Dipl. Ing. Arno Ziebart

Sales Manager Deutschland

Clusterday 2009 AEI

ClusterVision

clustervision

- Specialists in Compute, Storage & Graphic Clusters
- Unique position in Europe
- Offices in Amsterdam, London, Munich, Paris, Milan, Geneva, Barcelona, Oslo
- Strong partnerships with IBM and DELL across Europe
- Background in Science, Research, Engineering
- At forefront of clustering technology
- Award winning (Intermediair, Vosko, NBCC)
- Financially strong, profitable

Produkte

clustervision

High Performance Computing

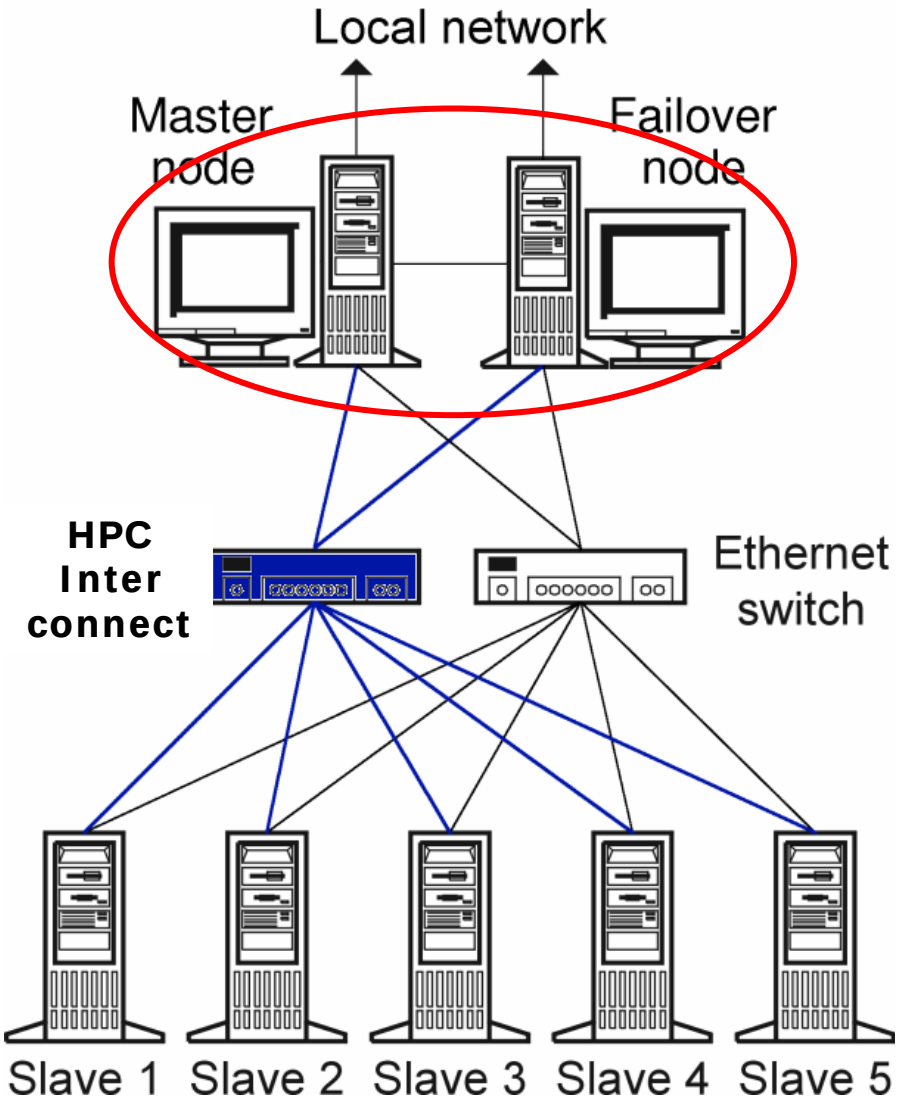
- Compute Cluster
- Storage Cluster
- Graphic GPU Cluster
- ClusterVisionOS
- HPC Server 2008
- Server & Racks
- Cooling and Aircondition

