用企鵝龍打造多人雲端實驗叢集

Building Multiuser Hadoop Testbed with DRBL

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

Powered by DRBL

Programmer v.s. System Admin.

Source:
http://www.funnyjunksite.com/wp-content/uploads/2007/08/programmer.jpg

Source:
http://www.sysadminday.com/images/people/136-3697.JPG

Agenda

PART 1 :

What is Cluster Computing ?

How to deploy PC cluster ?

PART 2 :

What is DRBL and Clonezilla ?

Can DRBL help to deploy Hadoop ?

PART 3 :

Live Demo of DRBL Live

and Clonezilla Live

PART 1 :

PC Cluster 101

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

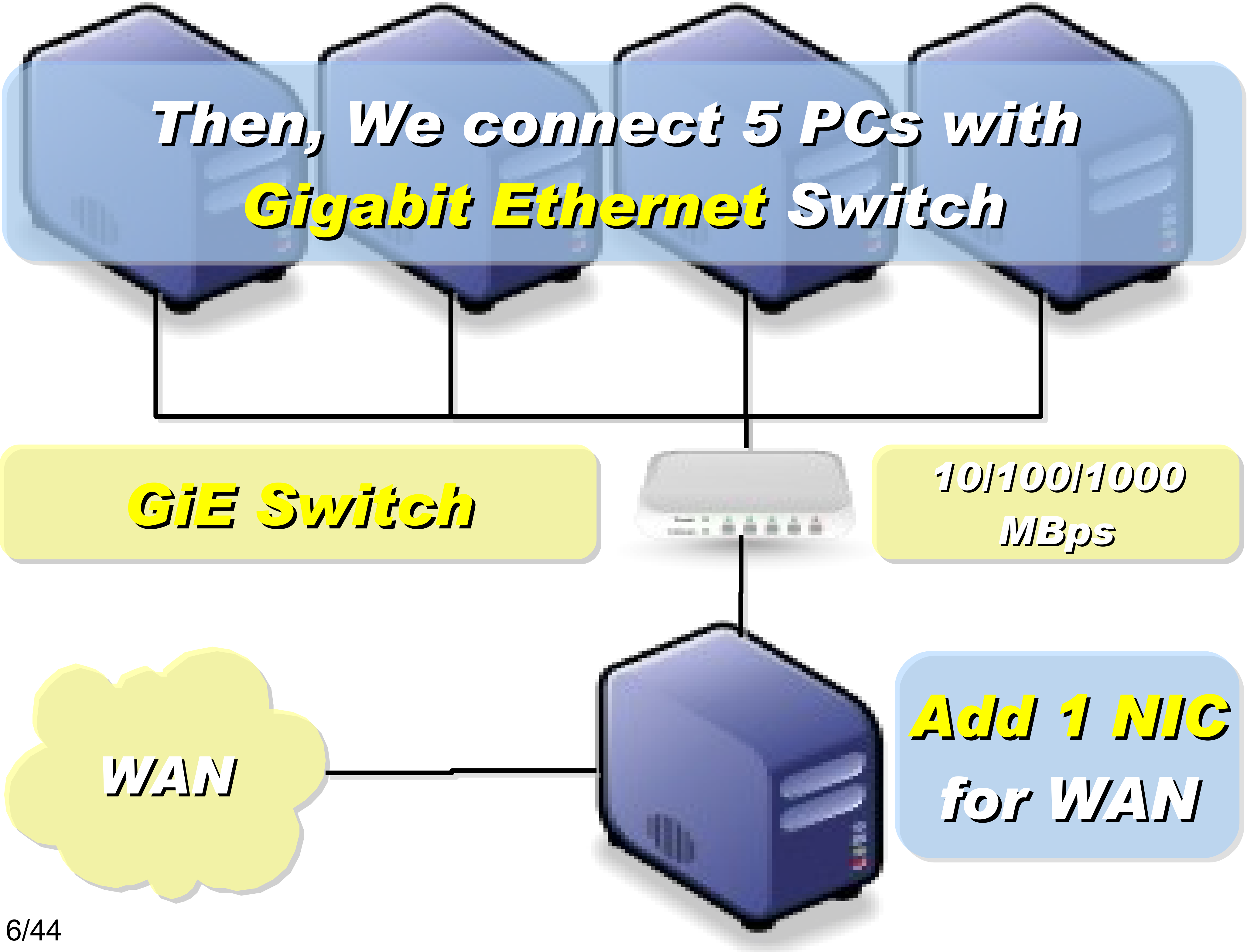
Powered by DRBL

At First, We have " 4 + 1 " PC Cluster

It'd better be 2^n

Manage Scheduler

