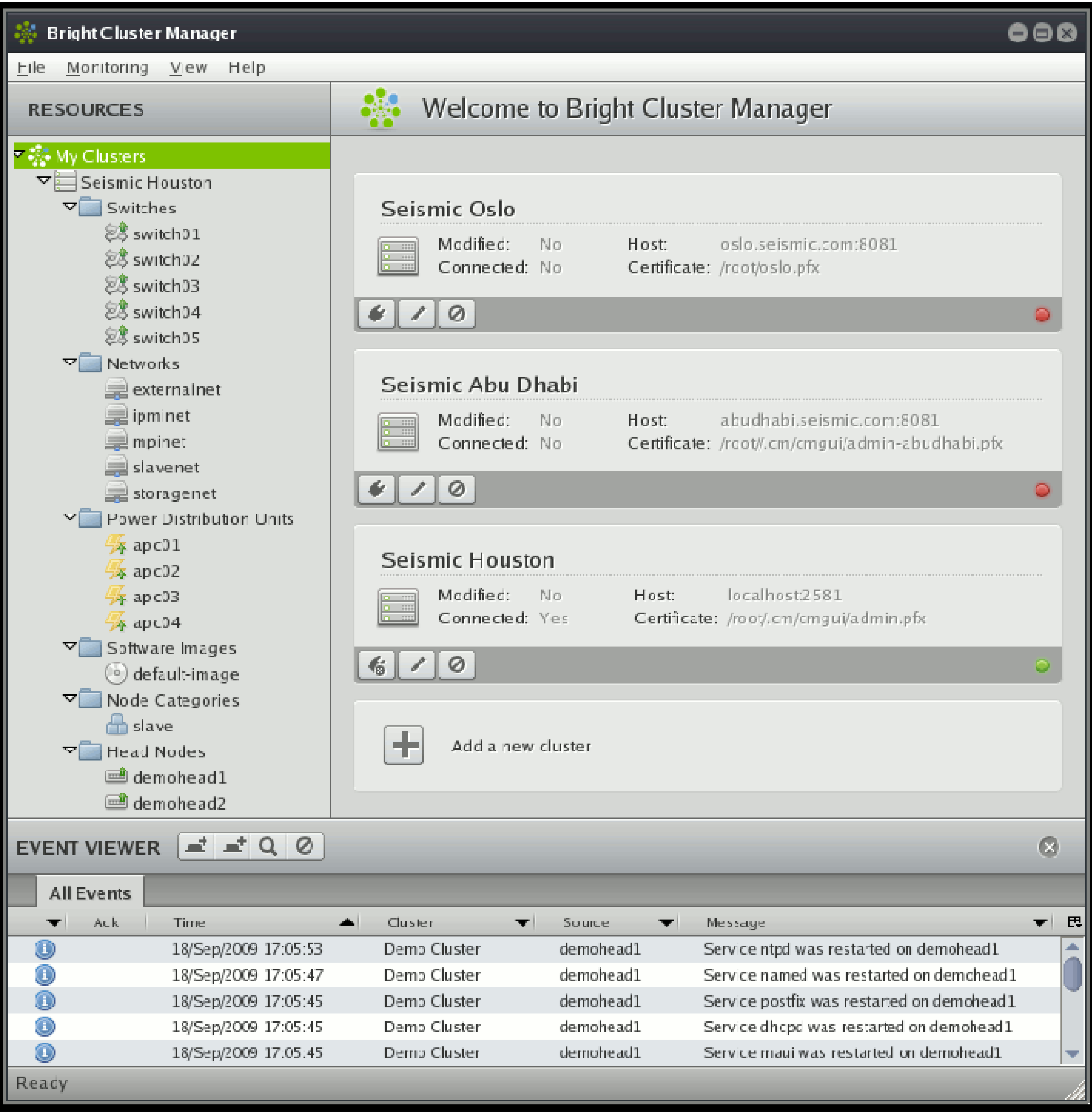
Installation Progress

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



Welcome to Bright Cluster Manager

Seismic Oslo

Modified: No Host: oslo.seismic.com:8081
Connected: No Certificate: /root/oslo.pfx

Seismic Abu Dhabi

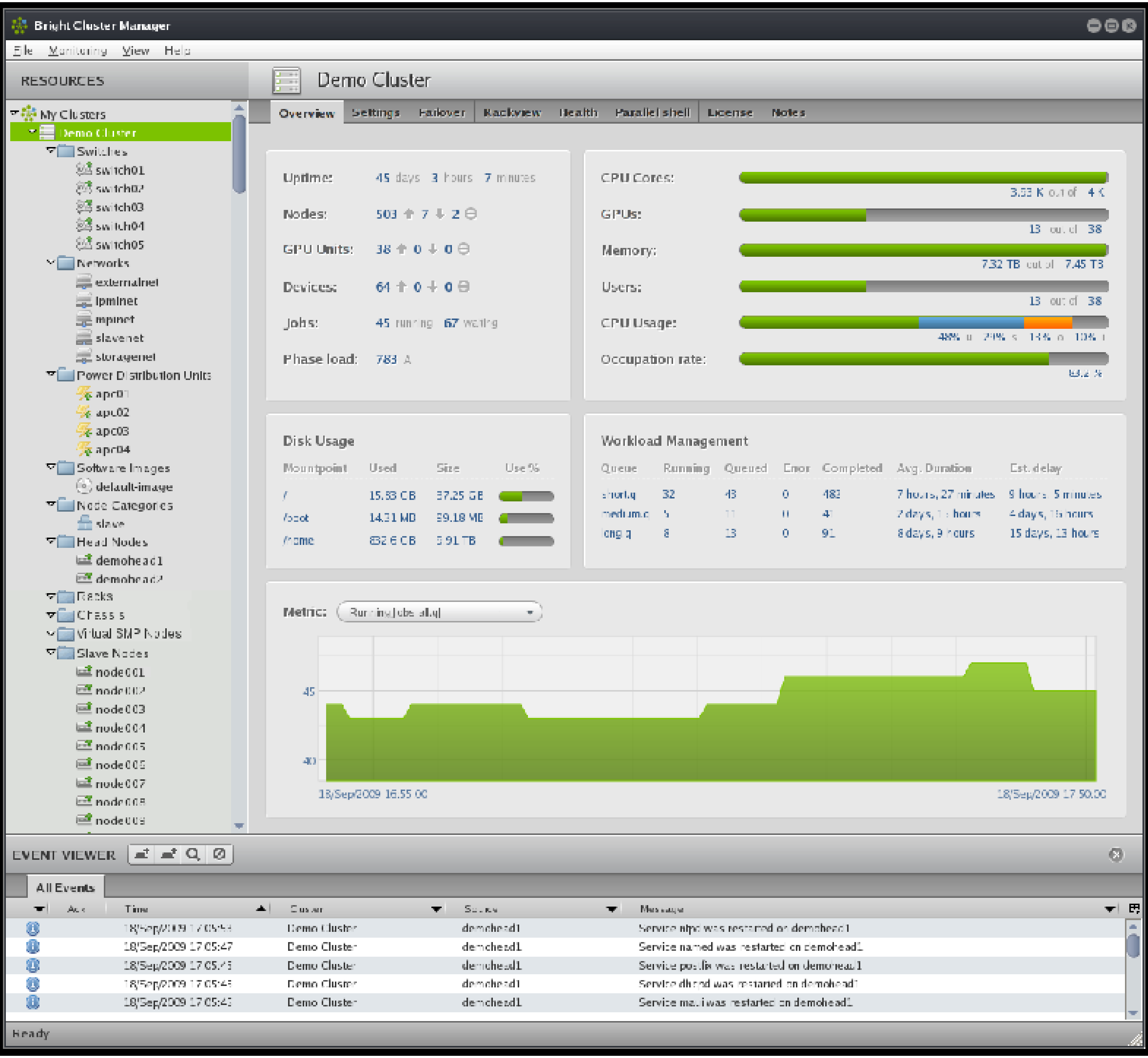
Modified: No Host: abudhabi.seismic.com:8081
Connected: No Certificate: /root/.cm/cmgui/admin-abudhabi.pfx

