Multicore Computing
Lecture23 – How to Build Your Own Cluster

SUNG KYUN KWAN UNIVERSITY

남 범 석

bnam@skku.edu

Essential Components of Cluster

- Authentication
  - NIS, LDAP
- Cluster File System
  - NFS, AFS
- Clock Synchronization
  - NTP

The Beginning

- Thomas Sterling and Donald Becker CESDIS, Goddard Space Flight Center, Greenbelt, MD
- Summer 1994: built an experimental cluster
- 16 x 486DX4, 100MHz processors
- 16MB of RAM each, 256MB in total
- Channel bonded Ethernet (2 x 10Mbps)
- Called their cluster Beowulf

What is a Beowulf ? Why Beowulf?

- What is a Beowulf?
  - Massively parallel computer
  - Runs a free operating system
  - Connected by high speed interconnect

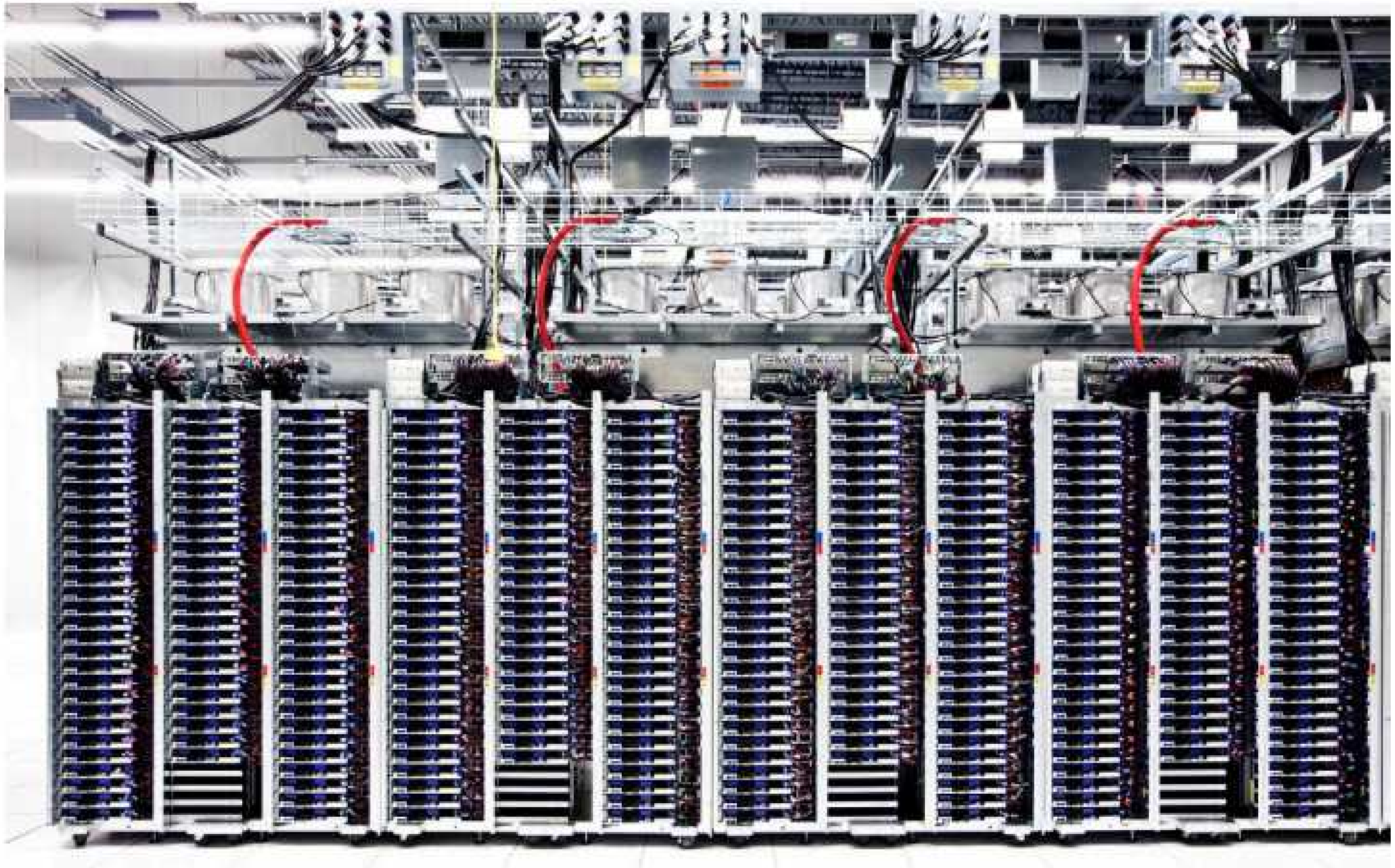
- Why Beowulf?
  - It's cheap! (Good for start-ups)
  - Reliability in software rather than in specialized hardware
  - Everything in a Beowulf is open-source and open standard
  - Easier to manage/upgrade