Then, We connect 5 PCs with Gigabit Ethernet Switch

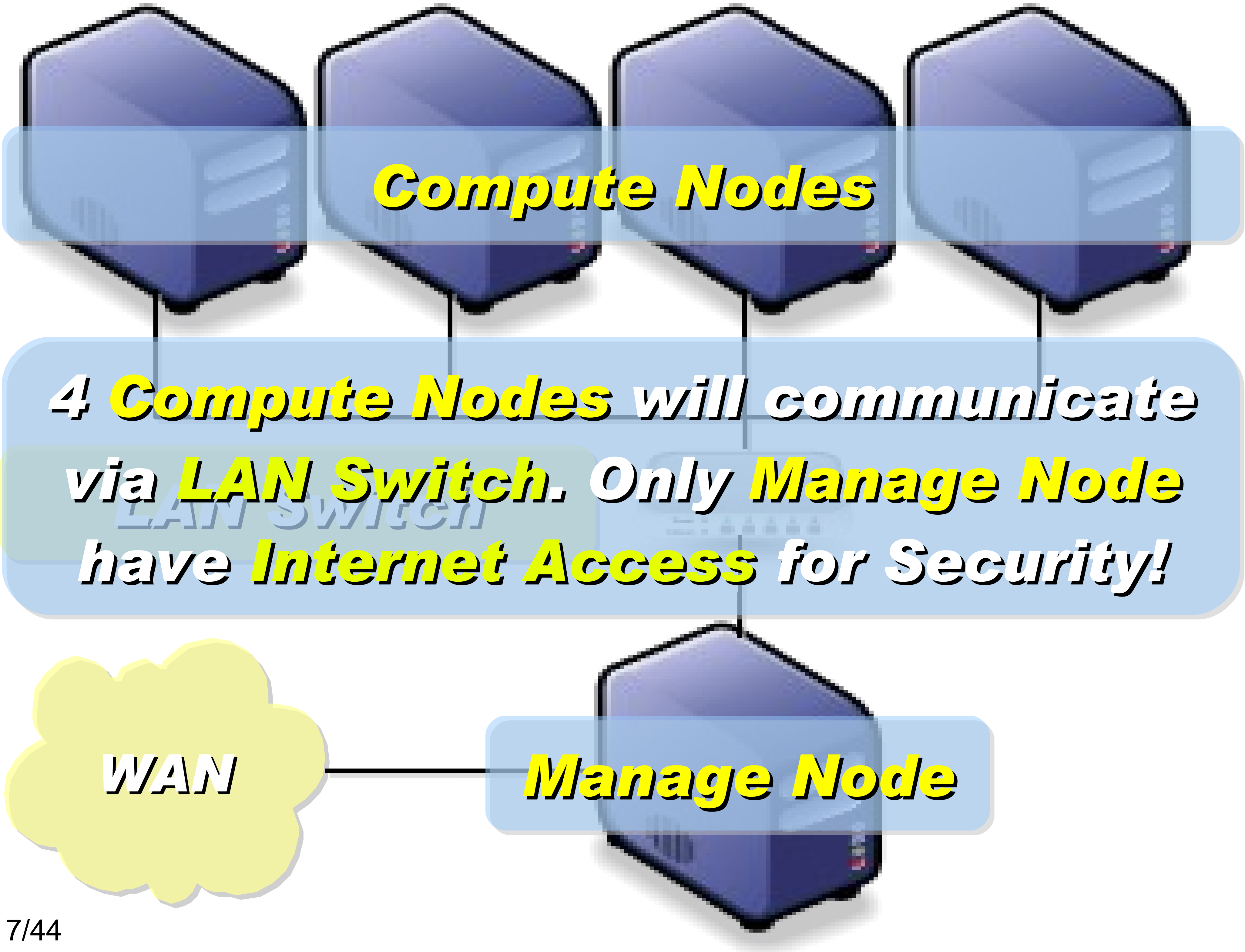


GiE Switch

10/100/1000 MBps

WAN

Add 1 NIC for WAN



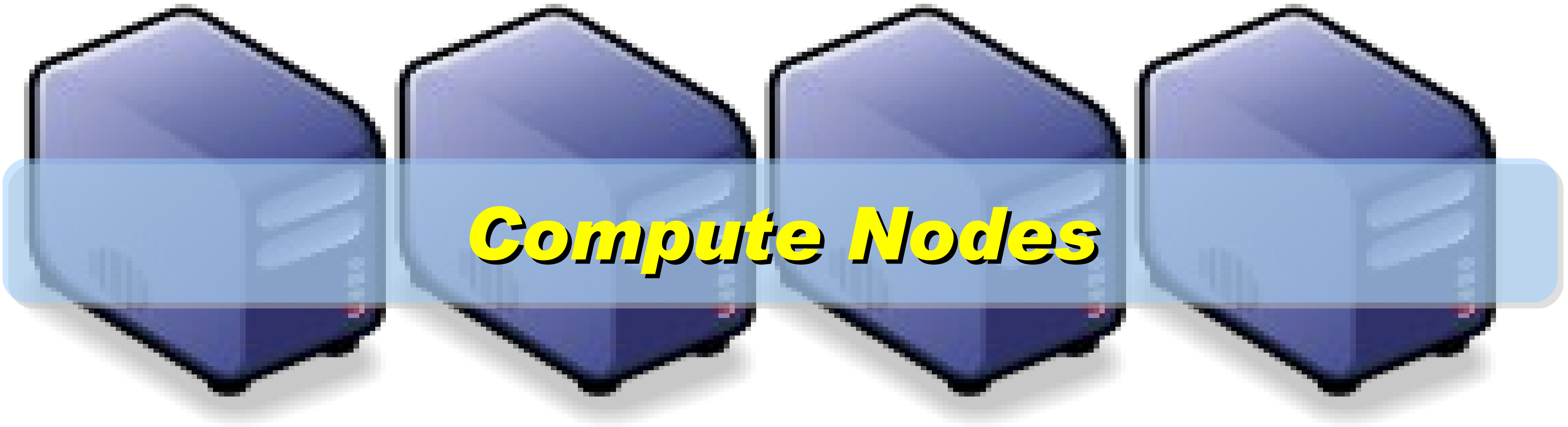
Compute Nodes

4 Compute Nodes will communicate via LAN Switch. Only Manage Node have Internet Access for Security!

LAN Switch

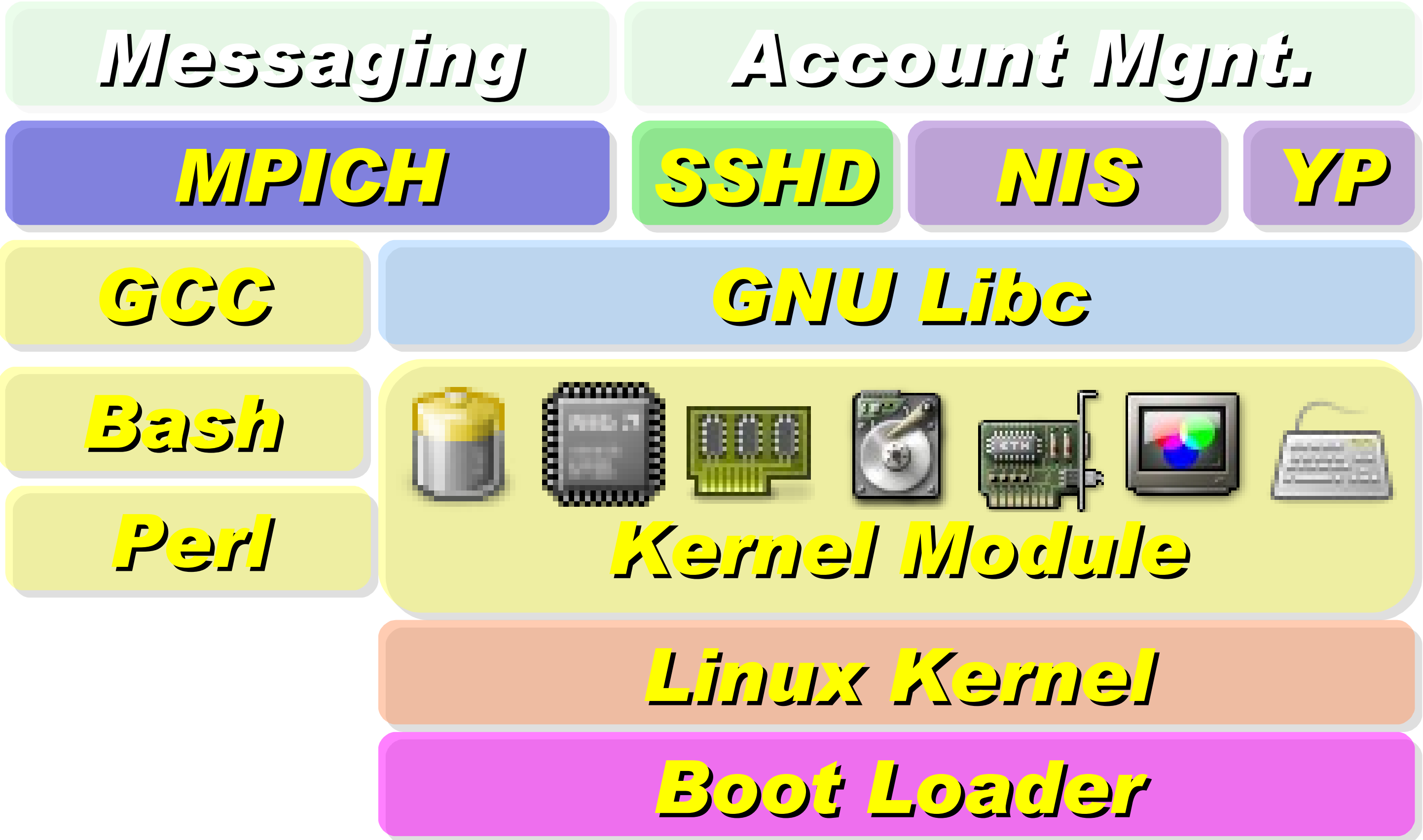
WAN

Manage Node



Compute Nodes

Basic System Setup for Cluster



Messaging

Account Mgnt.

MPICH

SSHD

NIS

YP

GCC

GNU Libc

Bash

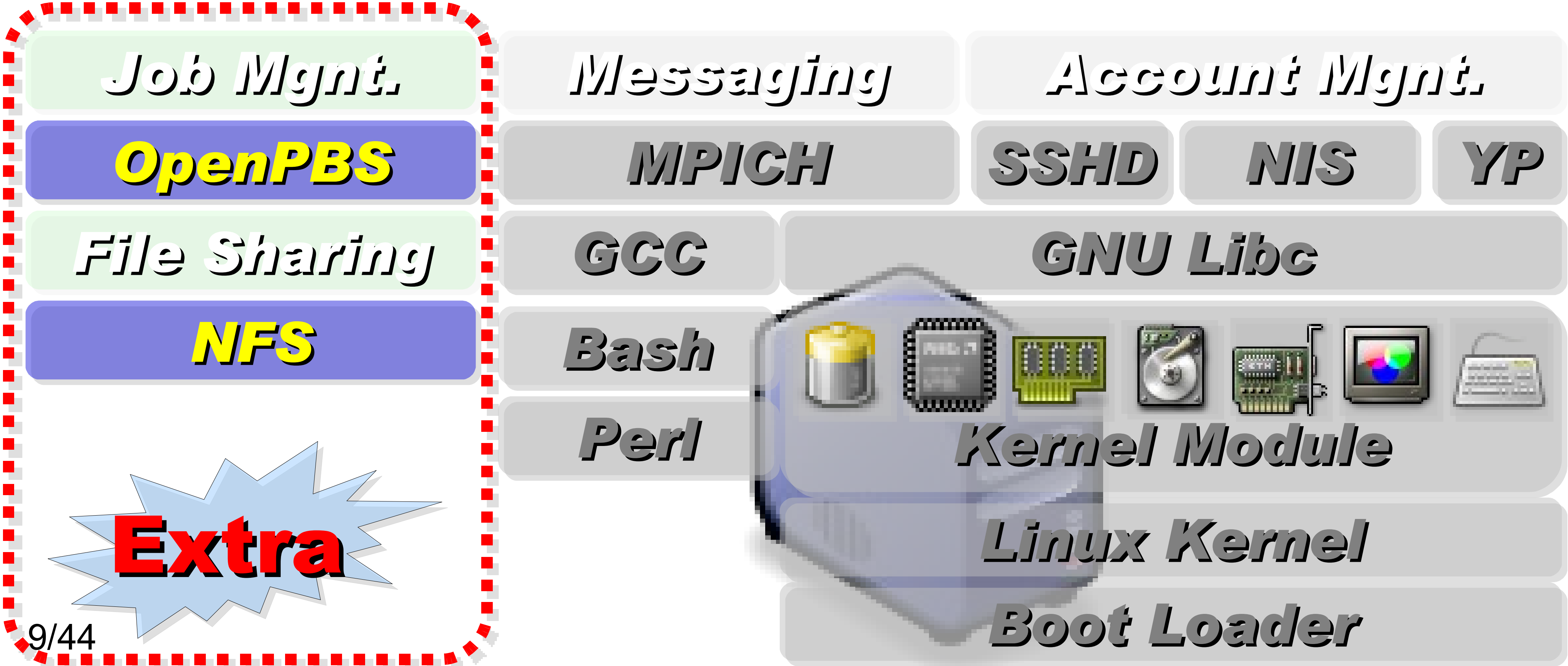
Perl

Kernel Module

Linux Kernel

Boot Loader

On Manage Node, We need to install Scheduler and Network File System for sharing Files with Compute Node



Job Mgnt. | Messaging | Account Mgnt.

