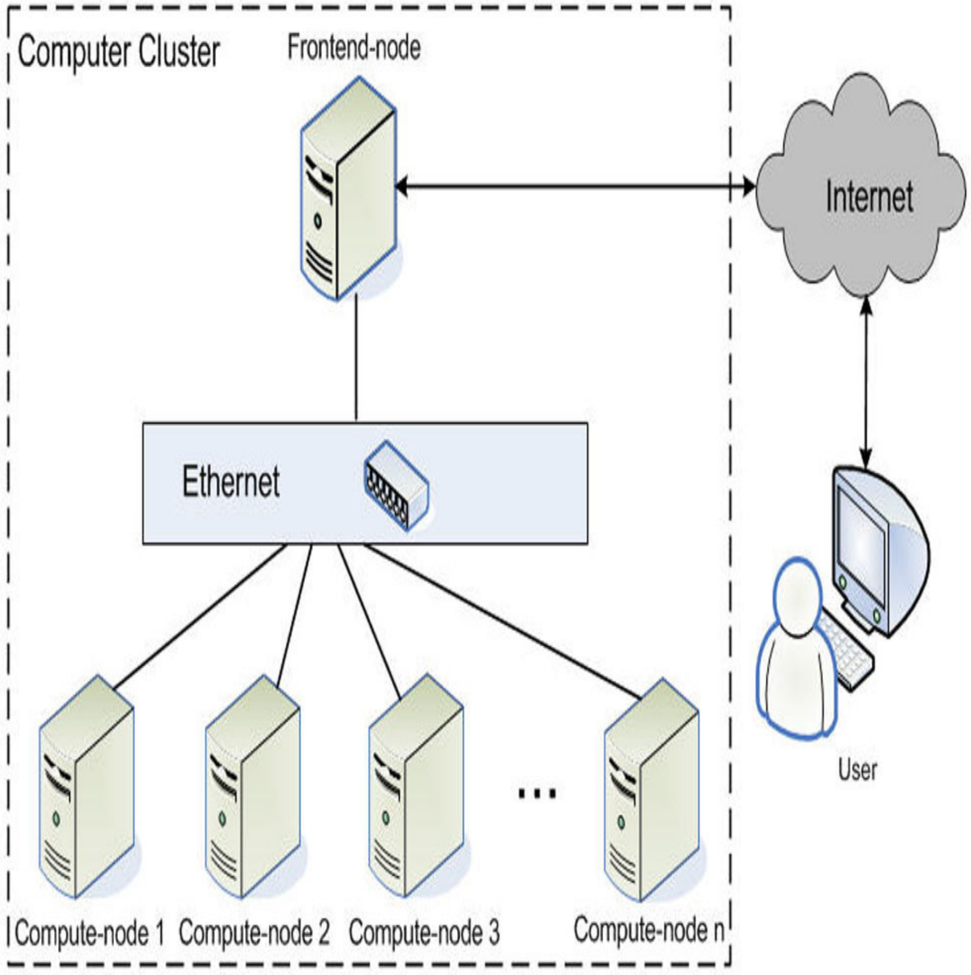
Installation and configuration of Linux Cluster

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

What is Cluster computer?

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



Types of clusters

- **Storage clusters** - provide a consistent file system image across servers in a cluster, allowing the servers to simultaneously read and write to a single shared file system

- **High-availability clusters** - provide continuous availability of services by eliminating single points of failure

- **Load-balancing clusters** - dispatch network service requests to multiple cluster nodes to balance the request load among the cluster nodes

- **High-performance clusters** - use cluster nodes to perform concurrent calculations by allowing application to work in parallel