Google Circa 1997

- google.stanford.edu
  - Eric and Sergey (Google Founders) volunteered to receive shipments of computers that other research groups order and hold on them for sometime.

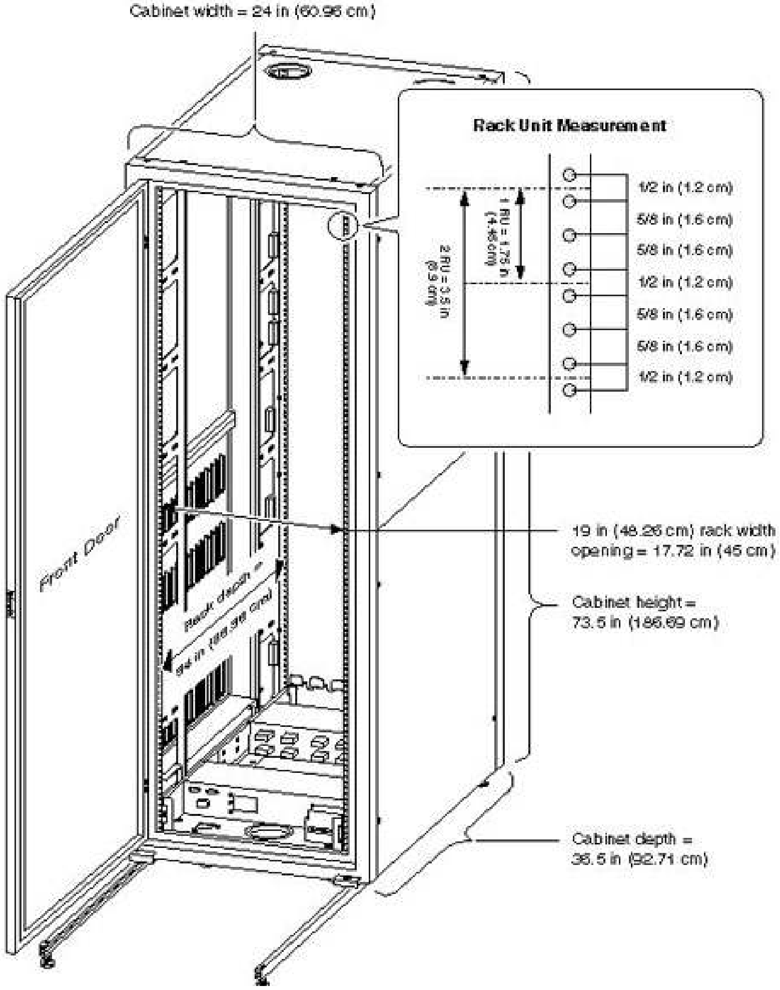
Google Datacenters

- Single aisle in Google data center (Circa 2012)