Grid

- Grid-enabled Clusters
- Grid Support
- Virtual Laboratory

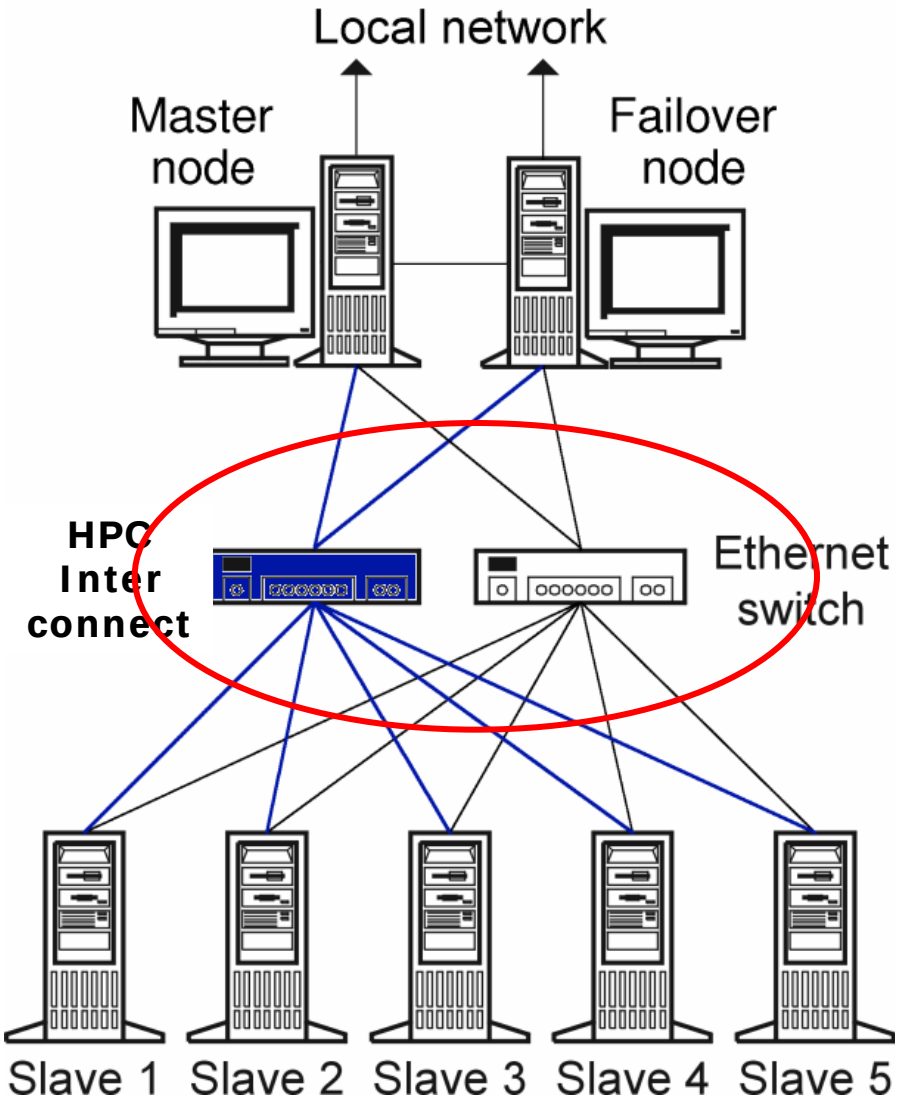
Cluster Trends

clustervision



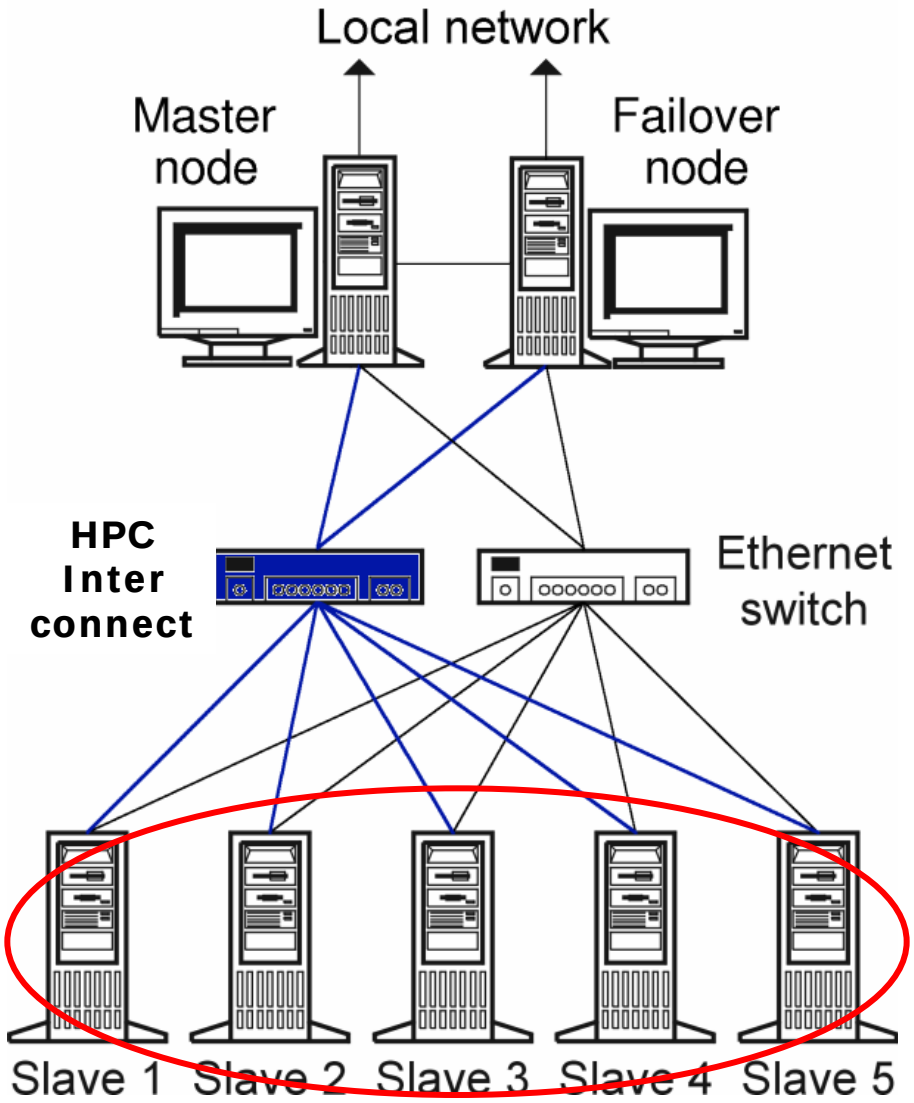
Cluster Trends

clustervision



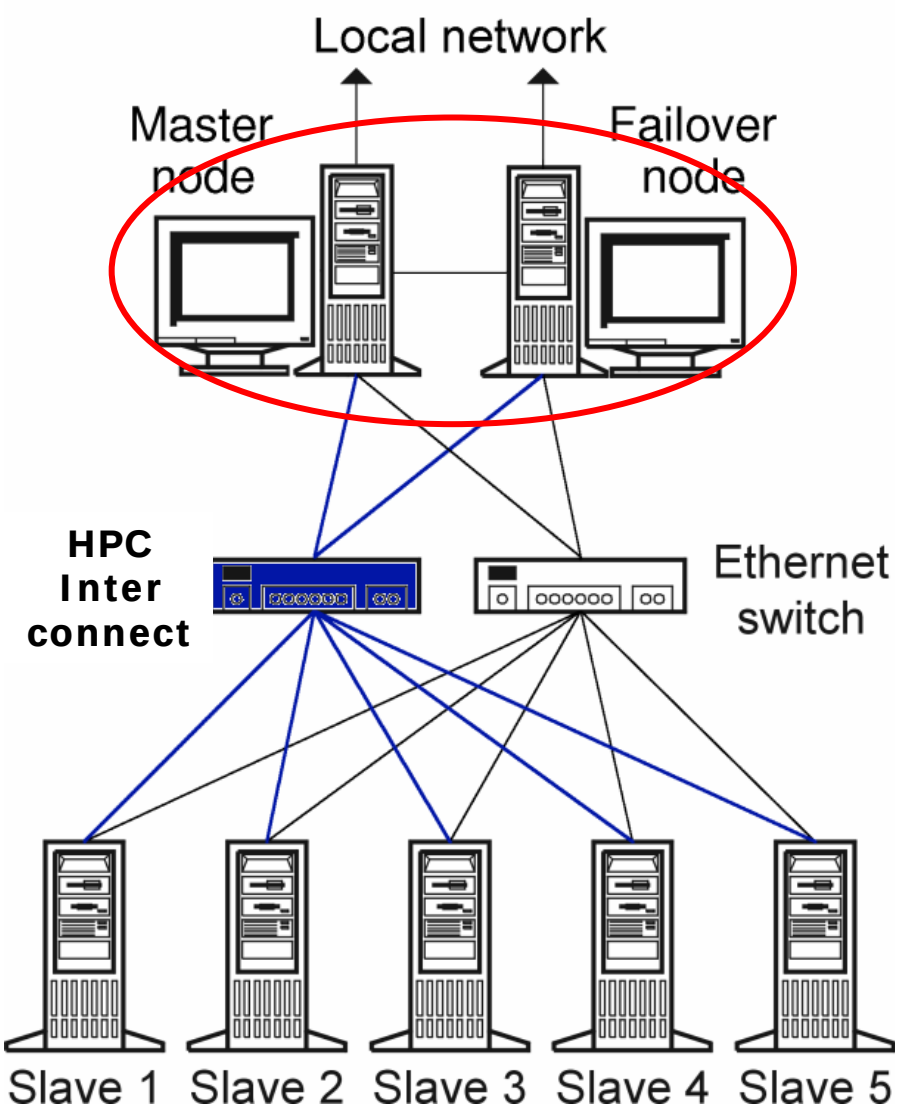
Cluster Trends

clustervision



Cluster Trends

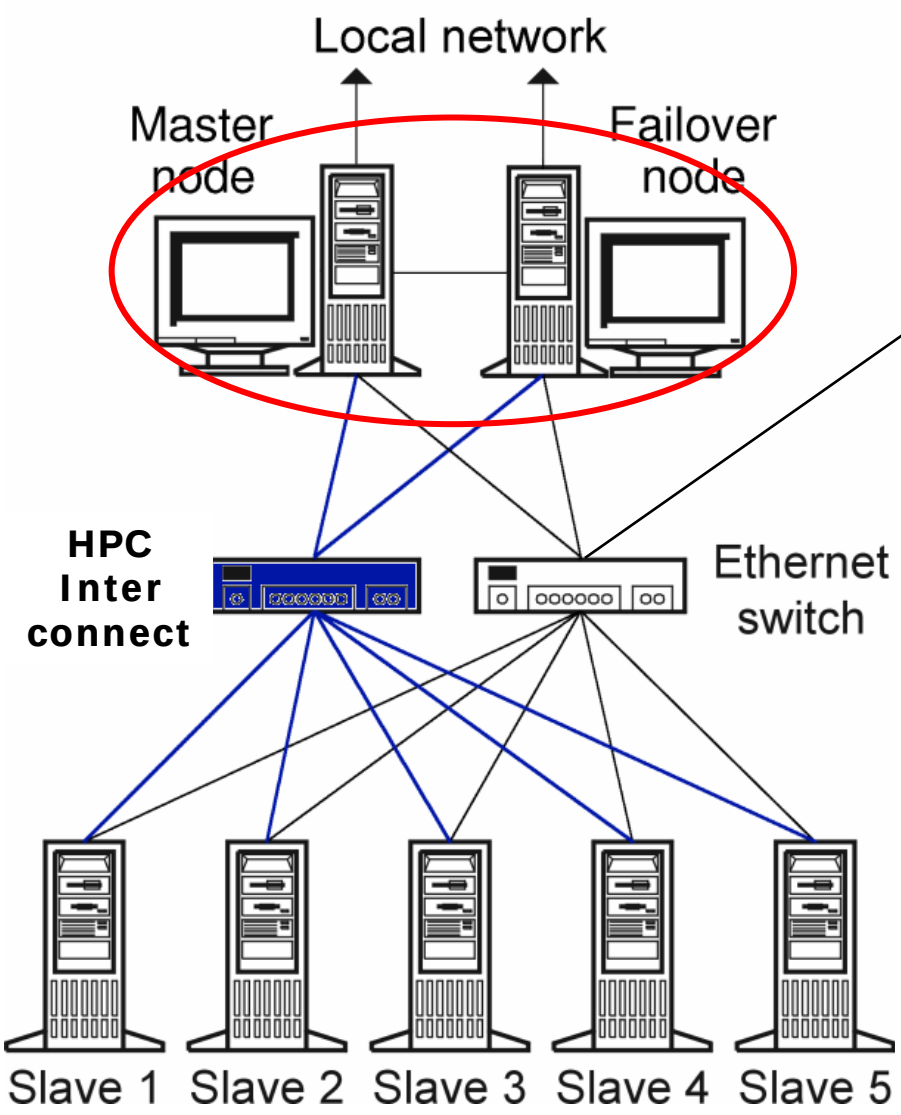
clustervision

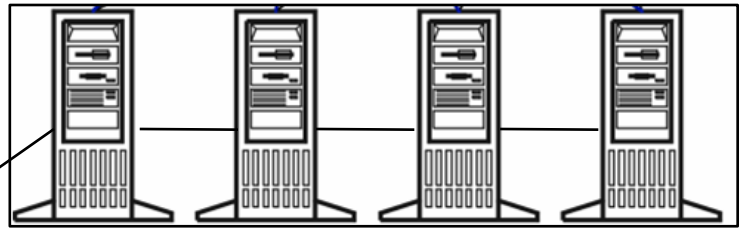


- Clustermanagement
  - Imagemanagement
  - Clusteradministration (Ldap,User,Security,Nodes)
  - Scheduling, Green-IT
  - Monitoring, Überwachung
  - SW-Development
  - User Moduls

Cluster Trends

clustervision



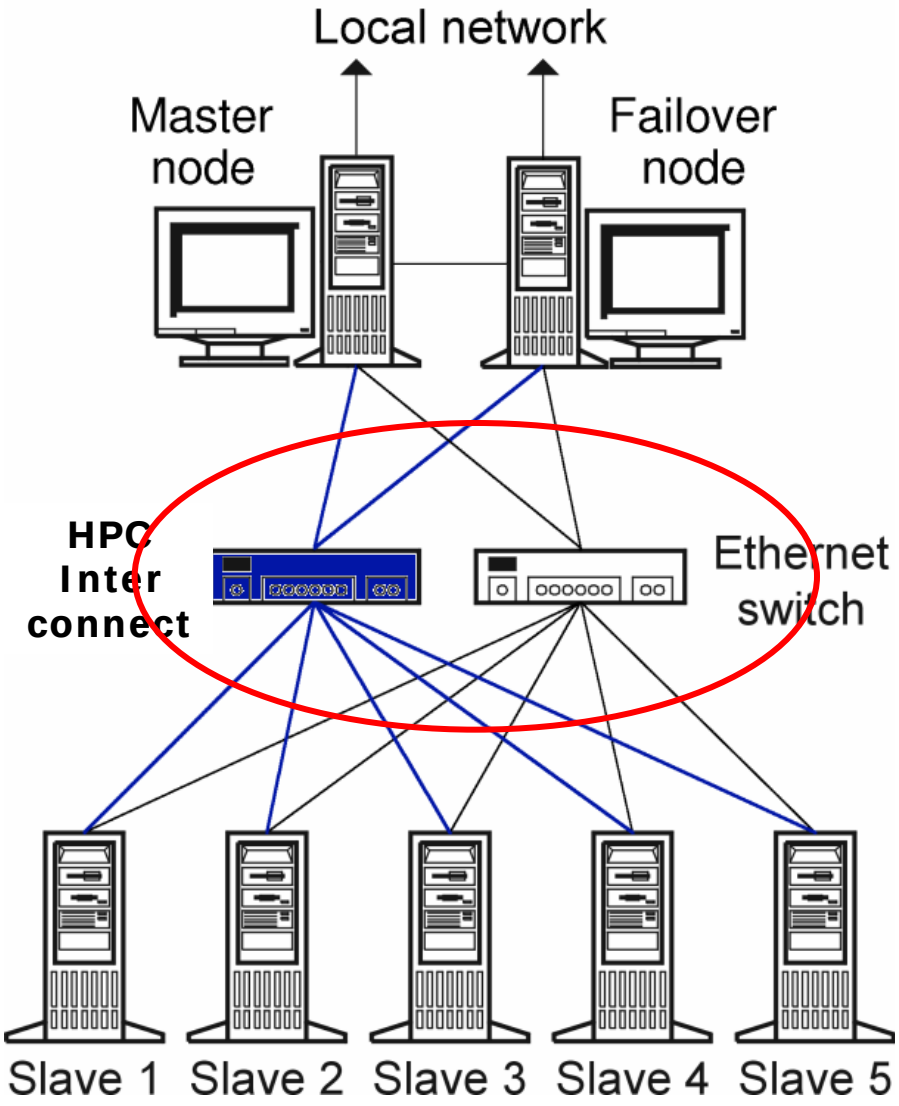


Monitoring OS Deployment

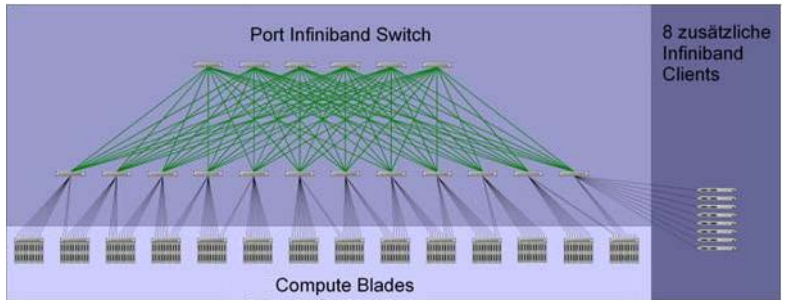
- Clustermanagement
  - Imagemanagement
  - Clusteradministration (Ldap,User,Security,Nodes)
  - Scheduling, Green-IT
  - Monitoring, Überwachung
  - SW-Development
  - User Moduls