Cluster Components

Hardware

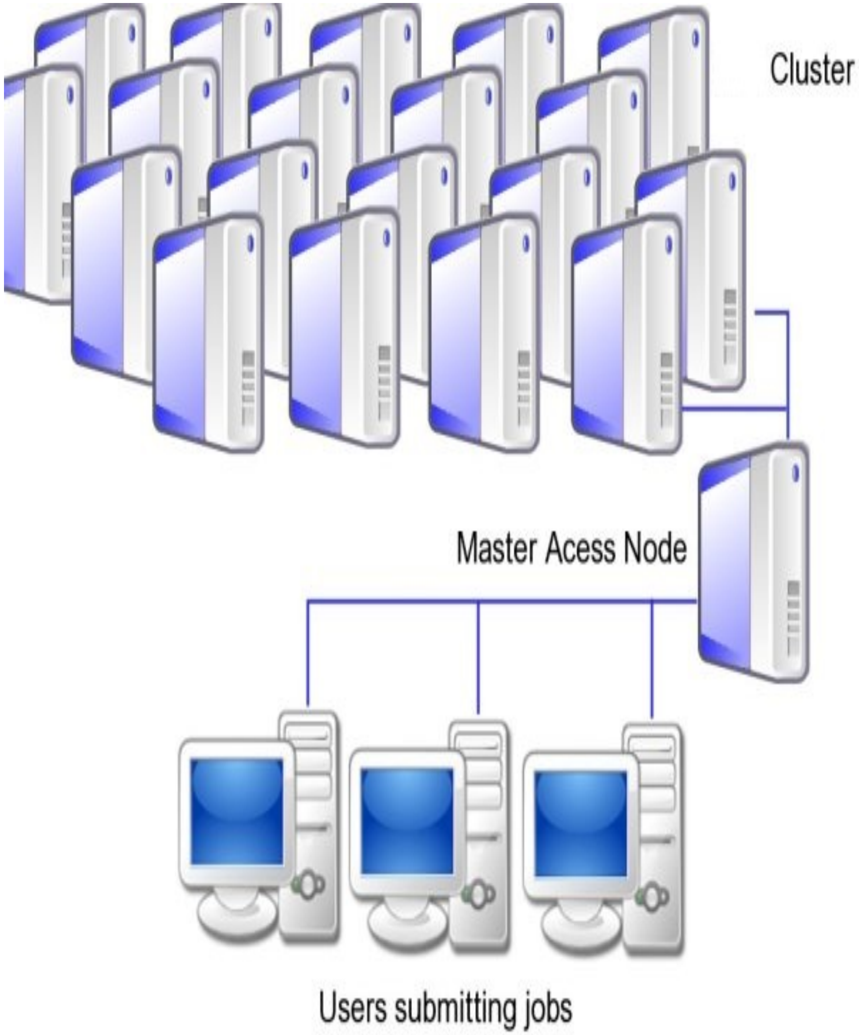
- Computers(Nodes)
- Disk array
- Network devices
- Backup device
- Admin front end
- UPS
- Rack units

Software

- Operating system
- MPI
- Compilers
- Scheduler

Nodes

- They are broadly classified as computing nodes and master node
- Master node handle the communication between the users and the computing nodes
- It also run a number of services to manage and control the executions in computing nodes
- Computing nodes are responsible for taking care of the actual numerical executions



Operating System

- Varies flavors of Linux can be used such as Centos, Red Hat, Debian and others
- First the master node should be installed and all the services for the cluster is configured
- Usually installation for the master node is interactive
- Non interactive way of installation and configuration is usually used for computing nodes
- One can use disk-based or disk-less computing nodes

Services running on Master nodes

- DNS(/etc/hosts)

  Used to resolve names dynamically
- DHCP

  Provide IP address dynamically
- TFTP

  Transfer data during the configurations of computing nodes
- NTP

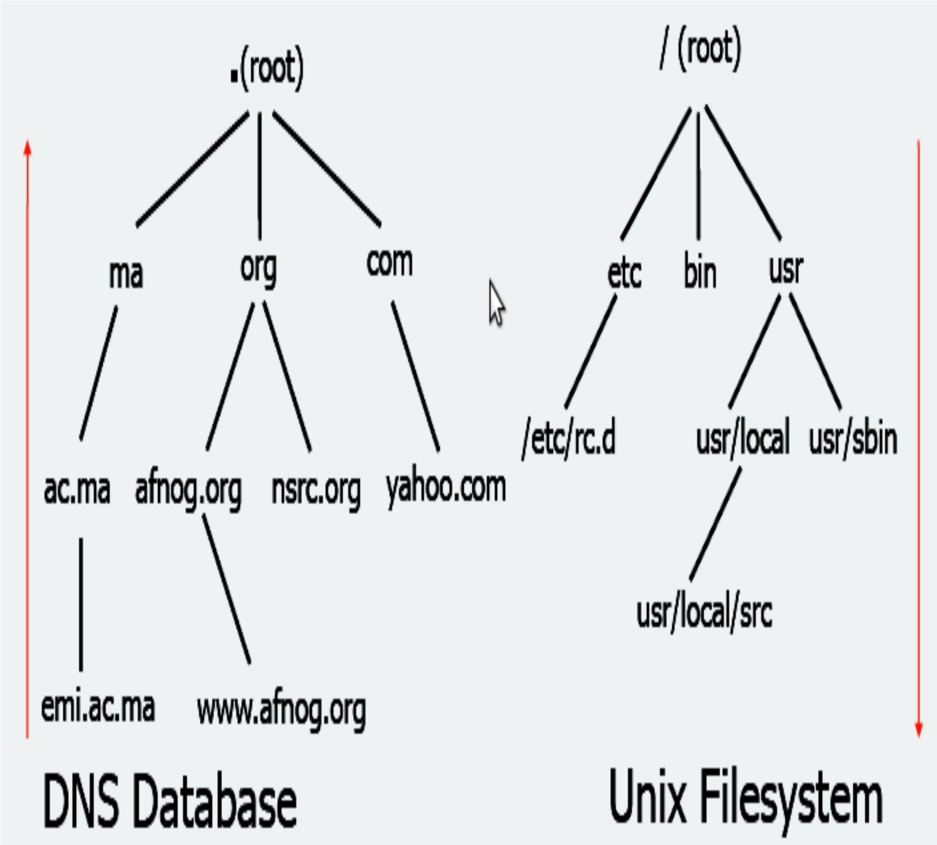
  Maintain time sync
- NFS

  Handle shared file systems
- SSH

  Remote login and file transfer
- Others...