Seismic Houston

Modified: No Host: localhost:2581
Connected: Yes Certificate: /root/.cm/cmgui/admin.pfx

Add a new cluster

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 maui was restarted on demohead1



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: 3.93 K out of 4 K

GPUs: 13 out of 38

Memory: 7.32 TB out of 7.45 TB

Users: 13 out of 38

CPU Usage: 48% u 29% s 13% o 10% i

Occupation rate: 83.2 %

Disk Usage

Mountpoint	Used	Size	Use %
/	15.83 GB	37.25 GB	
/boot	14.31 MB	39.18 MB	
/home	832.6 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, 11 hours	4 days, 16 hours
long.q	8	13	0	91	8 days, 9 hours	15 days, 13 hours

Metric: Running Jobs all.q

	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 17:05:43	Demo Cluster	demohead1	Service mail was restarted on demohead1

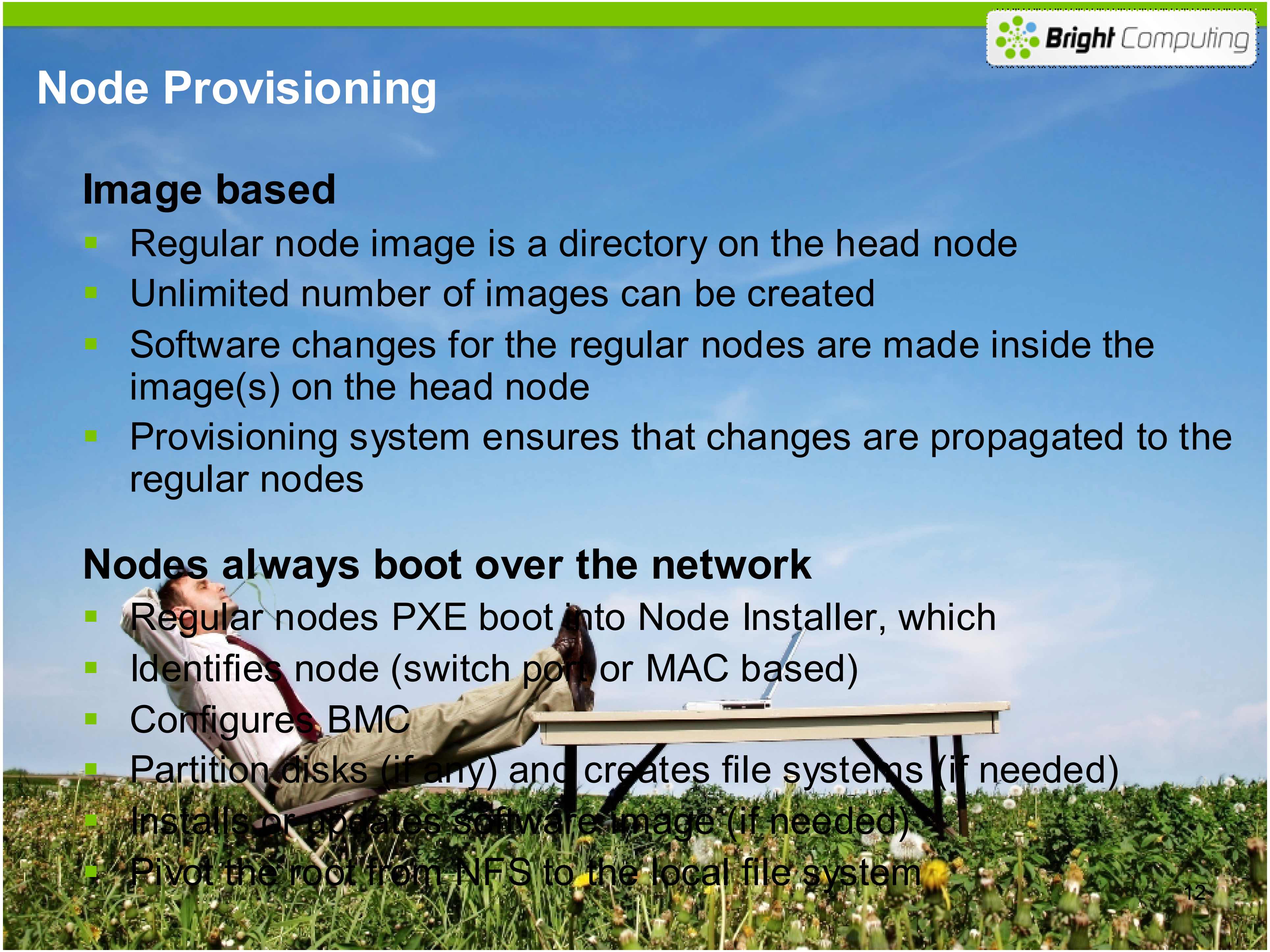
Node Provisioning

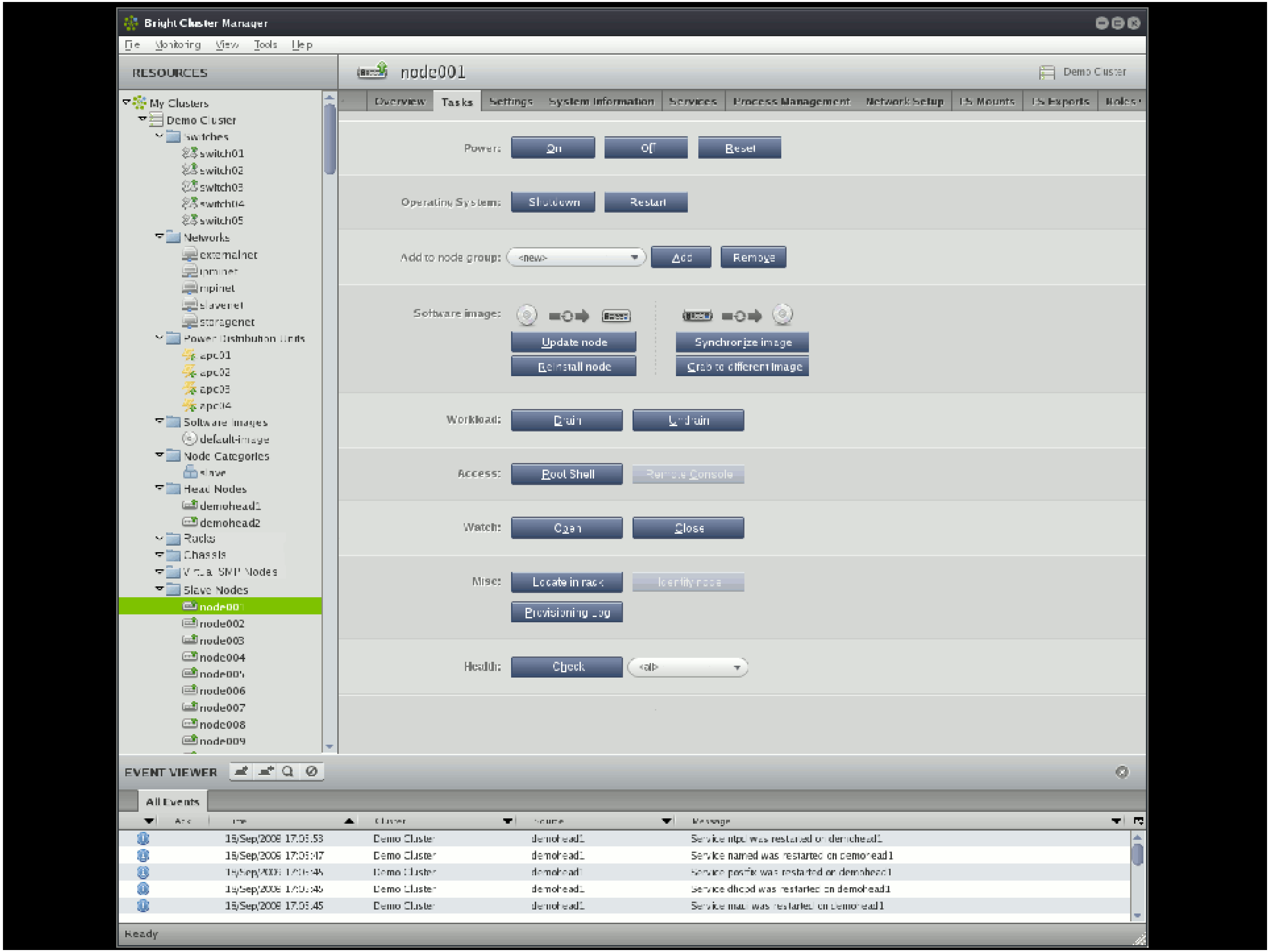
Image based

- 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

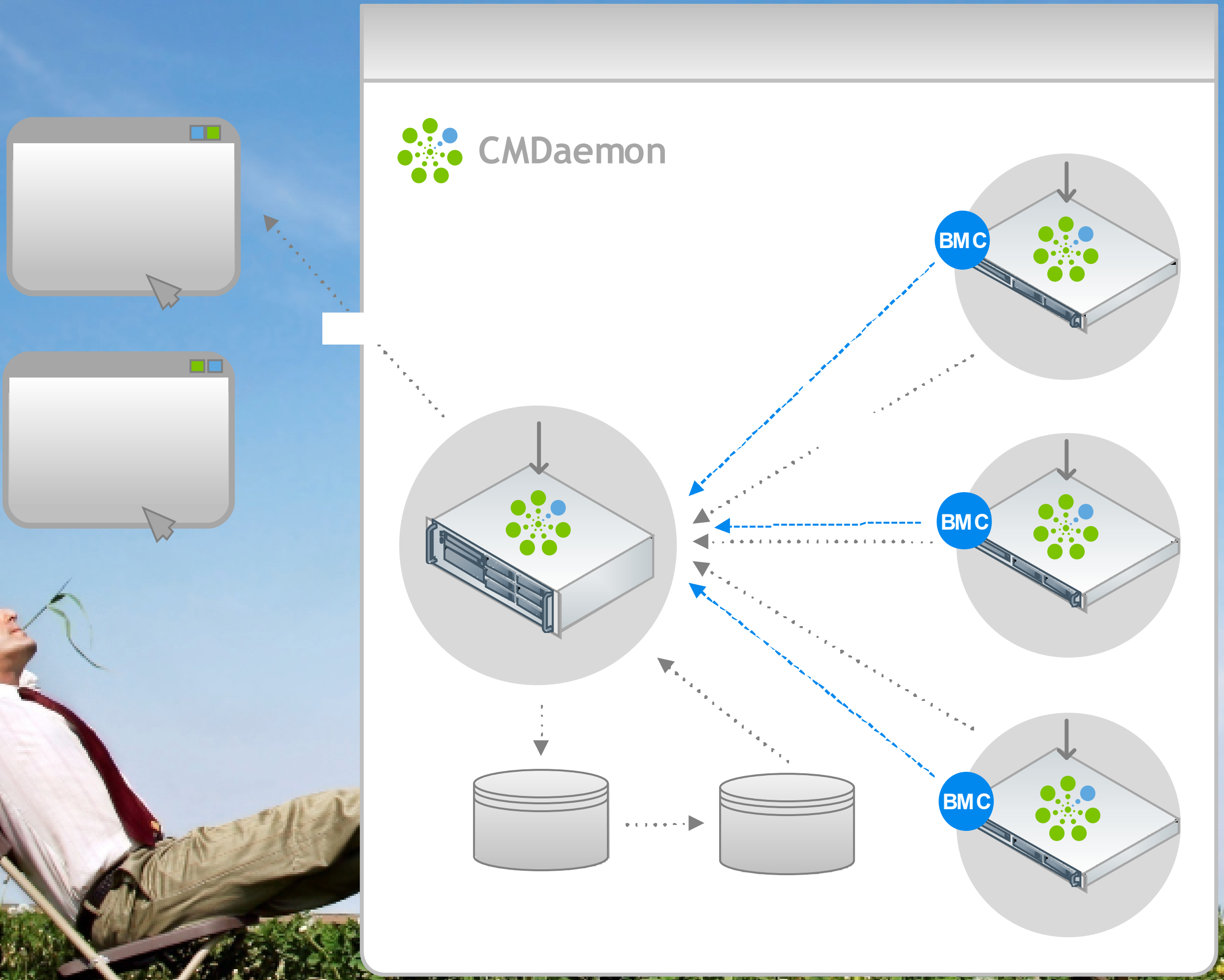


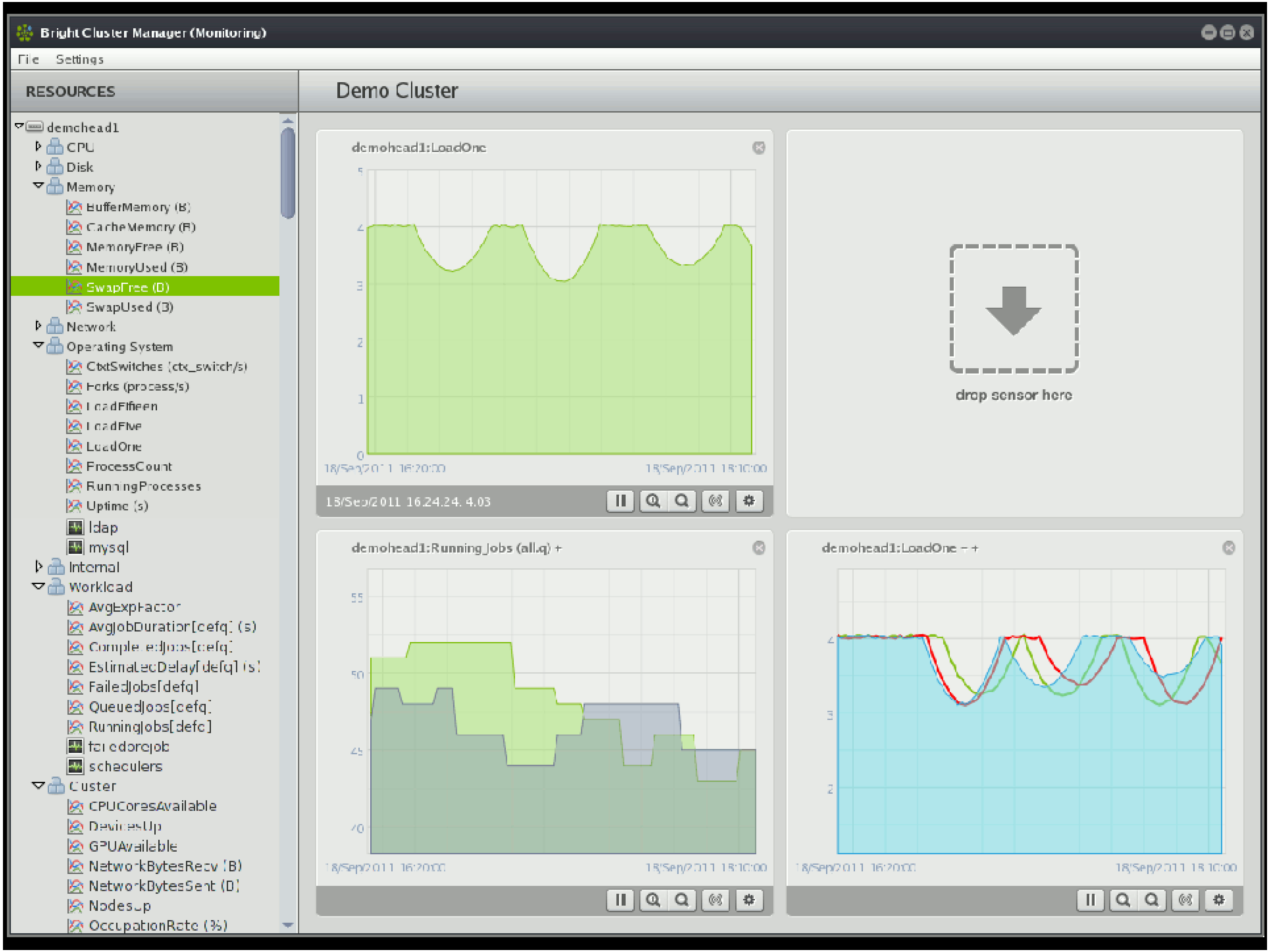


