Introduction to Commodity Linux clusters (part 2)

With examples from ICTP HPC cluster "Argo"

Maria Verina

Abdus Salam International Center for Theoretical Physics (ICTP)
Trieste, Italy

The Information and Communication Technology Section (**ICTS**)

Member of HPC team

Credits:

- Partially based on:

"Installation and configuration of Linux Cluster"

(smr2613, ICTP 2014)

Addisu Gezahegn
University of Trieste
ICTP,Trieste
asemie@ictp.it

- With additional elements from:

Clement Onime
Antonio Messina
Moreno Baricevic

From Function to Structure

- Introduction to Commodity Linux clusters, part 1

The problem/Function:

"make parallel/HPC apps run well for your users"

- Introduction to Commodity Linux clusters, part 2

Cluster Structure:

… fairly complex …

- But it CAN be done!

Why Complex?

- Number of components and technologies
- New to IT team: think big
- Commodity parts, no Blue Print:
  - Each cluster is unique
- Life cycle:
  - Build/operate daily
  - Initial/extensions (heterogeneous?)

- Parts “fit” together
- Best practices exist
- Standards/protocols exist
- Need broad range of expertise
- Team effort

What is Cluster computer?

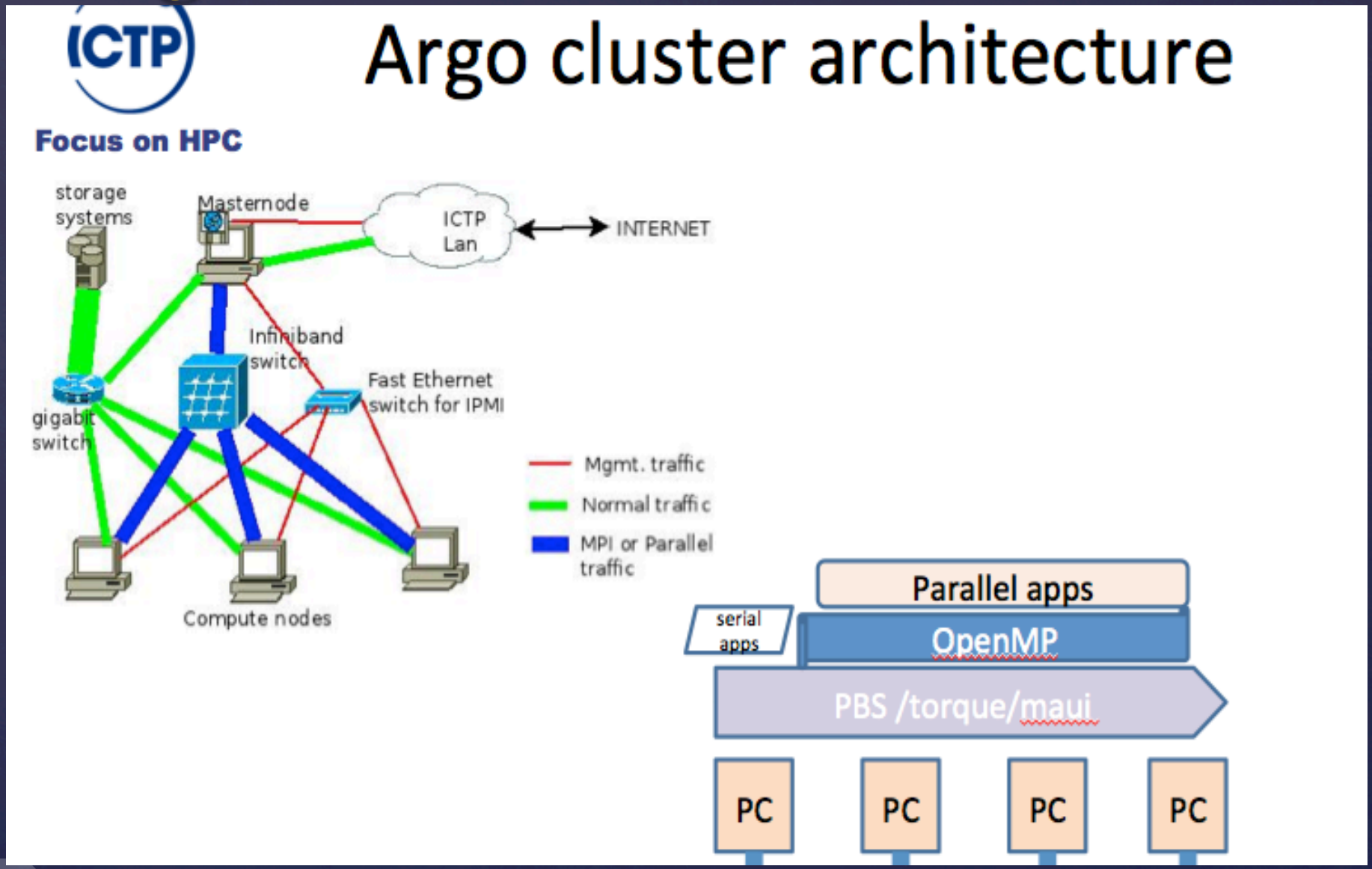
- It is a single logical unit consisting of multiple computers that are linked through a network

ICTP HPC cluster experience

- ICTP HPC cluster "Argo"
- 198 nodes
  - 2640 cores
  - Intel x86_64 +GPU
  - heterogeneous
- Total RAM: 6.8TB
- IB: 40 Gbps QDR
- Storage: ~300TB NFS
- racks: 5
- Users: ~200

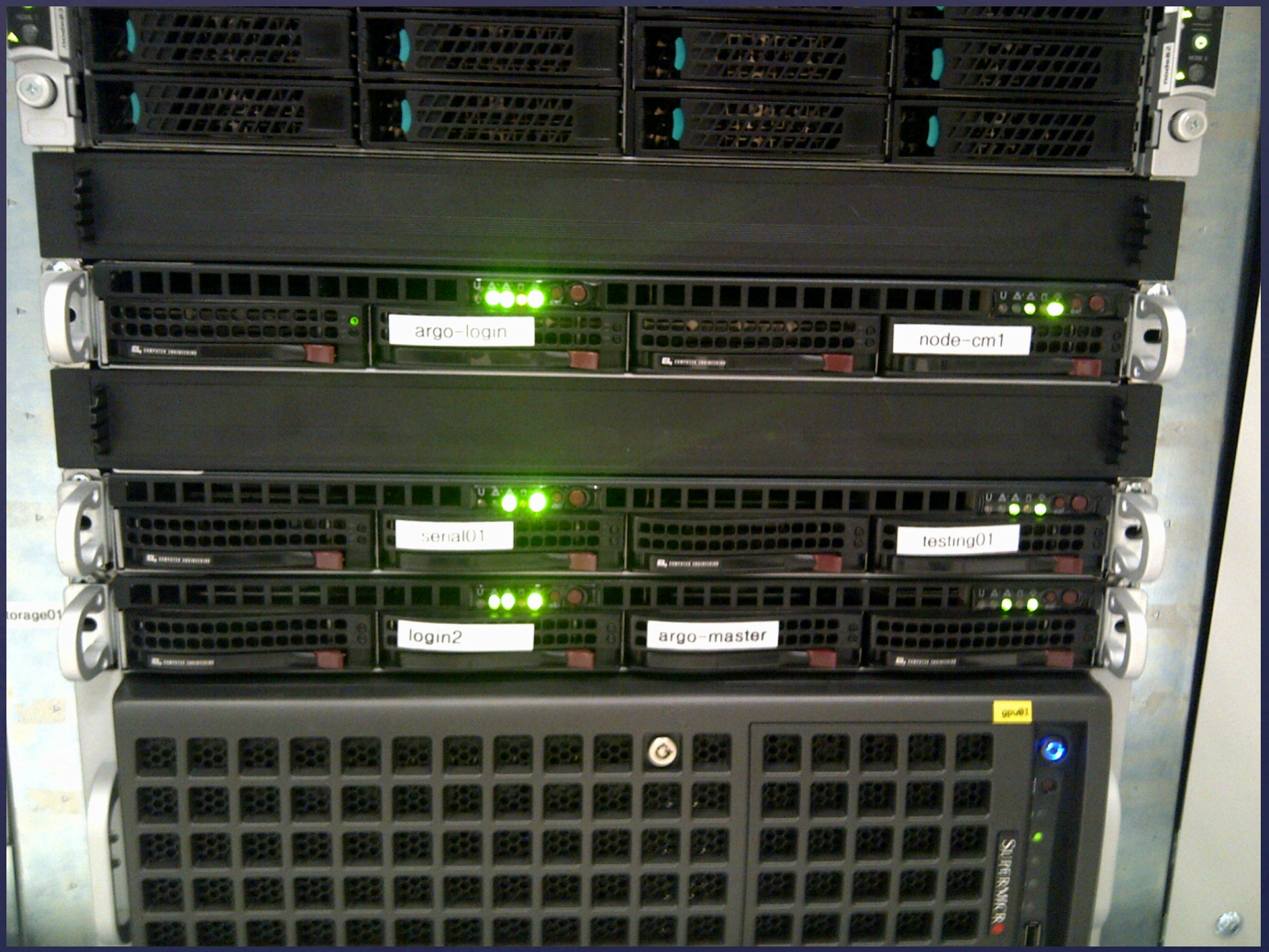
The story of growth.
Queues: dedicated + common
Storage: dedicated + common