Network Trends

clustervision



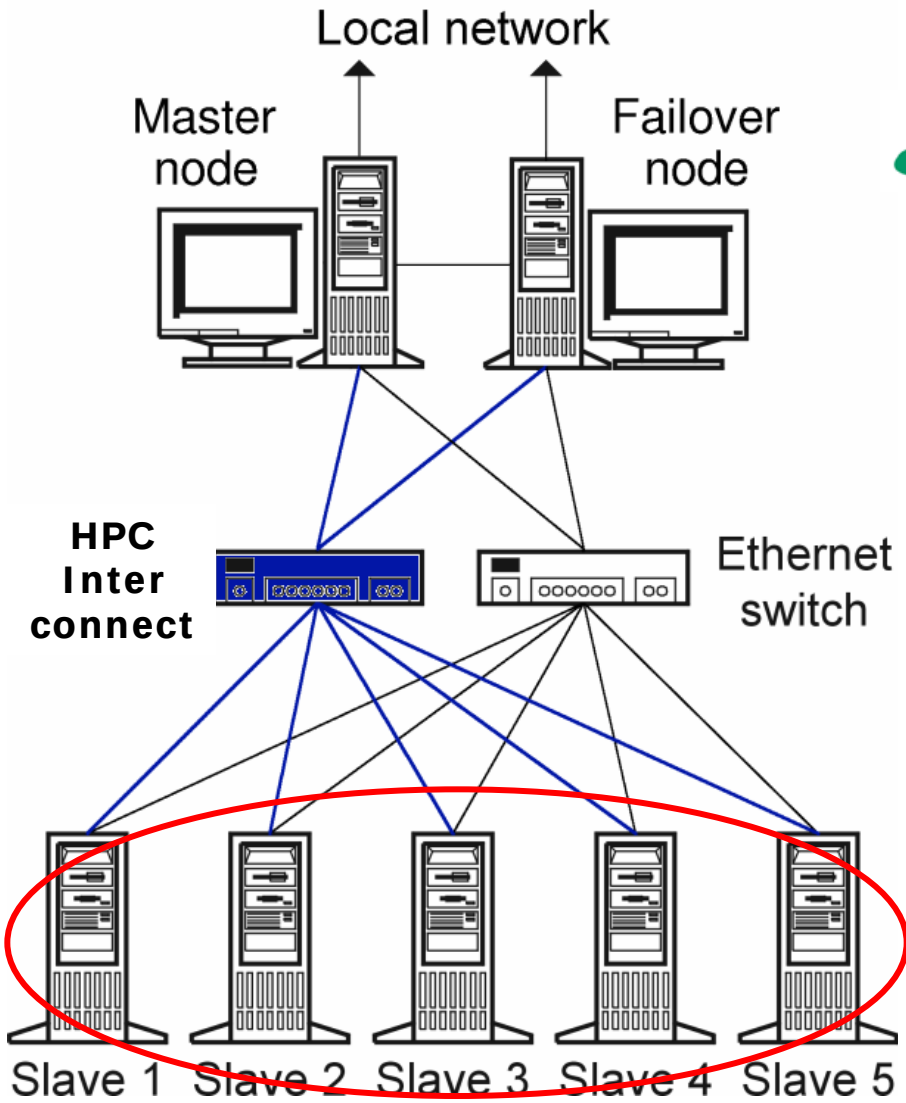
- Myri-10G, Myrinet 2000
- InfiniBand SDR/DDR/QDR
- InfiniBand ConnectX
- InfiniPath SDR/DDR



- New native 36 port asics
- 324 port switches
- Latency matters
- Convergence Ethernet10G

Cluster Trends

clustervision



AMD Opteron Shanghai”

Intel Xeon Nehalem

- GPU-FPU Coprocessors
  - Nividia GTX285,C1060,S1070
  - ATI/AMD Firestream
  - Clearspeed

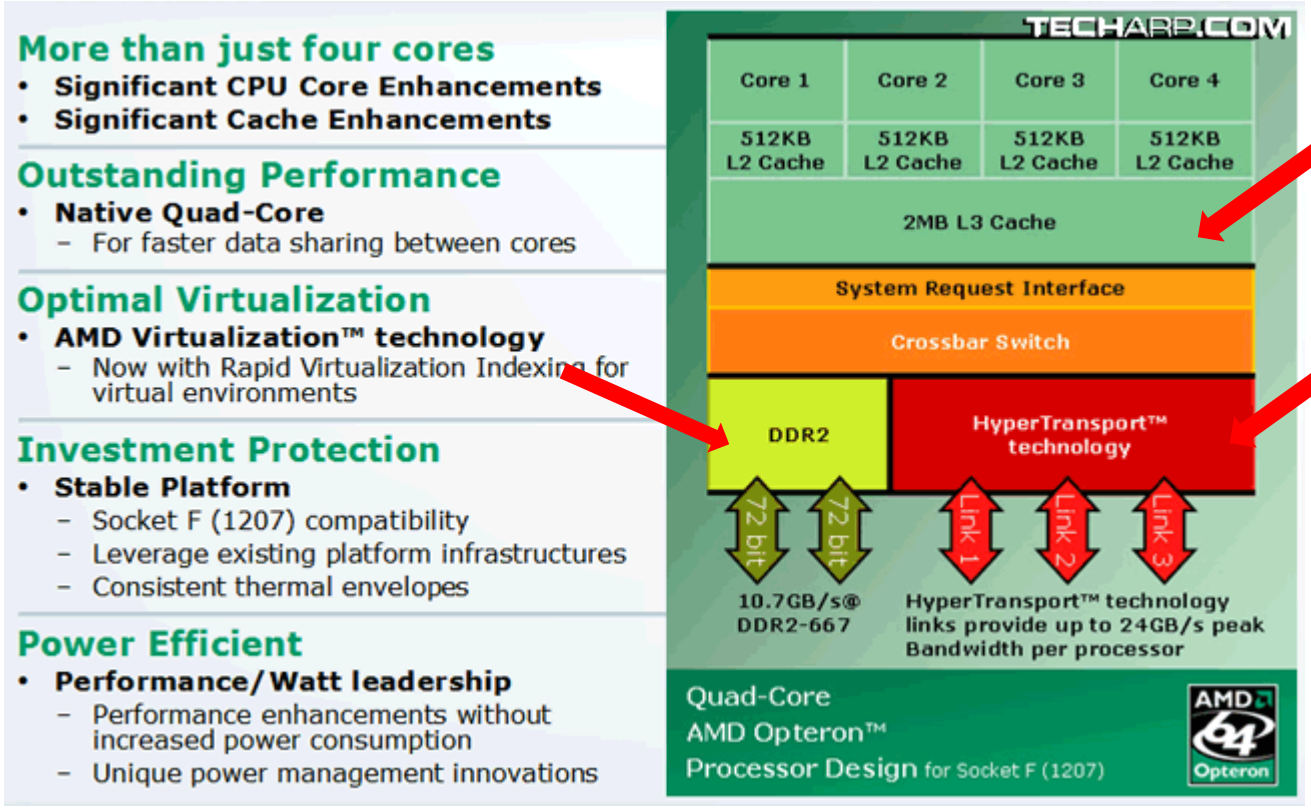
- OpenCL
  - Hardwareunabhängiges API
  - Multivendor
- Powerconsumption

VASP "Shanghai"

clustervision

Testsystem 1:

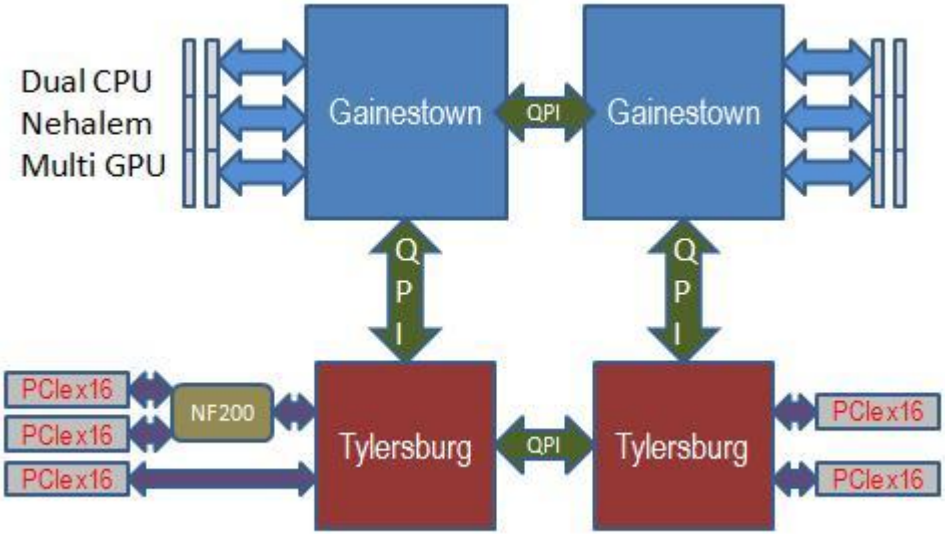
- 2 * Opteron 2380
- 8 Kerne à 2.6 GHz
- DDRII-667 Registered Memory



VASP "Nehalem"

Testsystem 2:

- 2 * Xeon 5540 „Nehalem"
- 8 Kerne à 2.53 GHz
- 2 Threads pro Kern
- DDRIII-1333 Registered Memory



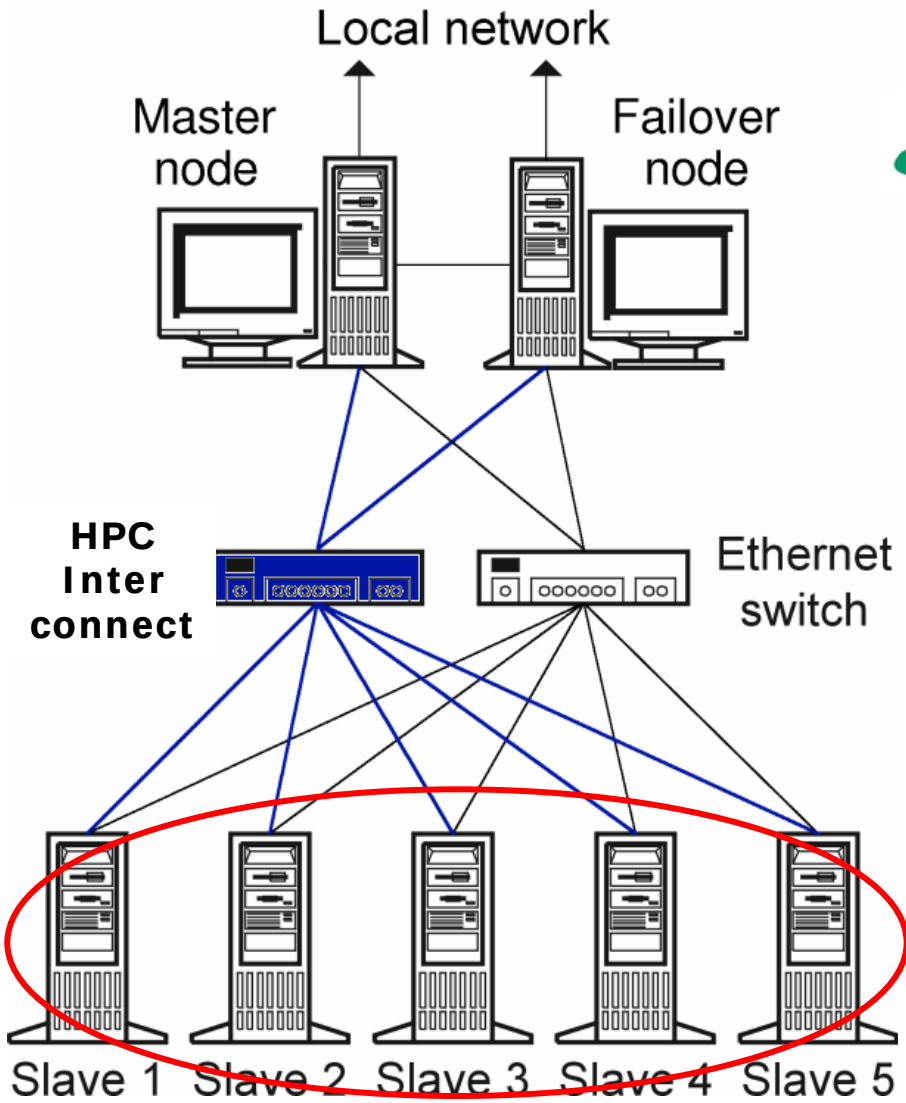
STREAM "Nehalem"

clustervision

Function	Rate (MB/s)	Avg time	Min time	Max time
Copy:	30619.1990	0.0209	0.0209	0.0210
Scale:	30271.4890	0.0212	0.0211	0.0212
Add:	29734.0962	0.0323	0.0323	0.0324
Triad:	29959.7601	0.0321	0.0320	0.0321