DNS(Domain Name system)

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

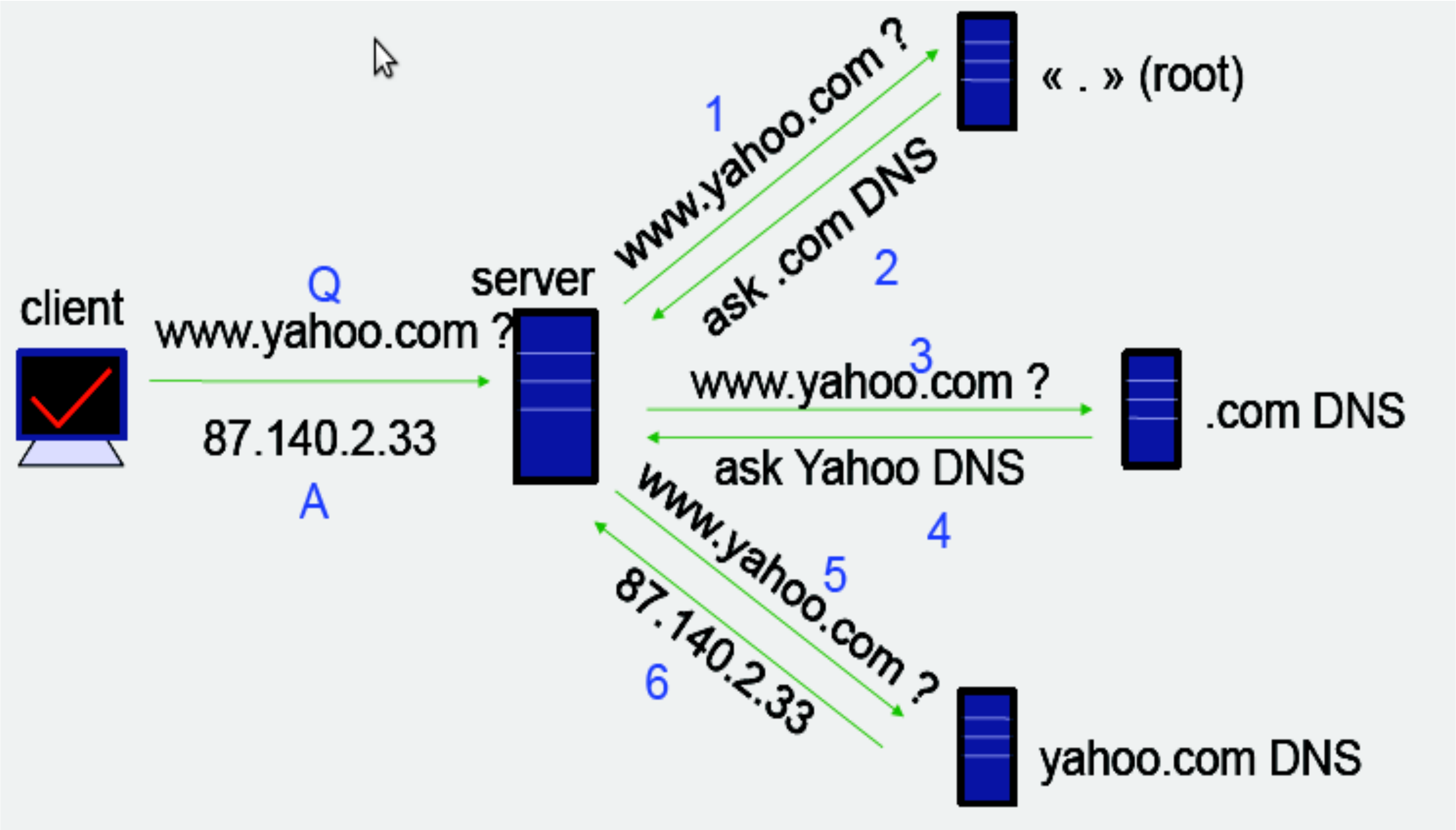


From Afnog Workshop

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 is referred to other servers

A DNS query


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

/etc/hosts

- It is a plain file used to map host names to IP addresses
- In the absence of name server, any network program on your system consults /etc/hosts file to determine the IP address that corresponds to the host name
- The leftmost column is the IP address to be resolved. The next column is for the host's name. Any subsequent columns are alias for that host

```
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
```