Argo architecture



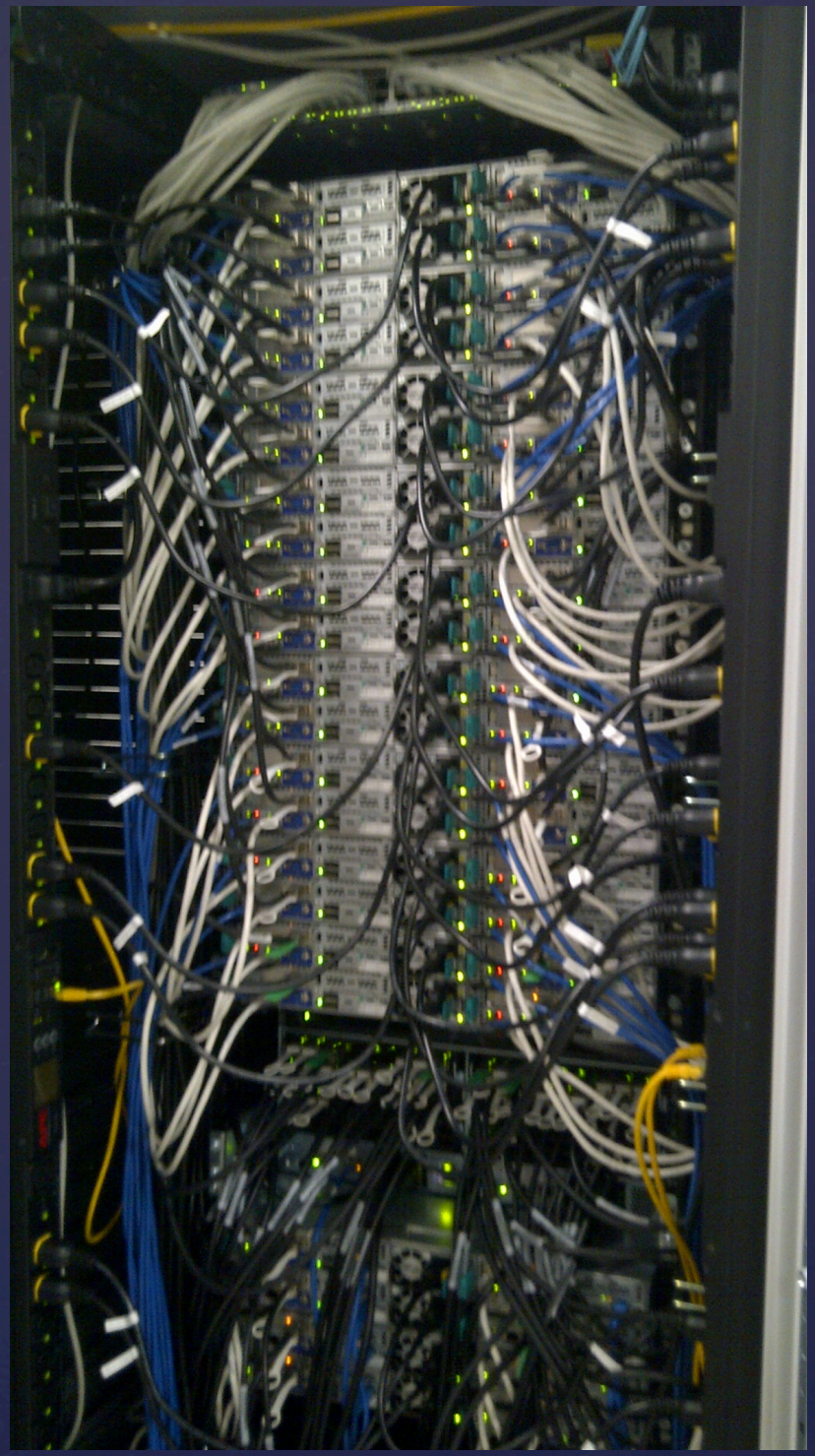
Taken from Clement's slide

ICTP HPC Cluster Argo in pictures


argo-login
node-cm1
serial01
testing01
login2
argo-master

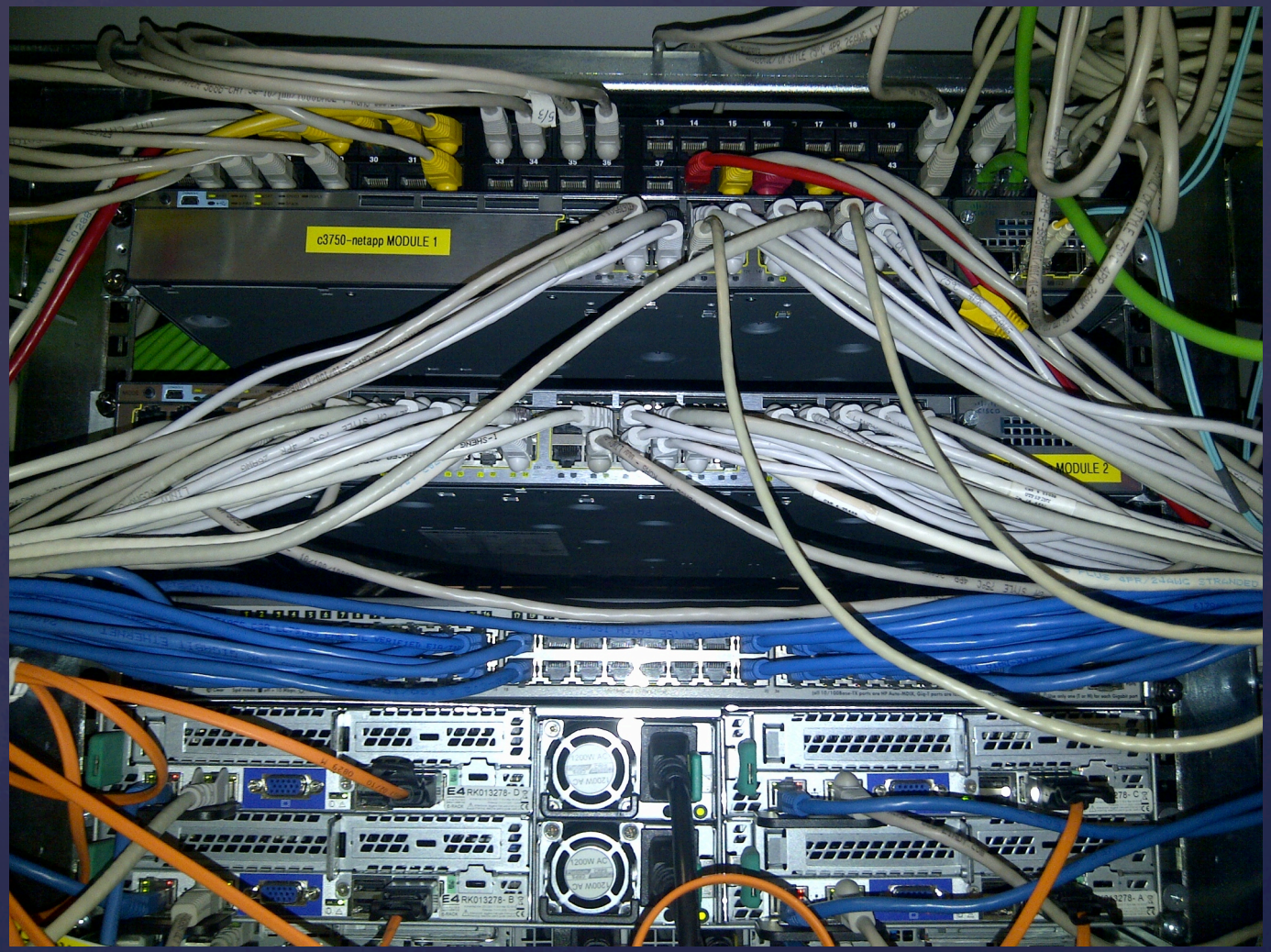

node30
QLOGIC

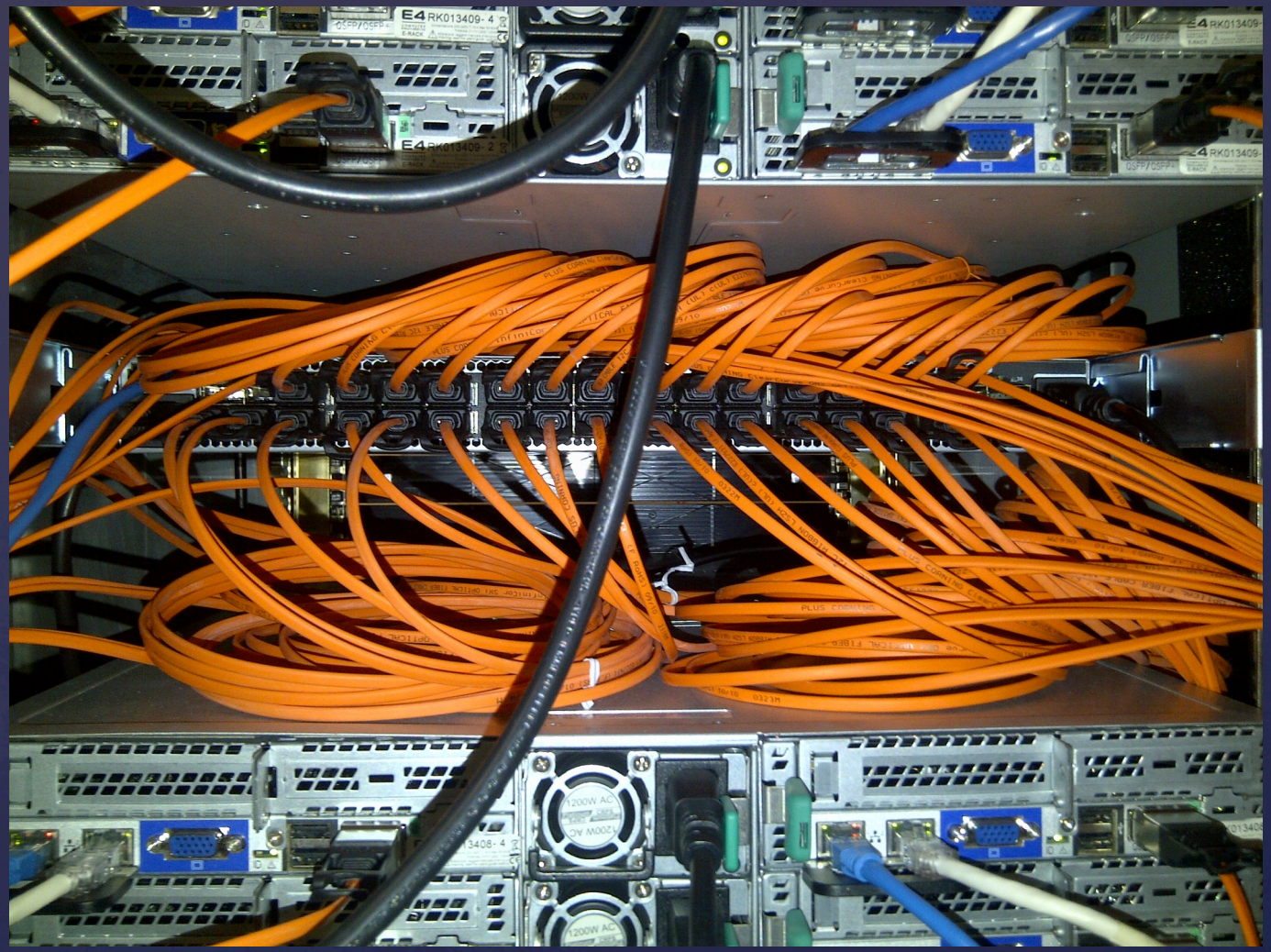

node01
node03




Warning!