Cluster Trends

clustervision



AMD Opteron Shanghai”

Intel Xeon Nehalem

- GPU-FPU Coprocessors
  - Nividia GTX285,C1060,S1070
  - ATI/AMD Firestream
  - Clearspeed

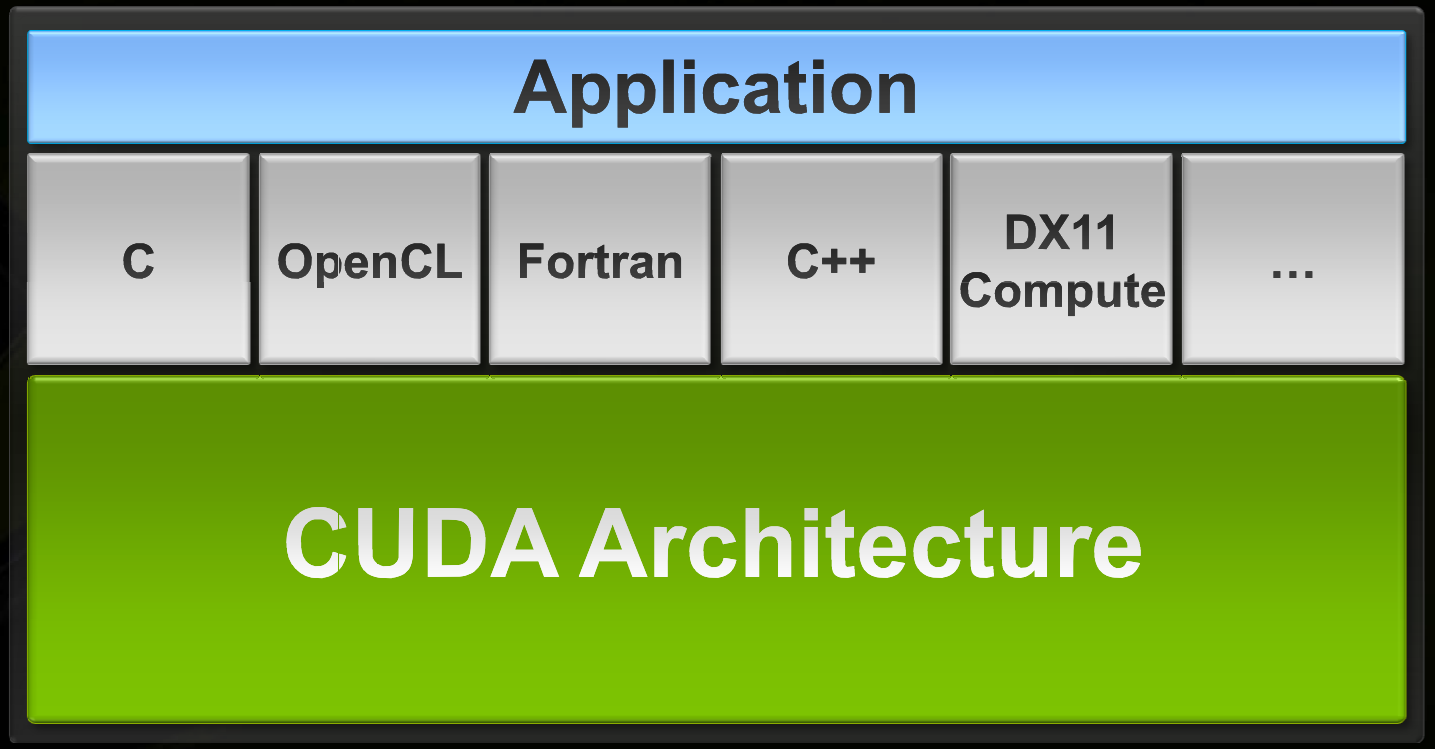
- OpenCL
  - Hardwareunabhängiges API
  - Multivendor
- Powerconsumption

OpenCL

- A new compute API for parallel programming of heterogeneous systems
- Allows developers to harness the compute power of BOTH the GPU and the CPU
- A multi-vendor standards effort managed through the Khronos Group

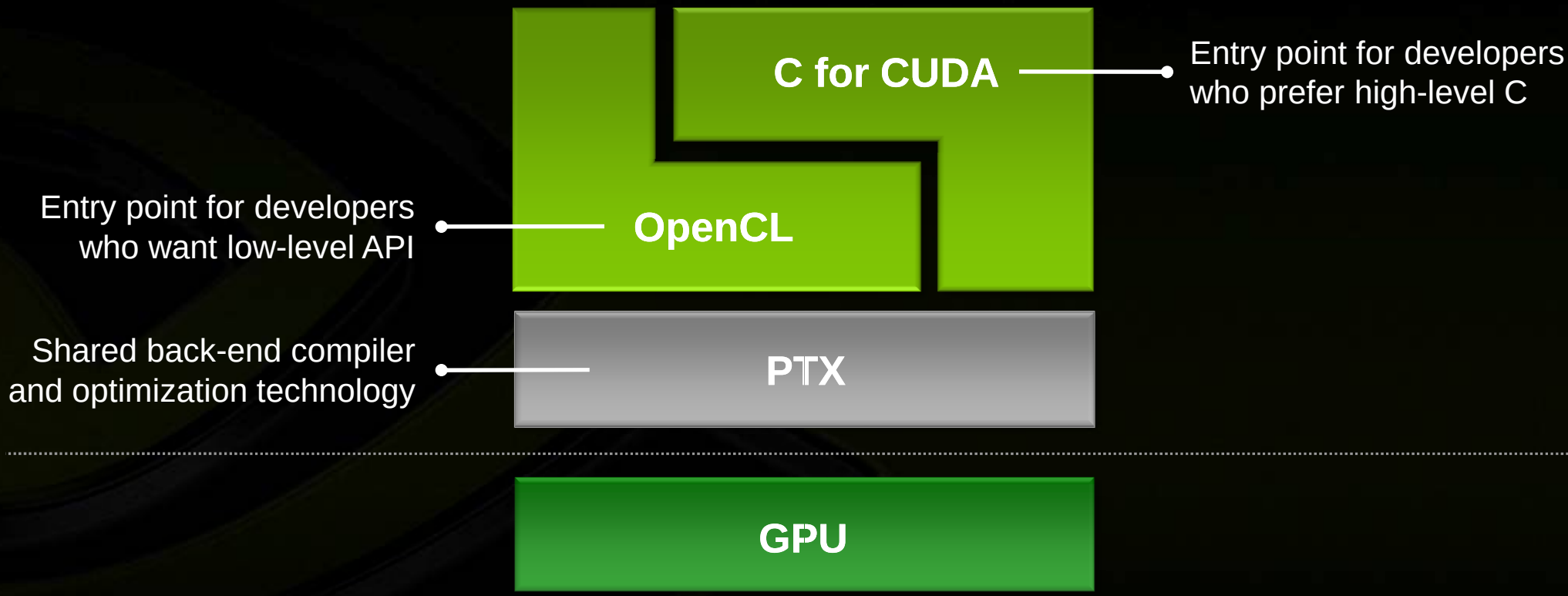
CUDA – NVIDIA's Revolutionary Parallel Computing Architecture

- ISA and hardware compute engine
- Includes a C-compiler & support for OpenCL and DX11 Compute
- Architected to natively support all computational interfaces (standard languages and APIs)



Option 1

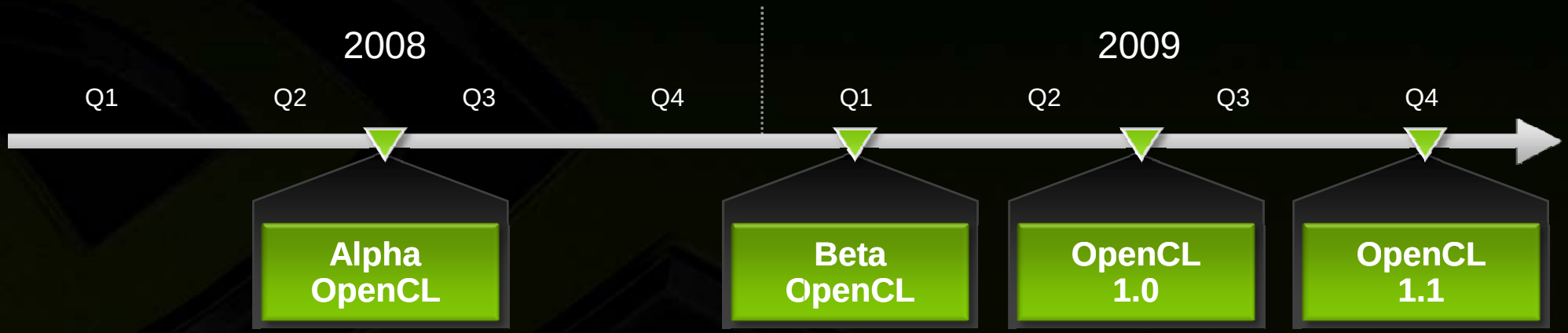
OpenCL and C for CUDA



Different Programming Styles

- C for CUDA
  - C with parallel keywords
  - C runtime that abstracts driver API
  - Memory managed by C runtime
  - Generates PTX
- OpenCL
  - Hardware API - similar to OpenGL
  - Programmer has complete access to hardware device
  - Memory managed by programmer
  - Generates PTX