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";
         }
```

DHPC 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 serve as 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" ; }
```

TFTP(Trival file transfer protocol)

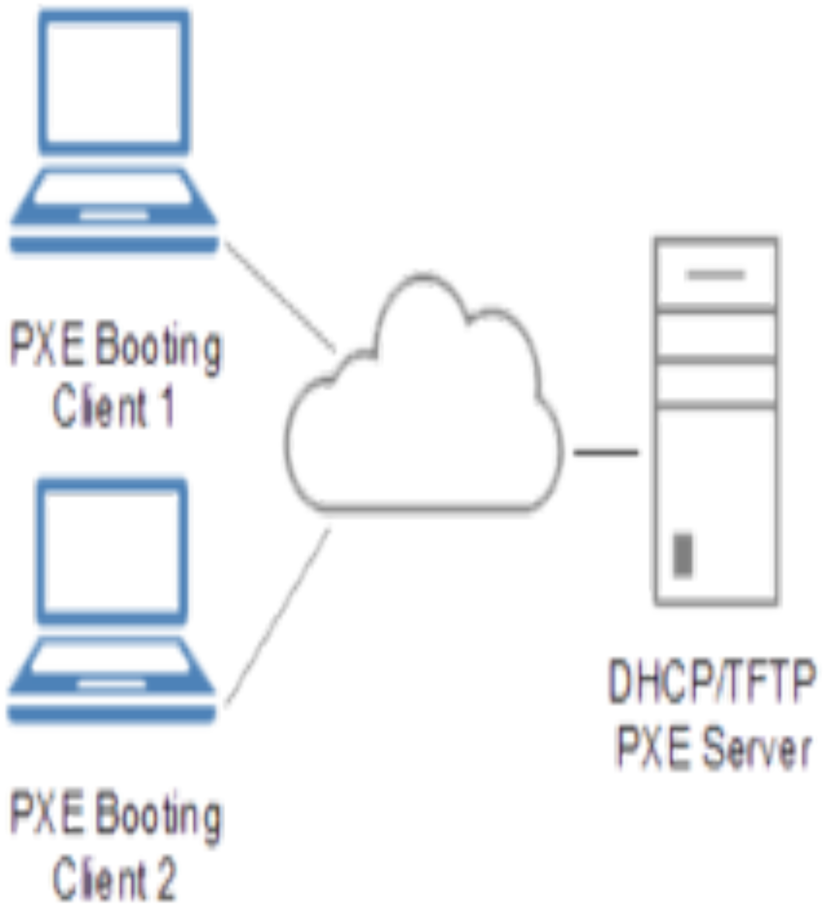
- The protocol is extensively used to support remote booting of diskless devices
- The server is normally started by inetd, but can also run standalone
- TFTP services does not require an account or password on the server system.
- Due to the lack of authentication information,tftpd will allow only publicly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

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



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

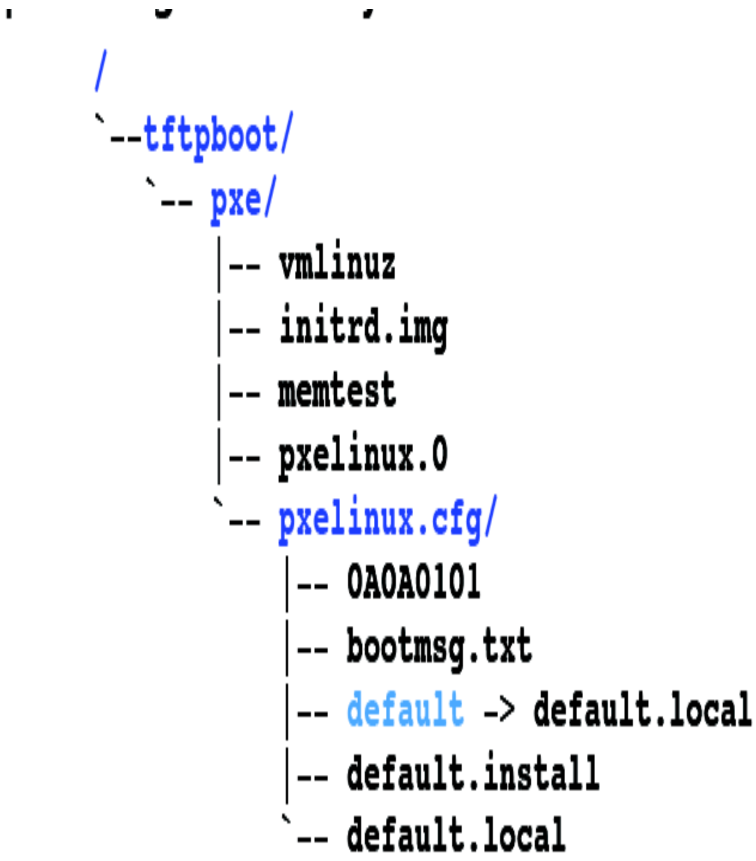
- Usually hexadecimal equivalent of pxe configuration file is used to install a node and the “default” is used when one wants to boot the node from the local hard disk