RK012565

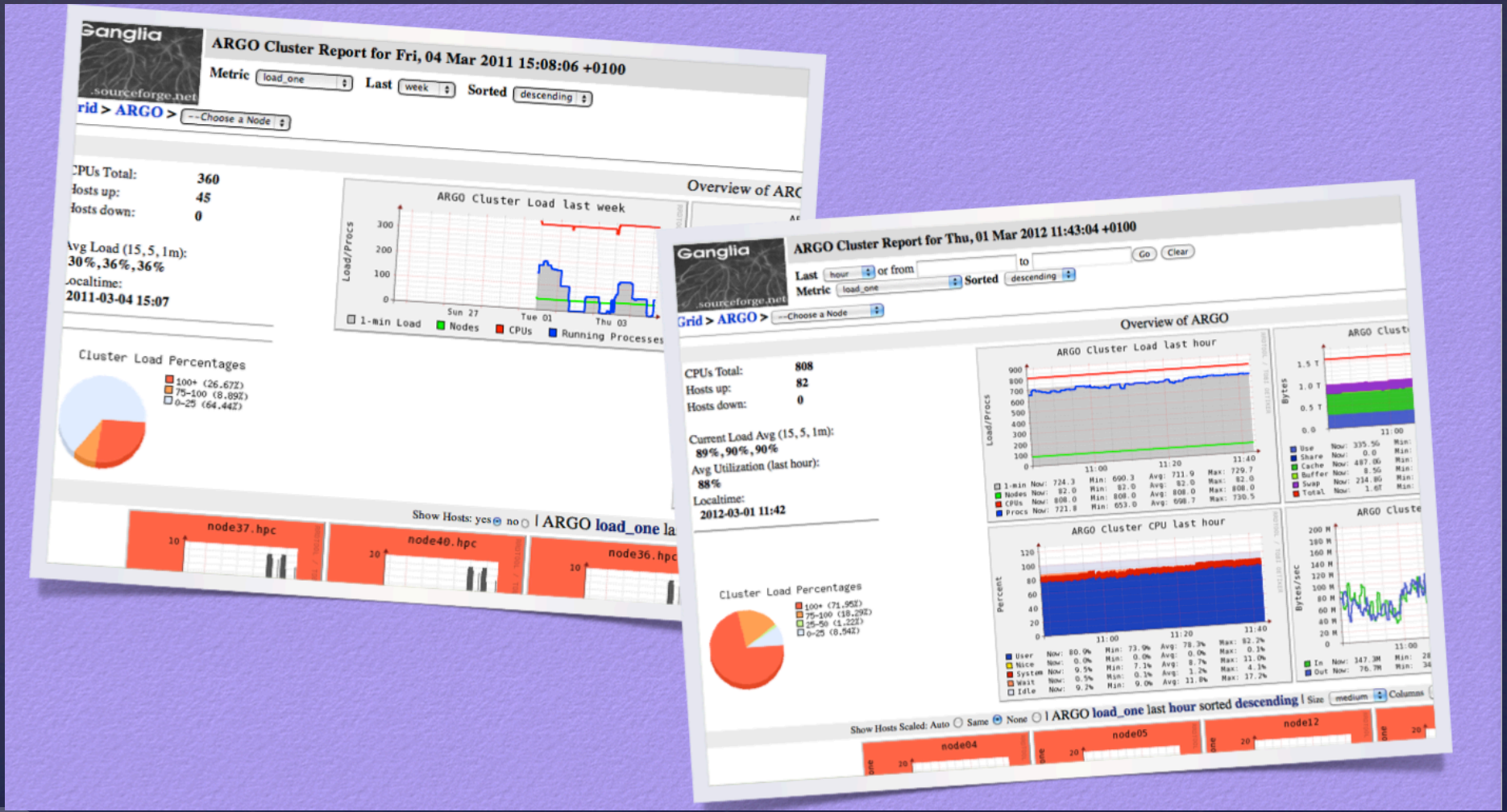

c3750-netapp MODULE 1
MODULE 2

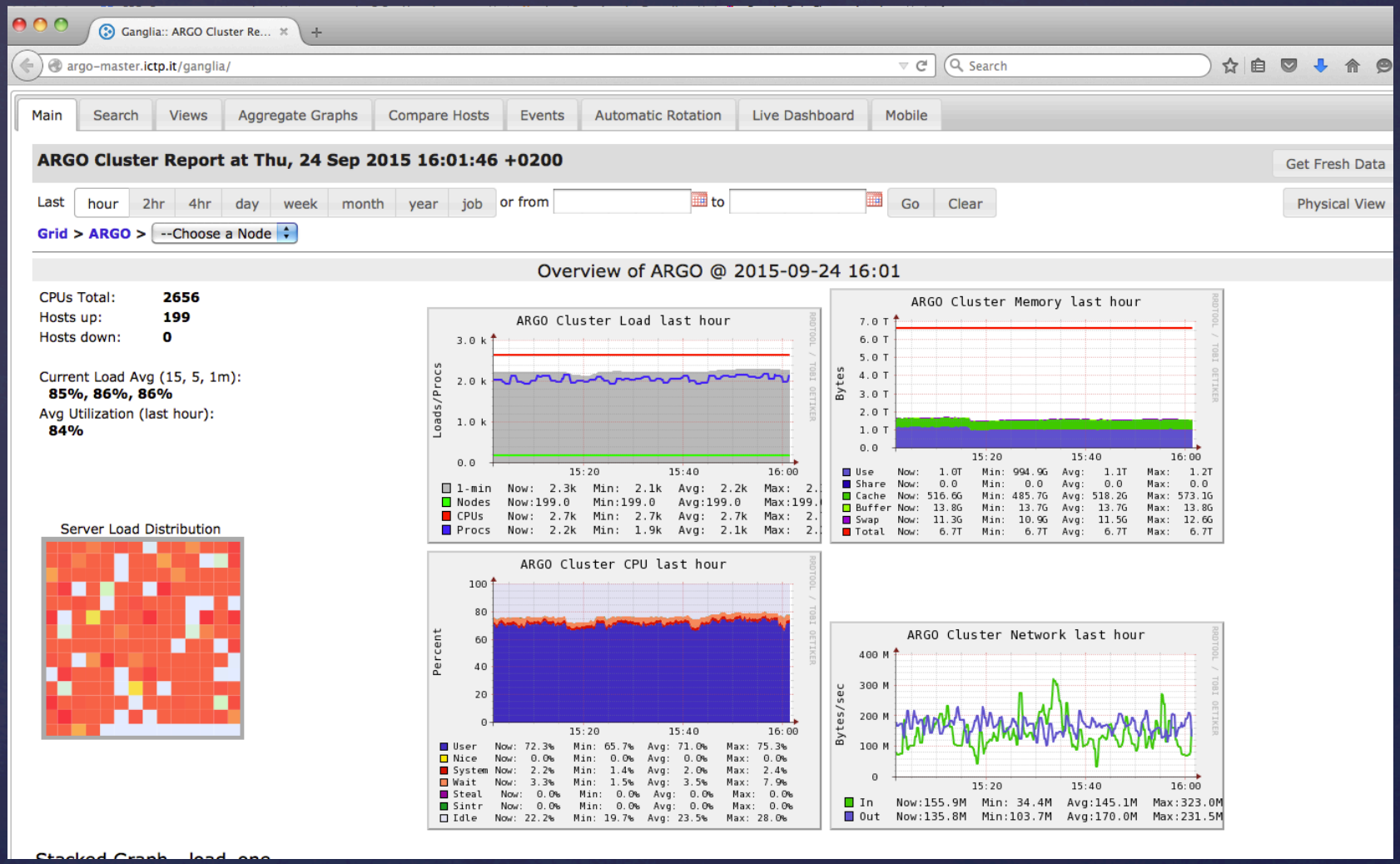


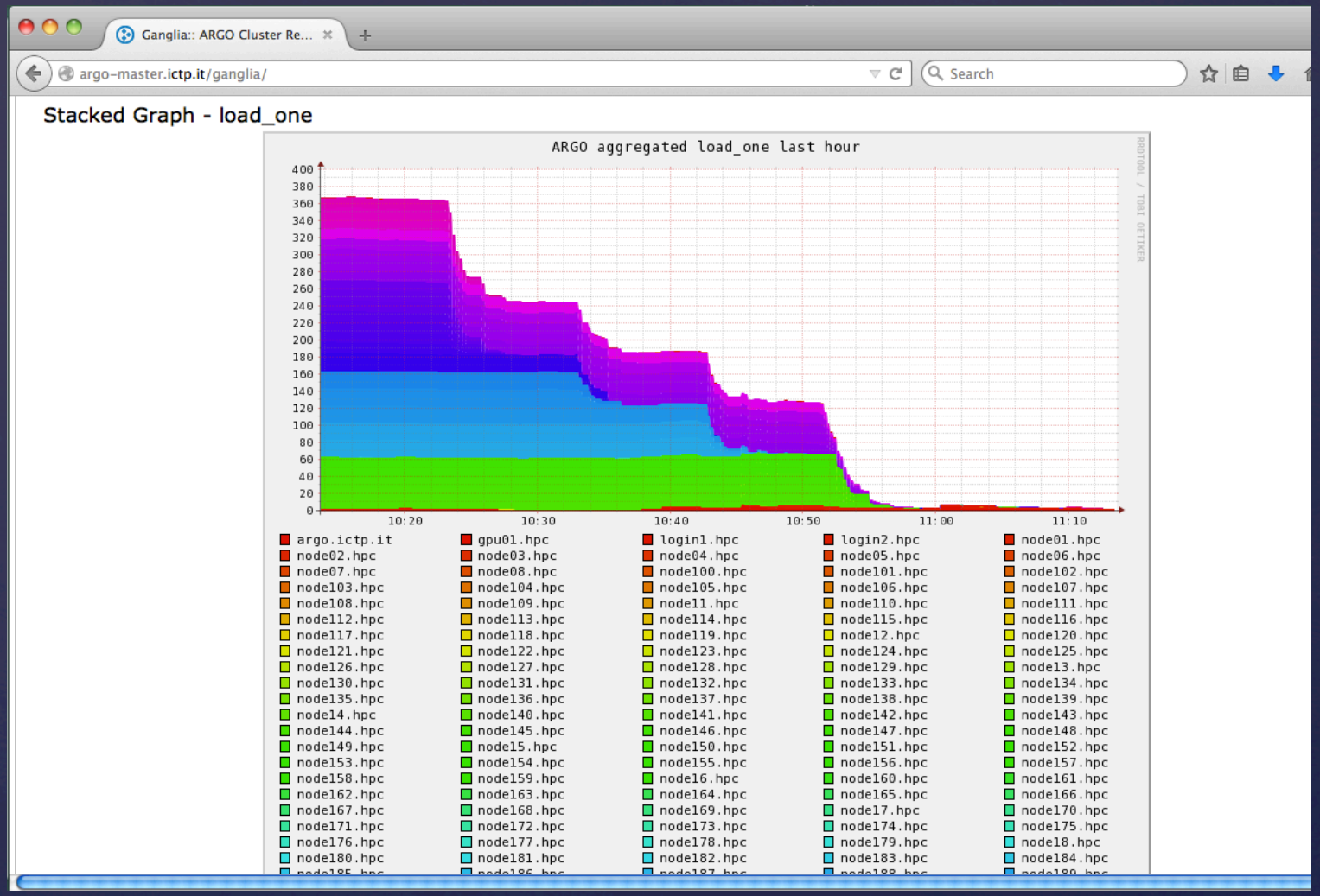

395
068
401
Phase 1
Phase 2
Phase 3

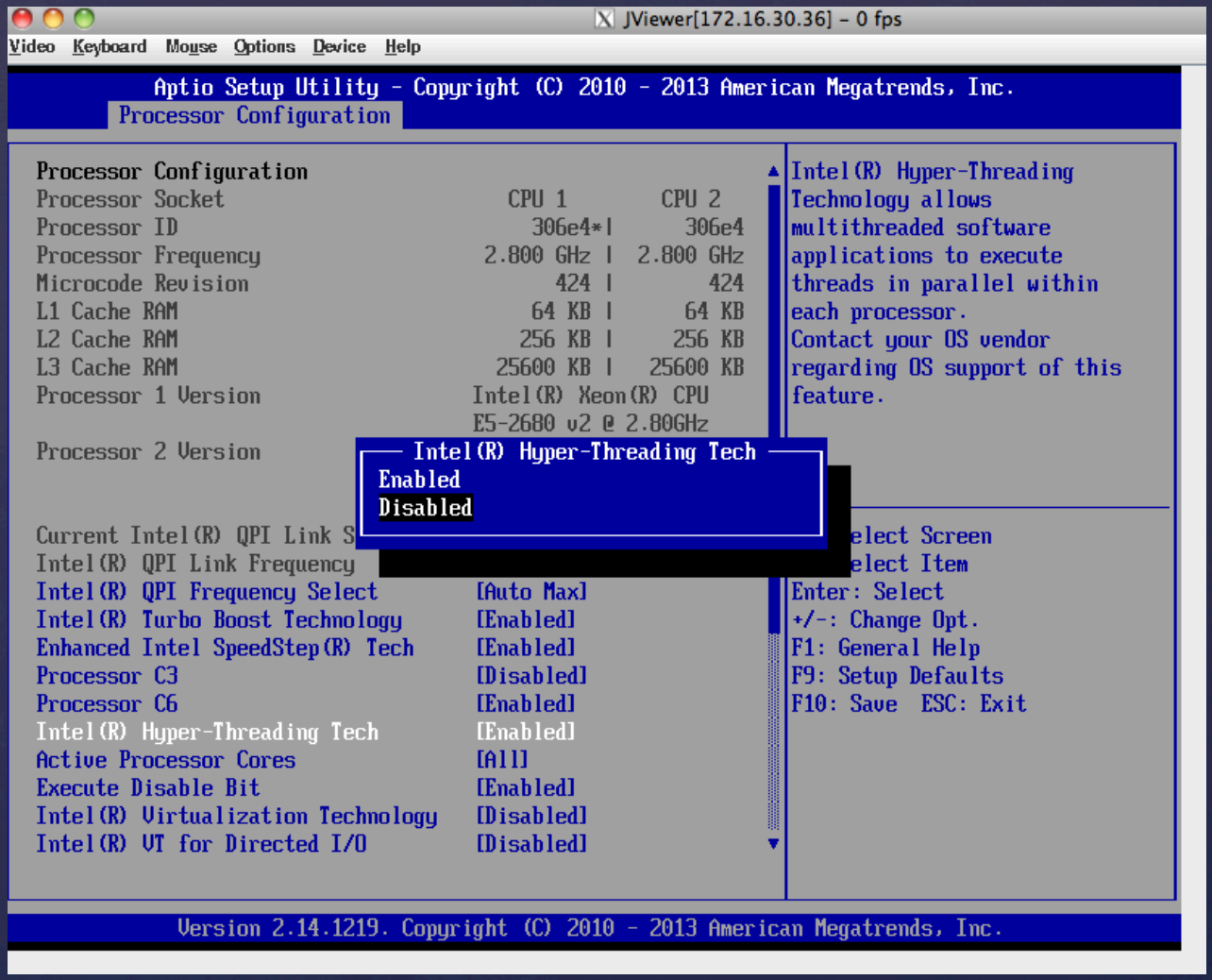

RITTAL
System OK
Home
22°C
Temperature
Server-In
15.6 kW
Cooling capacity

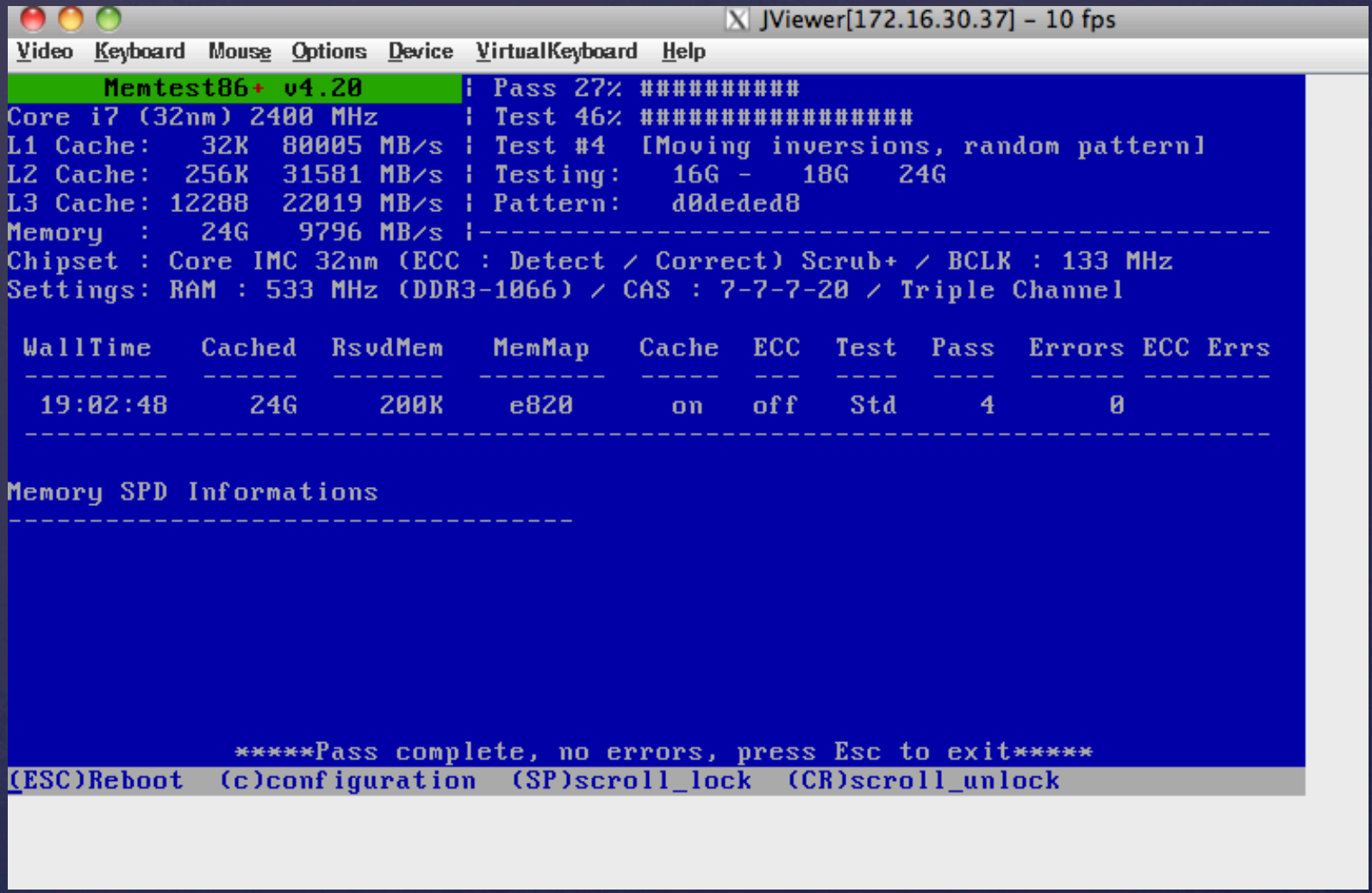
snapshots from Argo's life

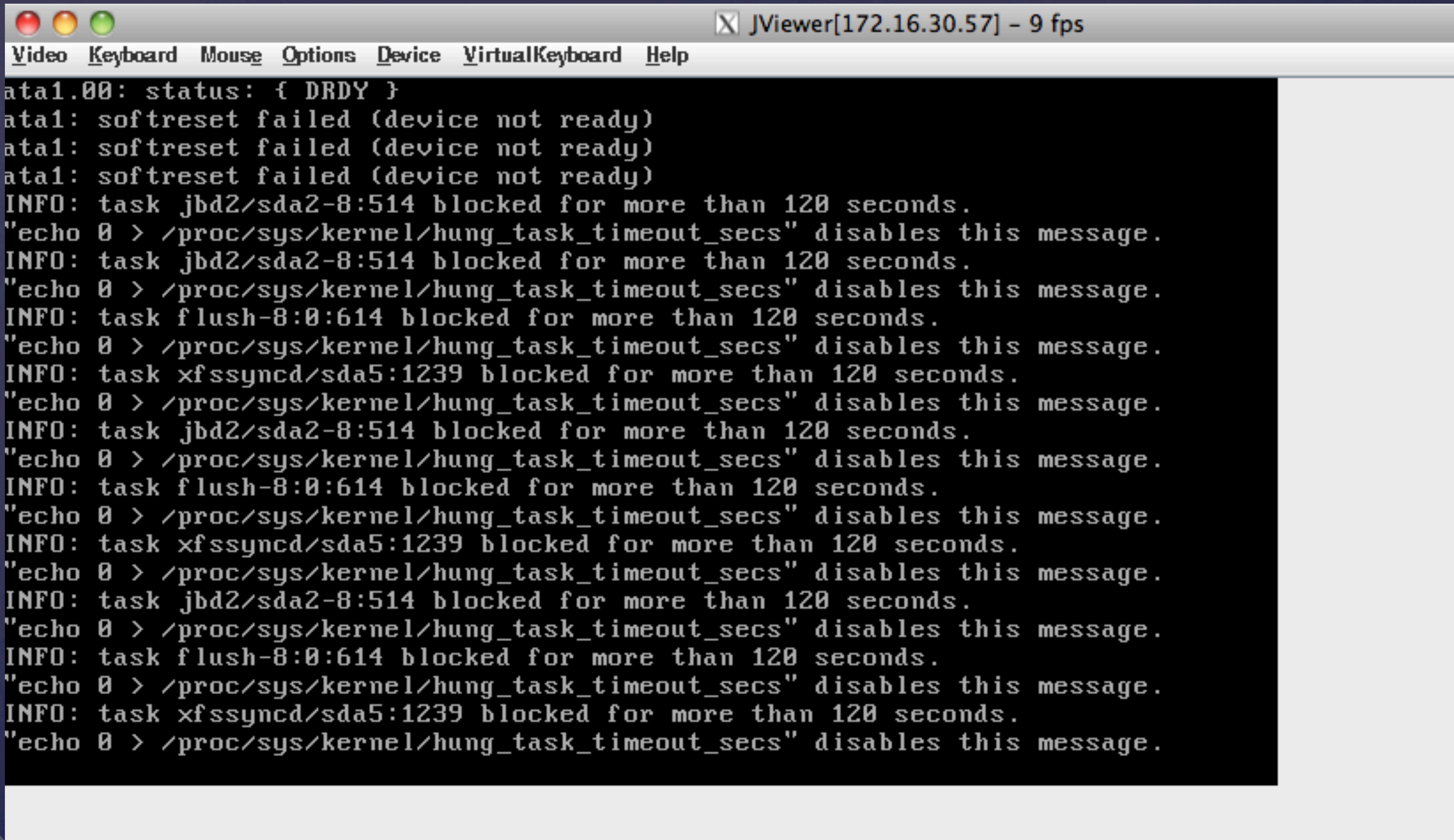

Ganglia
.sourceforge.net
ARGO Cluster Report for Fri, 04 Mar 2011 15:08:06 +0100
Metric load_one
Last week
Sorted descending
rid > ARGO > --Choose a Node
CPUs Total: 360
Hosts up: 45
Hosts down: 0
Avg Load (15,5,1m): 30%,36%,36%
Localtime: 2011-03-04 15:07
ARGO Cluster Load last week
Load/Procs
1-min Load
Nodes
CPUs
Running Processes
Cluster Load Percentages
100+ (26.67%)
75-100 (8.89%)
0-25 (64.44%)
Show Hosts: yes no | ARGO load_one la
node37.hpc
node40.hpc
node36.hpc
Ganglia
.sourceforge.net
ARGO Cluster Report for Thu, 01 Mar 2012 11:43:04 +0100
Last hour or from to Go Clear
Metric load_one
Sorted descending
Grid > ARGO > --Choose a Node
Overview of ARGO
CPUs Total: 808
Hosts up: 82
Hosts down: 0
Current Load Avg (15,5,1m): 89%,90%,90%
Avg Utilization (last hour): 88%
Localtime: 2012-03-01 11:42
ARGO Cluster Load last hour
ARGO Cluster CPU last hour
Cluster Load Percentages
100+ (71.95%)
75-100 (18.29%)
25-50 (1.22%)
0-25 (8.54%)
Show Hosts Scaled: Auto Same None | ARGO load_one last hour sorted descending | Size medium Columns
node04
node05
node12

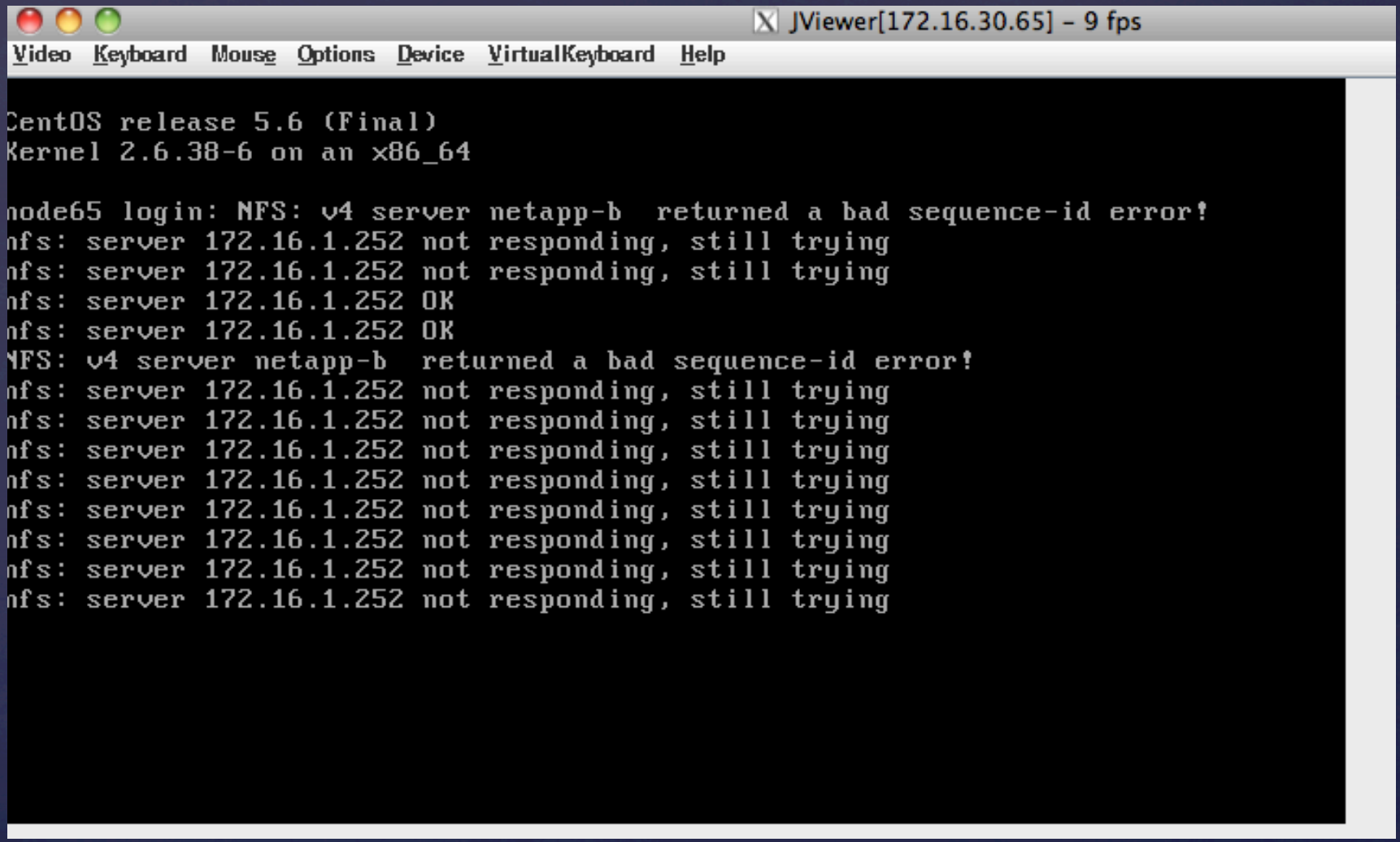

Ganglia:: ARGO Cluster Re...
argo-master.ictp.it/ganglia/
Search
Main
Search
Views
Aggregate Graphs
Compare Hosts
Events
Automatic Rotation
Live Dashboard
Mobile
ARGO Cluster Report at Thu, 24 Sep 2015 16:01:46 +0200
Get Fresh Data
Last hour 2hr 4hr day week month year job or from to Go Clear
Physical View
Grid > ARGO > --Choose a Node
Overview of ARGO @ 2015-09-24 16:01
CPUs Total: 2656
Hosts up: 199
Hosts down: 0
Current Load Avg (15, 5, 1m):
85%, 86%, 86%
Avg Utilization (last hour):
84%
Server Load Distribution
ARGO Cluster Load last hour
Loads/Procs
ARGO Cluster Memory last hour
Bytes
ARGO Cluster CPU last hour
Percent
ARGO Cluster Network last hour
Bytes/sec