OpenPBS | MPICH | SSHD | NIS | YP

File Sharing | GCC | GNU Libc

NFS | Bash | Kernel Module

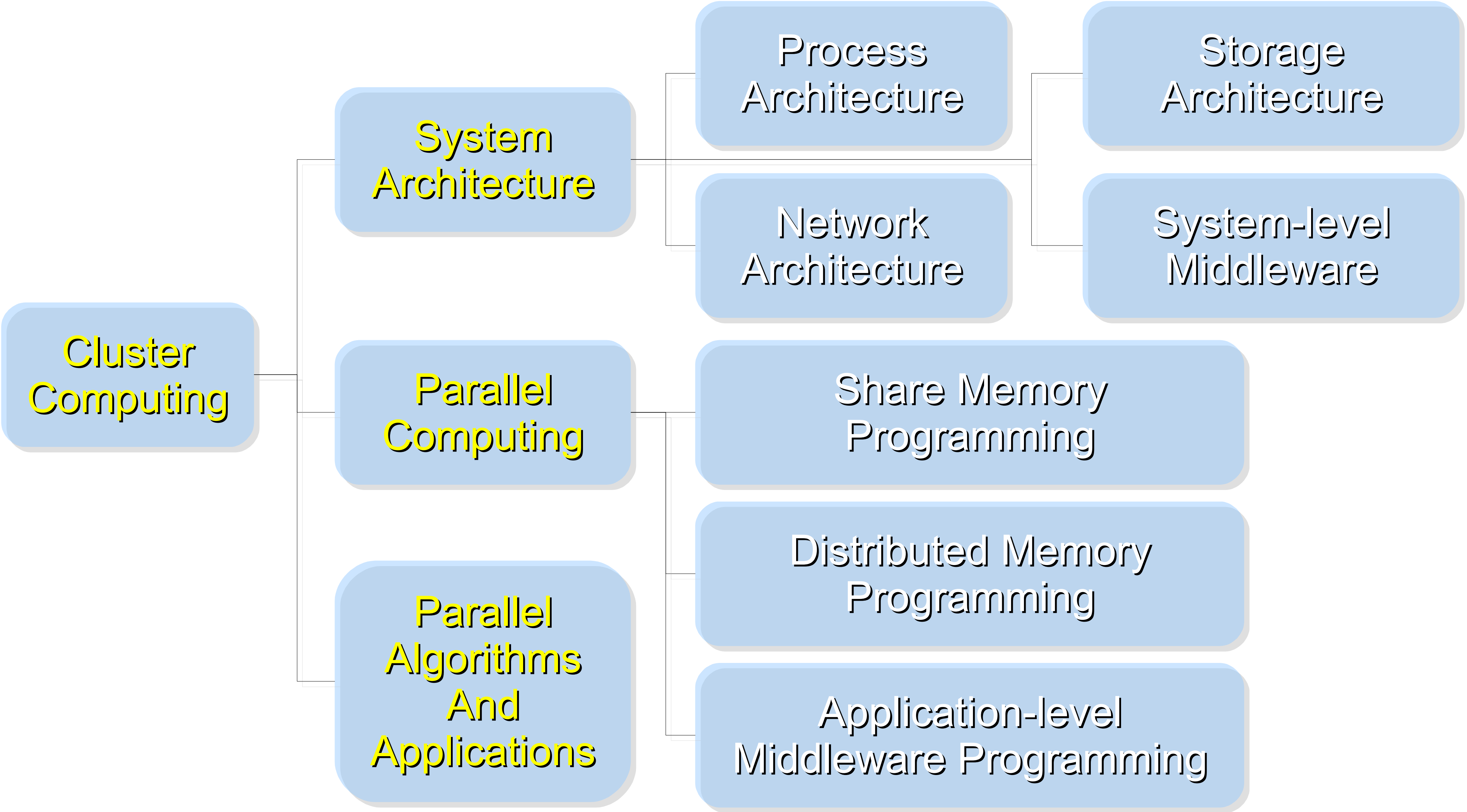
Perl

Extra

Linux Kernel

Boot Loader

Research topics about PC Cluster



Cluster Computing

System Architecture

Process Architecture

Network Architecture

Storage Architecture

System-level Middleware

Parallel Computing

Share Memory Programming

Distributed Memory Programming

Application-level Middleware Programming

Parallel Algorithms And Applications

Ref: Cluster Computing in the Classroom: Topics, Guidelines, and Experiences
http://www.gridbus.org/papers/CC-Edu.pdf

Challenges of Cluster Computing

- ***Hardware***
  - ***Ethernet Speed / PC Density***
  - ***Power / Cooling / Heat***
  - ***Network and Storage Architecture***
- ***Software***
  - ***Job Scheduler (Cluster level)***
  - ***Account Management***
  - ***File Sharing / Package Management***
- ***Limitation***
  - ***Shared Memory***
  - ***Global Memory Management***

Common Method to deploy Cluster

1. Setup one Template machine

2. Cloning to multiple machine

3. Configure Settings

↓

4. Install Job Scheduler

↓

5. Running Benchmark

Challenges of Common Method

Add New User Account ?

Upgrade Software ?

How to share user data ?

Configuration Syncronization

How to deploy 4000+ Nodes ????

資料標題：Scaling Hadoop to 4000 nodes at Yahoo!
資料日期：September 30, 2008

Total Nodes	**4000**
Total cores	**30000**
Data	**16PB**

	500-node cluster		**4000-node cluster**	
	write	**read**	**write**	**read**
number of files	990	990	14,000	14,000
file size (MB)	320	320	360	360
total MB processes	316,800	316,800	5,040,000	5,040,000
tasks per node	2	2	4	4
avg. throughput (MB/s)	**5.8**	**18**	**40**	**66**

Advanced Methods to deploy Cluster

- ***SSI (Single System Image)***
  - ***Multiple PCs as Single Computing Resources***
  - ***Image-based***
    - ***homogeneous***
    - ***ex. SystemImager, OSCAR, Kadeploy***
  - ***Package-based***
    - ***heterogeneous***
    - ***easy update and modify packages***
    - ***ex. FAI, DRBL***
- ***Other deploy tools***
  - ***Rocks : RPM only***
  - ***cfengine : configuration engine***

Comparison of Cluster Deploy Tools

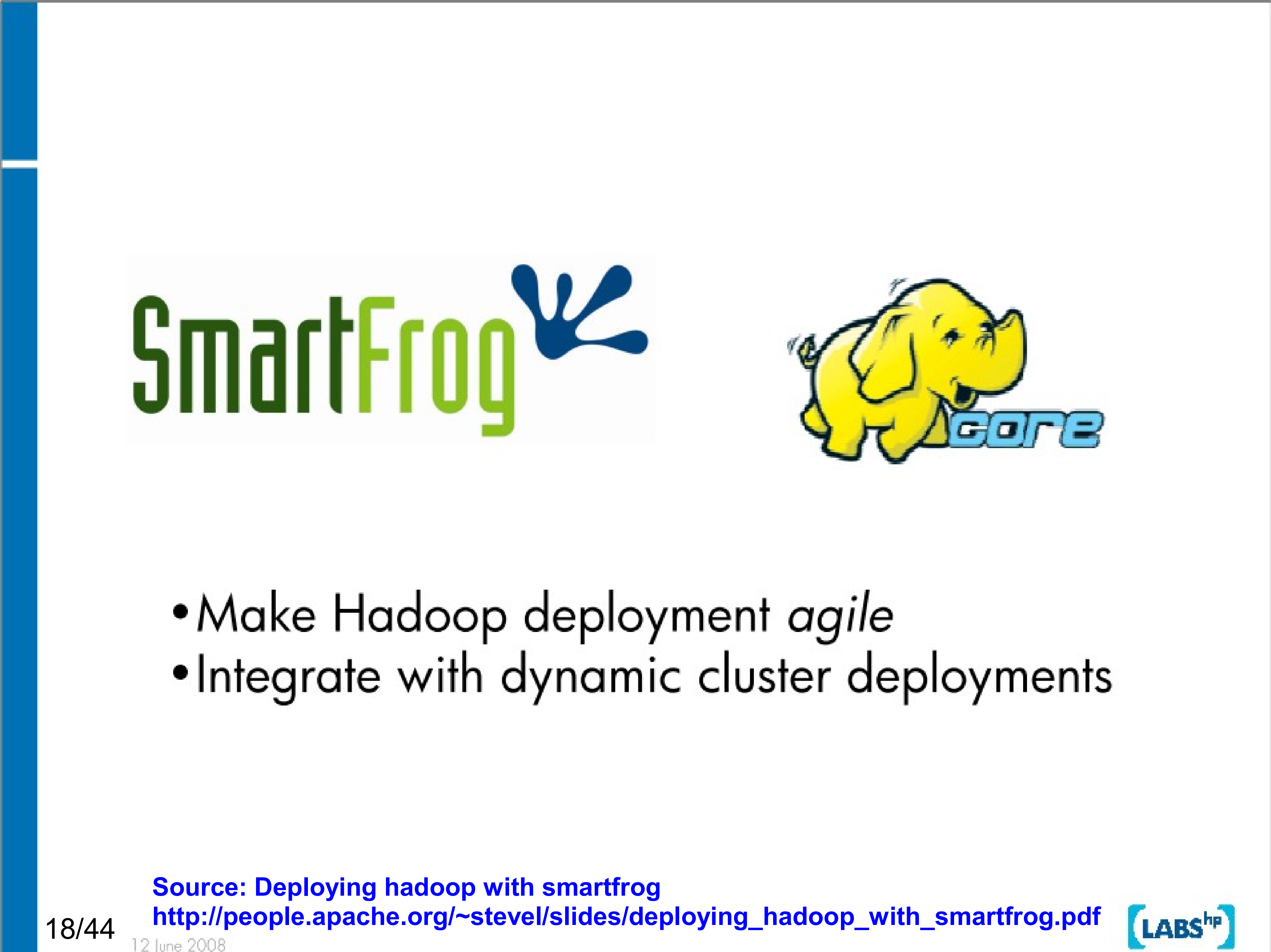
	Distribution	Support Diskless/ Sysmless	Type	Node configuration tools	Cluster management tools	Database installation
System Imager	ALL	Yes	Image	Yes	No	No
OSCAR	RPM-based	Yes	Image	Yes	Yes	No
Kadeploy	ALL	No	Image	Yes	Yes	Yes
DRBL	ALL	Yes	Package	Yes	Yes	No
FAI	Debian-Based	Yes	Package	Yes	No	No