NVIDIA's OpenCL Roadmap

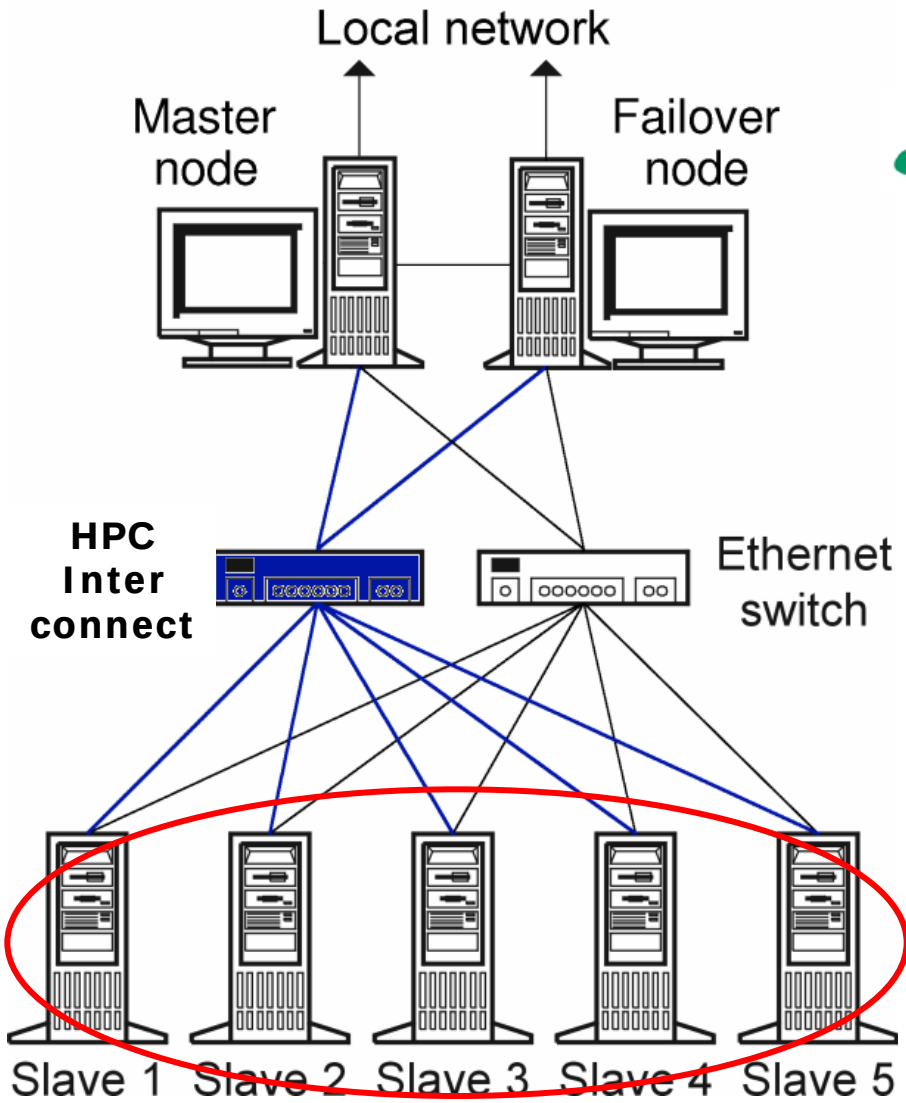


Summary

- OpenCL is fabulous for developers, the industry and consumers
- The addition of OpenCL and DX11 compute to C for CUDA offers a fantastic array of GPU Computing choices for developers
- C for CUDA:
  - Is the only runtime C environment for GPUs today
  - Has 25,000+ users and 100+ applications
  - Will evolve and co-exist with OpenCL and DX11 compute

Cluster Trends

clustervision



AMD Opteron Shanghai”

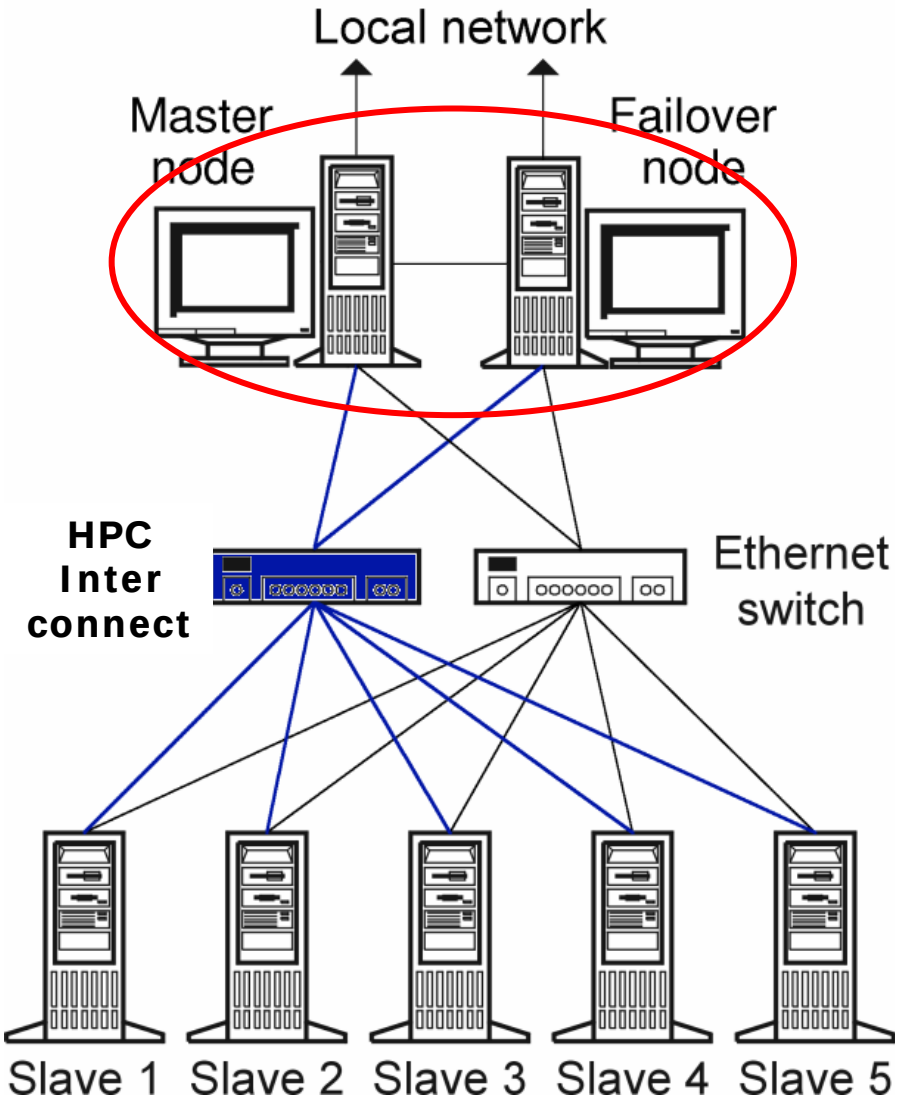
Intel Xeon Nehalem

- GPU-FPU Coprocessors
  - Nividia GTX285,C1060,S1070
  - ATI/AMD Firestream
  - Clearspeed

- OpenCL
  - Hardwareunabhängiges API
  - Multivendor
- Powerconsumption

Management Trends

clustervision



- Linux Base Distribution
- Workload Management System
- Parallel Filesystem
- Development Tool
- Parallel Middleware
- Interconnect Drivers
- Software Libraries
- Cluster Management &
- Monitoring Tools
- Node Provisioning System
- Years of HPC expertise

Architecture

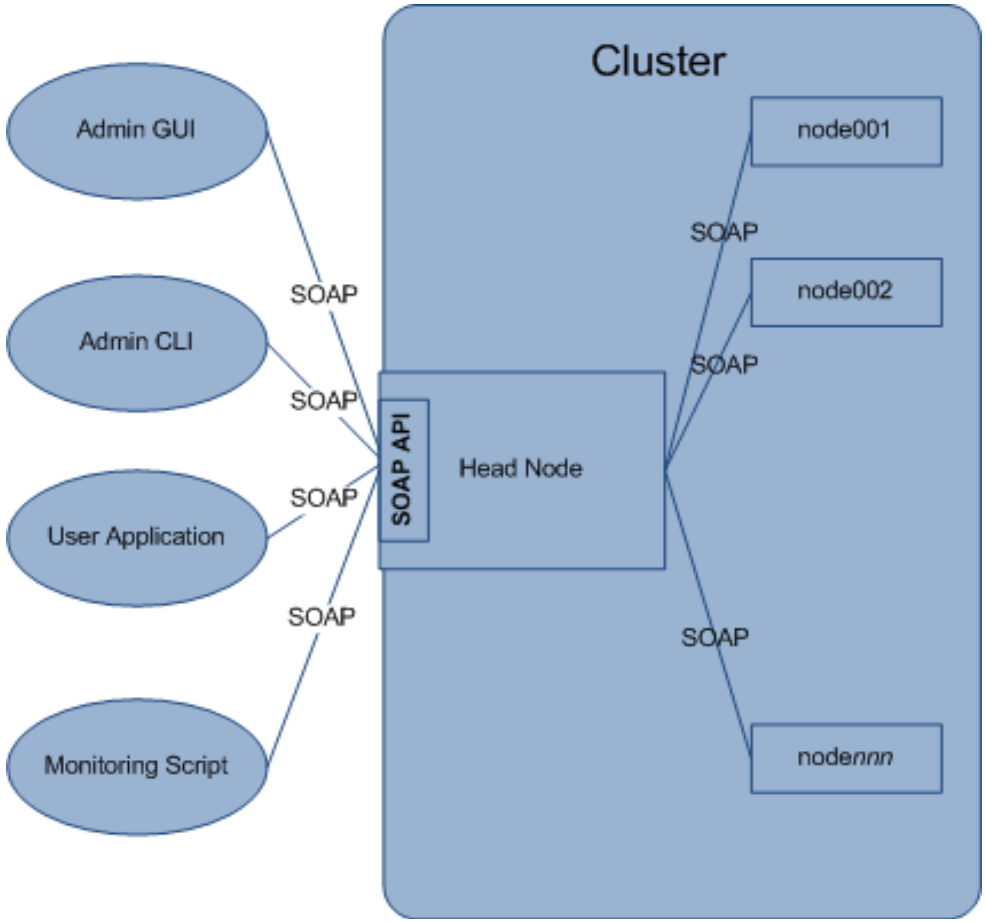
clustervision

Cluster exposes SOAP API to outside world to facilitate integration.