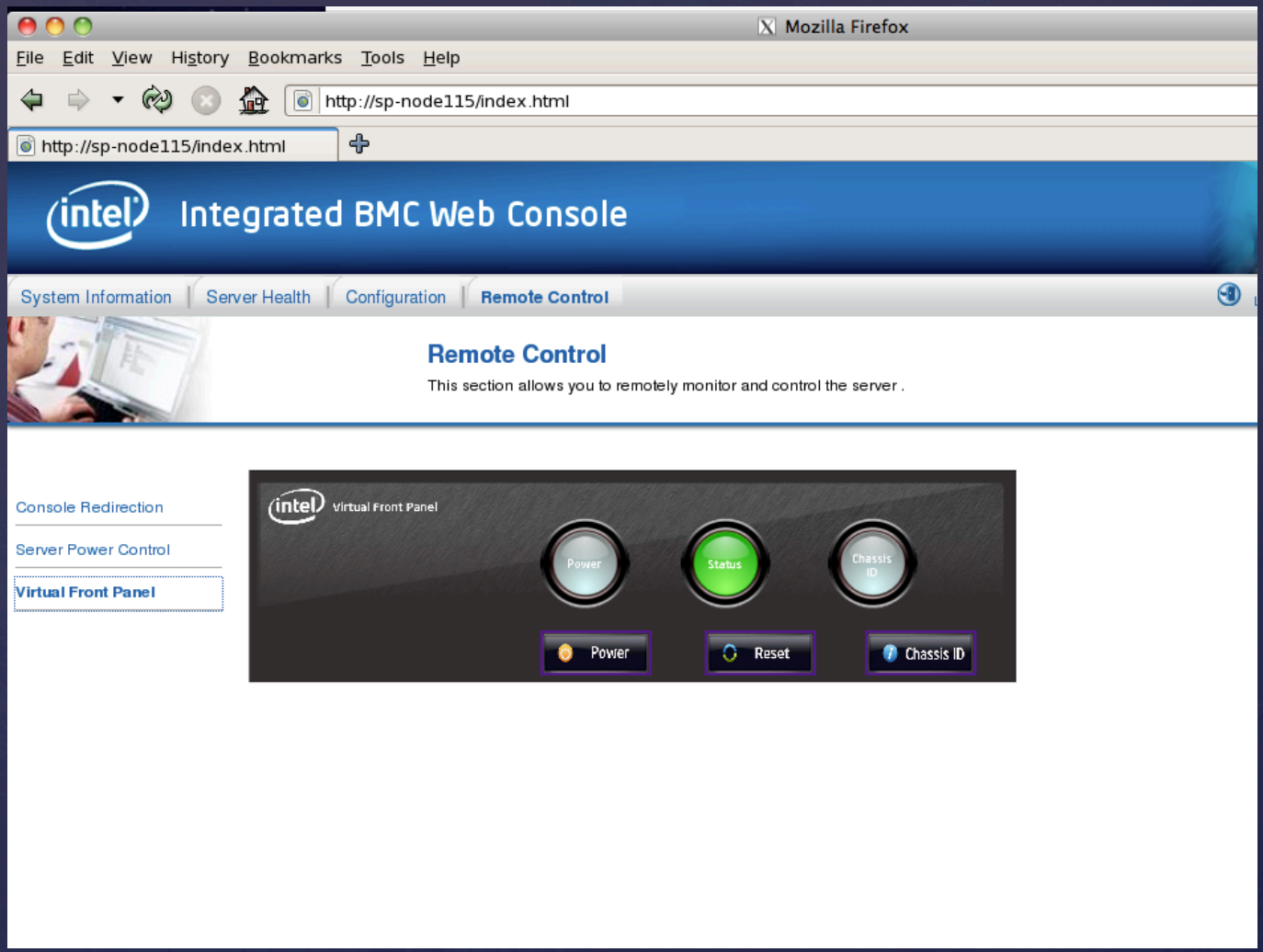

Ganglia:: ARGO Cluster Re...
argo-master.ictp.it/ganglia/
Search
Stacked Graph - load_one
ARGO aggregated load_one last hour
400
380
360
340
320
300
280
260
240
220
200
180
160
140
120
100
80
60
40
20
0
10:20
10:30
10:40
10:50
11:00
11:10
argo.ictp.it
gpu01.hpc
login1.hpc
login2.hpc
node01.hpc
node02.hpc
node03.hpc
node04.hpc
node05.hpc
node06.hpc
node07.hpc
node08.hpc
node100.hpc
node101.hpc
node102.hpc
node103.hpc
node104.hpc
node105.hpc
node106.hpc
node107.hpc
node108.hpc
node109.hpc
node11.hpc
node110.hpc
node111.hpc
node112.hpc
node113.hpc
node114.hpc
node115.hpc
node116.hpc
node117.hpc
node118.hpc
node119.hpc
node12.hpc
node120.hpc
node121.hpc
node122.hpc
node123.hpc
node124.hpc
node125.hpc
node126.hpc
node127.hpc
node128.hpc
node129.hpc
node13.hpc
node130.hpc
node131.hpc
node132.hpc
node133.hpc
node134.hpc
node135.hpc
node136.hpc
node137.hpc
node138.hpc
node139.hpc
node14.hpc
node140.hpc
node141.hpc
node142.hpc
node143.hpc
node144.hpc
node145.hpc
node146.hpc
node147.hpc
node148.hpc
node149.hpc
node15.hpc
node150.hpc
node151.hpc
node152.hpc
node153.hpc
node154.hpc
node155.hpc
node156.hpc
node157.hpc
node158.hpc
node159.hpc
node16.hpc
node160.hpc
node161.hpc
node162.hpc
node163.hpc
node164.hpc
node165.hpc
node166.hpc
node167.hpc
node168.hpc
node169.hpc
node17.hpc
node170.hpc
node171.hpc
node172.hpc
node173.hpc
node174.hpc
node175.hpc
node176.hpc
node177.hpc
node178.hpc
node179.hpc
node18.hpc
node180.hpc
node181.hpc
node182.hpc
node183.hpc
node184.hpc

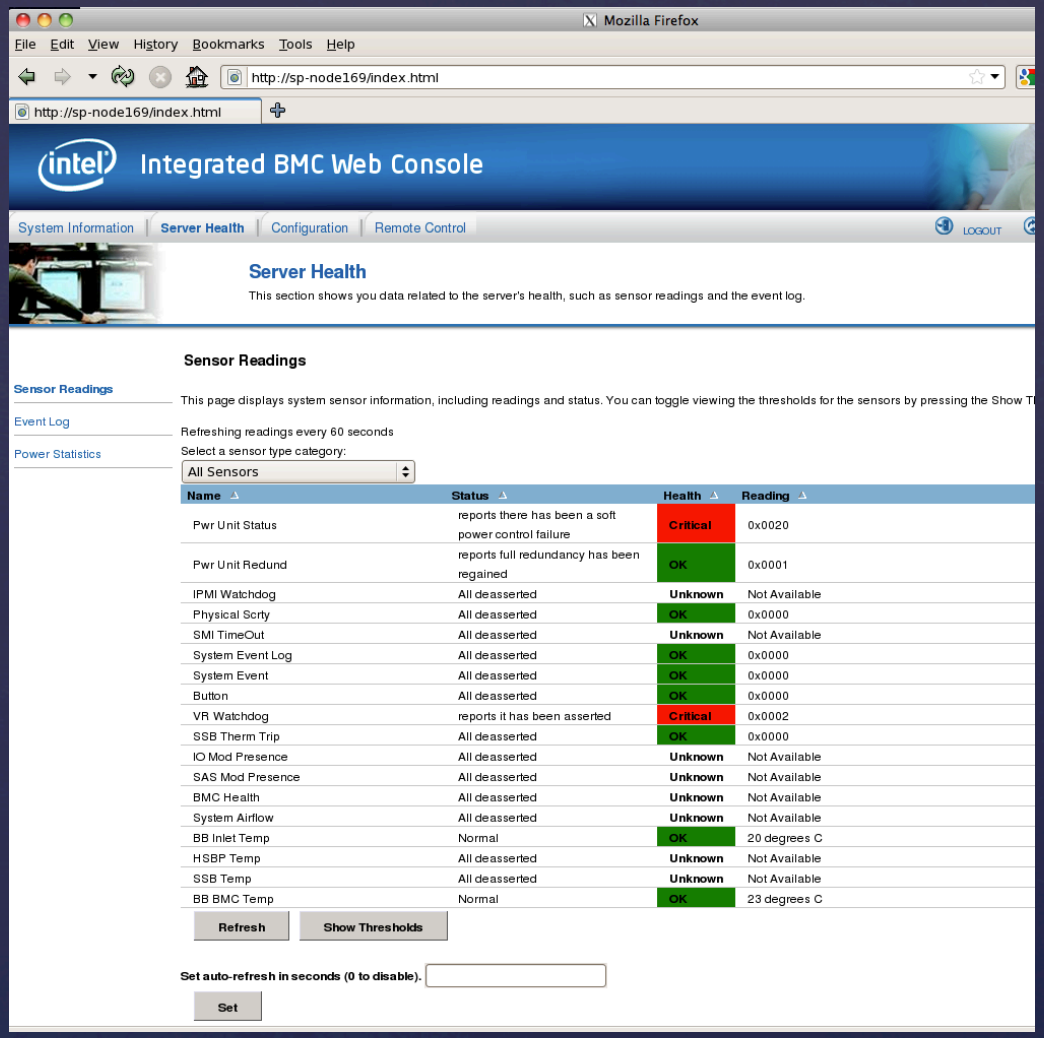

JViewer[172.16.30.36] - 0 fps
Video Keyboard Mouse Options Device Help
Aptio Setup Utility - Copyright (C) 2010 - 2013 American Megatrends, Inc.
Processor Configuration
Processor Configuration
Processor Socket CPU 1 CPU 2
Processor ID 306e4* 306e4
Processor Frequency 2.800 GHz 2.800 GHz
Microcode Revision 424 424
L1 Cache RAM 64 KB 64 KB
L2 Cache RAM 256 KB 256 KB
L3 Cache RAM 25600 KB 25600 KB
Processor 1 Version Intel(R) Xeon(R) CPU E5-2680 v2 @ 2.80GHz
Processor 2 Version
Intel(R) Hyper-Threading Tech
Enabled
Disabled
Current Intel(R) QPI Link S
Intel(R) QPI Link Frequency
Intel(R) QPI Frequency Select [Auto Max]
Intel(R) Turbo Boost Technology [Enabled]
Enhanced Intel SpeedStep(R) Tech [Enabled]
Processor C3 [Disabled]
Processor C6 [Enabled]
Intel(R) Hyper-Threading Tech [Enabled]
Active Processor Cores [All]
Execute Disable Bit [Enabled]
Intel(R) Virtualization Technology [Disabled]
Intel(R) VT for Directed I/O [Disabled]
Intel(R) Hyper-Threading Technology allows multithreaded software applications to execute threads in parallel within each processor.
Contact your OS vendor regarding OS support of this feature.
elect Screen
elect Item
Enter: Select
+/-: Change Opt.
F1: General Help
F9: Setup Defaults
F10: Save ESC: Exit
Version 2.14.1219. Copyright (C) 2010 - 2013 American Megatrends, Inc.

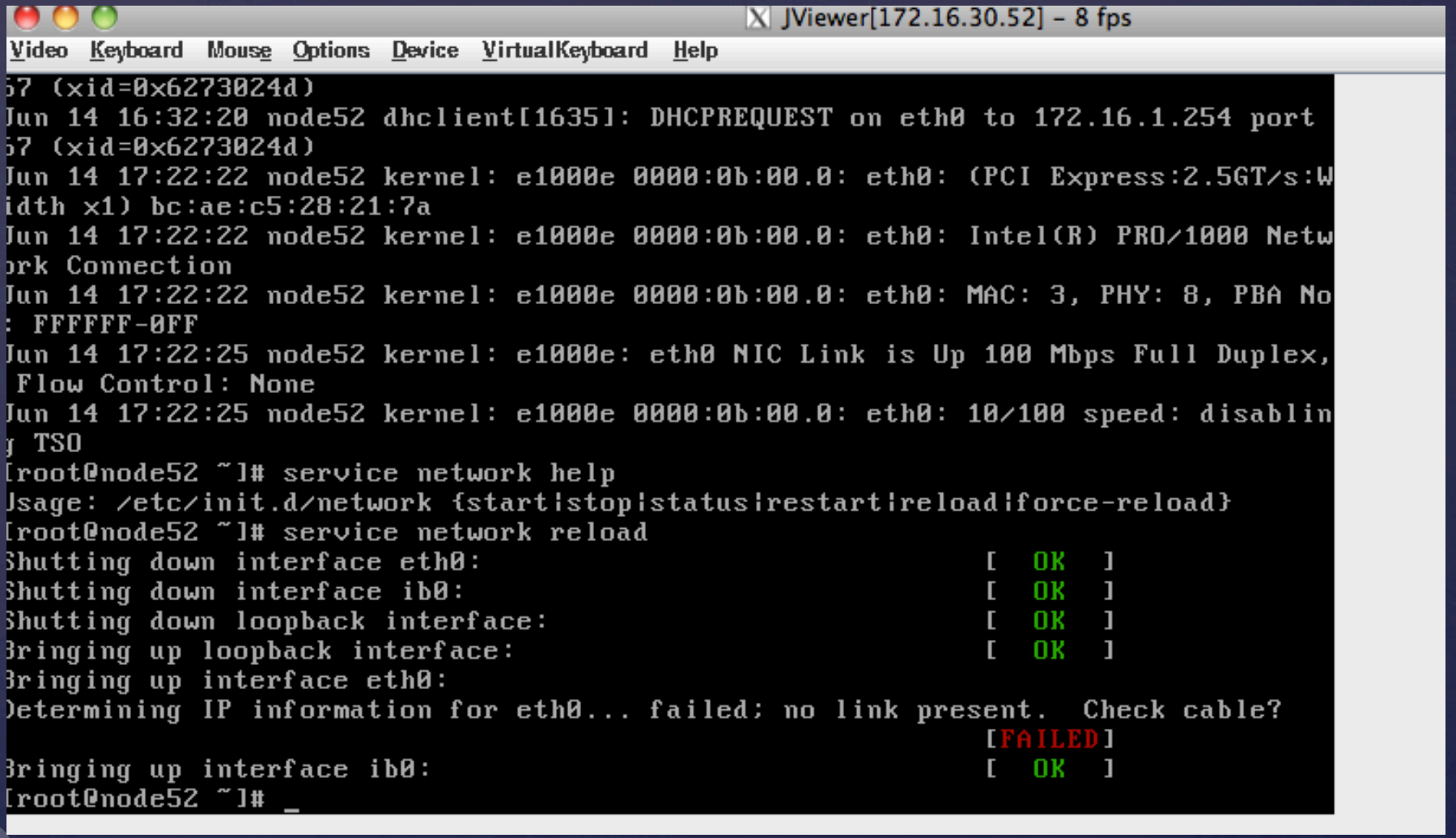

JViewer[172.16.30.37] - 10 fps
Video Keyboard Mouse Options Device VirtualKeyboard Help
Memtest86+ v4.20
Core i7 (32nm) 2400 MHz
L1 Cache: 32K 80005 MB/s
L2 Cache: 256K 31581 MB/s
L3 Cache: 12288 22019 MB/s
Memory : 24G 9796 MB/s
Pass 27% ##########
Test 46% ##################
Test #4 [Moving inversions, random pattern]
Testing: 16G - 18G 24G
Pattern: d0deded8
Chipset : Core IMC 32nm (ECC : Detect / Correct) Scrub+ / BCLK : 133 MHz
Settings: RAM : 533 MHz (DDR3-1066) / CAS : 7-7-7-20 / Triple Channel
WallTime Cached RsvdMem MemMap Cache ECC Test Pass Errors ECC Errs
19:02:48 24G 200K e820 on off Std 4 0
Memory SPD Informations
*****Pass complete, no errors, press Esc to exit*****
(ESC)Reboot (c)configuration (SP)scroll_lock (CR)scroll_unlock