PART 2-1 :

Hadoop Deployment Tool

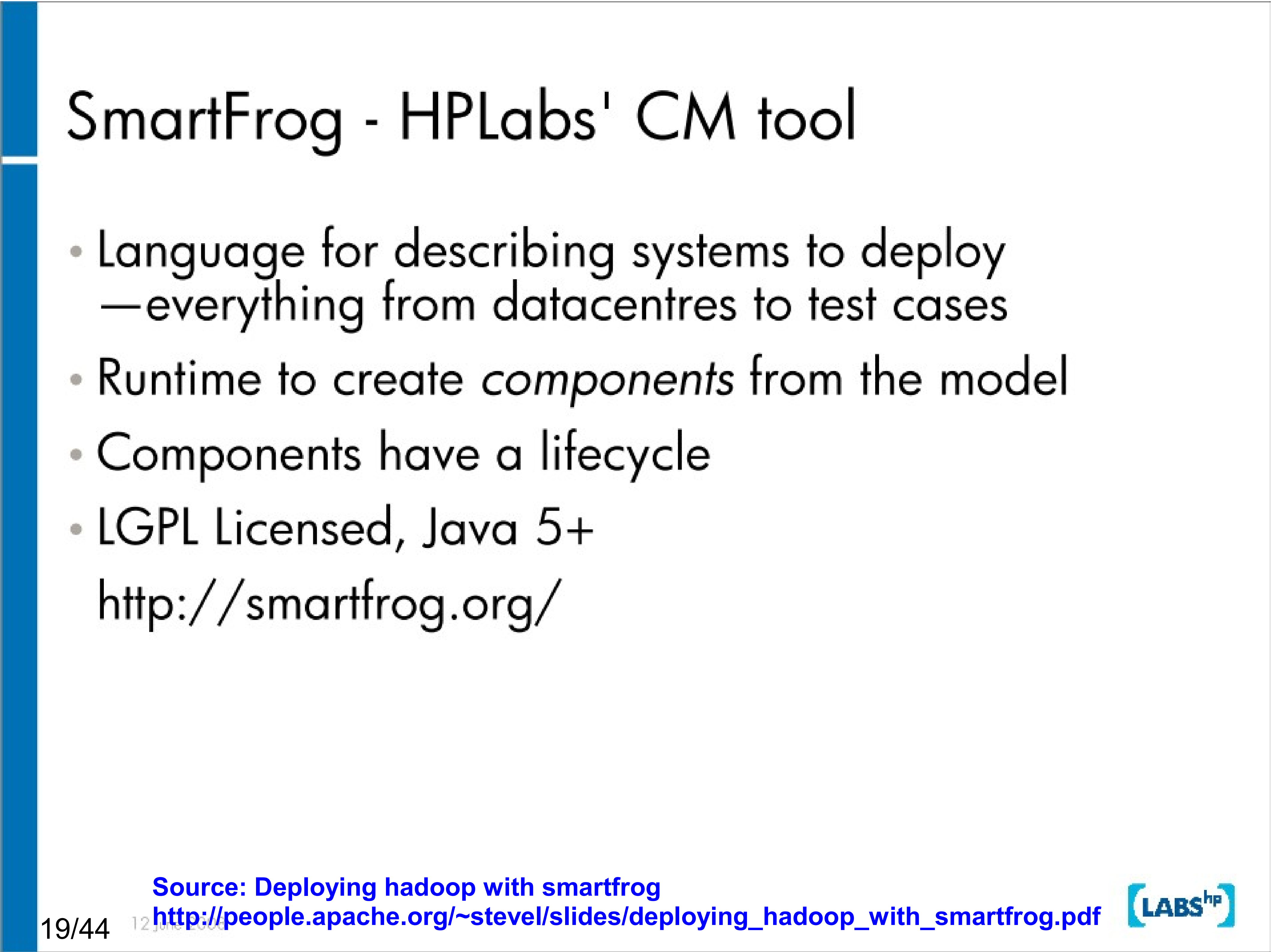
Jazz Wang
Yao-Tsung Wang
jazz@nchc.org.tw

Powered by DRBL



- *Make Hadoop deployment agile*
- Integrate with dynamic cluster deployments

Source: Deploying hadoop with smartfrog
http://people.apache.org/~stevel/slides/deploying_hadoop_with_smartfrog.pdf



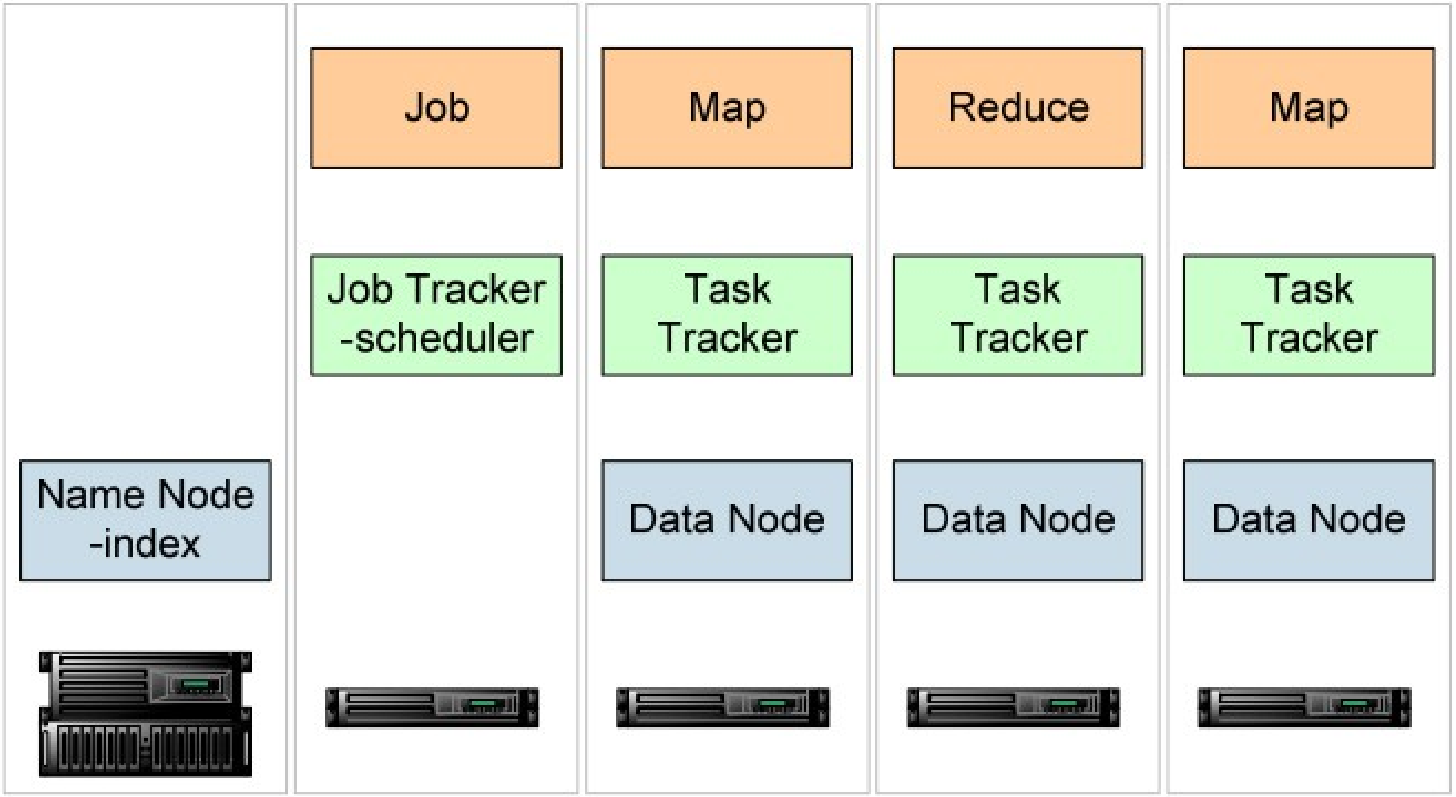
SmartFrog - HPLabs' CM tool

- Language for describing systems to deploy —everything from datacentres to test cases
- Runtime to create *components* from the model
- Components have a lifecycle
- LGPL Licensed, Java 5+

http://smartfrog.org/

Source: Deploying hadoop with smartfrog
http://people.apache.org/~stevel/slides/deploying_hadoop_with_smartfrog.pdf