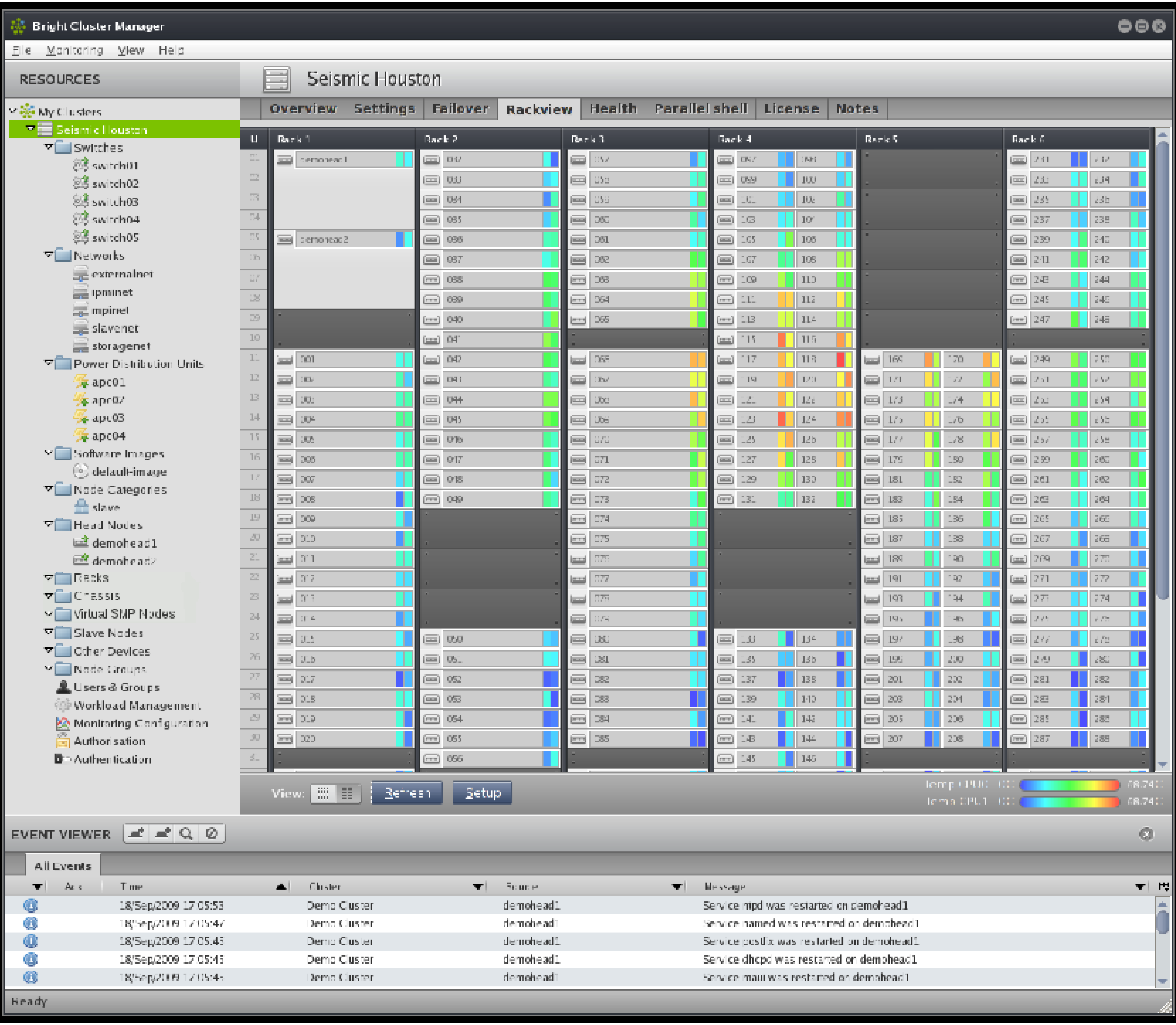
Ack	Time	Cluster	Source	Message
	15/Sep/2008 17:05:53	Demo Cluster	demohead1	Service ntpd was restarted on demohead1
	15/Sep/2008 17:05:47	Demo Cluster	demohead1	Service named was restarted on demohead1
	15/Sep/2008 17:05:45	Demo Cluster	demohead1	Service postfix was restarted on demohead1
	15/Sep/2008 17:05:45	Demo Cluster	demohead1	Service dhcpd was restarted on demohead1
	15/Sep/2008 17:05:45	Demo Cluster	demohead1	Service mail was restarted on demohead1