JViewer[172.16.30.57] – 9 fps
Video Keyboard Mouse Options Device VirtualKeyboard Help
ata1.00: status: { DRDY }
ata1: softreset failed (device not ready)
ata1: softreset failed (device not ready)
ata1: softreset failed (device not ready)
INFO: task jbd2/sda2-8:514 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task jbd2/sda2-8:514 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task flush-8:0:614 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task xfssyncd/sda5:1239 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task jbd2/sda2-8:514 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task flush-8:0:614 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task xfssyncd/sda5:1239 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task jbd2/sda2-8:514 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task flush-8:0:614 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task xfssyncd/sda5:1239 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.

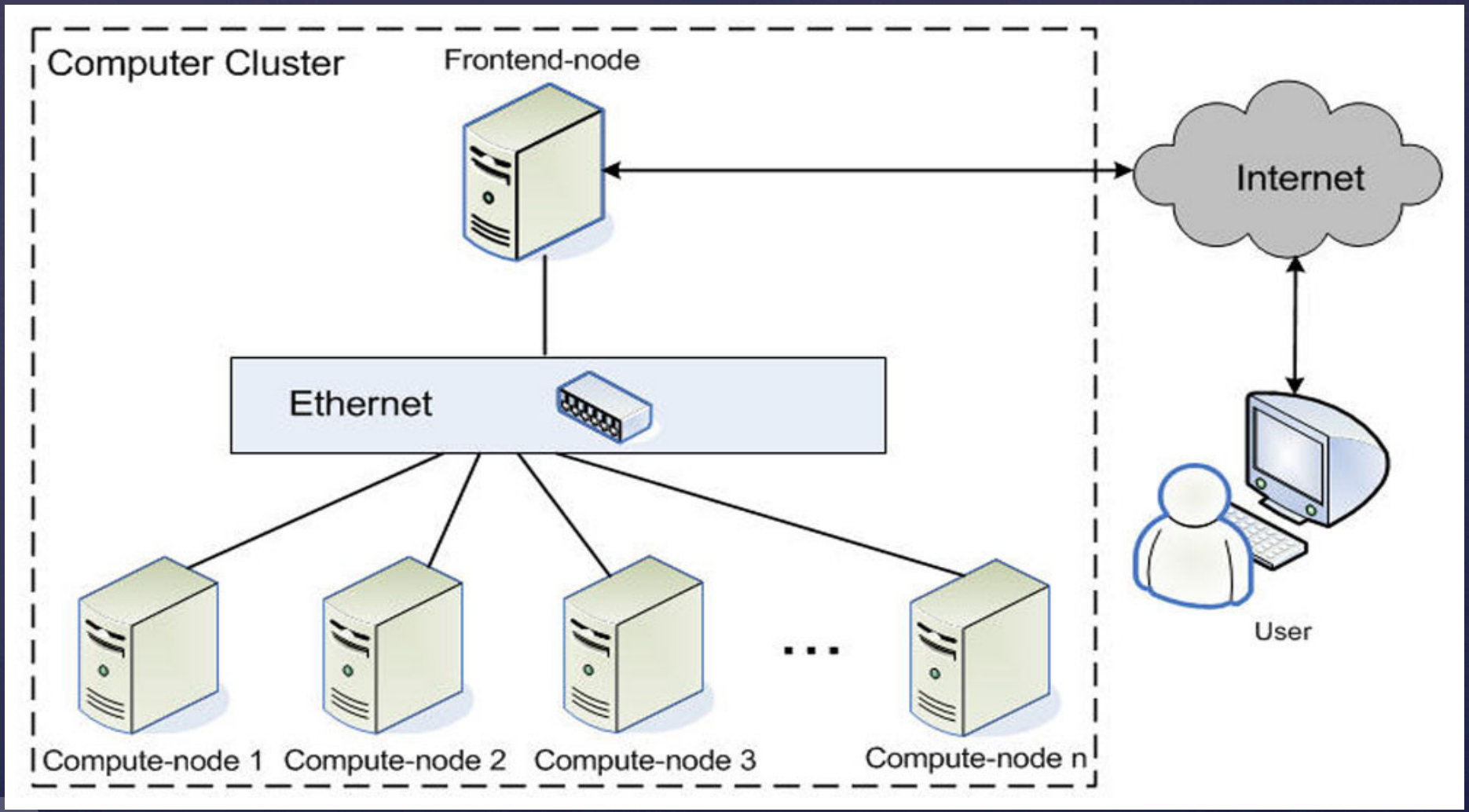

JViewer[172.16.30.65] - 9 fps
Video Keyboard Mouse Options Device VirtualKeyboard Help
CentOS release 5.6 (Final)
Kernel 2.6.38-6 on an x86_64
node65 login: NFS: v4 server netapp-b returned a bad sequence-id error!
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 OK
nfs: server 172.16.1.252 OK
NFS: v4 server netapp-b returned a bad sequence-id error!
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying


Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://sp-node115/index.html
http://sp-node115/index.html
intel
Integrated BMC Web Console
System Information
Server Health
Configuration
Remote Control
Remote Control
This section allows you to remotely monitor and control the server .
Console Redirection
Server Power Control
Virtual Front Panel
intel
Virtual Front Panel
Power
Status
Chassis ID
Power
Reset
Chassis ID


Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://sp-node169/index.html
http://sp-node169/index.html
Integrated BMC Web Console
System Information
Server Health
Configuration
Remote Control
LOGOUT
Server Health
This section shows you data related to the server's health, such as sensor readings and the event log.
Sensor Readings
Event Log
Power Statistics
Sensor Readings
This page displays system sensor information, including readings and status. You can toggle viewing the thresholds for the sensors by pressing the Show T
Refreshing readings every 60 seconds
Select a sensor type category:
All Sensors
Name
Status
Health
Reading
Pwr Unit Status
reports there has been a soft power control failure
Critical
0x0020
Pwr Unit Redund
reports full redundancy has been regained
OK
0x0001
IPMI Watchdog
All deasserted
Unknown
Not Available
Physical Scrty
All deasserted
OK
0x0000
SMI TimeOut
All deasserted
Unknown
Not Available
System Event Log
All deasserted
OK
0x0000
System Event
All deasserted
OK
0x0000
Button
All deasserted
OK
0x0000
VR Watchdog
reports it has been asserted
Critical
0x0002
SSB Therm Trip
All deasserted
OK
0x0000
IO Mod Presence
All deasserted
Unknown
Not Available
SAS Mod Presence
All deasserted
Unknown
Not Available
BMC Health
All deasserted
Unknown
Not Available
System Airflow
All deasserted
Unknown
Not Available
BB Inlet Temp
Normal
OK
20 degrees C
HSBP Temp
All deasserted
Unknown
Not Available
SSB Temp
All deasserted
Unknown
Not Available
BB BMC Temp
Normal
OK
23 degrees C
Refresh
Show Thresholds
Set auto-refresh in seconds (0 to disable).
Set


JViewer[172.16.30.52] - 8 fps
Video Keyboard Mouse Options Device VirtualKeyboard Help
67 (xid=0x6273024d)
Jun 14 16:32:20 node52 dhclient[1635]: DHCPREQUEST on eth0 to 172.16.1.254 port
67 (xid=0x6273024d)
Jun 14 17:22:22 node52 kernel: e1000e 0000:0b:00.0: eth0: (PCI Express:2.5GT/s:W
idth x1) bc:ae:c5:28:21:7a
Jun 14 17:22:22 node52 kernel: e1000e 0000:0b:00.0: eth0: Intel(R) PRO/1000 Netw
ork Connection
Jun 14 17:22:22 node52 kernel: e1000e 0000:0b:00.0: eth0: MAC: 3, PHY: 8, PBA No
: FFFFFF-0FF
Jun 14 17:22:25 node52 kernel: e1000e: eth0 NIC Link is Up 100 Mbps Full Duplex,
Flow Control: None
Jun 14 17:22:25 node52 kernel: e1000e 0000:0b:00.0: eth0: 10/100 speed: disablin
g TSO
[root@node52 ~]# service network help
Usage: /etc/init.d/network {start|stop|status|restart|reload|force-reload}
[root@node52 ~]# service network reload
Shutting down interface eth0: [OK]
Shutting down interface ib0: [OK]
Shutting down loopback interface: [OK]
Bringing up loopback interface: [OK]
Bringing up interface eth0:
Determining IP information for eth0... failed; no link present. Check cable?
[FAILED]
Bringing up interface ib0: [OK]
[root@node52 ~]# _

Cluster structure



Cluster Structure

- strictly speaking:

+ hardware (nodes)
+ network
+ software
+ applications running on it!

=cluster

- broadly speaking:

There is an "eco-system" around it:
power, UPS
cooling,
rack space,
documentation,
health monitoring,
backup
user support

Cluster view

- User view (limited structure view)
  - "Power" user can obtain more performance

Common language:
job, jobID,
node name,
"why?",
running, producing output,
job script, working directory

- HPC team view: all the architectural details, known issues

Cluster Components, Hardware

- Computers(Nodes)
- Network devices
- Storage
- …

Nodes

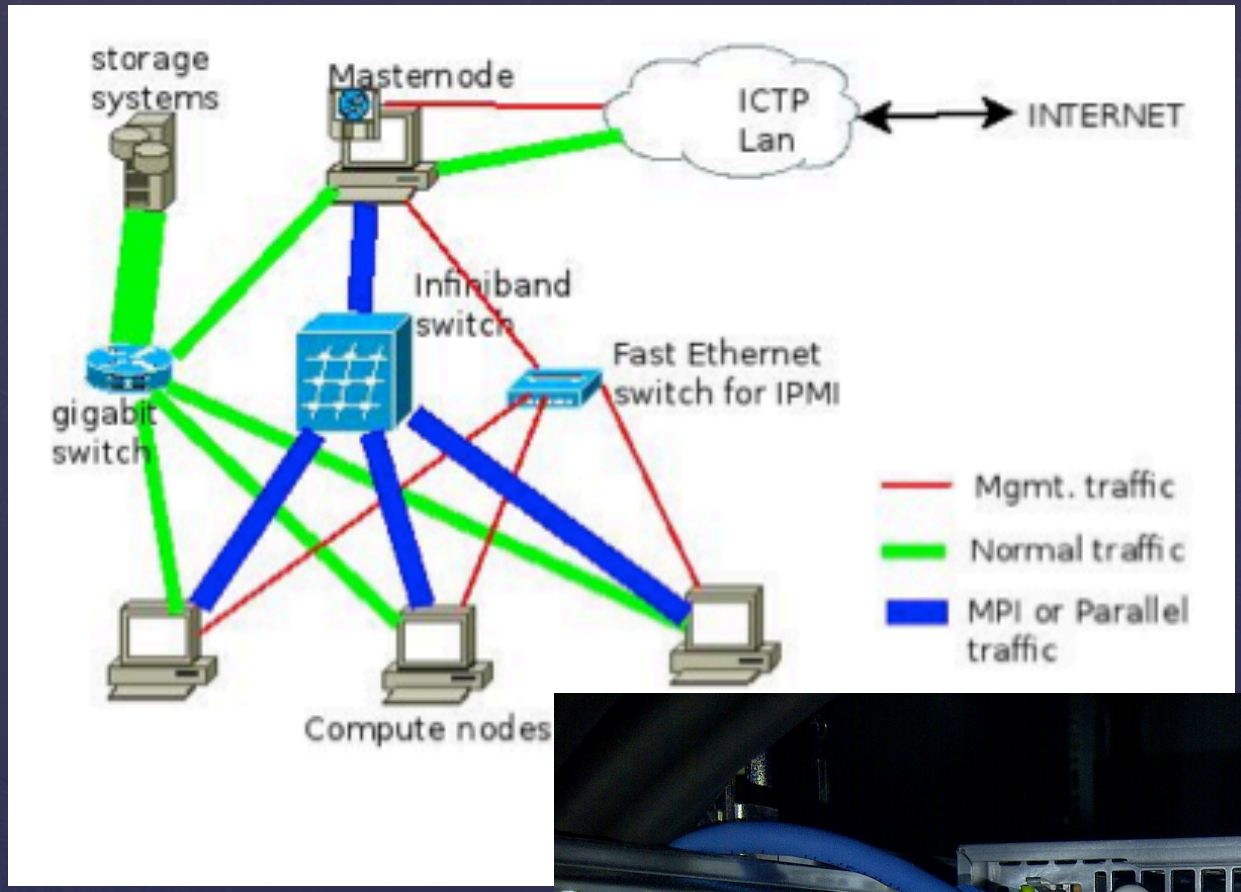
- Computing nodes/Worker nodes
  - homogeneous (initially)
  - heterogeneous: Not in one queue!
- Master node
- Specialised nodes (login, gpu,storage)

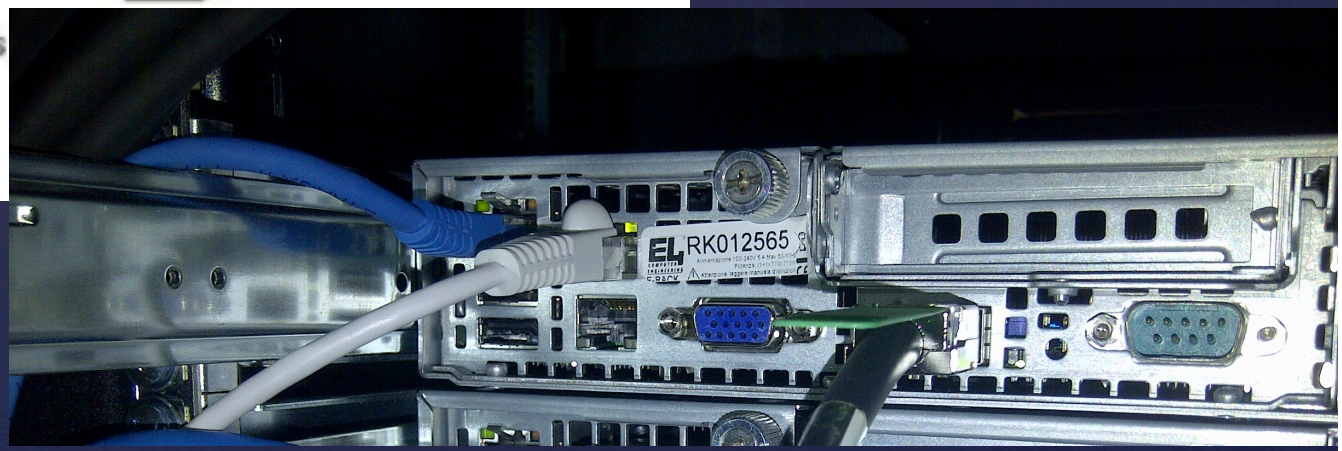
- compute node model
  - CPUs
  - Memory!
  - IB port
  - rack mountable
  - Service Processor (IPMI)
- master node model
  - differs