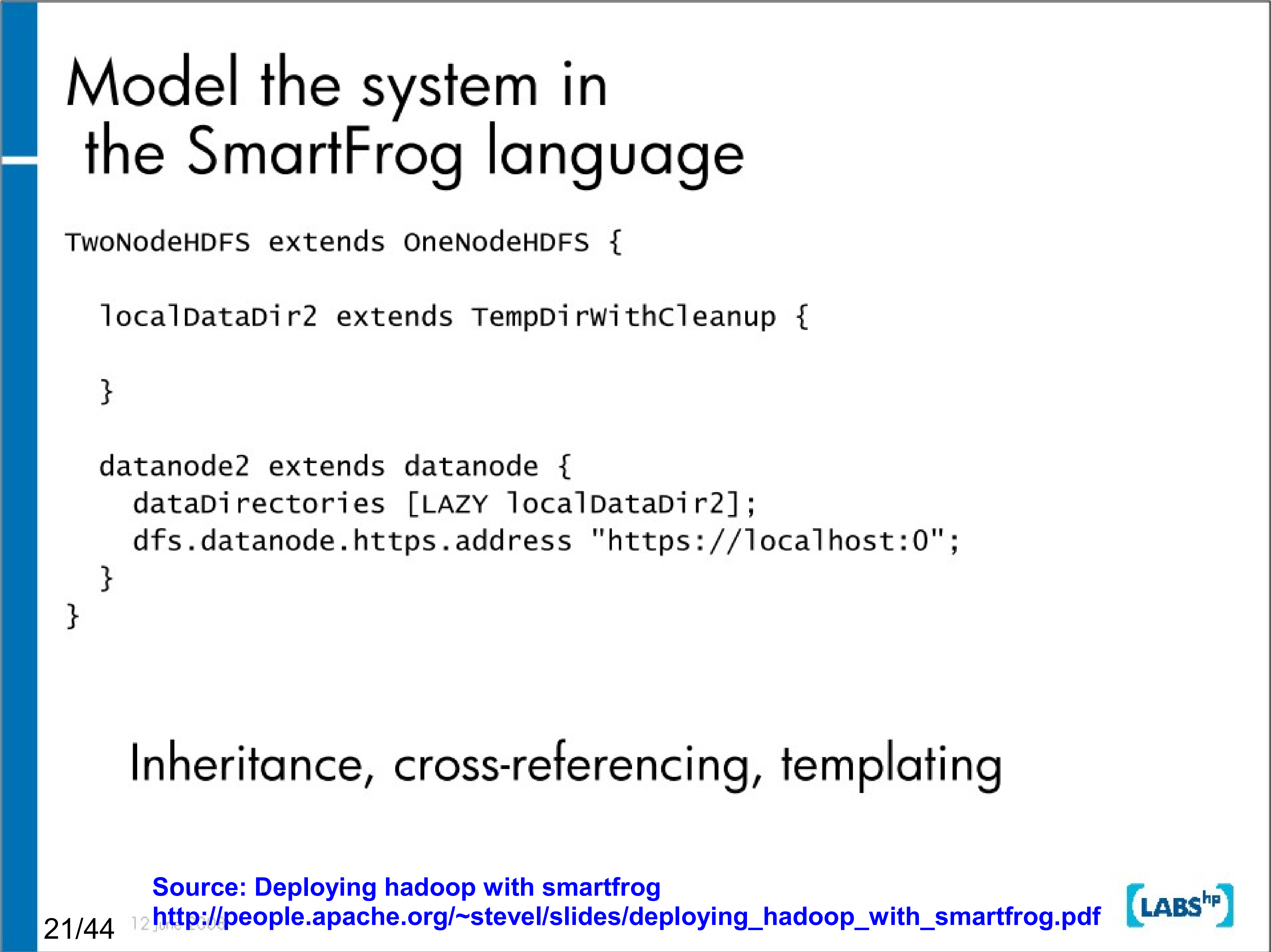
Basic problem: deploying Hadoop



User Job		Job	Map	Reduce	Map
Map/ Reduce		Job Tracker -scheduler	Task Tracker	Task Tracker	Task Tracker
HDFS	Name Node -index		Data Node	Data Node	Data Node
Hardware + OS					

one namenode, 1+ Job Tracker, many data nodes and task trackers

Source: Deploying hadoop with smartfrog
http://people.apache.org/~stevel/slides/deploying_hadoop_with_smartfrog.pdf



Model the system in the SmartFrog language

```
TwoNodeHDFS extends OneNodeHDFS {

  localDataDir2 extends TempDirWithCleanup {

  }

  datanode2 extends datanode {
    dataDirectories [LAZY localDataDir2];
    dfs.datanode.https.address "https://localhost:0";
  }
}
```

Inheritance, cross-referencing, templating

Source: Deploying hadoop with smartfrog
http://people.apache.org/~stevel/slides/deploying_hadoop_with_smartfrog.pdf

PART 2-2 :

企鵝龍與再生龍
DRBL and Clonezilla

Jazz Wang

Yao-Tsung Wang

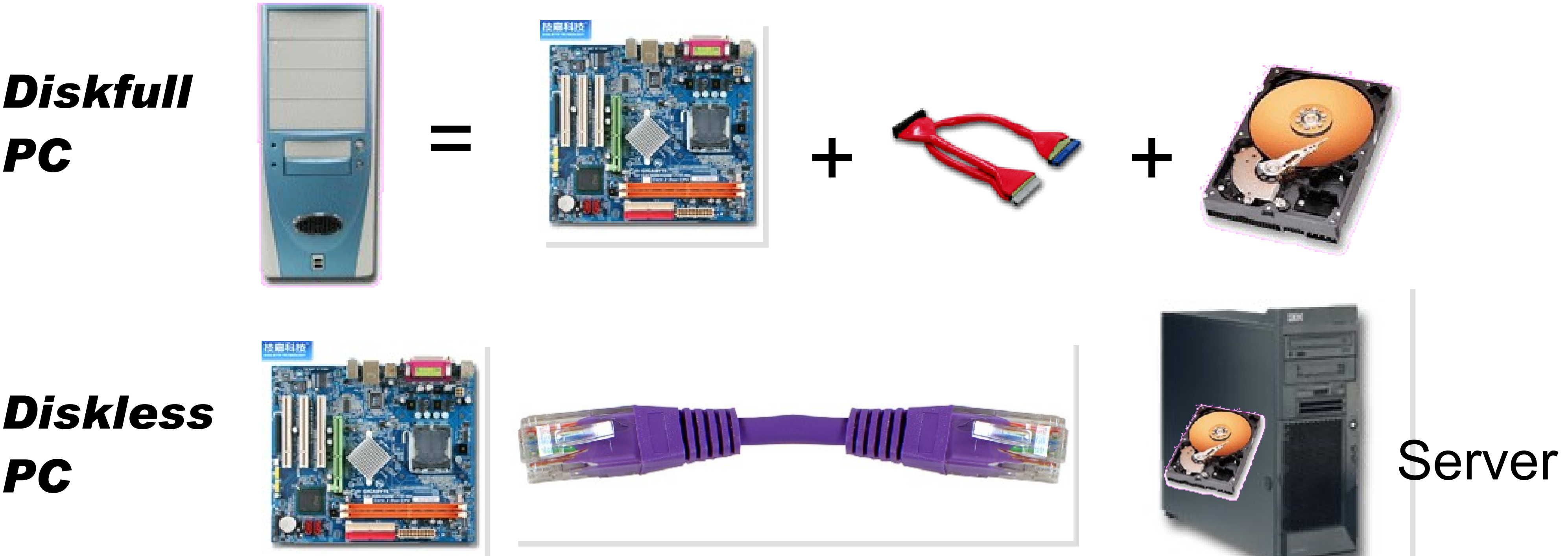
jazz@nchc.org.tw

Powered by **DRBL**

何謂企鵝龍 What is DRBL ??