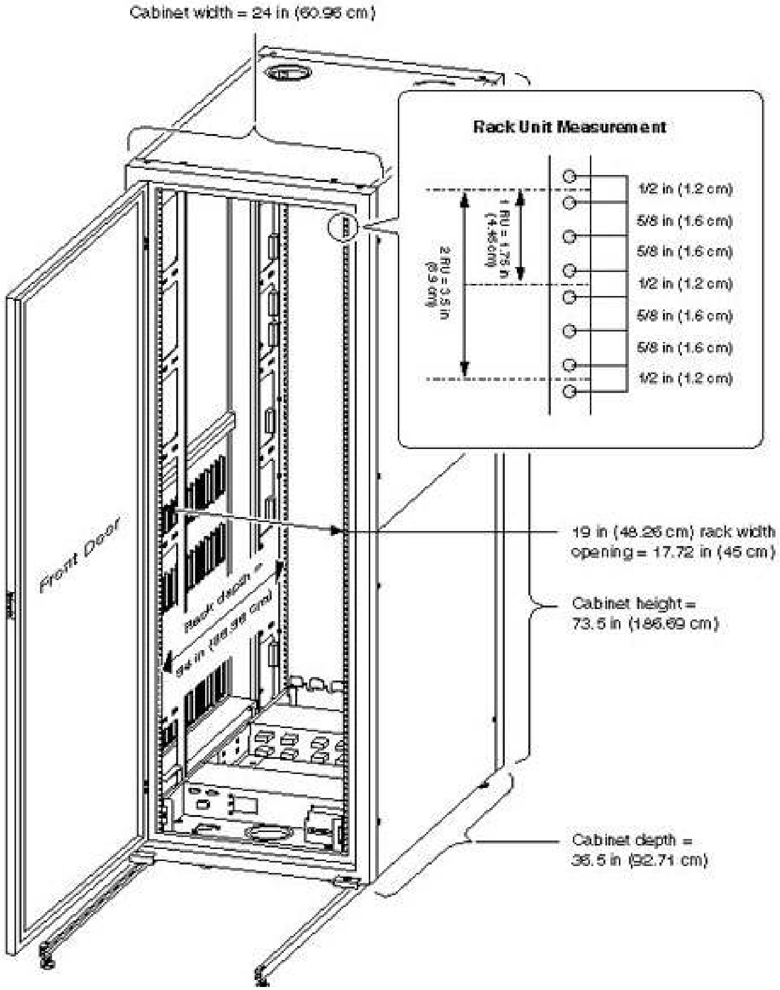
Rack Mountable Server (1U, 2U, 4U)

- Each node is a computer


Cabinet width = 24 in (60.96 cm)
Rack Unit Measurement
1 RU = 1.75 in (4.45 cm)
2 RU = 3.5 in (8.9 cm)
1/2 in (1.2 cm)
5/8 in (1.6 cm)
5/8 in (1.6 cm)
1/2 in (1.2 cm)
5/8 in (1.6 cm)
5/8 in (1.6 cm)
1/2 in (1.2 cm)
Front Door
Rack depth = 34 in (86.36 cm)
19 in (48.26 cm) rack width opening = 17.72 in (45 cm)
Cabinet height = 73.5 in (186.69 cm)
Cabinet depth = 36.5 in (92.71 cm)

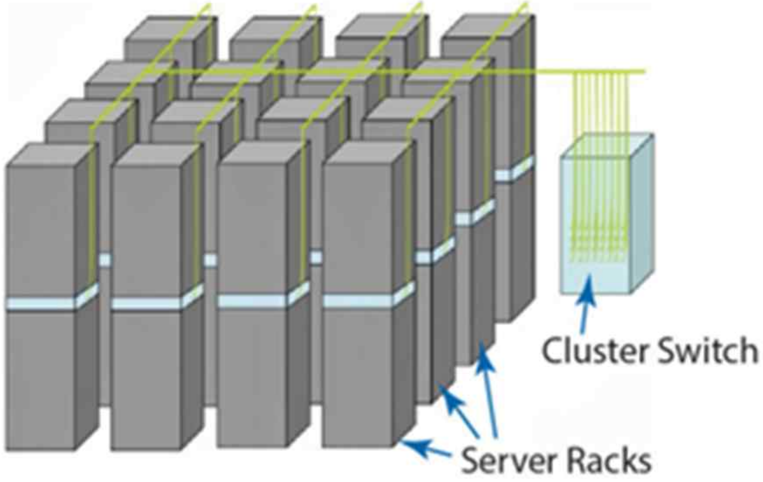
Rack Mountable Server (1U, 2U, 4U)