- /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

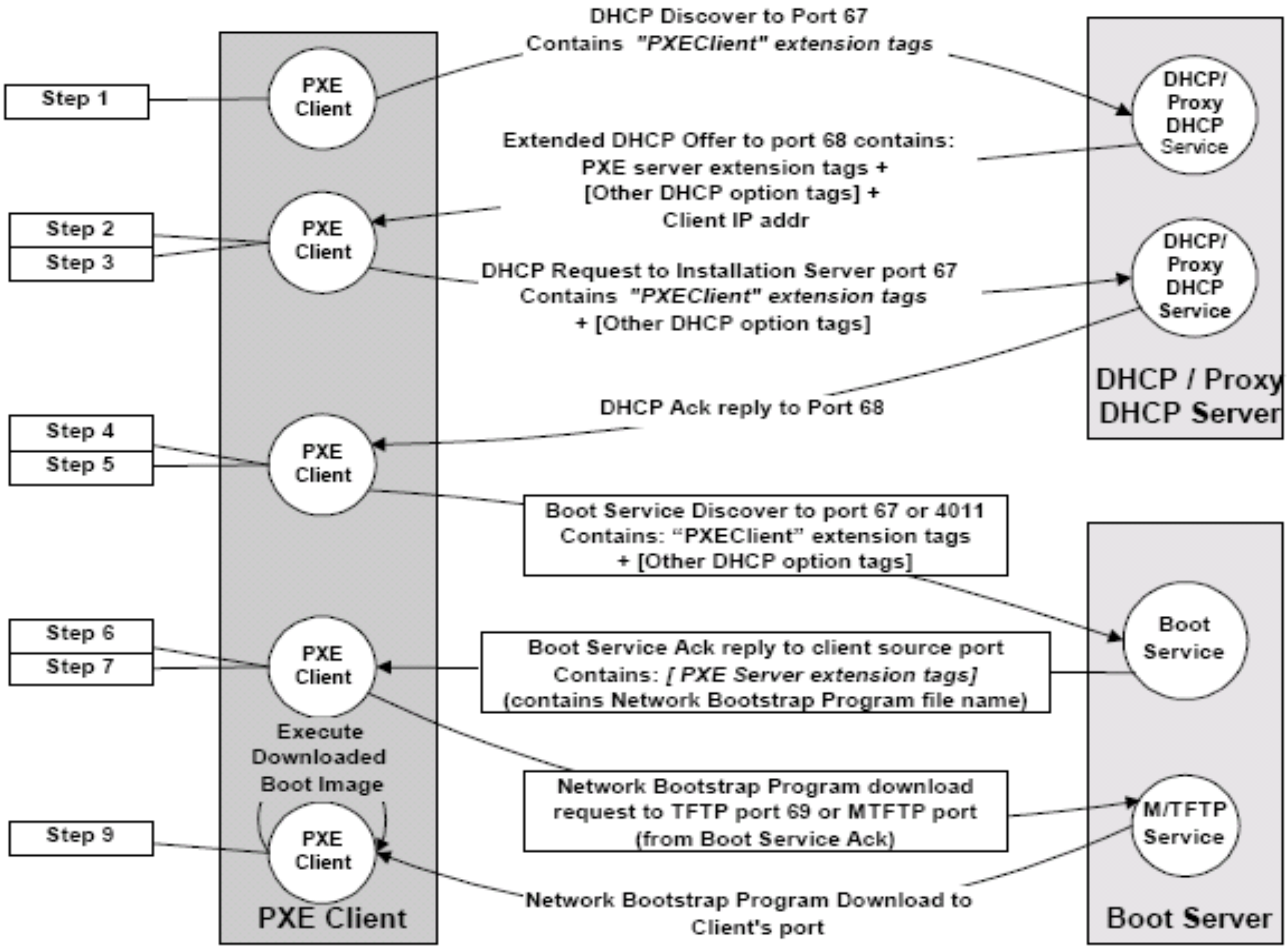
- ➢Gethostip - converts the given hostname or IP address into complete hexadecimal representation for a given IP address, or a partial hexadecimal representation to match a range of IP addresses
- ➢Make sure the TFTP directory includes files shown in diagram



```
/
`--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....



Step 1
PXE Client
DHCP Discover to Port 67
Contains "PXEClient" extension tags
DHCP/ Proxy DHCP Service
Extended DHCP Offer to port 68 contains:
PXE server extension tags +
[Other DHCP option tags] +
Client IP addr
Step 2
Step 3
PXE Client
DHCP Request to Installation Server port 67
Contains "PXEClient" extension tags
+ [Other DHCP option tags]
DHCP/ Proxy DHCP Service
DHCP / Proxy DHCP Server
DHCP Ack reply to Port 68
Step 4
Step 5
PXE Client
Boot Service Discover to port 67 or 4011
Contains: "PXEClient" extension tags
+ [Other DHCP option tags]
Boot Service
Step 6
Step 7
PXE Client
Boot Service Ack reply to client source port
Contains: [PXE Server extension tags]
(contains Network Bootstrap Program file name)
Execute Downloaded Boot Image
Network Bootstrap Program download