- ***Diskless Remote Boot in Linux***
- ***Network is cheap, Man Hour is expansive.***
- ***In short, DRBL is ....***
  - ***Use network cable to replace SATA cable***
  - ***All student PCs are connected to one single server***

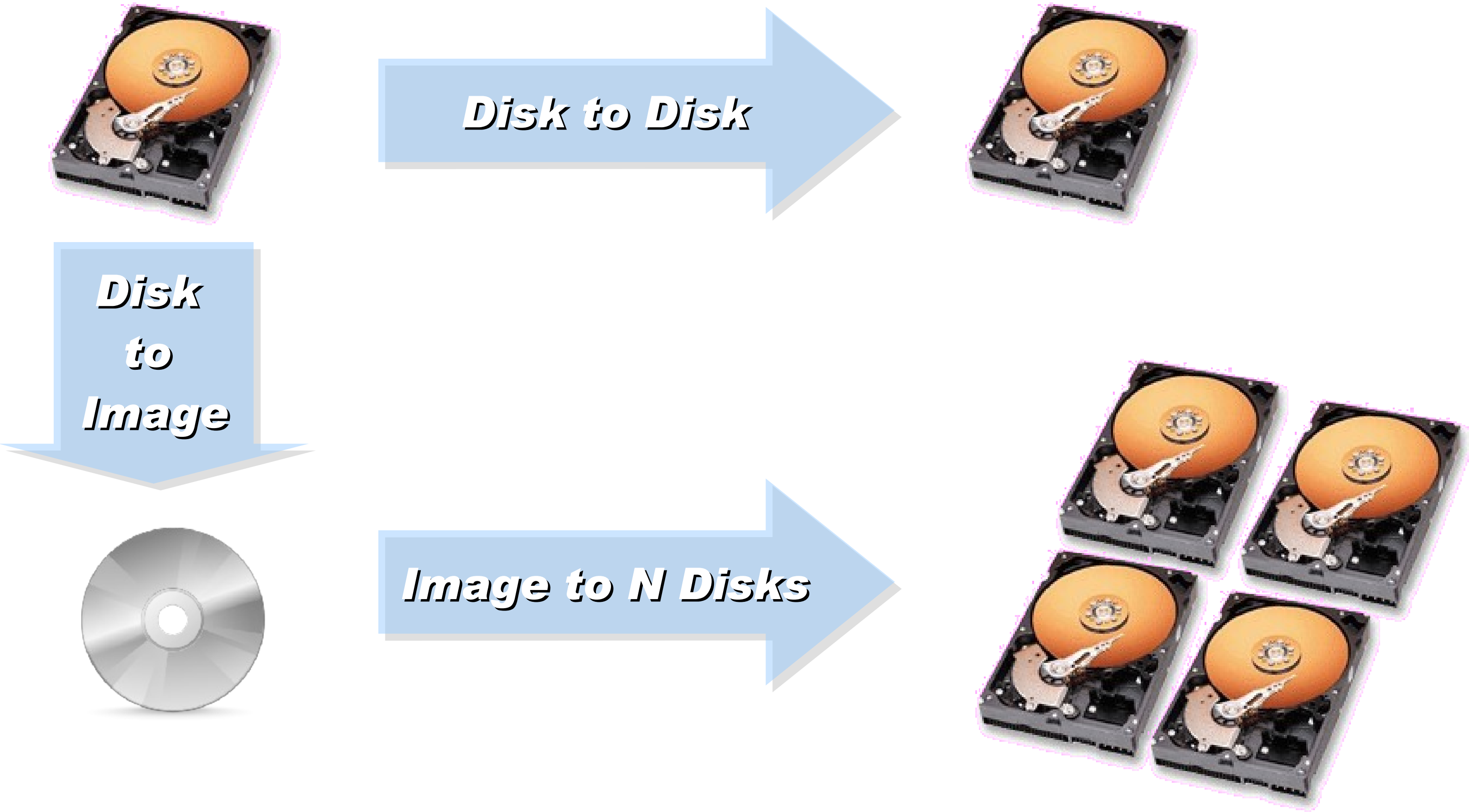
Powered by DRBL



source: http://www.mren.com.tw

何謂再生龍 What is Clonezilla ??

- **Clone (複製) + zilla = Clonezilla (再生龍)**
- **Open Source Alternative to Norton Ghost**
- **Support Windows, Linux and Mac**



Disk to Disk

Disk to Image

Image to N Disks

PART 2-3 :

企鵝龍的開機原理

How does DRBL work ?

Jazz Wang

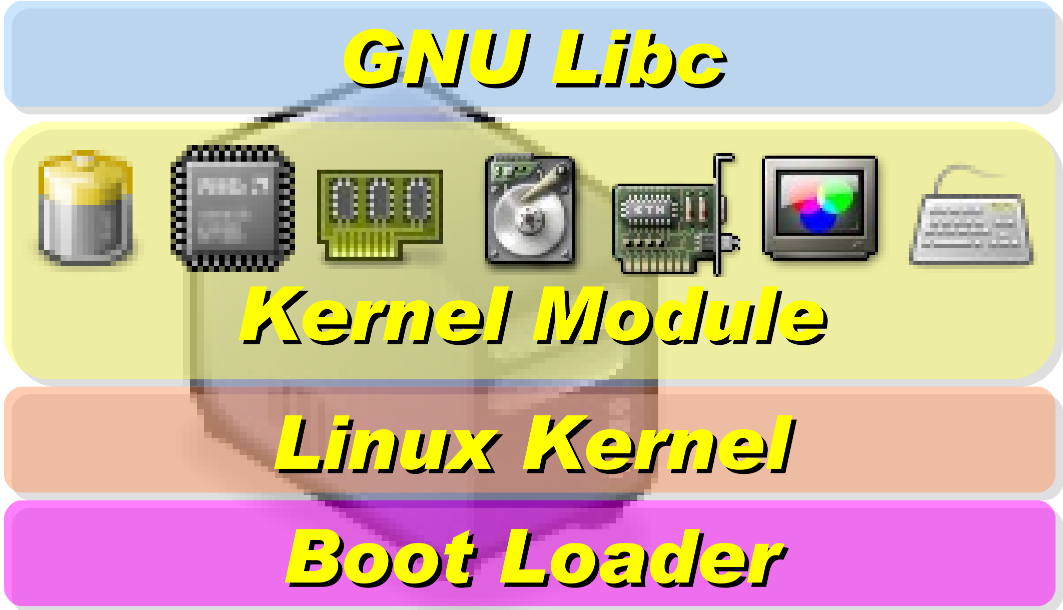
Yao-Tsung Wang

jazz@nchc.org.tw

Powered by DRBL

1st, We install Base System of GNU/ Linux on Management Node. You can choose:

Redhat, Fedora, CentOS, Mandriva, Ubuntu, Debian, ...



GNU Libc

Kernel Module

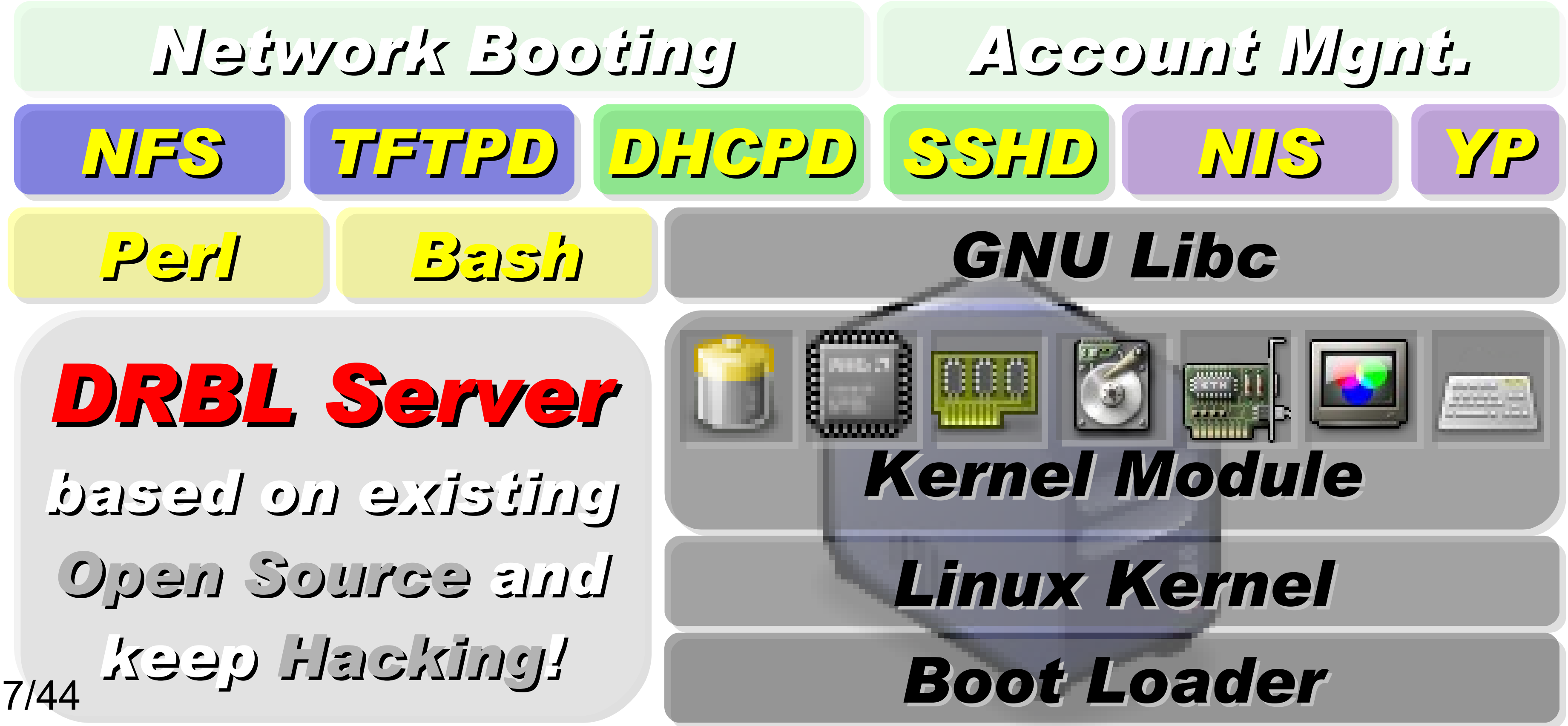
Linux Kernel

Boot Loader

2nd, We install DRBL package and configure it as DRBL Server.