Bright Computing

Architecture — Monitoring

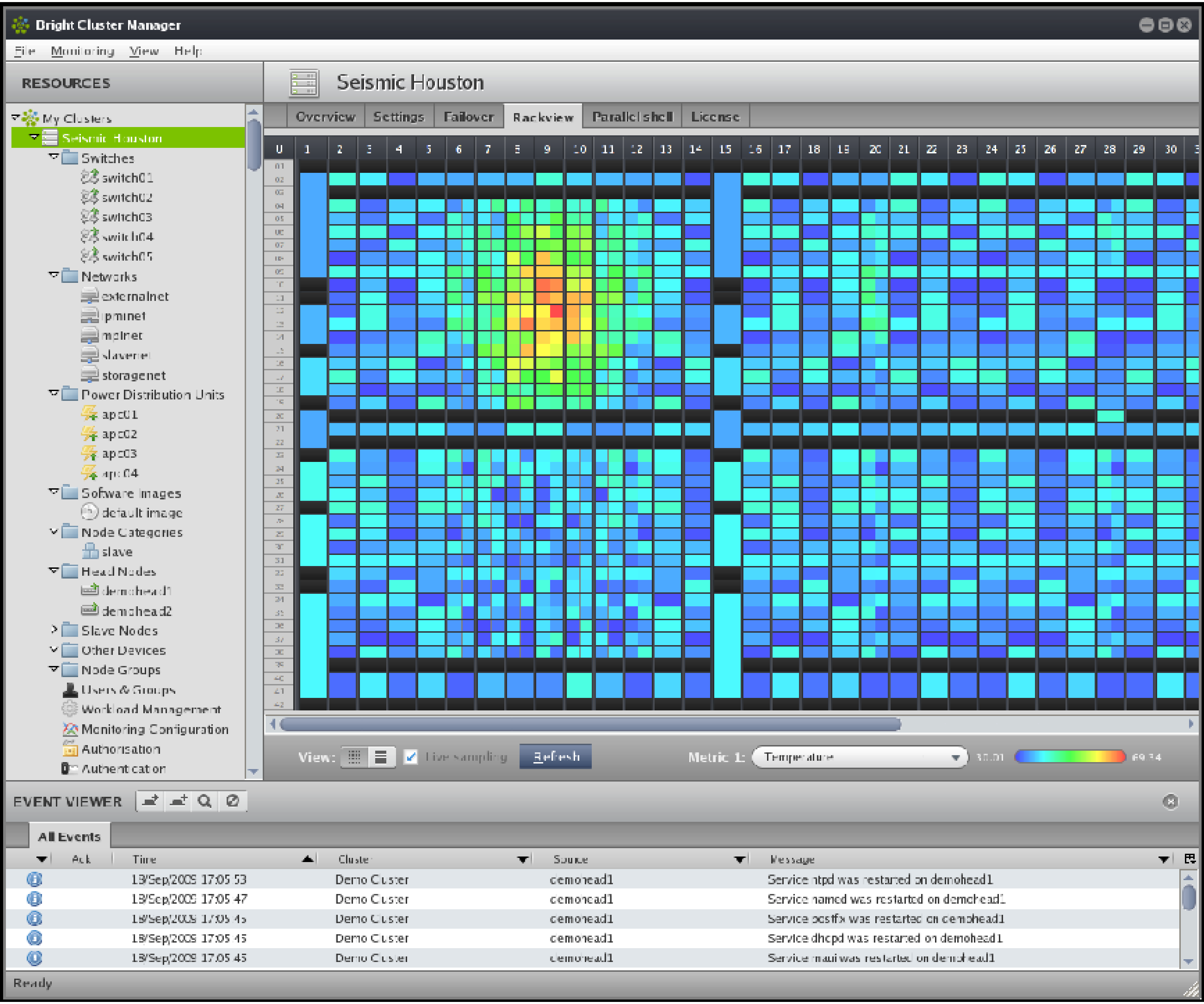



Bright Cluster Manager (Monitoring)
File
Settings
RESOURCES
demohead1
CPU
Disk
Memory
BufferMemory (B)
CacheMemory (B)
MemoryFree (B)
MemoryUsed (B)
SwapFree (B)
SwapUsed (B)
Network
Operating System
CtxtSwitches (ctx_switch/s)
Forks (process/s)
LoadFifteen
LoadFive
LoadOne
ProcessCount
RunningProcesses
Uptime (s)
ldap
mysql
Internal
Workload
AvgExpFactor
AvgJobDuration[defq] (s)
CompletedJobs[defq]
EstimatedDelay[defq] (s)
FailedJobs[defq]
QueuedJobs[defq]
RunningJobs[defq]
failedprejob
schedulers
Cluster
CPUCoresAvailable
DevicesUp
GPUAvailable
NetworkBytesRecv (B)
NetworkBytesSent (B)
NodesUp
OccupationRate (%)
Demo Cluster
demohead1:LoadOne
5
4
3
2
1
0
18/Sep/2011 16:20:00
18/Sep/2011 18:10:00
18/Sep/2011 16:24:24, 4.03
drop sensor here
demohead1:Running_Jobs (all.q) +
55
50
45
40
18/Sep/2011 16:20:00
18/Sep/2011 18:10:00
demohead1:LoadOne + +
4
3
2
18/Sep/2011 16:20:00
18/Sep/2011 18:10:00



- 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

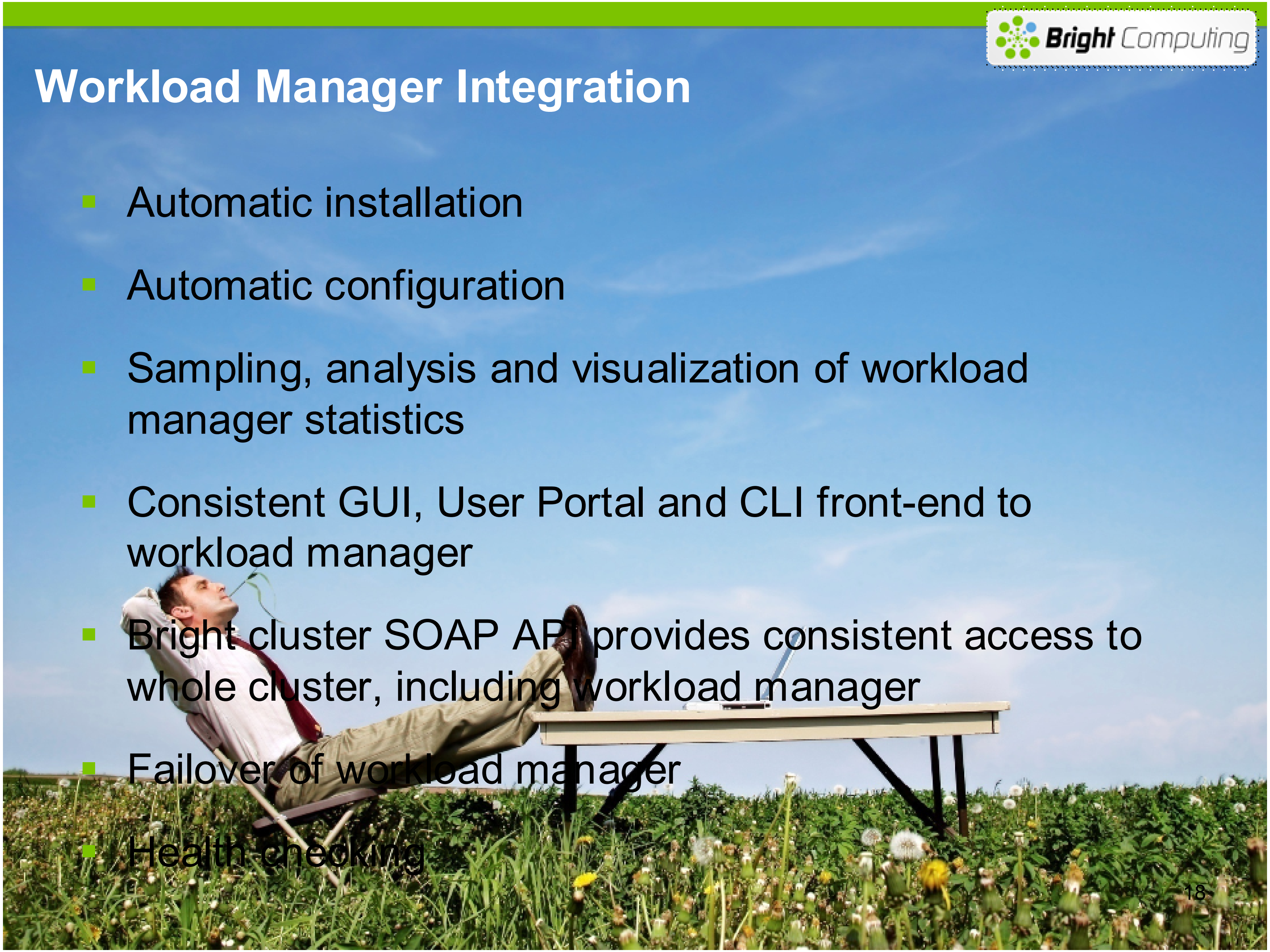
Ack	Time	Cluster	Source	Message
	18/Sep/2009 17:05:53	Demo Cluster	demohead1	Service mpd was restarted on demohead1
	18/Sep/2009 17:05:47	Demo Cluster	demohead1	Service named was restarted on demohead1
	18/Sep/2009 17:05:45	Demo Cluster	demohead1	Service postfix was restarted on demohead1
	18/Sep/2009 17:05:43	Demo Cluster	demohead1	Service dhcpd was restarted on demohead1
	18/Sep/2009 17:05:4[illegible]	Demo Cluster	demohead1	Service mail was restarted on demohead1