Cluster Components, Hardware

- Computers(Nodes)
- Network devices
- Storage
- …

Network devices

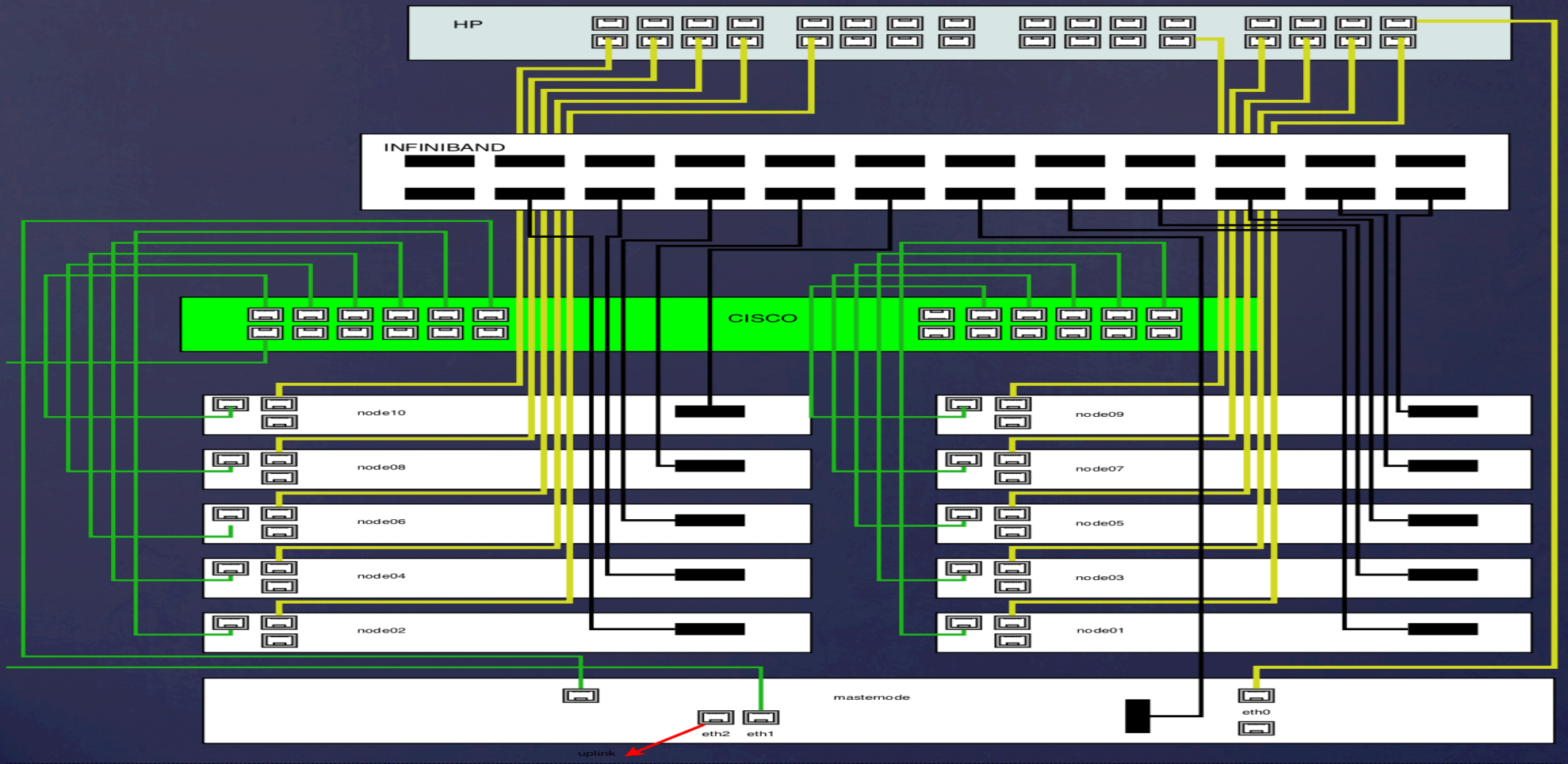




Network Devices

- GigaBit Ethernet network (GETH)
  - Ssh/ job submission
  - Data I/O traffic (NFS)
  - good bandwidth, NOT low latency for HPC
- Infiniband (or Myrinet) network
  - Message Passing Interface(MPI) for parallel computation
  - low latency and high bandwidth
- Fast Ethernet network
  - management traffic (IPMI to Service Processors)
- Switches and cables

Network diagram of Addis HPC



Taken from Antonio's presentation

Cluster Components, Hardware

- Computers(Nodes)
- Network devices
- Storage: from HDs to RAID group
- …

Redundant Array of Independent Disks(RAID)

Level	Useable capacity	Data protection
RAID0	$Size_{min}$ * n	None
RAID1	$Size_{min}$	Failure of one single disk
RAID5	$Size_{min}$ * (n – 1)	Concurrent failure of one single disk
RAID6	$Size_{min}$ * (n – 2)	Concurrent failure of two disks
RAID1+0	$Size_{min}$ * (n/2)	Concurrent failure of more than two disks

Taken from Clement's slide

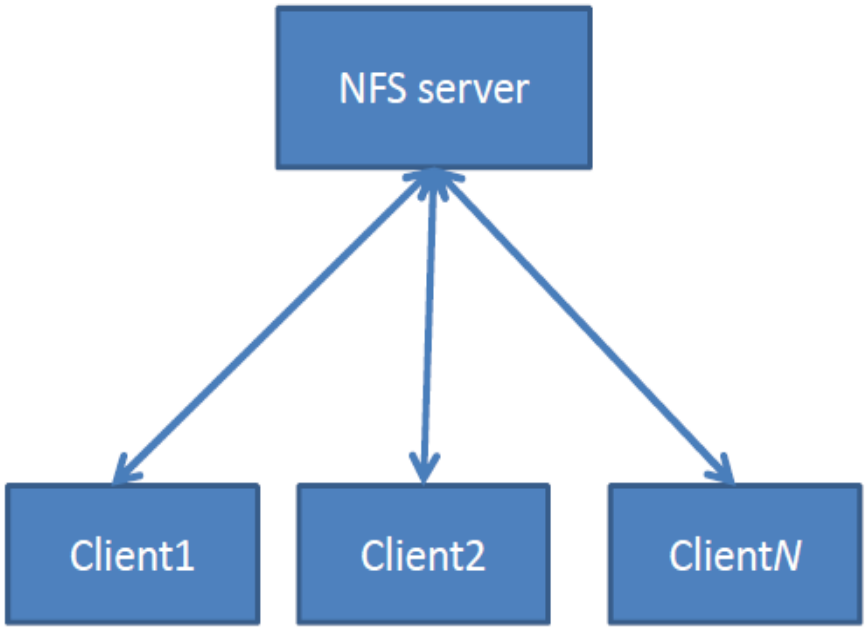
Storage: Disk array

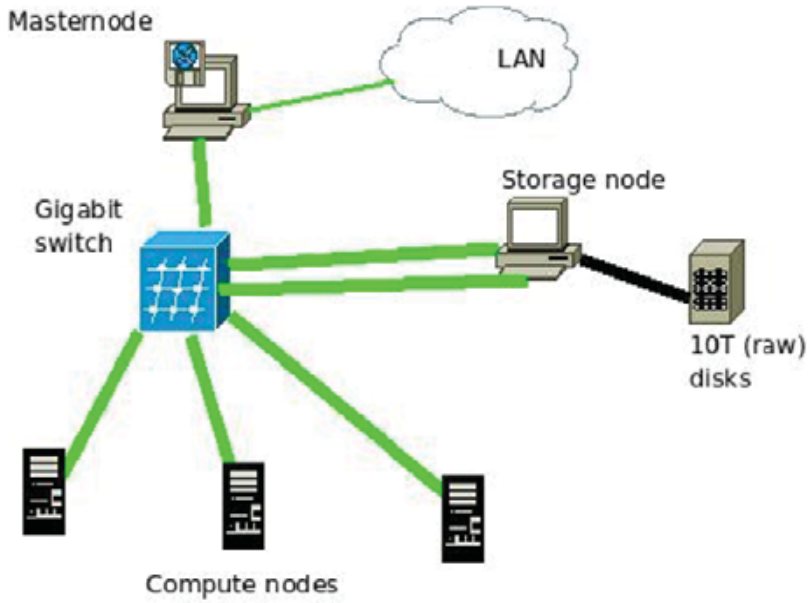
- a large group of hard disk drives (HDDs)
- RAID: to improve performance and dat a protection
- NFS: to mount these disks for the cluster.

Storage: Disk array

- HDDs work in parallel
- IOPS: the more the merrirer!
- SSD in some cases
- place the data as near as possible to the computing power
- know your app data traffic(size and IOPS)

NFS architecture





Taken from Clement's slide

NFS(Network File System)

- NFS Server/NAS exports.
- Clients ”mount” into their file system.
- Under the mount point, file access is transparent.

  (for the user)

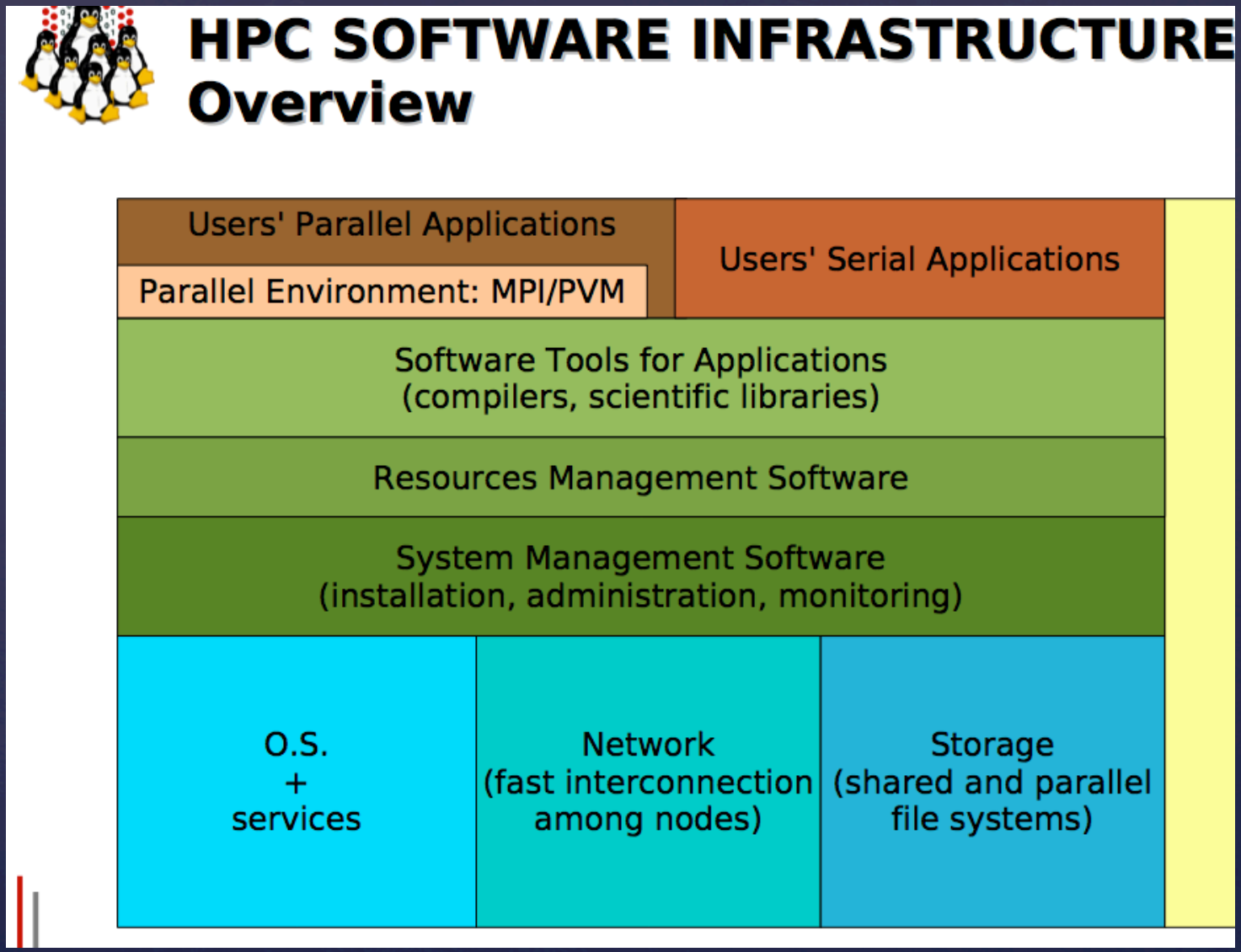
- /etc/export

```
/distro          10.1.0.0/16(ro,root_squash)
/distro/centos   10.1.0.0/16(ro,root_squash)
/home 10.1.0.0/16(rw,no_root_squash)
```

NFS(Network File System)

- Used for:
  - /home RW
  - /opt software RO
  - /distro install repository, RO
  - /scratch RW
  - /projects RW
- quotas
- auto-cleaning ? (agree!)
- data-plan
- /projects visible from Desktops! – out of cluster