There are lots of service needed:

SSHD, DHCPD, TFTPD, NFS Server,
NIS Server, YP Server ...



Network Booting

Account Mgnt.

NFS

TFTPD

DHCPD

SSHD

NIS

YP

Perl

Bash

GNU Libc

Kernel Module

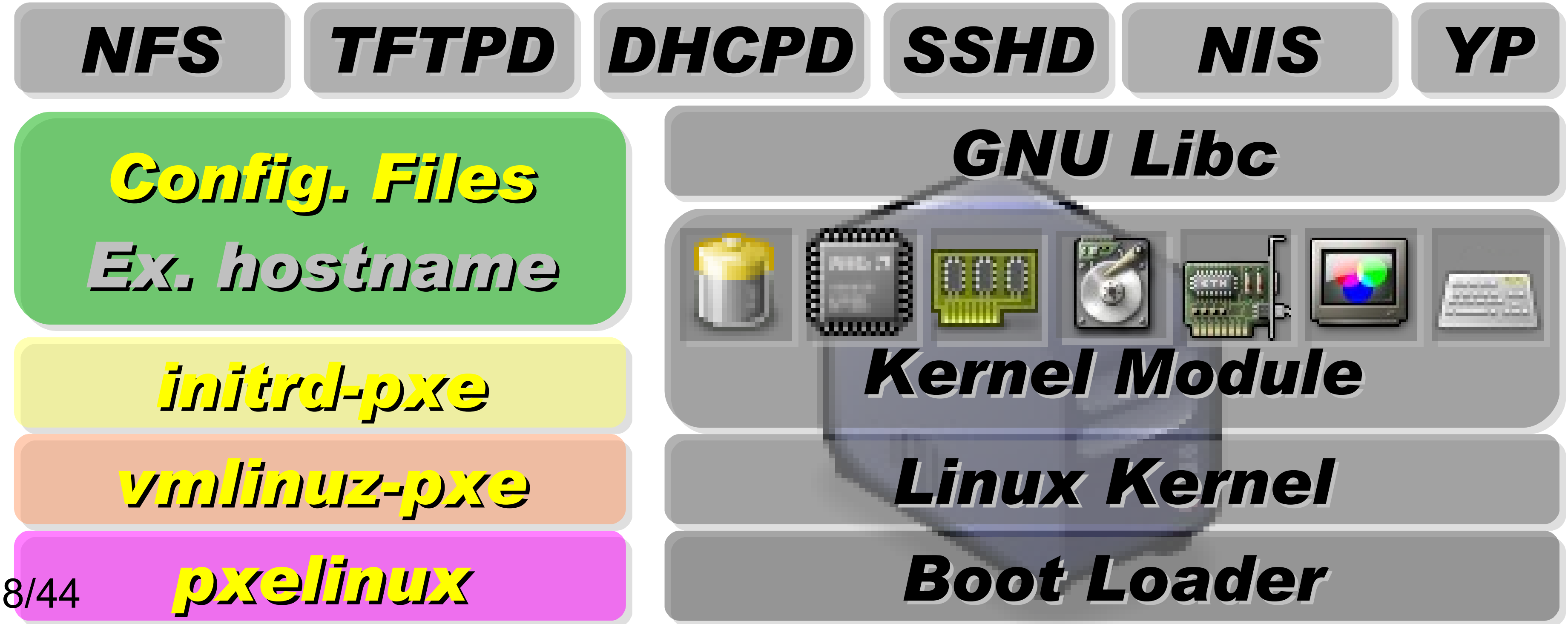
Linux Kernel

Boot Loader

DRBL Server

based on existing Open Source and keep Hacking!

After running "drblsrv -i" & "drblpush -i", there will be pxelinux, vmlinux-pex, initrd-pxe in TFTPROOT, and different configuration files for each Compute Node in NFSROOT



NFS

TFTPD

DHCPD

SSHD

NIS

YP

Config. Files

Ex. hostname

GNU Libc

Kernel Module

initrd-pxe

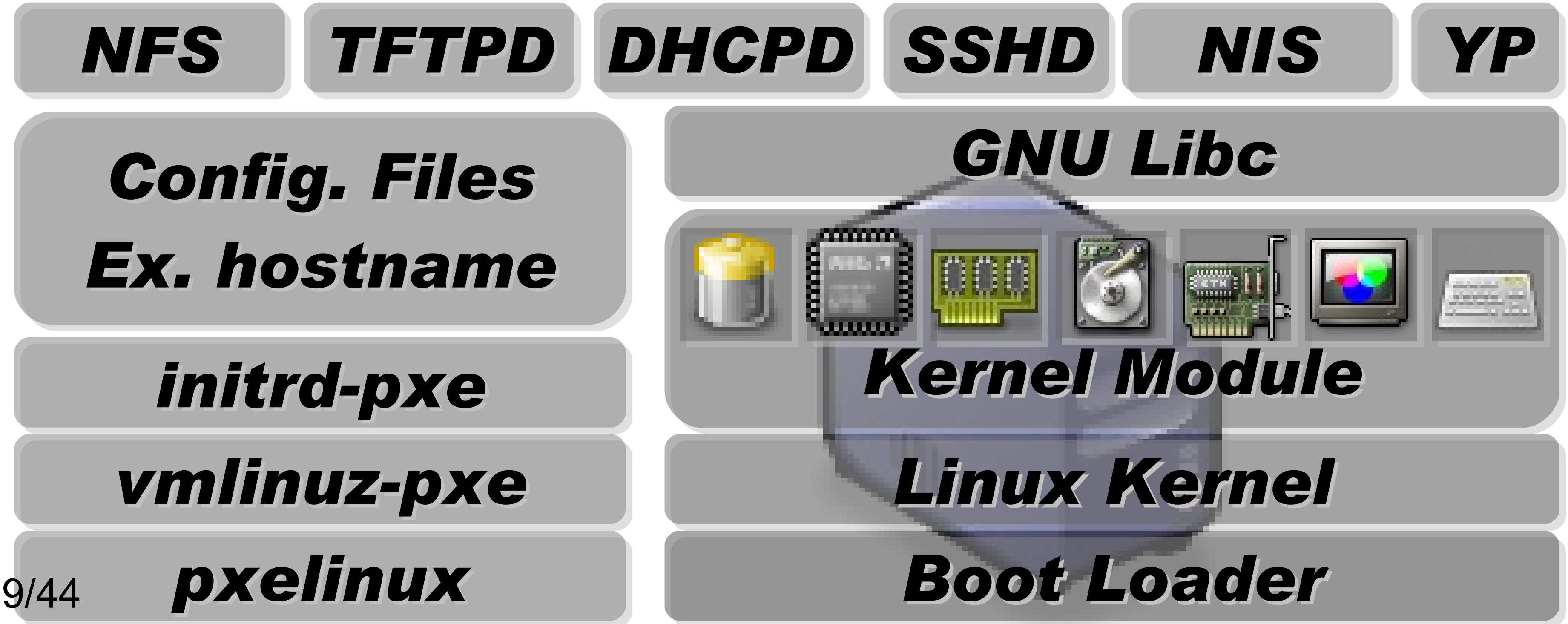
Linux Kernel

vmlinuz-pxe

Boot Loader

pxelinux

3nd, We enable **PXE** function in **BIOS** configuration.