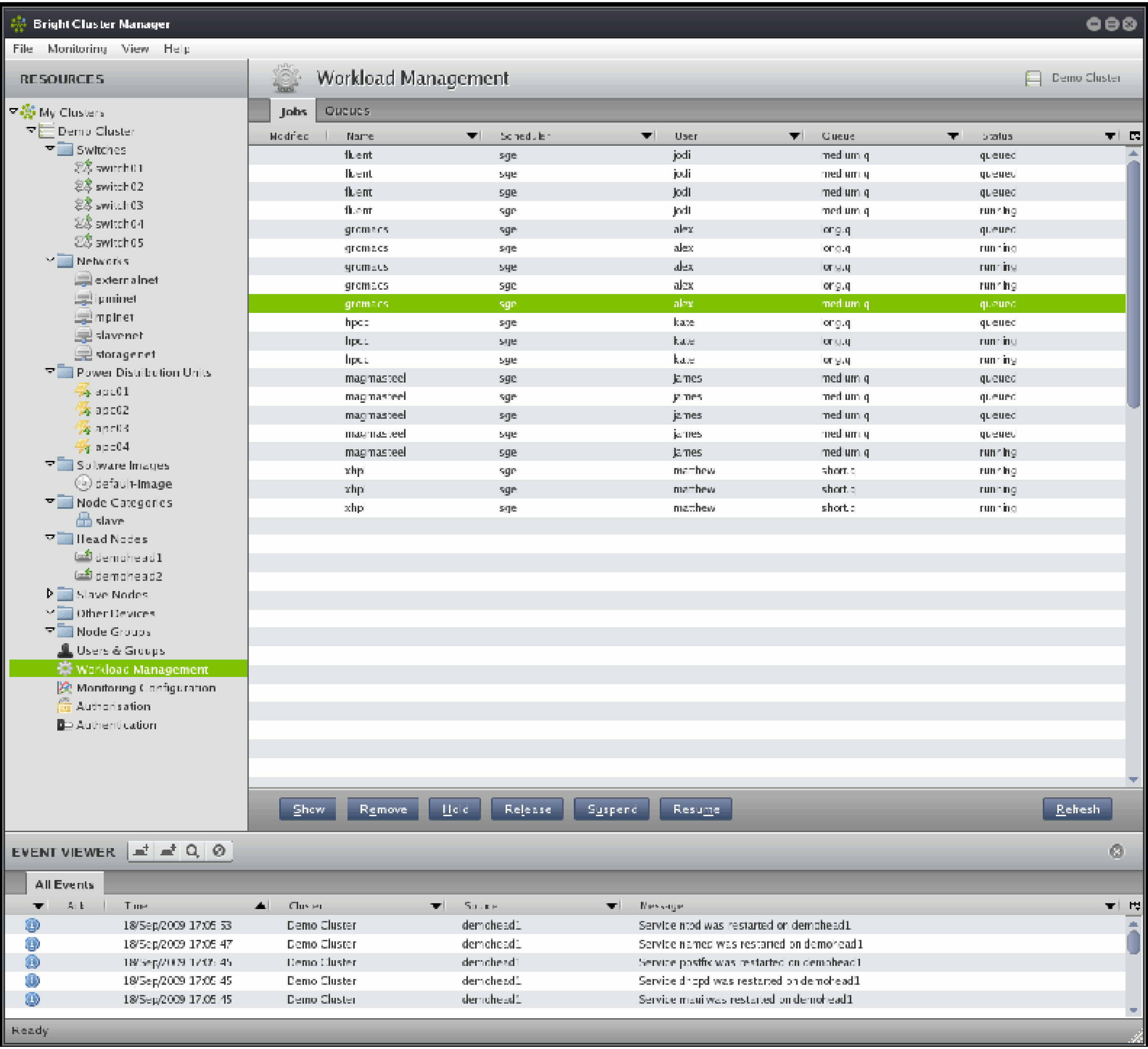
	Ack	Time	Cluster	Source	Message
●		18/Sep/2009 17:05:53	Demo Cluster	demohead1	Service ntpd was restarted on demohead1
●		18/Sep/2009 17:05:47	Demo Cluster	demohead1	Service named was restarted on demohead1
●		18/Sep/2009 17:05:45	Demo Cluster	demohead1	Service postfix was restarted on demohead1
●		18/Sep/2009 17:05:45	Demo Cluster	demohead1	Service dhcpd was restarted on demohead1
●		18/Sep/2009 17:05:45	Demo Cluster	demohead1	Service maui 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



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	org.q	queued
	gromacs	sge	alex	org.q	running
	gromacs	sge	alex	org.q	running
	gromacs	sge	alex	org.q	running
	gromacs	sge	alex	medium.q	queued
	hpcc	sge	kate	org.q	queued
	hpcc	sge	kate	org.q	running
	hpcc	sge	kate	org.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	mathew	short.q	running
	xhp	sge	mathew	short.q	running
	xhp	sge	mathew	short.q	running