request to TFTP port 69 or MTFTP port
(from Boot Service Ack)
M/TFTP Service
Step 9
PXE Client
Network Bootstrap Program Download to
Client's port
PXE Client
Boot Server

Kickstart Installations

- It is automated installation method to install operating system in our machine
- Kickstart installations can be performed using a local CD-ROM, a local hard drive, 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
- It can be created using the Kickstart Configurator application or by writing it from scratch.
- For detail go to http://fedoraproject.org/wiki/Anaconda/Kickstart

Sample kickstart file

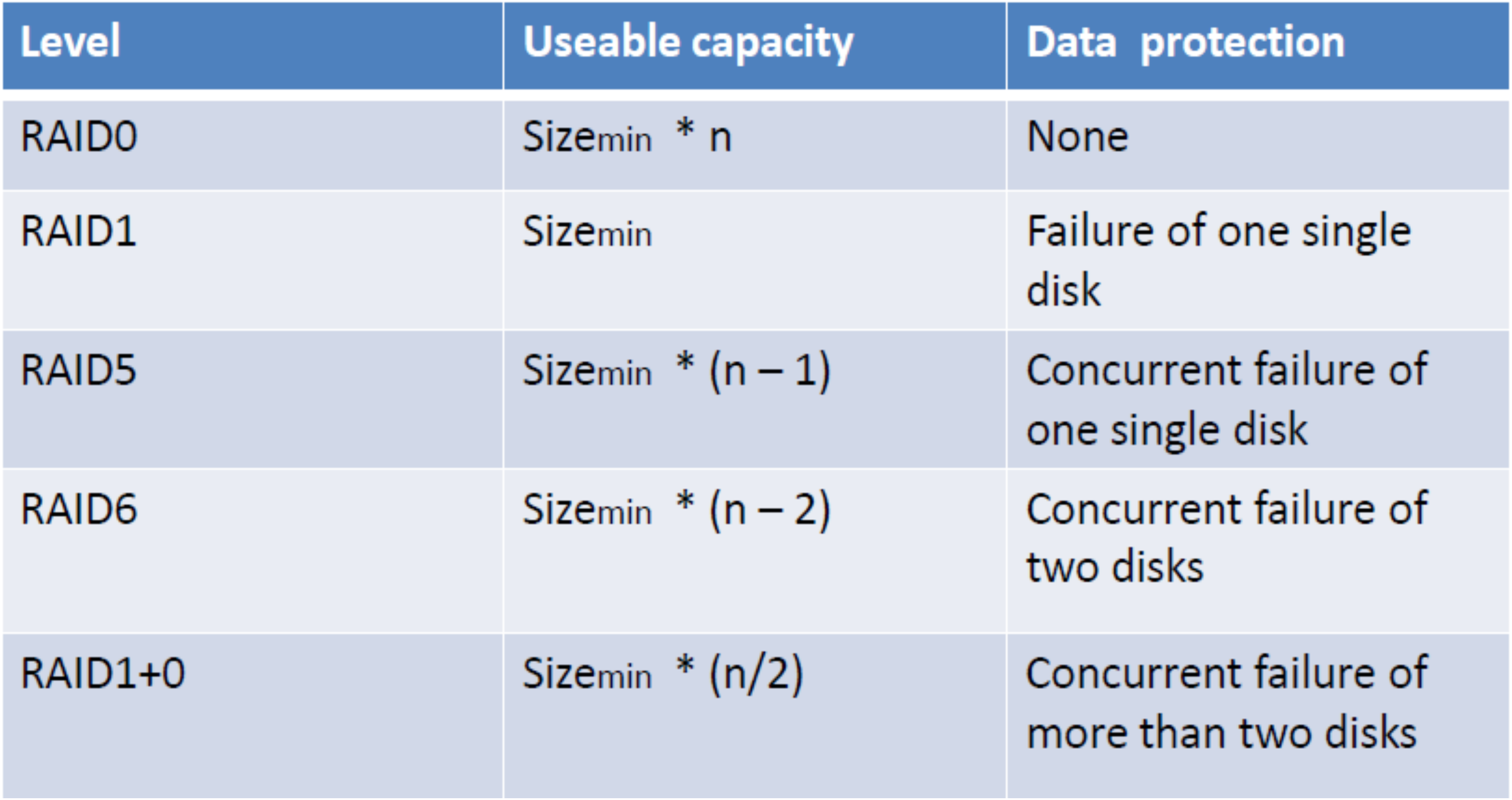
- Kickstart files allowing us to configure network, system configurations, HD partitioning and package selections
- In the pre-installations section one can choose hardware setup and configurations
- In the post-installations section one can include customizations and additional configurations.
- It 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
```

Disk array

- It is a hardware element that contains a large group of hard disk drives (HDDs)
- RAID is configured over these disks to improve performance and fault tolerance
- NFS is used to mount these disks for the cluster.