SSL for authentication & encryption

Can be used by:

- Cluster management software
- Desktop HPC applications
- Custom monitoring & management scripts

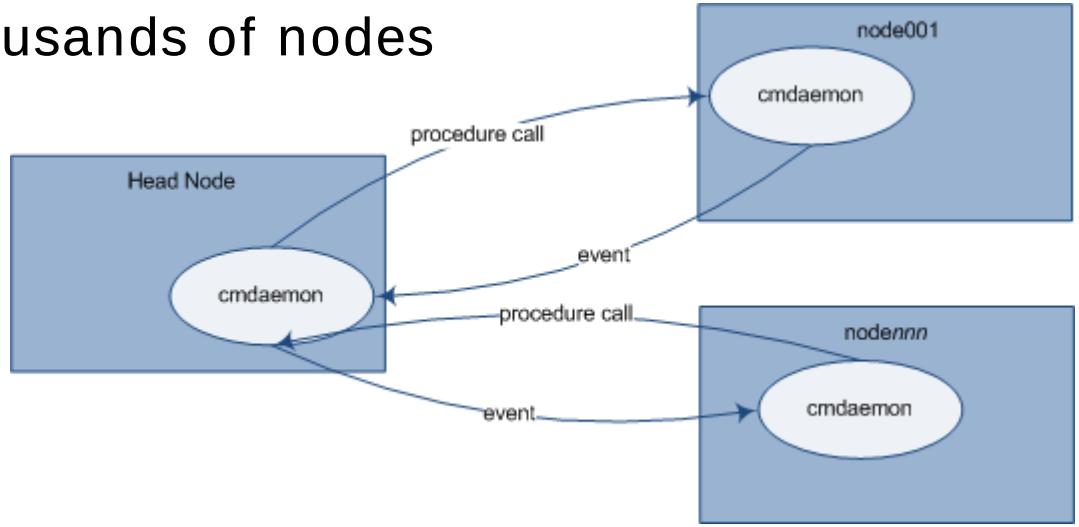


Single Daemon Architecture

clustervision

Cluster Management Daemon (CMDaemon)

- Runs on all nodes in cluster
- Optimized for low resource consumption
- Multithreaded
- Syncronized to the same timestamp
- Designed to scale to thousands of nodes
- Easily extendable to Multi-Headnode Services and Sub clustering
- Pull mechanism
- Designed for HPC needs



Cluster Monitoring

clustervision

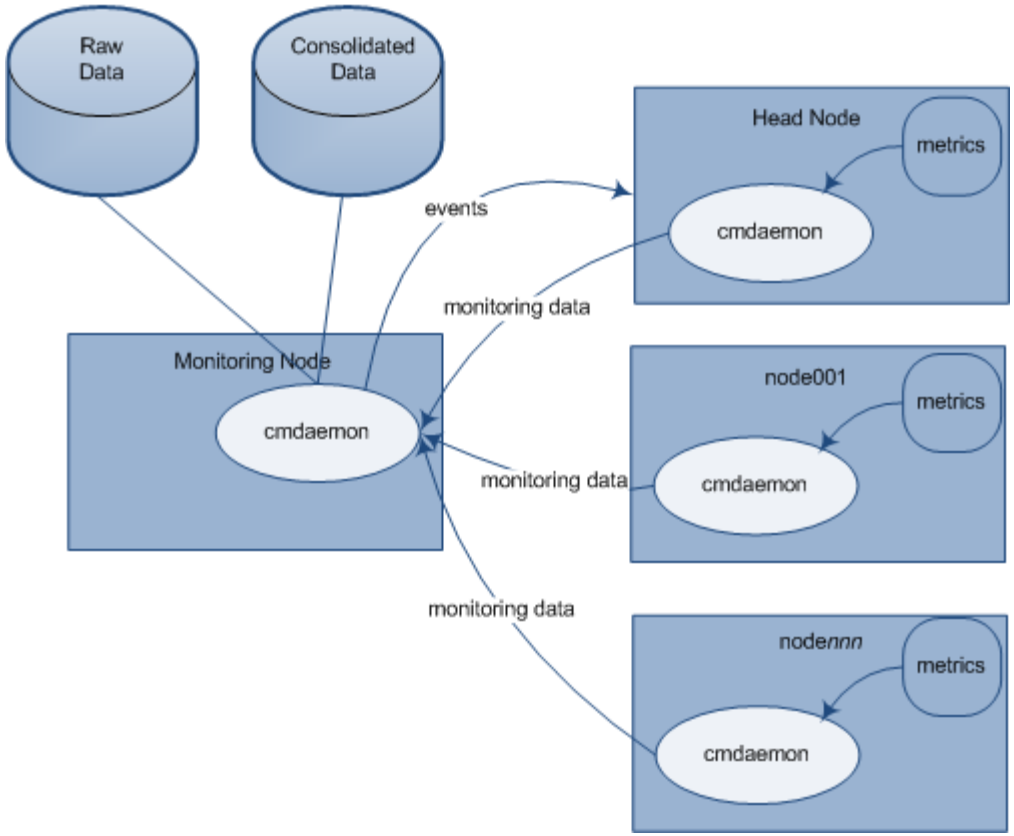
Monitoring data stored in two formats:

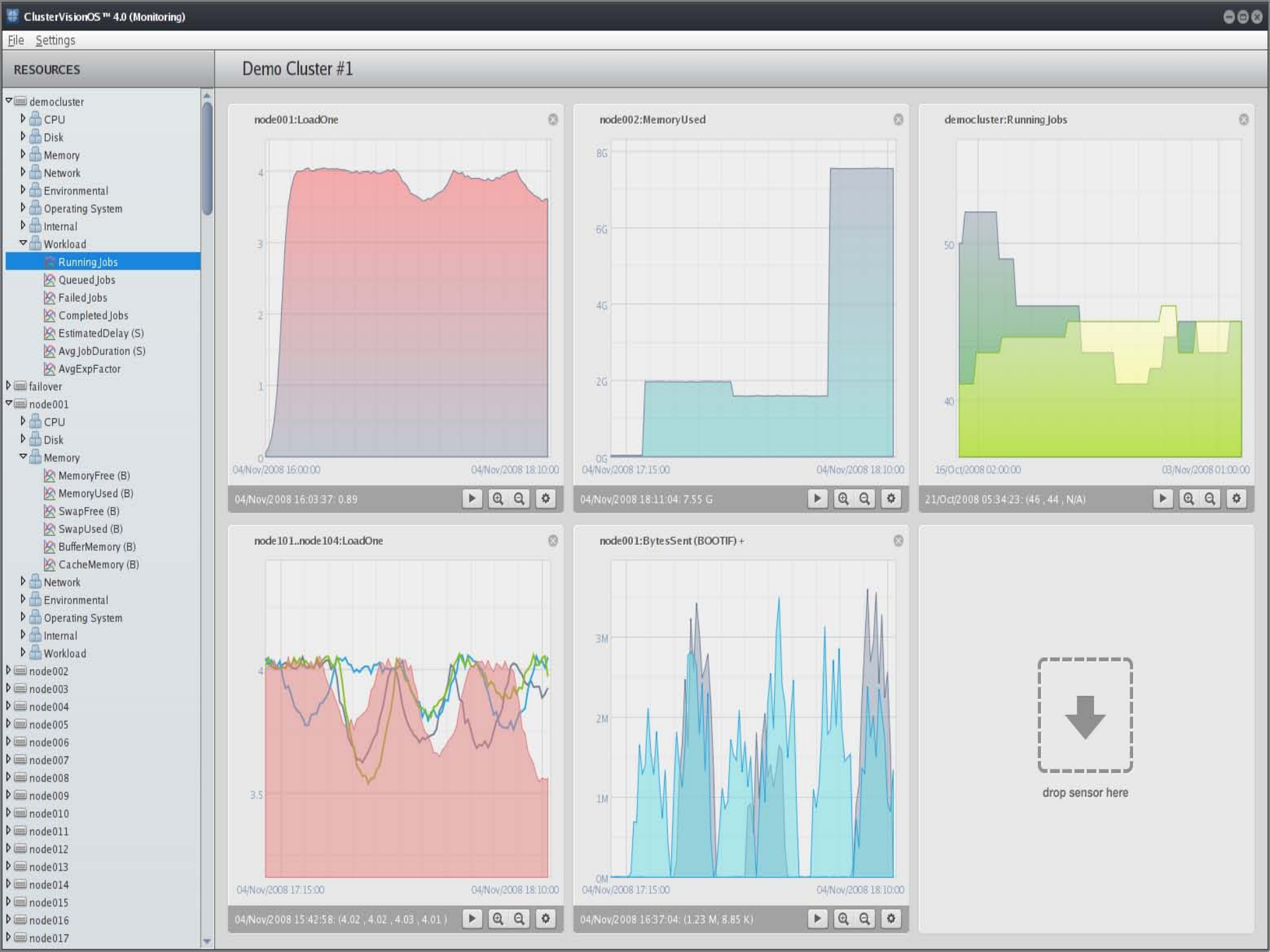
- Raw (limited time)
- Consolidated (configurable period & interval)

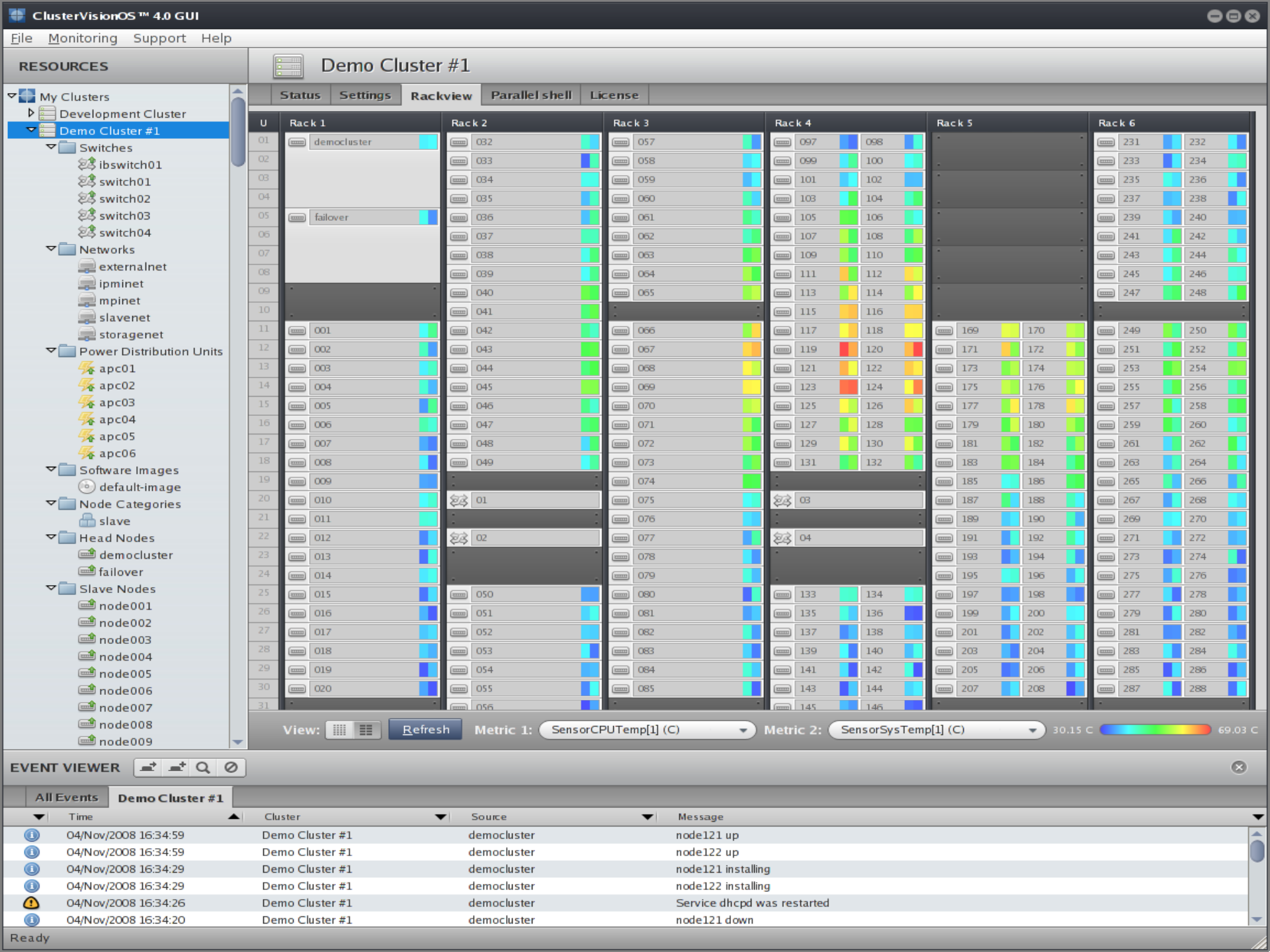
Consolidation example:

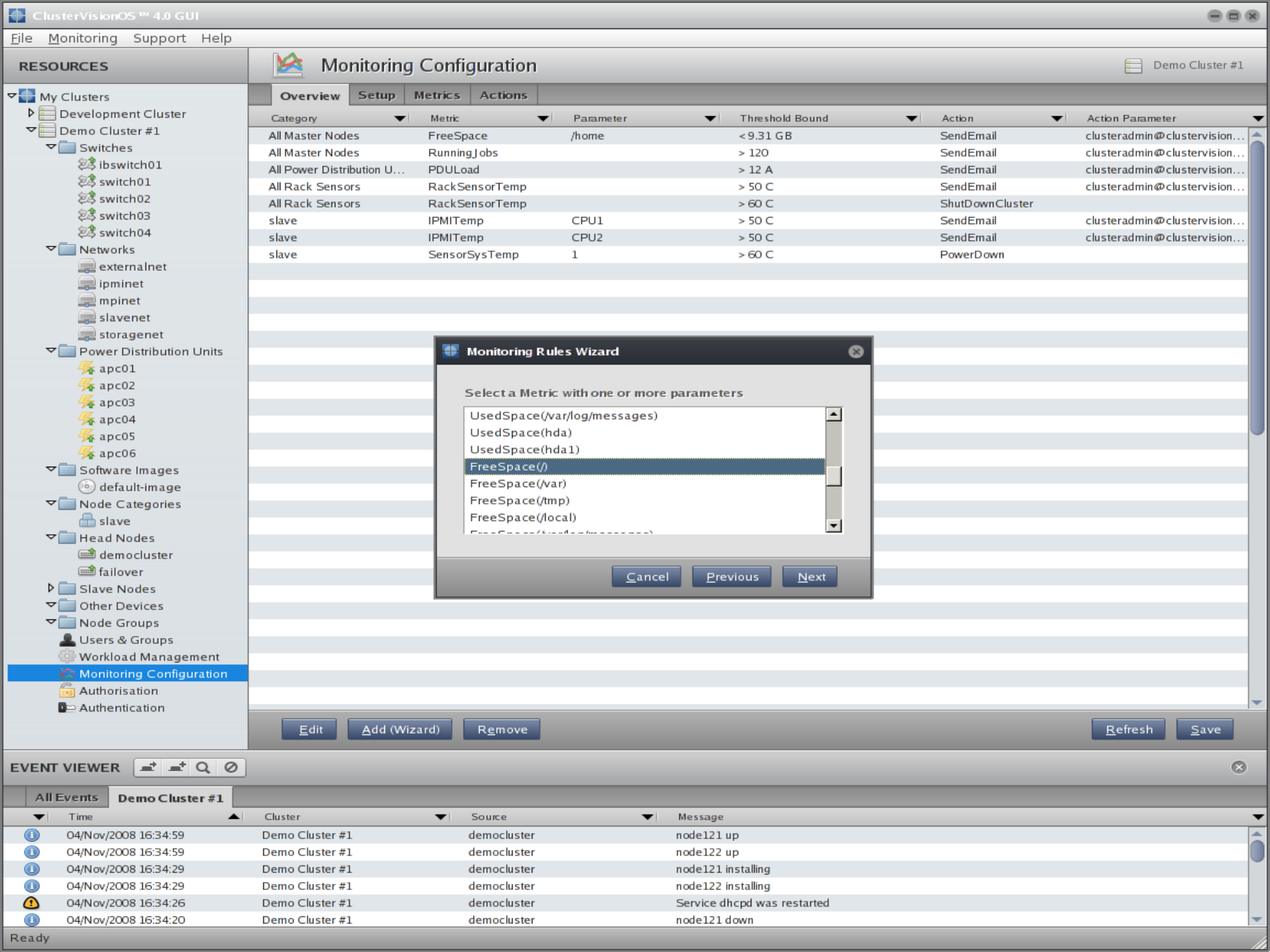
Network bandwidth utilization over last 3 years (period) averaged per 30 minutes (interval).

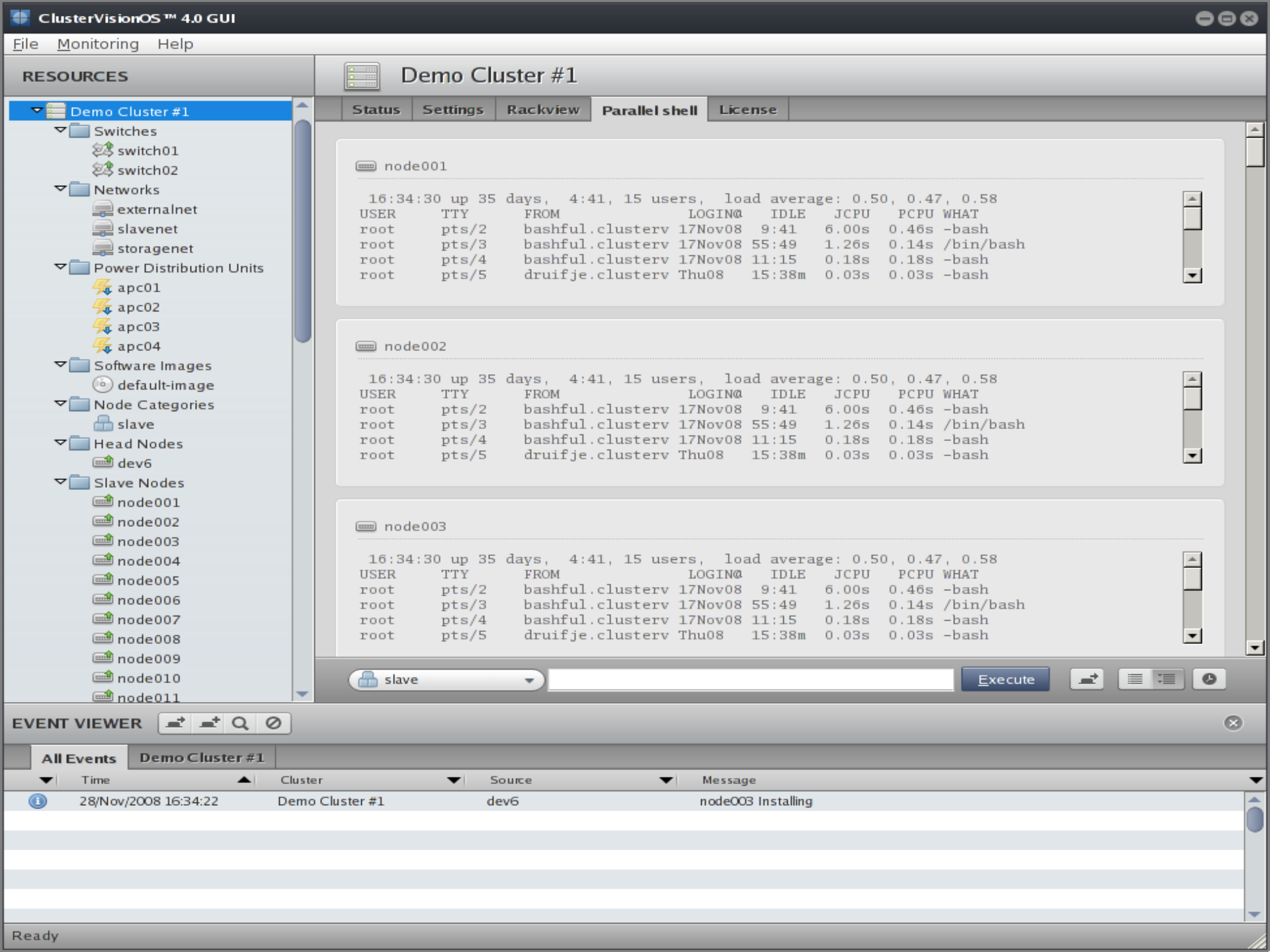
- Fine-grained control
- Configurable thresholds and actions