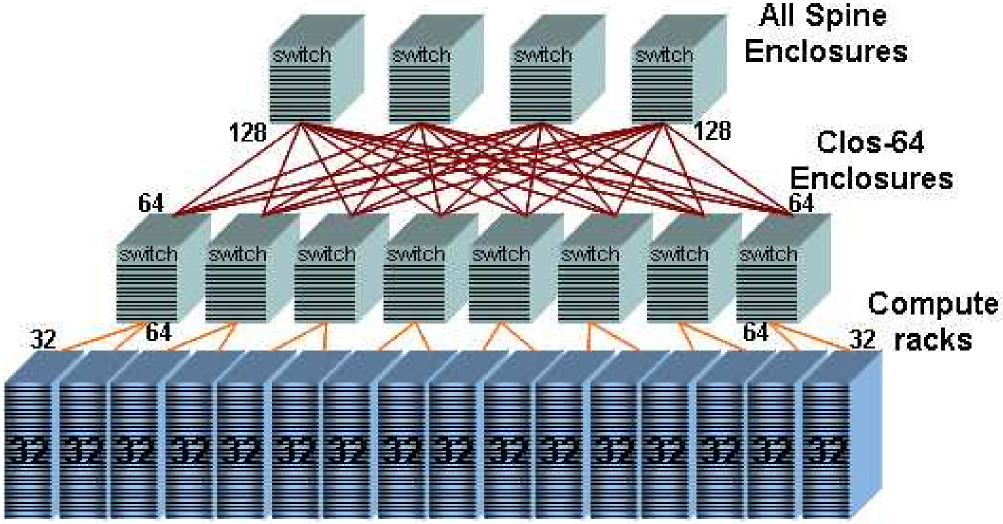
- Blades
  - Shared power supply, cooling, etc


Cabinet width = 24 in (60.96 cm)
Rack Unit Measurement
1 RU = 1.75 in (4.45 cm)
2 RU = 3.5 in (8.9 cm)
1/2 in (1.2 cm)
5/8 in (1.6 cm)
5/8 in (1.6 cm)
1/2 in (1.2 cm)
5/8 in (1.6 cm)
5/8 in (1.6 cm)
1/2 in (1.2 cm)
Front Door
Rack depth = 34 in (86.36 cm)
19 in (48.26 cm) rack width opening = 17.72 in (45 cm)
Cabinet height = 73.5 in (186.69 cm)
Cabinet depth = 36.5 in (92.71 cm)

Data Center

- Fat-Tree Interconnect in warehouse scale





Lonestar: 512 Compute Nodes

Software

- Three choices
  - Free cluster management software (a patched up Linux)
    - oneSIS, OpenHPC, Rocks, Stacki, Warewulf, etc
  - Commercial cluster management software
    - Are you rich?
  - ***Vanilla Linux (RedHat, CentOS, Ubuntu, FreeBSD, etc)***
    - ***Authentication: NIS, LDAP, etc***
    - ***Cluster File System: NFS, AFS, etc***

NIS

1. Create an user foo
on one machine.

Can the user foo login
on another machine?

2. User foo changes passwd
on one machine.

Can the user foo login
on another machine with the new passwd?

NFS

3. User foo create a file bar
on one machine.

Can the user foo access bar
on another machine?

NIS - Network Information System

- Released by Sun Microsystems in 1980
- Originally called Sun Yellow Pages
  - Commands have prefix “yp”
    - eg. ypcat, ypwhich, ypinit, ypdomainname, etc
  - Due to legal reasons, changed the name to NIS
- Client-server model
  - A master server maintains the authoritative copies of system files, such as passwd, and makes the contents available over the network

Network Information Service

- Databases are called NIS maps
  - /etc/passwd
  - /etc/group
  - /etc/netgroup
  - /etc/hosts
  - /etc/networks
  - /etc/protocols
  - /etc/services
  - /etc/aliases
  - /etc/auto_master
  - … …