Cluster Components, Software



Taken from Moreno's slide

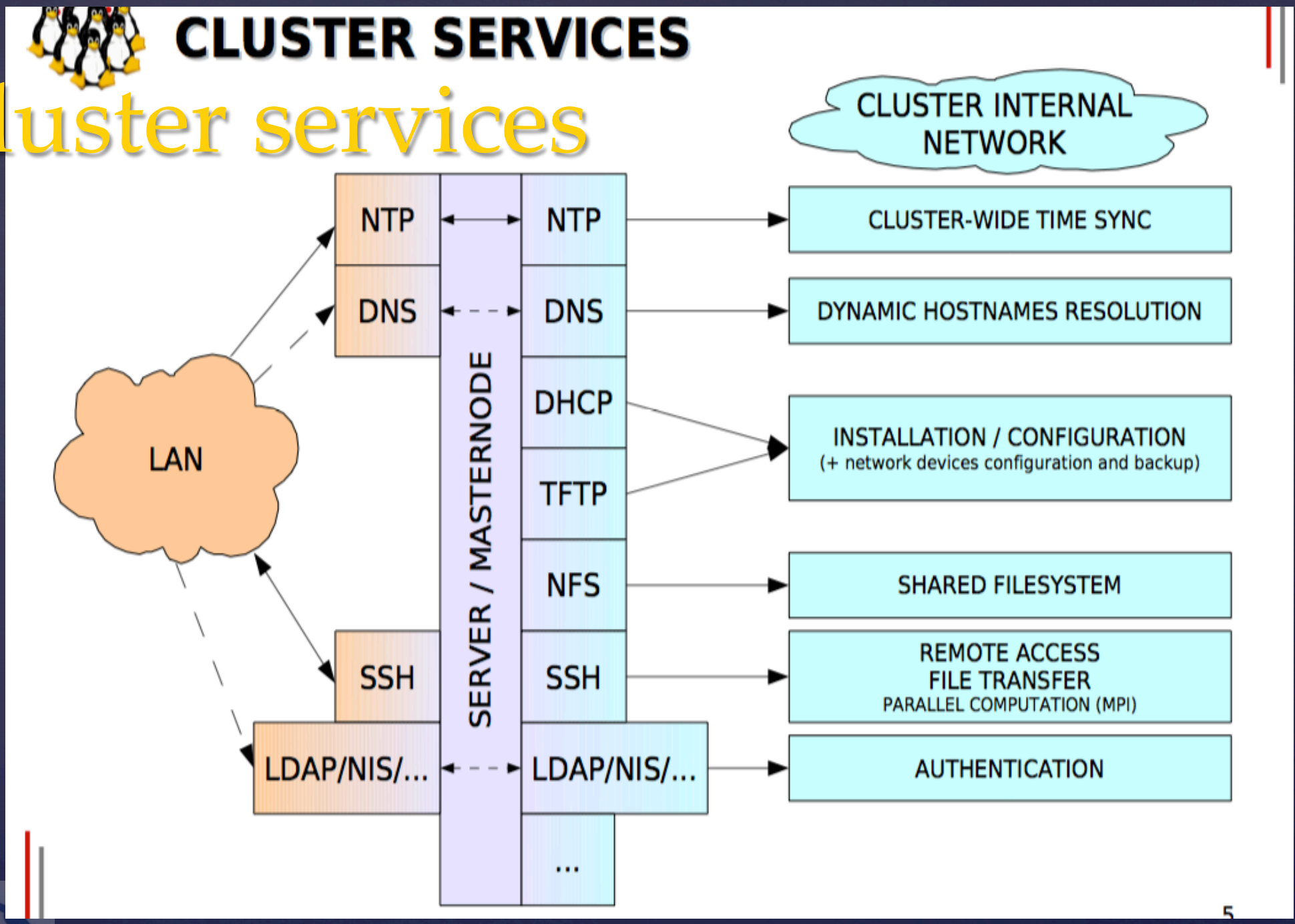
Cluster Components, Software

- On Master node:
  - OS
  - MPI ?
  - Resource manager
  - Scheduler
  - Installation services
  - Compilations
  - …
- what runs on Compute node?
  - OS
  - MPI
  - Resource manager client part
  - User apps
  - ganglia client

Operating System

- Linux: Centos, Red Hat, Debian …
- master node install: interactive (manual)
- Computing nodes: automated, hands-off mass install

Cluster services



Taken from Moreno's slide

Services running on Master node

- DNS(/etc/hosts)
- DHCP
- TFTP
- NTP
- NFS (opt)

- ssh
- Torque
- Maui
- Ganglia
- Syslog
- http (opt)
- …

Common view cluster-wide

- user accounts (/etc/passwd)
- hostnames (/etc/hosts)
- mount points (/etc/auto.home)
- ...

- Pushed to nodes at regular intervals
- and at boot time

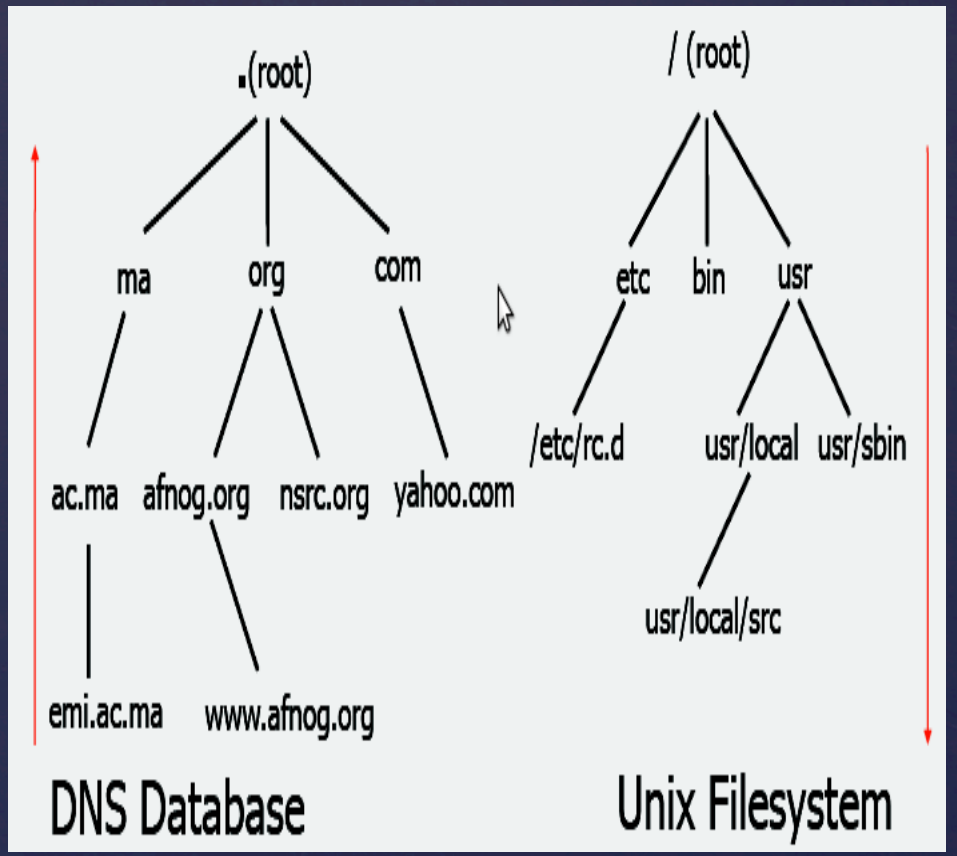
/etc/hosts

- plain file used to map host names to IP addresses
- In the absence of DNS name server, any network program on your system consults /etc/hosts
- can be rsynced to all nodes

```
127.0.0.1   localhost.localdomain localhost
10.1.0.1    master.cluster       master
10.1.1.1    node1.cluster        node1
10.1.1.2    node2.cluster        node2
```

DNS (Domain Name System)

- It is a network services that is used to convert domain names to IP addresses (and vice versa)
- DNS is hierarchical
- DNS administration is shared – no single central entity administrates all DNS data

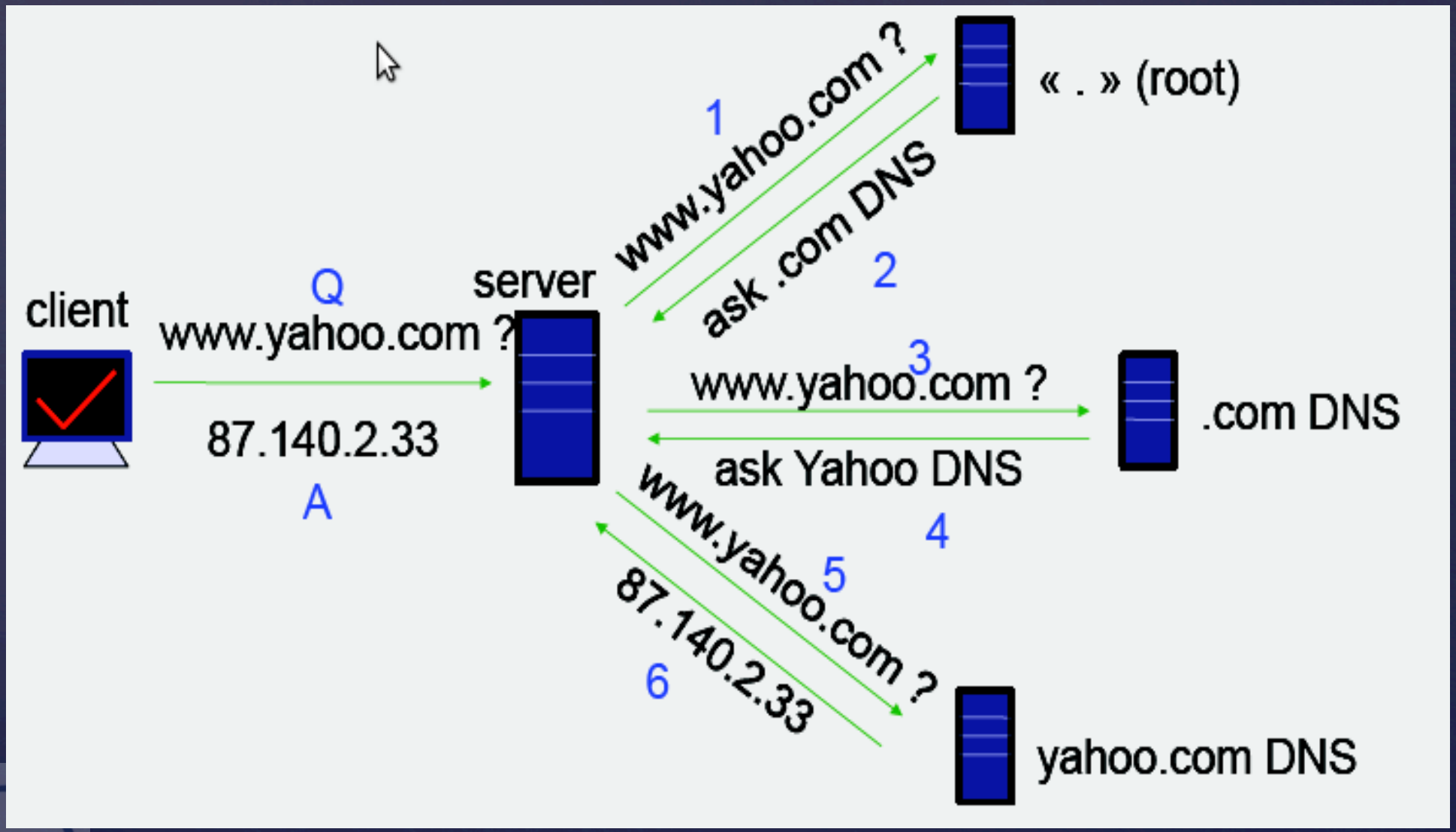


From Afnog Workshop

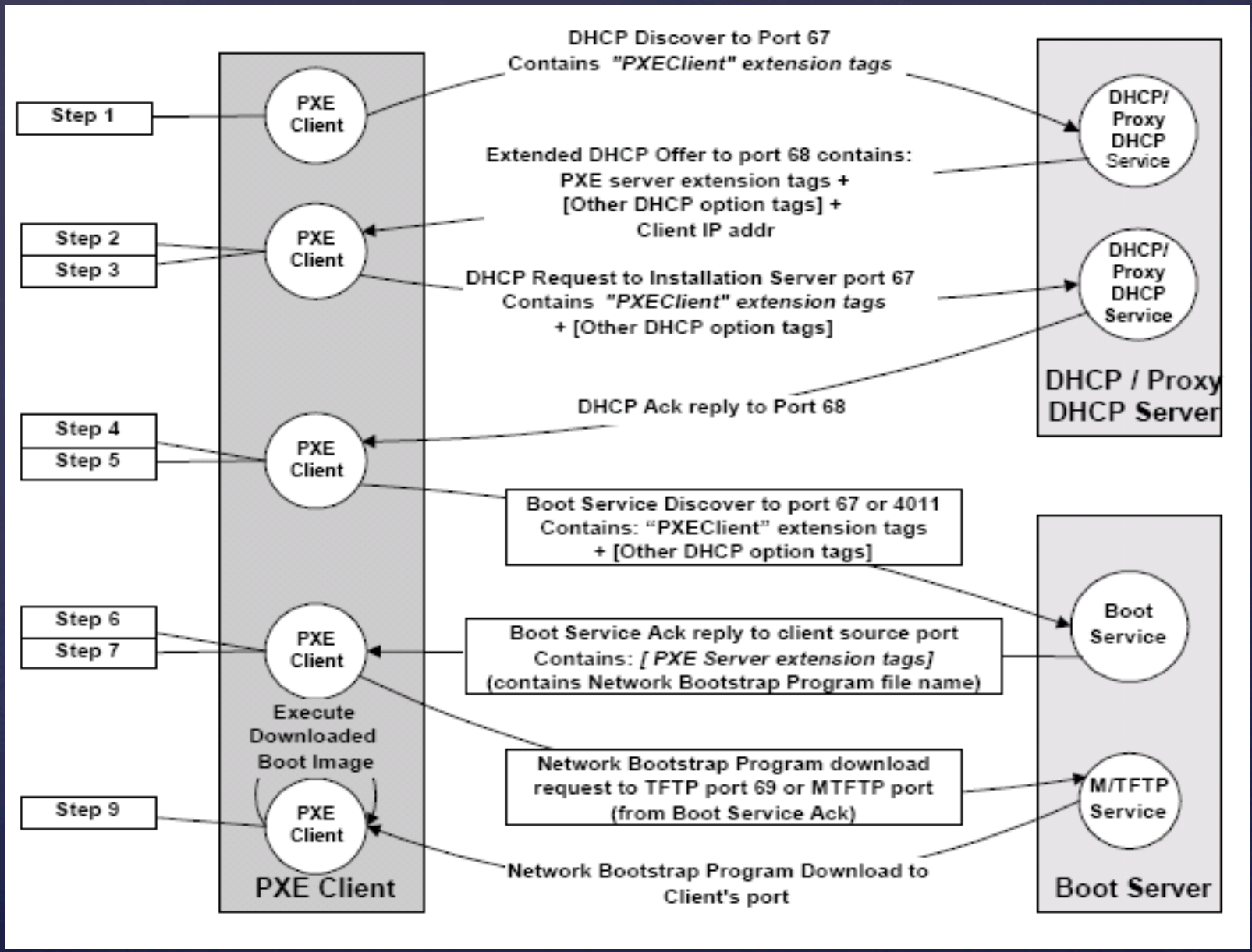
How does DNS work?