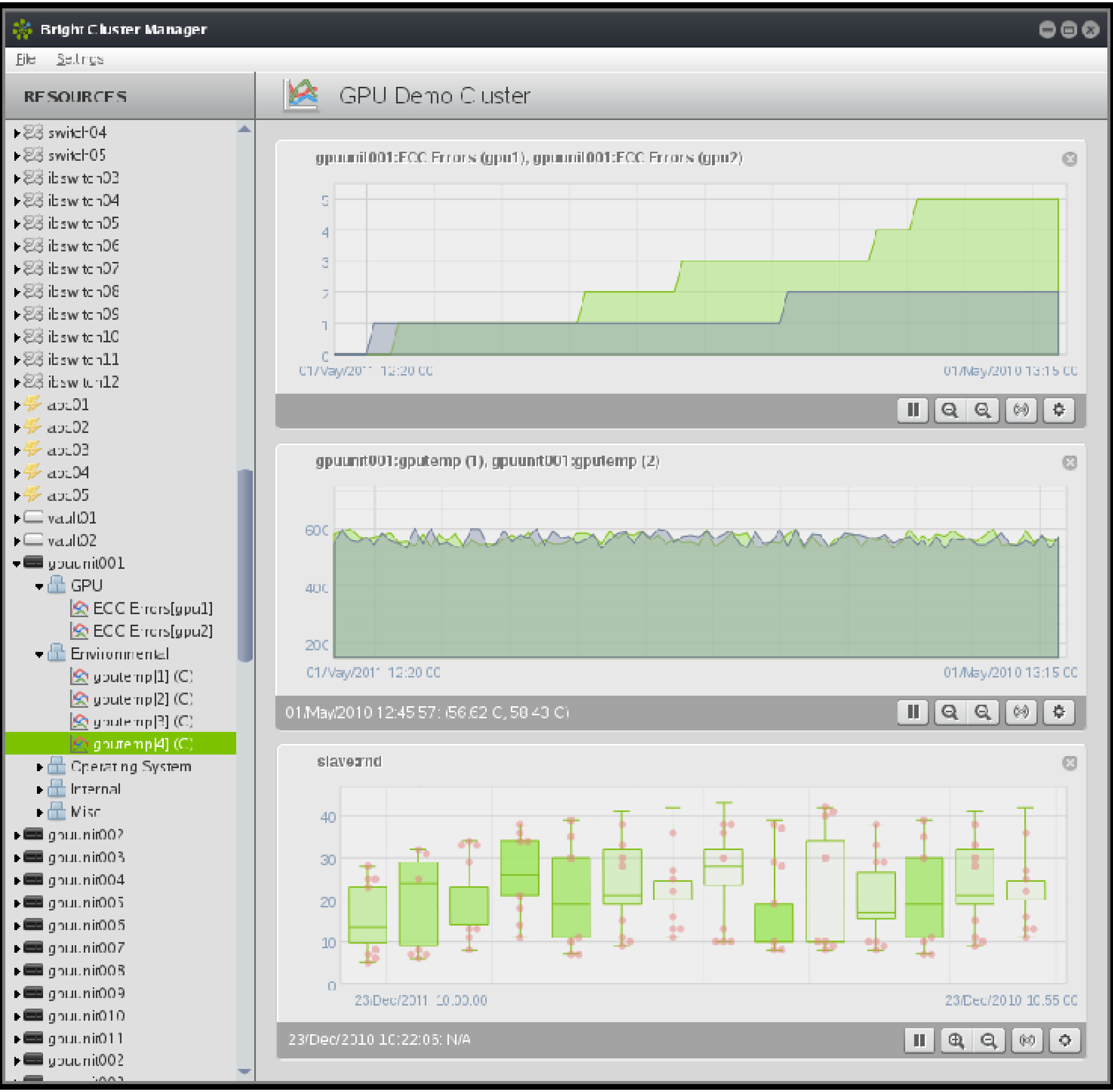
	Ack	Time	Cluster	Source	Message
ⓘ		18/Sep/2009 17:05:53	Demo Cluster	demohead1	Service ntpd was restarted on demohead1
ⓘ		18/Sep/2009 17:05:47	Demo Cluster	demohead1	Service named was restarted on demohead1
ⓘ		18/Sep/2009 17:05:45	Demo Cluster	demohead1	Service postfix was restarted on demohead1
ⓘ		18/Sep/2009 17:05:45	Demo Cluster	demohead1	Service dhcpd was restarted on demohead1
ⓘ		18/Sep/2009 17:05:45	Demo Cluster	demohead1	Service mail was restarted on demohead1

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

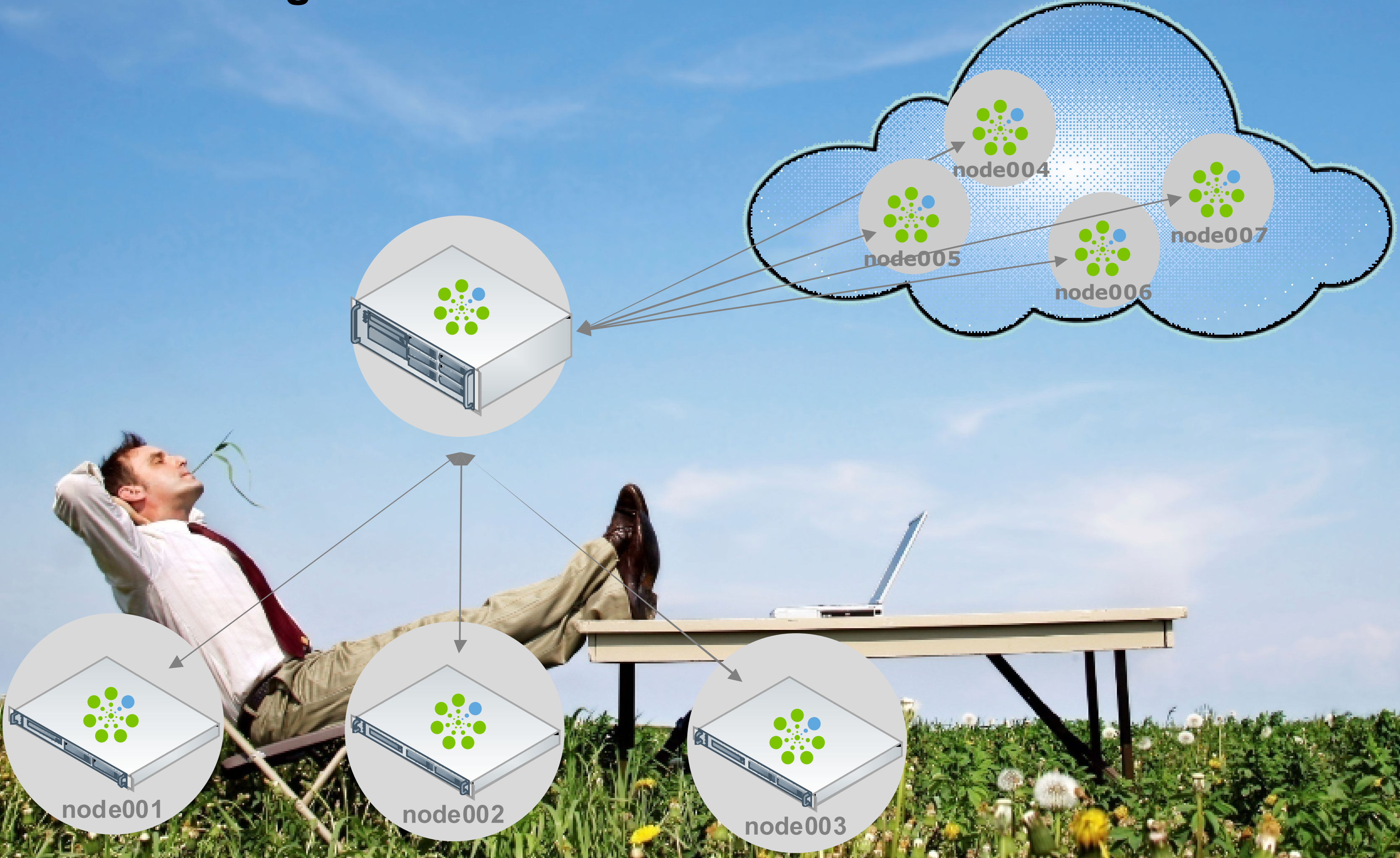


01/May/2010 12:45:57: (56.62 C, 58.43 C)