How NIS works

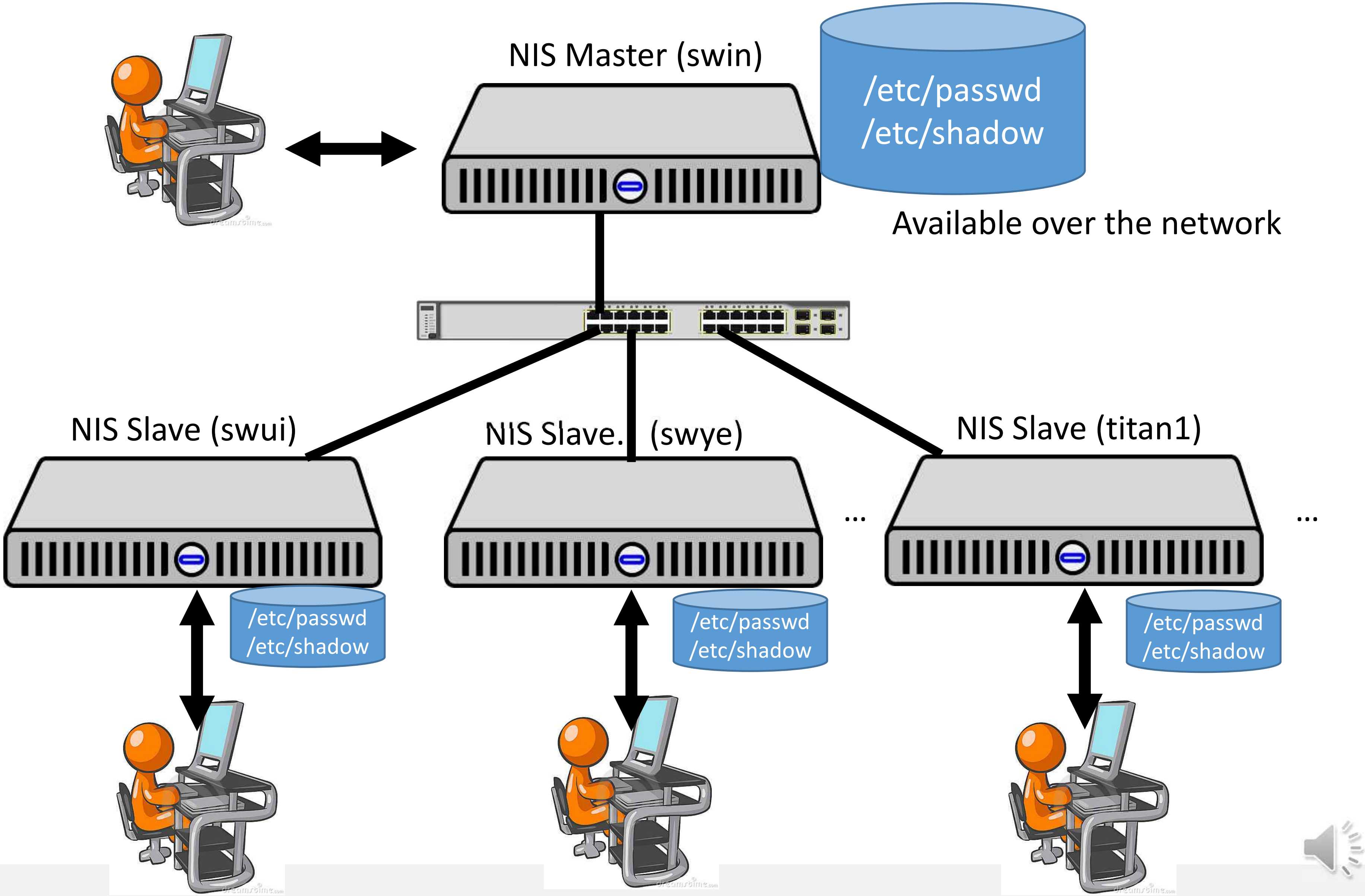
- How NIS works
  - NIS's data files are stored in one directory
    - Usually /var/yp
  - Each NIS map is stored in a hashed format in a subdirectory named for the NIS domain
    - Exact Map files names depends on the hashing library being used.
    - For example:
      - On swin, under /var/yp/inuiyeji, there are ndbm files:

```
bnam@swin:/var/yp/inuiyeji$ ls
group.bygid     netgroup.byuser    rpc.byname
group.byname    netid.byname       rpc.bynumber
hosts.byaddr    passwd.byname      services.byname
hosts.byname    passwd.byuid       services.byservicename
netgroup        protocols.byname   ypservers
netgroup.byhost  protocols.bynumber
bnam@swin:/var/yp/inuiyeji$
```

How NIS works

- There is a makefile under /var/yp
  - Which calls makedbm to generate NIS maps from flat files.
  - After you modify a system file, cd to NIS dir which is /var/yp, and run make.
    - Make checks the modification date and rebuild some maps.
- Slave server run ypxfr command regularly as cron to pull the map copies
- Server runs yppush to instruct each slave server to execute ypxfr.

How NIS Works



How to Setup NIS?