- The client (web browser, mail program, ...) use the OS's resolver to find the IP address - this is called a query
- The server being queried will try to find the answer on behalf of the client
- The server functions recursively, from top (the root) to bottom, until it finds the answer, asking other servers along the way - the server gets referred to other servers

A DNS query


« . » (root)
1
www.yahoo.com ?
ask .com DNS
2
client
Q
www.yahoo.com ?
server
3
www.yahoo.com ?
.com DNS
ask Yahoo DNS
4
87.140.2.33
A
www.yahoo.com ?
5
87.140.2.33
6
yahoo.com DNS

Network booting timeline



DHCP (Dynamic Host Configuration Protocol)

- DHCP allows hosts on a TCP/IP network to request and be assigned IP addresses, and also to discover information about the network to which they are attached
- DHCP provide a mechanism whereby the server can provide the client with information about how to configure its network interface (e.g., subnet mask), and also how the client can access various network services (e.g., DNS, IP routers, and so on).

```
subnet 239.252.197.0 netmask
    255.255.255.0 {
  range 239.252.197.10
      239.252.197.250;
  default-lease-time 86400 max-
      lease-time 172800;
  option subnet-mask
      255.255.255.0;
  option broadcast-address
      239.252.197.255;
  option routers 239.252.197.1;
  option domain-name-servers
      239.252.197.2,
      239.252.197.3;
  option domain-name "isc.org";
      }
```

DHCP configuration

- From master node we configure DHCP server which receives clients (computing nodes) requests and replies to them
- DHCP clients sends configuration requests to the server
- For Network booting, PXE actsas a DHCP client

```
ddns-update-style none;
ddns-updates off;
authoritative;
subnet 10.1.0.0 netmask 255.255.0.0 {
option domain-name "clusterXY";
option domain-name-servers 10.1.0.1;
option ntp-servers 10.1.0.1;
option subnet-mask 255.255.0.0;
option broadcast-address 10.1.255.255;
filename "/pxe/pxelinux.0";
next-server 10.1.0.1;
}
host node1 { hardware ethernet ..:..:..:..:..:.. ;
fixed-address 10.1.1.1 ; option host-name
"node1" ; }
host node2 { hardware ethernet ..:..:..:..:..:.. ;
fixed-address 10.1.1.2 ; option host-name
"node2" ; }
```

PXE(Preboot eXecution environment)

- The specification describes a standardized client-server environment that boots a software assembly, retrieved from a network, on PXE-enabled clients.
- On the client side it requires only a PXE-capable network interface controller (NIC), and uses network protocols such as DHCP and TFTP.
- TFTP server has to provide the following files for the clients to initialize the network booting:
  - ✓ pxelinux.0 - is use to load the operating system that is required to execute the assigned preboot work
  - ✓ vmlinuz – is a Linux kernel executable
  - ✓ Initrd.img – initial ramdisk contains various executables and drivers that permit the real root file system to be mounted

TFTP(Trival File Transfer protocol)

- The protocol is extensively used to support remote booting of diskless devices
- The server is normally started by inetd, but can also run standalone
- TFTP services does not require an account or password on the server system.
- Due to the lack of authentication information, tftpd will allow only publicly readable files

- /etc/xinetd.d/tftp file

```
service tftp
{
        disable                 =  no
        socket_type             =  dgram
        protocol                =  udp
        wait                    =  yes
        user                    =  root
        server                  =  /usr/sbin/in.tftpd
        server_args             =  -s /tftpboot -vvv
        per_source              =  11
        cps                     =  100 2
        flags                   =  IPv4
}
```

Network booting…

- One other thing that should be transferred through tftp is PXE configuration file that defines the menu displayed to the target host
- These configuration files can be stored /tftpboot/pxe/pxelinux.cfg directory in one of the following form, for a given ip and mac

- /tftproot/pxe/pxelinux.cfg/01-88-99-aa-bb-cc-dd
  - If the mac address is 01-88-99-aa-bb-cc-dd
  - /tftpboot/pxe/pxelinux.cfg/C000025B
  - hexadecimal equivalent of ip address 192.0.2.91
  - /tftpboot/pxe/pxelinux.cfg/default
  - Usually hexadecimal equivalent of pxe configuration file is used to install a node and the “default” is used when one wants to boot the node from the local hard disk

- /tftpboot/pxe/pxelinux.cfg/default

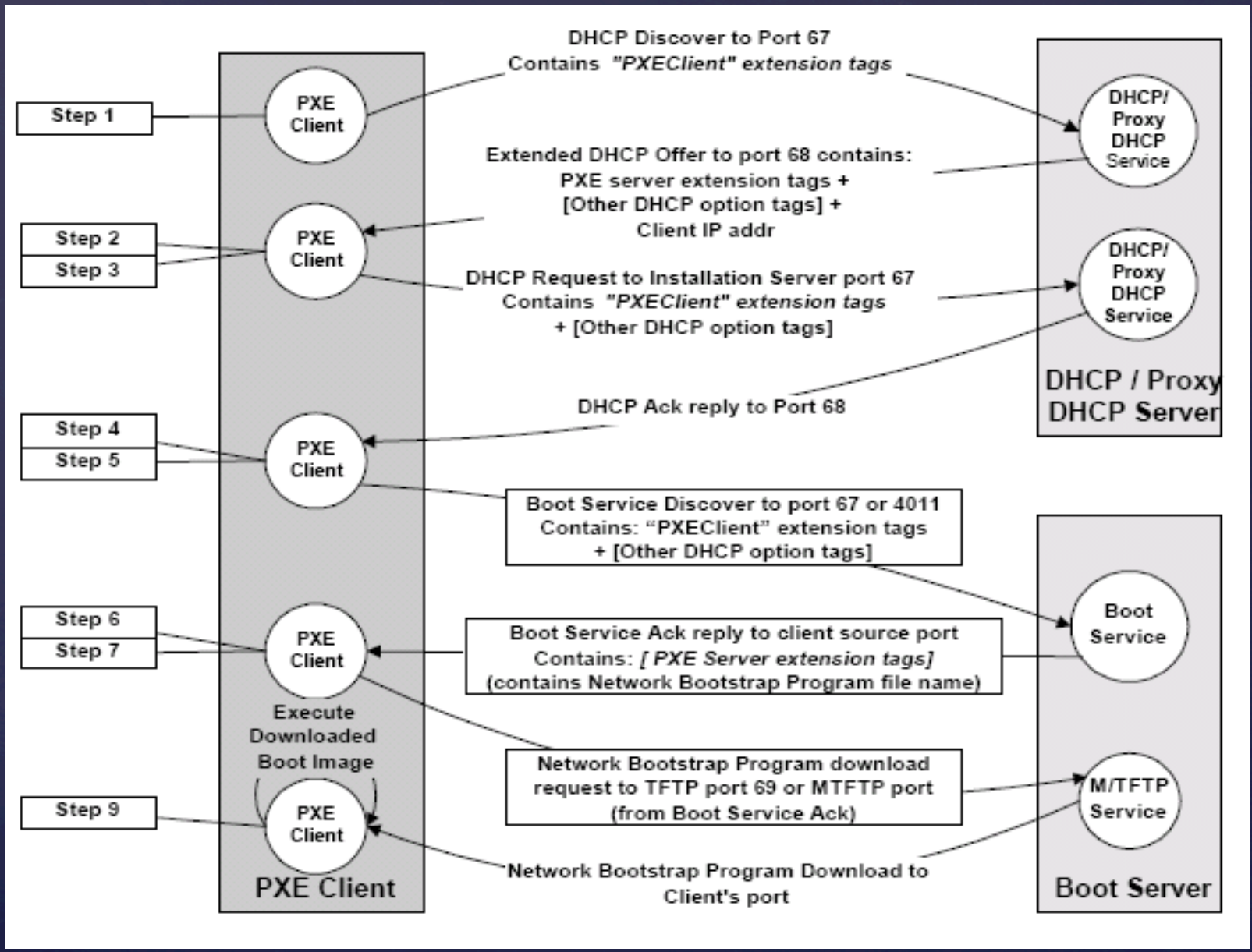
```
prompt 1
timeout 100
default local
label local
LOCALBOOT 0
label install
kernel vmlinuz
append initrd=initrd.img network
    ip=dhcp \
 ksdevice=eth0     ks=nfs:10.1.0.1:/
    distro/ks/ks.cfg \
    load_ramdisk=1
    prompt_ramdisk=0 \
    ramdisk_size=16384 \
    vga=normal selinux=0
```

Network booting “plan”

```
/
`--tftpboot/
    `-- pxe/
        |-- vmlinuz
        |-- initrd.img
        |-- memtest
        |-- pxelinux.0
        `-- pxelinux.cfg/
            |-- 0A0A0101
            |-- bootmsg.txt
            |-- default -> default.local
            |-- default.install
            `-- default.local
```

Taken from M. Baricevic, 2013 slide

Network booting timeline



Kickstart Installations

- automated installation (all answers provided, no keyboard input needed)
- Install media: local CD-ROM, or via NFS, FTP, or HTTP
- The kickstart file is a simple text file, containing a set of instruction on how to install the OS + post install steps
- Kickstart Configurator application or by writing it from scratch
- http://fedoraproject.org/wiki/Anaconda/Kickstart

Sample kickstart file

- Kickstart files allowing us to configure network, system configurations, HD partitioning and package selections
- In the pre-installations section one can choose hardware setup and configurations
- In the post-installation section one can include customizations and additional configurations.
- It is also possible to add lines of instructions that can stop the automated installation

```
#platform=x86, AMD64, or Intel EM64T
# System authorization information
auth  --useshadow  --enablemd5
# System bootloader configuration
bootloader --location=mbr
# Clear the Master Boot Record
zerombr
# Partition clearing information
clearpart --all --initlabel
# Use text mode install
text
# Firewall configuration
firewall --disabled
# Run the Setup Agent on first boot
firstboot --disable
# System keyboard
keyboard us
# System language
lang en_US
```

Cluster Management tools

- Cluster wide commands
- Password-less environment
- Cluster wide file distribution and gathering
- Appropriate access privilege

C3 tools

- Cluster Command Control (C3) tools are a suite of cluster tools that are useful for both administration and application support
- It includes tools for cluster-wide command execution, file distribution and gathering, process termination, remote shutdown and restart, and system image updates
- Example: cexec, cget, ckill , cpush and others

Code execution Environment, modules

- It provides dynamic modification of a user's environment
- It allows a group of related environment variables to be made or removed dynamically
- It enable us to avoid errors that can be caused from changing $PATH environment
- Some of frequently used module commands:
  - module avail
  - module list
  - module load/unload <package>
  - module purge

Other Important Software

- Open MPI is a Message Passing Interface (MPI) library project combining technologies and resources from several other projects (FT-MPI, LA-MPI, LAM/MPI, and PACX-MPI).
- Compilers(GNU, Portland Group and Intel)
- Scientific libraries
- Resource manager and Scheduler

  PBS/Torque and Maui scheduler
- Ganglia for metrics collection ad presentation

Health monitoring

- Internal monitoring:
  - services on master node up and running
  - node health (before the job)
- External monitoring: nagios

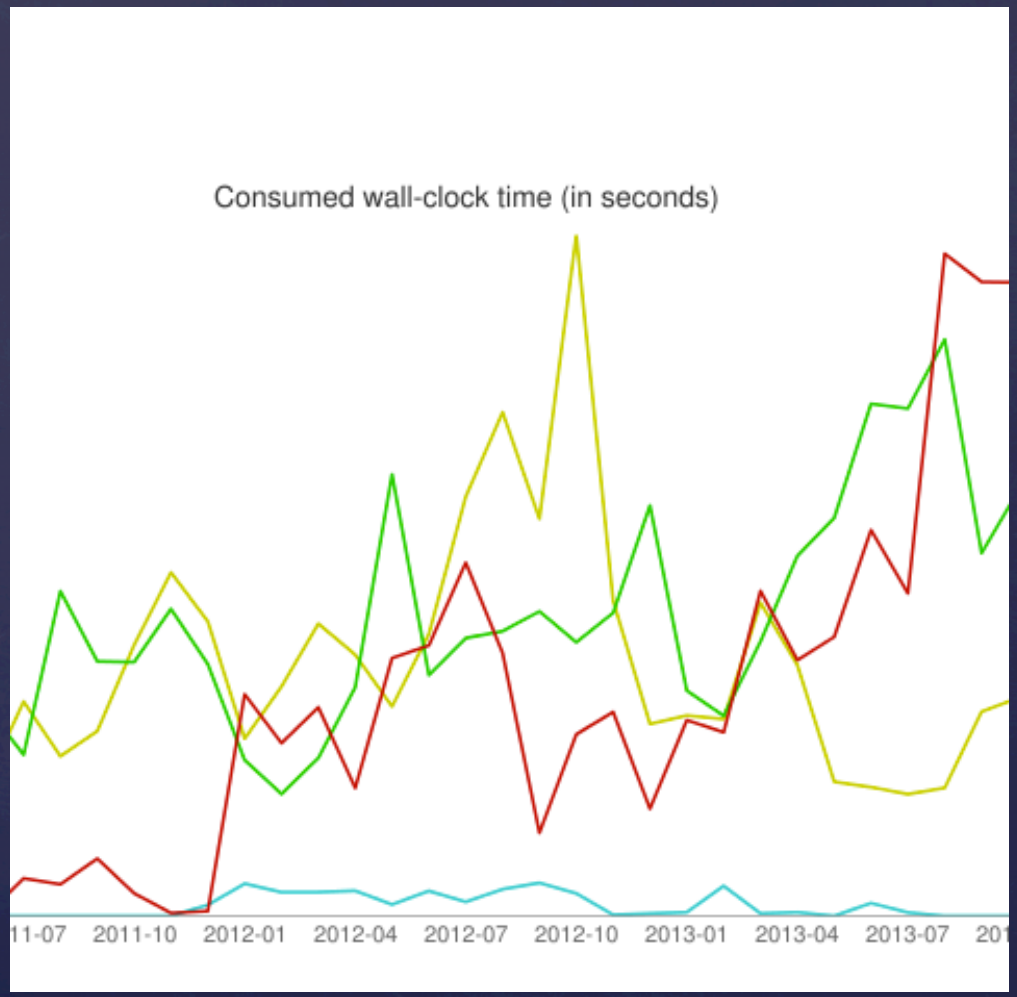
Service Status Details For Host 'argo-master'

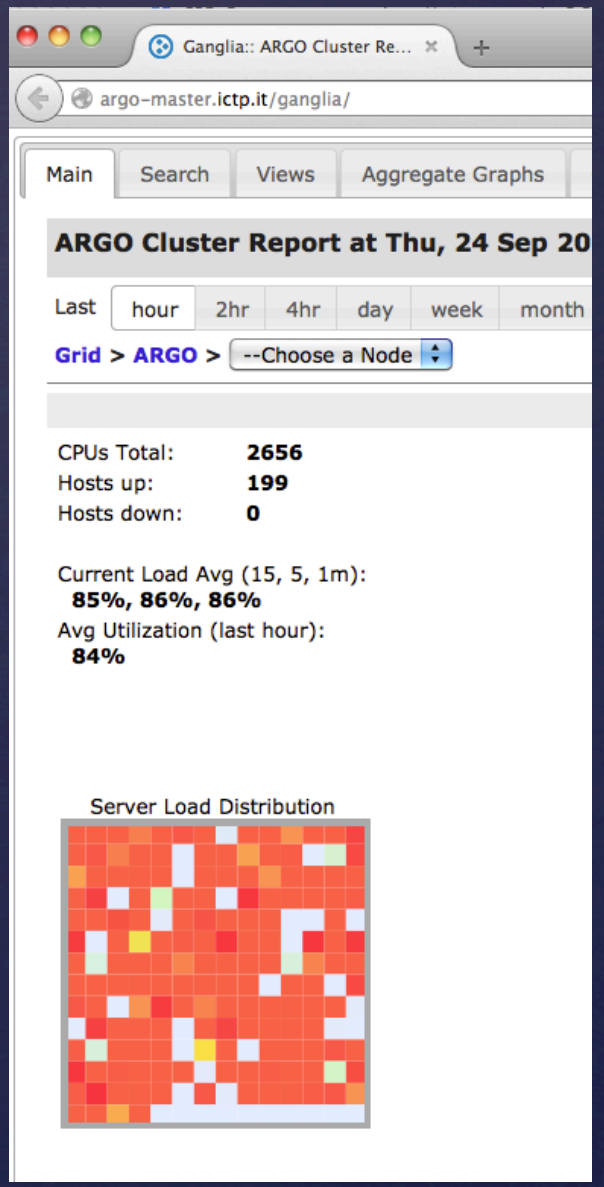
Host	Service	Status	Last Check	Duration	Attempt	Status Information
argo-master	2-DISKS-HEALTH	OK	09-29-2015 15:50:50	17d 6h 6m 52s	1/3	OK sda=PASSED sdb=PASSED
	HTTP	OK	09-29-2015 15:56:02	17d 6h 5m 13s	1/3	HTTP OK: HTTP/1.1 200 OK - 4104 bytes in 0.149 second response time
	HTTP-GANGLIA	OK	09-29-2015 15:55:02	9d 4h 5m 33s	1/3	HTTP OK: HTTP/1.1 200 OK - 133368 bytes in 2.986 second response time
	PBSNODES-DOWN	OK	09-29-2015 15:55:00	0d 1h 32m 29s	1/3	OK: pbsnodes down:0
	SSH-Check-load-8	OK	09-29-2015 15:54:16	10d 21h 53m 19s	1/3	OK: load (0.30) is below threshold (8/12) - load=0.30
	SSH_Disk_Free	OK	09-29-2015 15:53:35	44d 21h 4m 0s	1/3	OK: All Filesystems are below threshold (85/90%) [/=32% /var=57% /boot=32%]

Usage monitoring

- User view: cluster free ?
- Admin view: clusterutilised?

Usage monitoring





User support

- documentation
- presentations

- mailing list
- helpdesk/trouble ticket

- HPC team of people
  - general Linux,
  - shell scripting,
  - network management,
  - storage management,
  - batch system configuration,
  - compilation of scientific software

Conclusion:

commodity cluster can be built and operated with success

The problem/Function:

"make parallel/HPC apps run well for your users"

Cluster Structure:

… fairly complex …

- But it CAN be done!

The sum: hw + sw + motivated team

will make your cluster run

Thank You!

Questions ?