23/Dec/2010 10:22:05: N/A

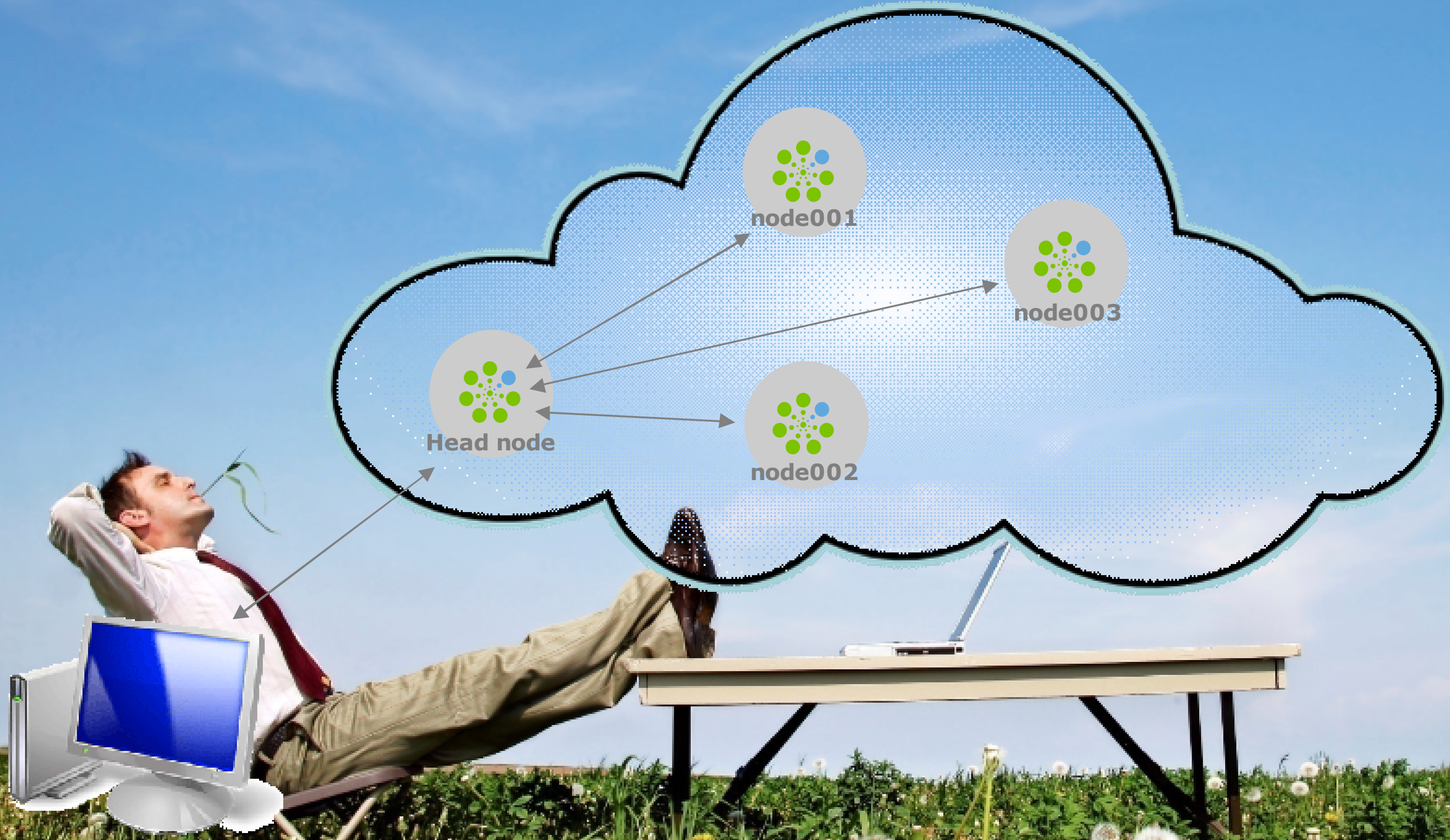
The Future

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

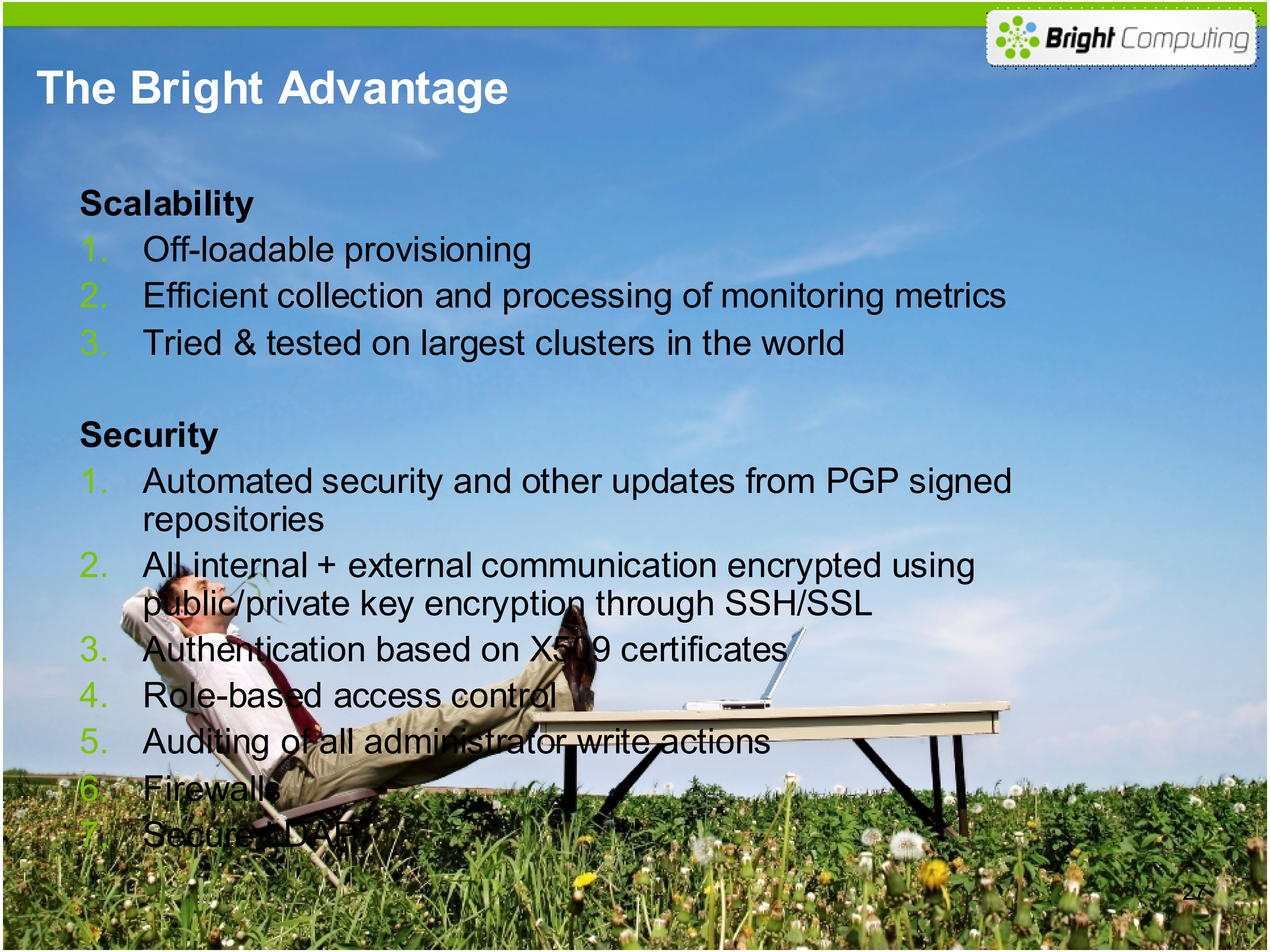
Performance

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