- NIS Master & Slaves
  - $ apt install nis
  - modify /etc/yp.conf, /etc/nsswitch.conf
    - For details, refer to https://www.server-world.info
  - $ systemctl restart rpcbind nis

  - To push accounts and passwords to slaves
    - $ cd /var/yp
    - $ make
      - /etc/passwd is translated into two different NIS Maps
      - passwd.byname
      - passwd.byuid

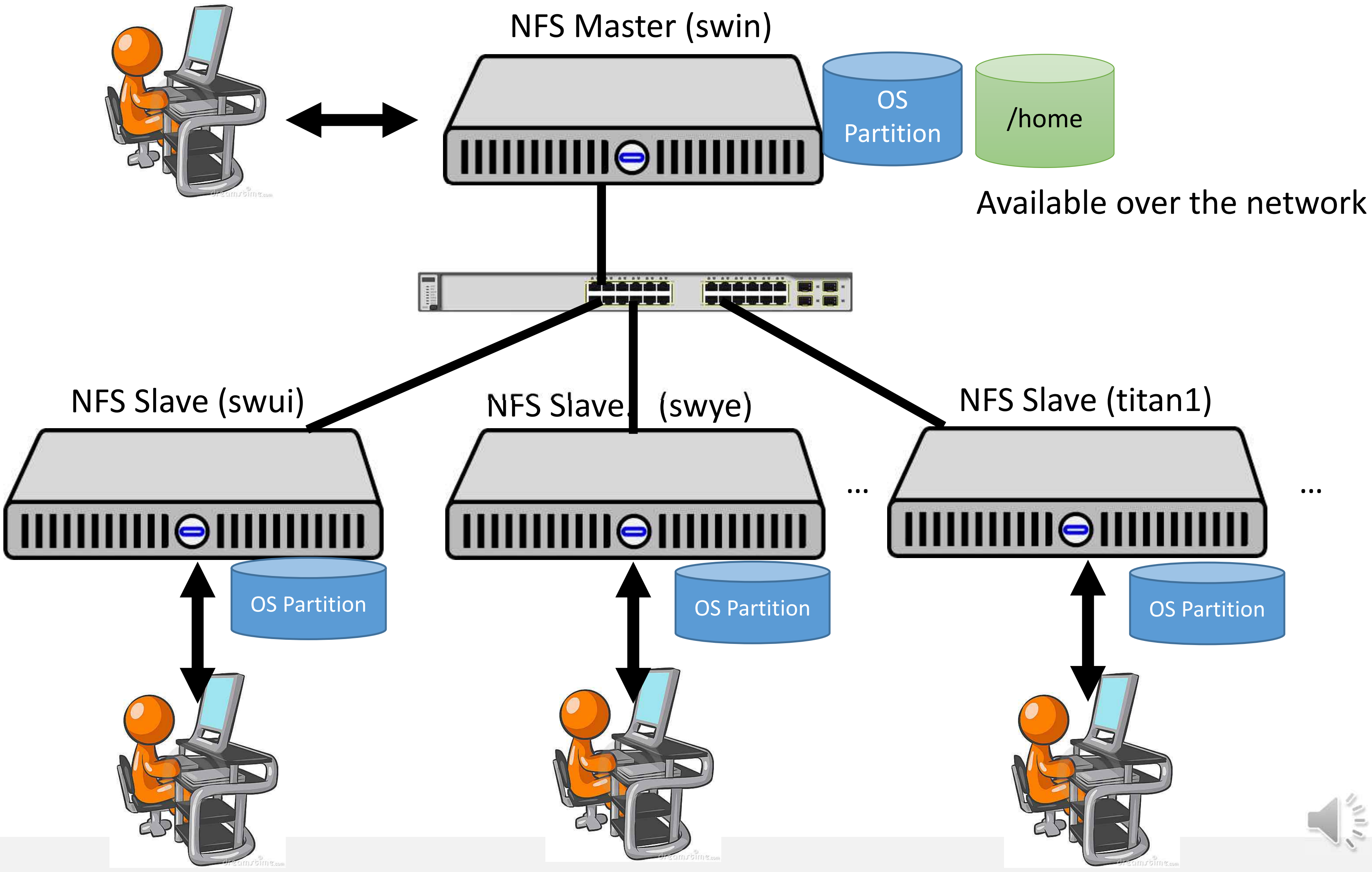
  - To change a user password
    - $ yppasswd

NFS (Network File System)

- NFS (Network File System) allows you to share file systems
- Reasons for sharing file system
  - Transparent to user
    - User can keep use their familiar commands
    - Access the same file from multiple nodes.
  - NFS was introduced by Sun Microsystems in 1985