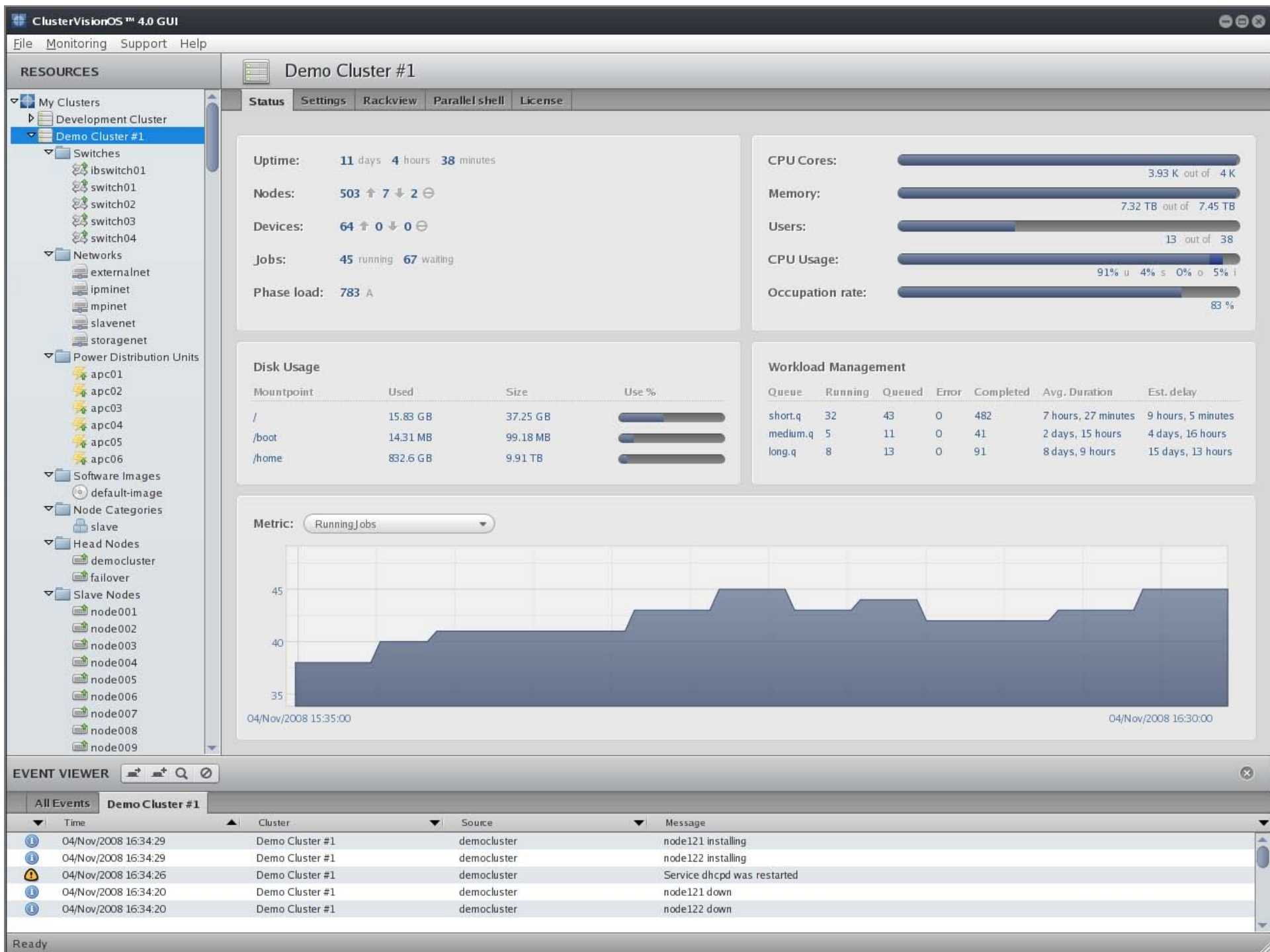



ClusterVisionOS™ 4.0 (Monitoring)
File Settings
RESOURCES
Demo Cluster #1
democluster
CPU
Disk
Memory
Network
Environmental
Operating System
Internal
Workload
Running Jobs
Queued Jobs
Failed Jobs
Completed Jobs
EstimatedDelay (S)
Avg JobDuration (S)
AvgExpFactor
failover
node001
CPU
Disk
Memory
MemoryFree (B)
MemoryUsed (B)
SwapFree (B)
SwapUsed (B)
BufferMemory (B)
CacheMemory (B)
Network
Environmental
Operating System
Internal
Workload
node002
node003
node004
node005
node006
node007
node008
node009
node010
node011
node012
node013
node014
node015
node016
node017
node001:LoadOne
04/Nov/2008 16:00:00
04/Nov/2008 18:10:00
04/Nov/2008 16:03:37: 0.89
node002:MemoryUsed
8G
6G
4G
2G
0G
04/Nov/2008 17:15:00
04/Nov/2008 18:10:00
04/Nov/2008 18:11:04: 7.55 G
democluster:Running Jobs
50
40
16/Oct/2008 02:00:00
03/Nov/2008 01:00:00
21/Oct/2008 05:34:23: (46 , 44 , N/A)
node101..node104:LoadOne
04/Nov/2008 17:15:00
04/Nov/2008 18:10:00
04/Nov/2008 15:42:58: (4.02 , 4.02 , 4.03 , 4.01)
node001:BytesSent (BOOTIF) +
3M
2M
1M
0M
04/Nov/2008 17:15:00
04/Nov/2008 18:10:00
04/Nov/2008 16:37:04: (1.23 M, 8.85 K)
drop sensor here


ClusterVisionOS ™ 4.0 GUI
File Monitoring Support Help
RESOURCES
Demo Cluster #1
My Clusters
Development Cluster
Demo Cluster #1
Switches
ibswitch01
switch01
switch02
switch03
switch04
Networks
externalnet
ipminet
mpinet
slavenet
storagenet
Power Distribution Units
apc01
apc02
apc03
apc04
apc05
apc06
Software Images
default-image
Node Categories
slave
Head Nodes
democluster
failover
Slave Nodes
node001
node002
node003
node004
node005
node006
node007
node008
node009
Status Settings Rackview Parallel shell License
U Rack 1 Rack 2 Rack 3 Rack 4 Rack 5 Rack 6
democluster
failover
View: Refresh Metric 1: SensorCPUTemp[1] (C) Metric 2: SensorSysTemp[1] (C) 30.15 C 69.03 C
EVENT VIEWER
All Events Demo Cluster #1
Time Cluster Source Message
04/Nov/2008 16:34:59 Demo Cluster #1 democluster node121 up
04/Nov/2008 16:34:59 Demo Cluster #1 democluster node122 up
04/Nov/2008 16:34:29 Demo Cluster #1 democluster node121 installing
04/Nov/2008 16:34:29 Demo Cluster #1 democluster node122 installing
04/Nov/2008 16:34:26 Demo Cluster #1 democluster Service dhcpd was restarted
04/Nov/2008 16:34:20 Demo Cluster #1 democluster node121 down
Ready


ClusterVisionOS ™ 4.0 GUI
File Monitoring Support Help
RESOURCES
Monitoring Configuration
Demo Cluster #1
My Clusters
Development Cluster
Demo Cluster #1
Switches
ibswitch01
switch01
switch02
switch03
switch04
Networks
externalnet
ipminet
mpinet
slavenet
storagenet
Power Distribution Units
apc01
apc02
apc03
apc04
apc05
apc06
Software Images
default-image
Node Categories
slave
Head Nodes
democluster
failover
Slave Nodes
Other Devices
Node Groups
Users & Groups
Workload Management
Monitoring Configuration
Authorisation
Authentication
Overview
Setup
Metrics
Actions
Category
Metric
Parameter
Threshold Bound
Action
Action Parameter
All Master Nodes FreeSpace /home < 9.31 GB SendEmail clusteradmin@clustervision...
All Master Nodes RunningJobs > 120 SendEmail clusteradmin@clustervision...
All Power Distribution U... PDULoad > 12 A SendEmail clusteradmin@clustervision...
All Rack Sensors RackSensorTemp > 50 C SendEmail clusteradmin@clustervision...
All Rack Sensors RackSensorTemp > 60 C ShutDownCluster
slave IPMITemp CPU1 > 50 C SendEmail clusteradmin@clustervision...
slave IPMITemp CPU2 > 50 C SendEmail clusteradmin@clustervision...
slave SensorSysTemp 1 > 60 C PowerDown
Monitoring Rules Wizard
Select a Metric with one or more parameters
UsedSpace(/var/log/messages)
UsedSpace(hda)
UsedSpace(hda1)
FreeSpace(/)
FreeSpace(/var)
FreeSpace(/tmp)
FreeSpace(/local)
Cancel
Previous
Next
Edit
Add (Wizard)
Remove
Refresh
Save
EVENT VIEWER
All Events
Demo Cluster #1
Time
Cluster
Source
Message
04/Nov/2008 16:34:59 Demo Cluster #1 democluster node121 up
04/Nov/2008 16:34:59 Demo Cluster #1 democluster node122 up
04/Nov/2008 16:34:29 Demo Cluster #1 democluster node121 installing
04/Nov/2008 16:34:29 Demo Cluster #1 democluster node122 installing
04/Nov/2008 16:34:26 Demo Cluster #1 democluster Service dhcpd was restarted
04/Nov/2008 16:34:20 Demo Cluster #1 democluster node121 down
Ready


ClusterVisionOS ™ 4.0 GUI
File Monitoring Help
RESOURCES
Demo Cluster #1
Switches
switch01
switch02
Networks
externalnet
slavenet
storagenet
Power Distribution Units
apc01
apc02
apc03
apc04
Software Images
default-image
Node Categories
slave
Head Nodes
dev6
Slave Nodes
node001
node002
node003
node004
node005
node006
node007
node008
node009
node010
node011
Demo Cluster #1
Status Settings Rackview Parallel shell License
node001
16:34:30 up 35 days, 4:41, 15 users, load average: 0.50, 0.47, 0.58
USER TTY FROM LOGIN@ IDLE JCPU PCPU WHAT
root pts/2 bashful.clusterv 17Nov08 9:41 6.00s 0.46s -bash
root pts/3 bashful.clusterv 17Nov08 55:49 1.26s 0.14s /bin/bash
root pts/4 bashful.clusterv 17Nov08 11:15 0.18s 0.18s -bash
root pts/5 druifje.clusterv Thu08 15:38m 0.03s 0.03s -bash
node002
16:34:30 up 35 days, 4:41, 15 users, load average: 0.50, 0.47, 0.58
USER TTY FROM LOGIN@ IDLE JCPU PCPU WHAT
root pts/2 bashful.clusterv 17Nov08 9:41 6.00s 0.46s -bash
root pts/3 bashful.clusterv 17Nov08 55:49 1.26s 0.14s /bin/bash
root pts/4 bashful.clusterv 17Nov08 11:15 0.18s 0.18s -bash
root pts/5 druifje.clusterv Thu08 15:38m 0.03s 0.03s -bash
node003
16:34:30 up 35 days, 4:41, 15 users, load average: 0.50, 0.47, 0.58
USER TTY FROM LOGIN@ IDLE JCPU PCPU WHAT
root pts/2 bashful.clusterv 17Nov08 9:41 6.00s 0.46s -bash
root pts/3 bashful.clusterv 17Nov08 55:49 1.26s 0.14s /bin/bash
root pts/4 bashful.clusterv 17Nov08 11:15 0.18s 0.18s -bash
root pts/5 druifje.clusterv Thu08 15:38m 0.03s 0.03s -bash
slave
Execute
EVENT VIEWER
All Events
Demo Cluster #1
Time
Cluster
Source
Message
28/Nov/2008 16:34:22
Demo Cluster #1
dev6
node003 Installing
Ready


ClusterVisionOS™ 4.0 GUI
File Monitoring Support Help
RESOURCES
Demo Cluster #1
Status Settings Rackview Parallel shell License
My Clusters
Development Cluster
Demo Cluster #1
Switches
ibswitch01
switch01
switch02
switch03
switch04
Networks
externalnet
ipminet
mpinet
slavenet
storagenet
Power Distribution Units
apc01
apc02
apc03
apc04
apc05
apc06
Software Images
default-image
Node Categories
slave
Head Nodes
democluster
failover
Slave Nodes
node001
node002
node003
node004
node005
node006
node007
node008
node009
Uptime: 11 days 4 hours 38 minutes
Nodes: 503 7 2
Devices: 64 0 0
Jobs: 45 running 67 waiting
Phase load: 783 A
CPU Cores: 3.93 K out of 4 K
Memory: 7.32 TB out of 7.45 TB
Users: 13 out of 38
CPU Usage: 91% u 4% s 0% o 5% i
Occupation rate: 83 %
Disk Usage
Mountpoint Used Size Use %
/ 15.83 GB 37.25 GB
/boot 14.31 MB 99.18 MB
/home 832.6 GB 9.91 TB
Workload Management
Queue Running Queued Error Completed Avg. Duration Est. delay
short.q 32 43 0 482 7 hours, 27 minutes 9 hours, 5 minutes
medium.q 5 11 0 41 2 days, 15 hours 4 days, 16 hours
long.q 8 13 0 91 8 days, 9 hours 15 days, 13 hours
Metric: Running Jobs
45
40
35
04/Nov/2008 15:35:00
04/Nov/2008 16:30:00
EVENT VIEWER
All Events
Demo Cluster #1
Time Cluster Source Message
04/Nov/2008 16:34:29 Demo Cluster #1 democluster node121 installing
04/Nov/2008 16:34:29 Demo Cluster #1 democluster node122 installing
04/Nov/2008 16:34:26 Demo Cluster #1 democluster Service dhcpd was restarted
04/Nov/2008 16:34:20 Demo Cluster #1 democluster node121 down
04/Nov/2008 16:34:20 Demo Cluster #1 democluster node122 down
Ready

Conclusion

clustervision

- Proven track-record in cluster computing
- Own cluster IP software stack
- 100% committed to cluster computing
- Cluster turn key/taylor made upon customer requirements
- Best in price, quality and time

clustervision.com

Thanks!

ClusterVision Deutschland
Feringastraße 6
85774 München
Germany
Tel: +49 899 921 6433
Fax: +49 899 218 5311
infode@clustervision.com

Stolze ClusterVision Kunden

clustervision

NUMECA