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

NFS(Network File System)

- It allows client computers to access files over a network in the same way as local storage is accessed
- In our configuration we used it to create local repository by copying the RPM package to exported directory
- One can change the configuration of NFS by editing the /etc/export file

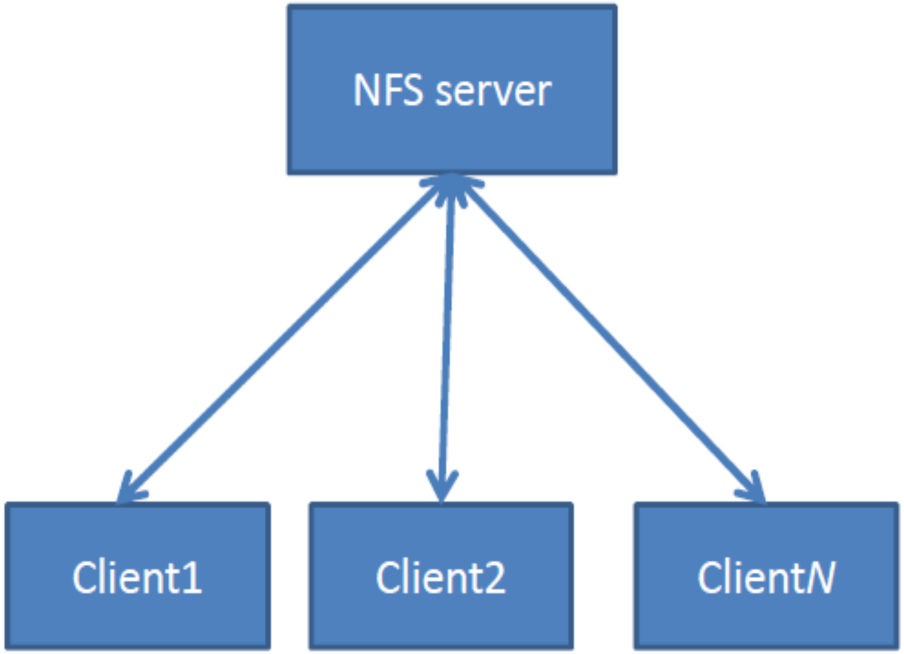
- /etc/export

```
/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 architecture

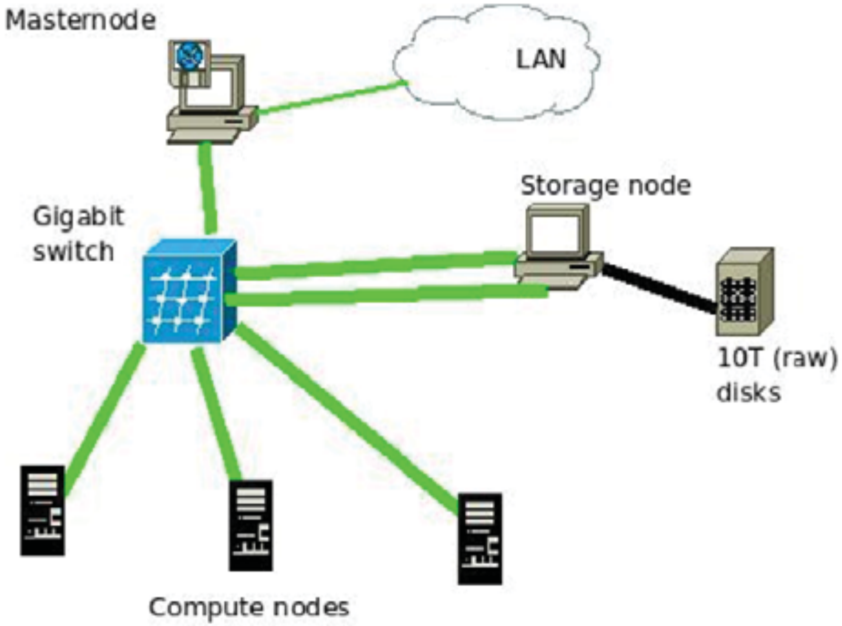


NFS server

Client1

Client2

Client*N*

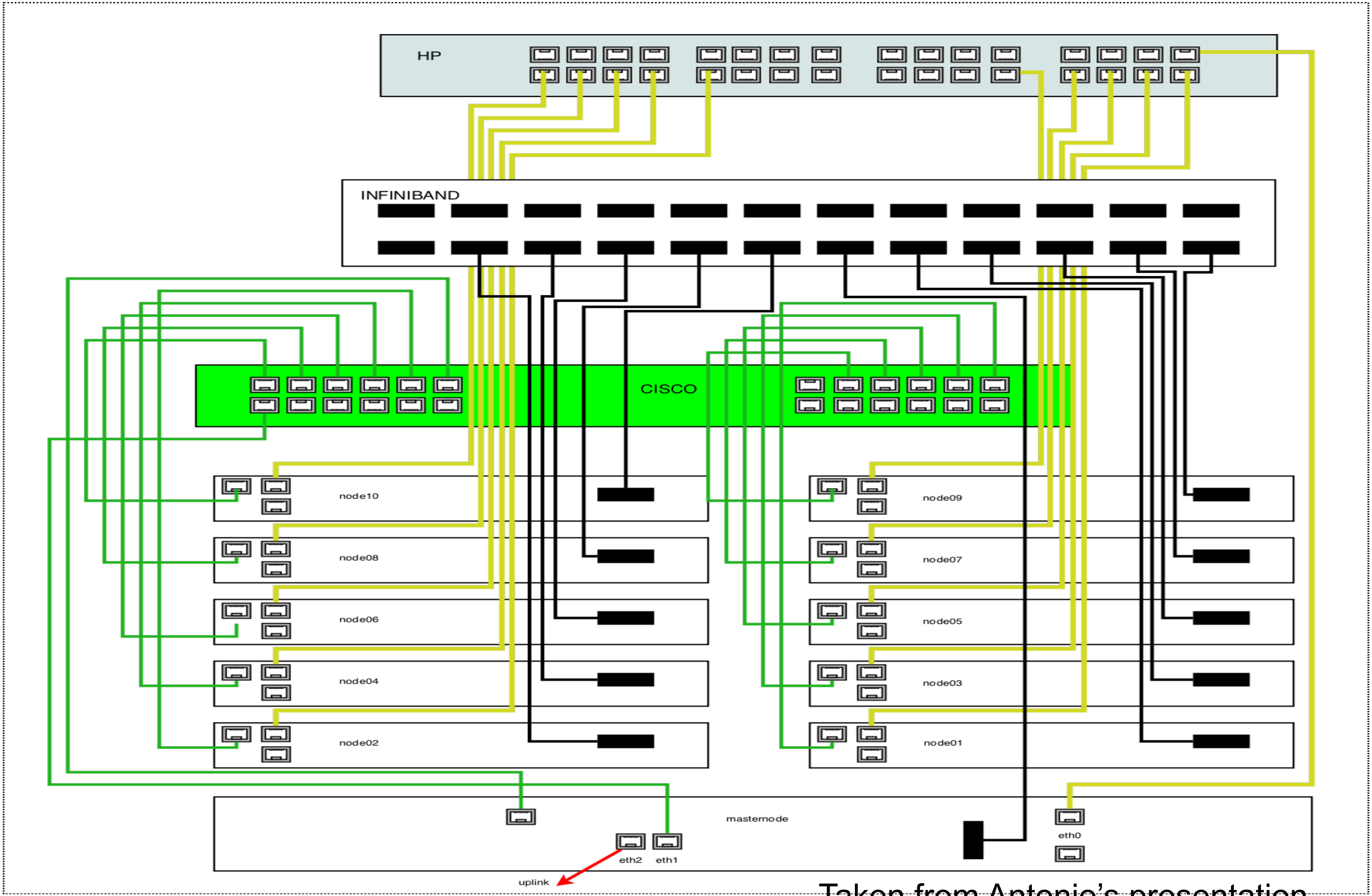


Taken from Clement’s slide

Network Devices

- Infiniband and Myrinet are used for high speed network mainly for parallel computation because they provide both low latency and high bandwidth
- GigaBit switch is used for I/O network(NFS), it has good bandwidth but with high latency
- Fast Ethernet is mainly used to handle management traffic

Network diagram of Addis HPC



Taken from Antonio's presentation

Admin front end

- Console(keyboard, monitor and mouse)
- KVM switches
- KVM cables

M@ITUO
8 PORT USB KVM SWITCH
PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8
MT-801 UK
CONSOLE
POWER
USB1 USB2 USB3
PC8 PC7 PC6 PC5 PC4 PC3 PC2 PC1

Cluster Management tools

- For this tools to run properly it requires the following:
- 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

Software Environment Management, 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 command

➢ module avail

➢ module list

➢ module load/unload <packages>

➢ module purge and others

Software Environment Management, 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 command
- ➢ module avail
- ➢ module list
- ➢ module load/unload <packages>
- ➢ module purge and others

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)
- Resource manager and Scheduler

  PBS/Torque and Maui scheduler

Thank You

&

Let’s go to Hands On