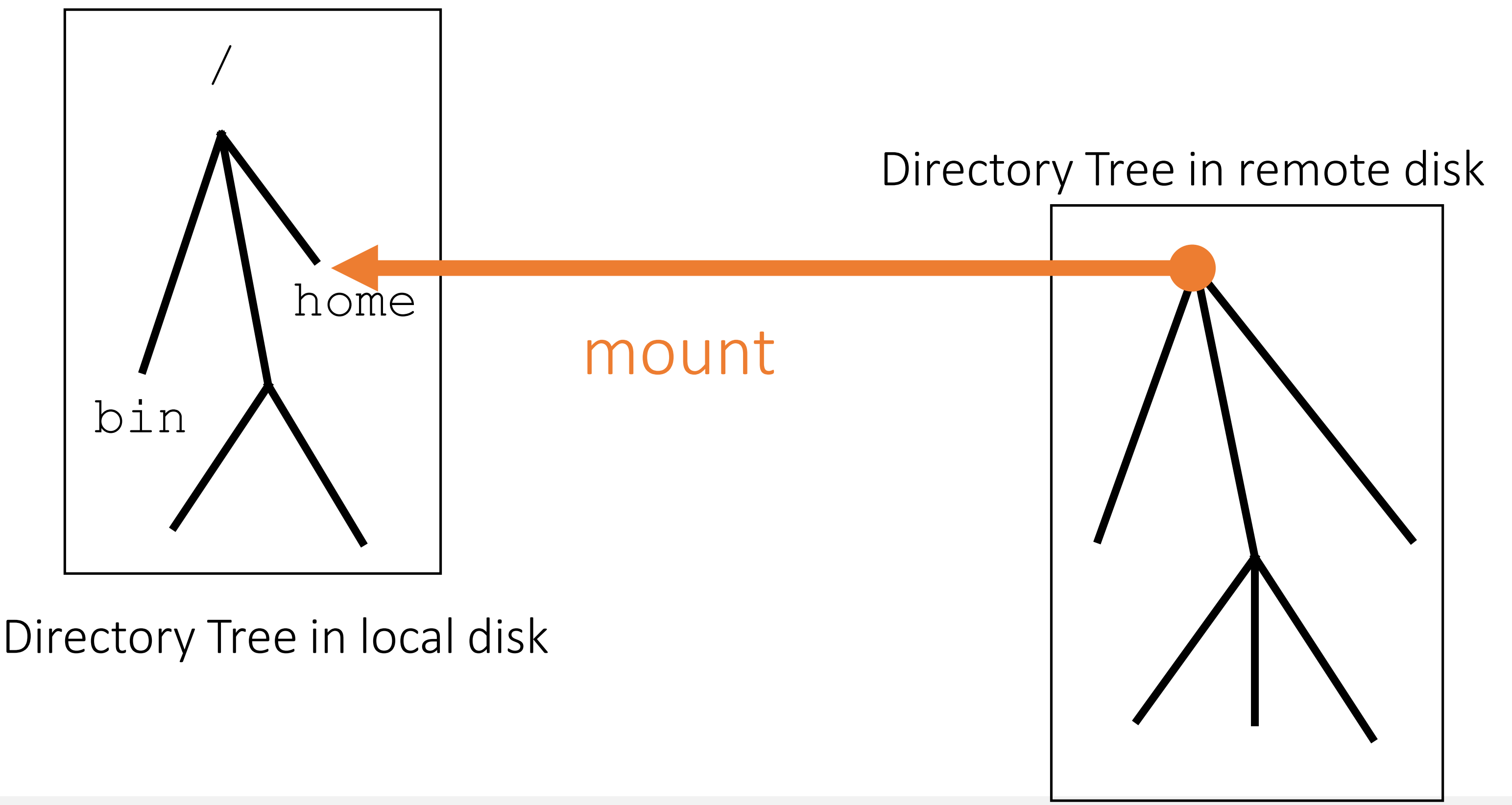
- NFS runs on top of Sun's RPC (Remote Procedure Call) protocol
  - RPC provides a system-independent way for processes to communicate in a client-server fashion over a network.

How NFS Works



Mounting Disks

- **`mount`**
  - Use mount command to establish temporary network mounts



Mounting Disks

- **`mount`**
  - Automatic mounting services such as automount
  - List in /etc/fstab or /etc/vfstabb
  - mount hostname:remote-dir local_dir
    - hostname is the NFS server
    - local_dir must exist already

- Using umount command

  umount local_dir[remote_dir] -l

How to Setup NFS?

- Master
  - $ apt-get install nfs-kernel-server
  - modify /etc/idmapd.conf, /etc/exports
  - $ systemctl restart nfs-server

- Slaves
  - $ apt-get install nfs-common
  - modify /etc/idmapd.conf, /etc/fstab
  - E.g.)

```
bnam@titan2:~$ cat /etc/fstab

/dev/mapper/centos-root /                        xfs     defaults        0 0
UUID=ed3bfc50-4cdd-49f3-88a3-f89c6a9f8291 /boot                xfs     defaults        0 0
/dev/mapper/centos-home /home                    xfs     defaults        0 0
/dev/mapper/centos-swap swap                     swap    defaults        0 0
swin.skku.edu:/home                              /home  nfs  defaults   0  0
```

Distributed time

- Premise
  - The notion of time is well-defined (and measurable) at each single location
  - But the relationship between time at different locations is unclear
    - Can minimize discrepancies, but never eliminate them
- Reality
  - Stationary GPS receivers can get global time with < 1μs error
  - Few systems designed to use this

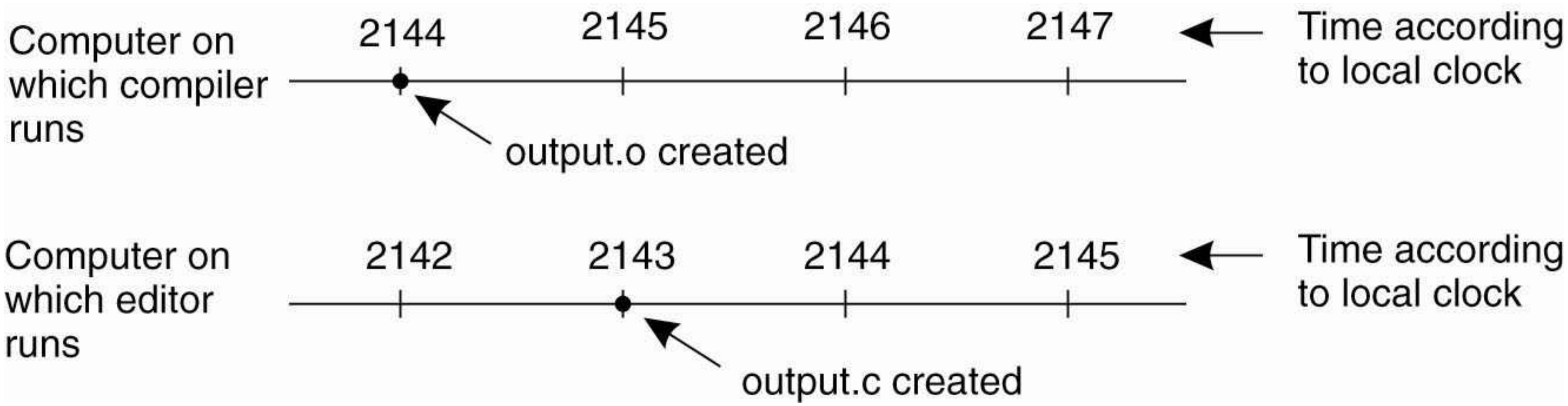
Clock Synchronization Issue

- **`make`** recompiles if *foo.c* is newer than *foo.o*

- Scenario
  - Current time on swin: 1:05pm
  - Current time on swji: 1:00pm
  - **`make`** on "*swin*" to build *foo.o on NFS*
  - Test on *"swji"*, find and fix a bug in *foo.c*
  - Re-run **`make`** on "swji"
  - *Nothing happens!*
- Why?

Clock Synchronization

- Time is unambiguous in centralized systems
  - System clock keeps time, all entities use this for time
- Distributed systems: each node has own system clock
  - Problem: An event that occurred after another may be assigned an earlier time


Computer on which compiler runs
2144
2145
2146
2147
Time according to local clock
output.o created
Computer on which editor runs
2142
2143
2144
2145
Time according to local clock
output.c created

How to Setup NTP?

- NTP server
  - $ apt-get install ntp
  - modify /etc/ntp.conf
  - $ systemctl restart ntp

- NTP client
  - $ apt-get install ntpdate
  - $ ntpdate *ntp.server.hostname*