NFS

TFTPD

DHCPD

SSHD

NIS

YP

Config. Files
Ex. hostname

GNU Libc

Kernel Module

initrd-pxe

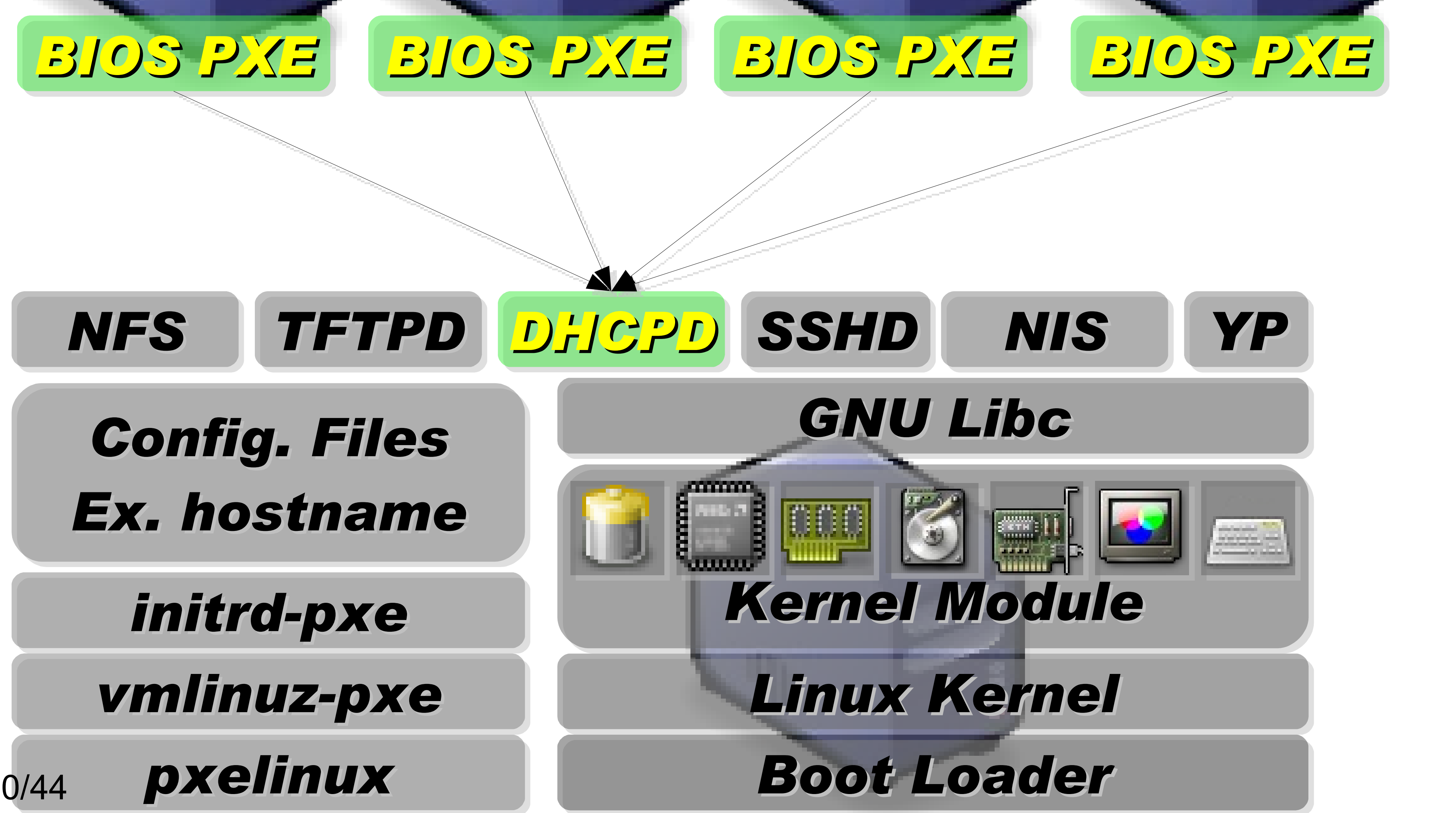
vmlinuz-pxe

Linux Kernel

pxelinux

Boot Loader

While Booting, PXE will query IP address from DHCPD.



BIOS PXE

BIOS PXE

BIOS PXE

BIOS PXE

NFS

TFTPD

DHCPD

SSHD

NIS

YP

Config. Files

Ex. hostname

GNU Libc

Kernel Module

initrd-pxe

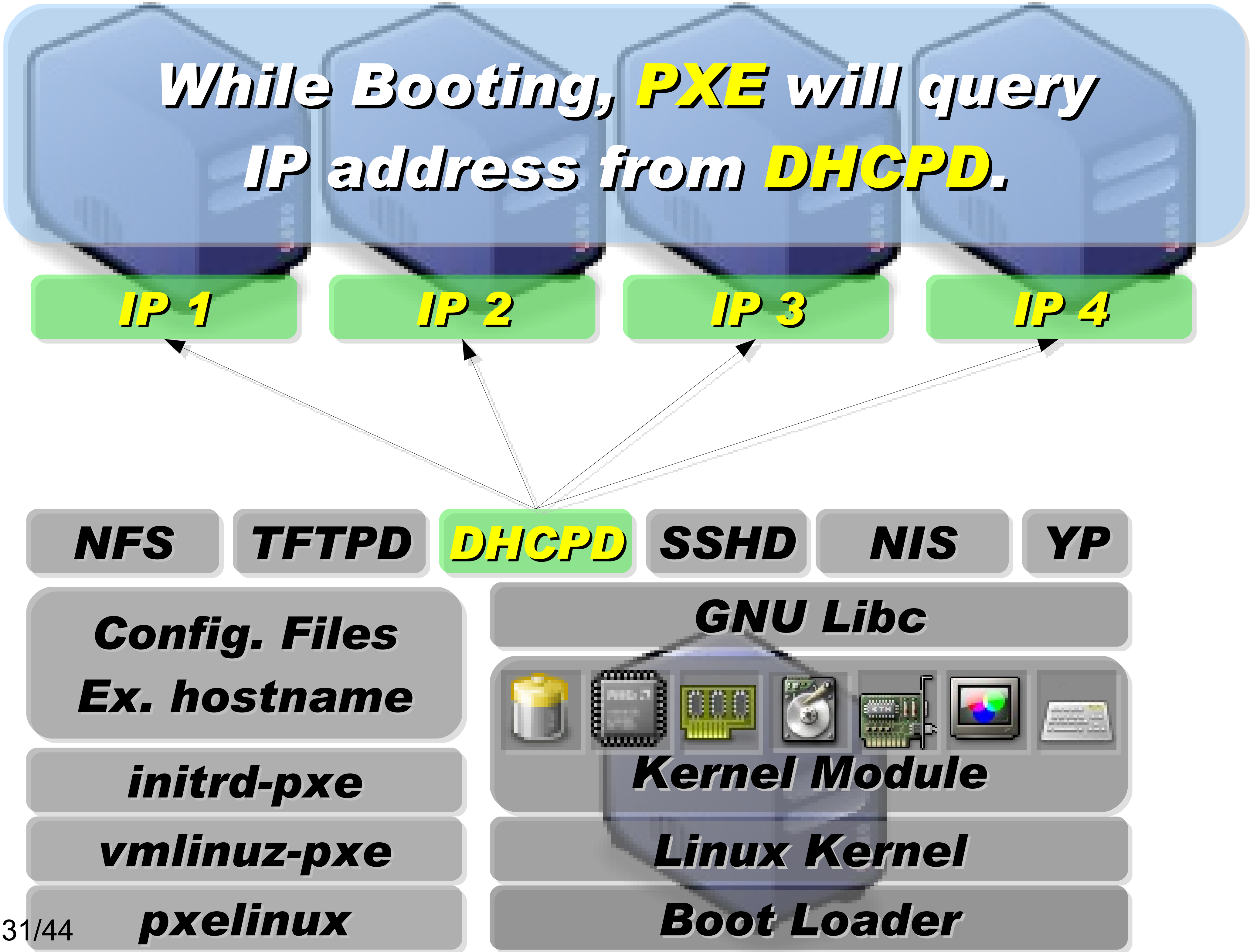
vmlinuz-pxe

Linux Kernel

pxelinux

Boot Loader

While Booting, PXE will query IP address from DHCPD.



IP 1

IP 2

IP 3

IP 4

NFS

TFTPD

DHCPD

SSHD

NIS

YP

Config. Files
Ex. hostname

GNU Libc

initrd-pxe

Kernel Module

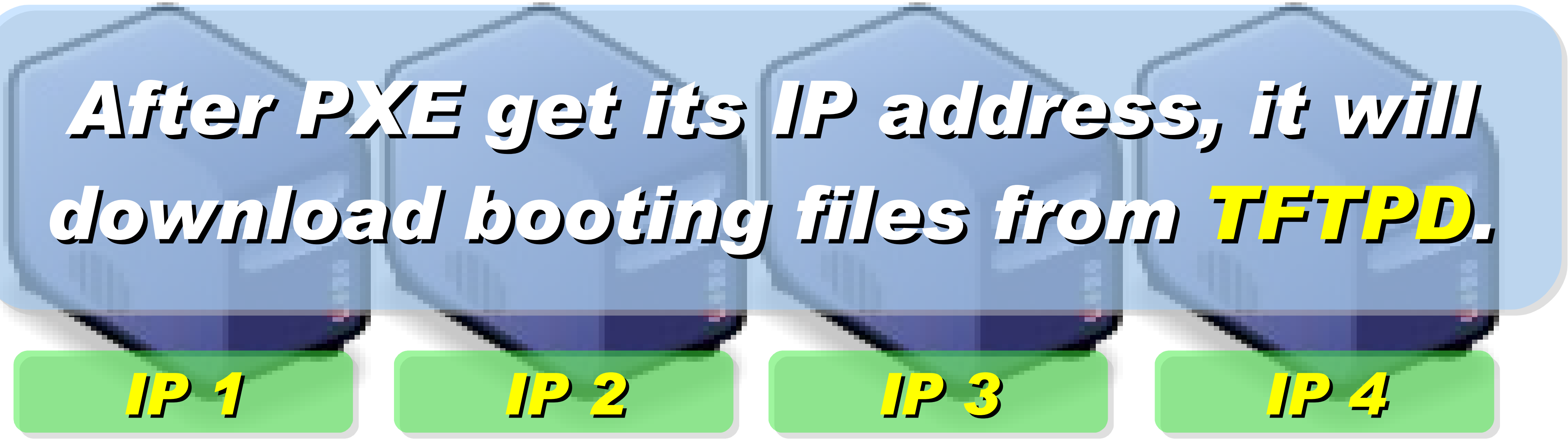
vmlinuz-pxe

Linux Kernel

pxelinux

Boot Loader

After PXE get its IP address, it will download booting files from TFTPD.

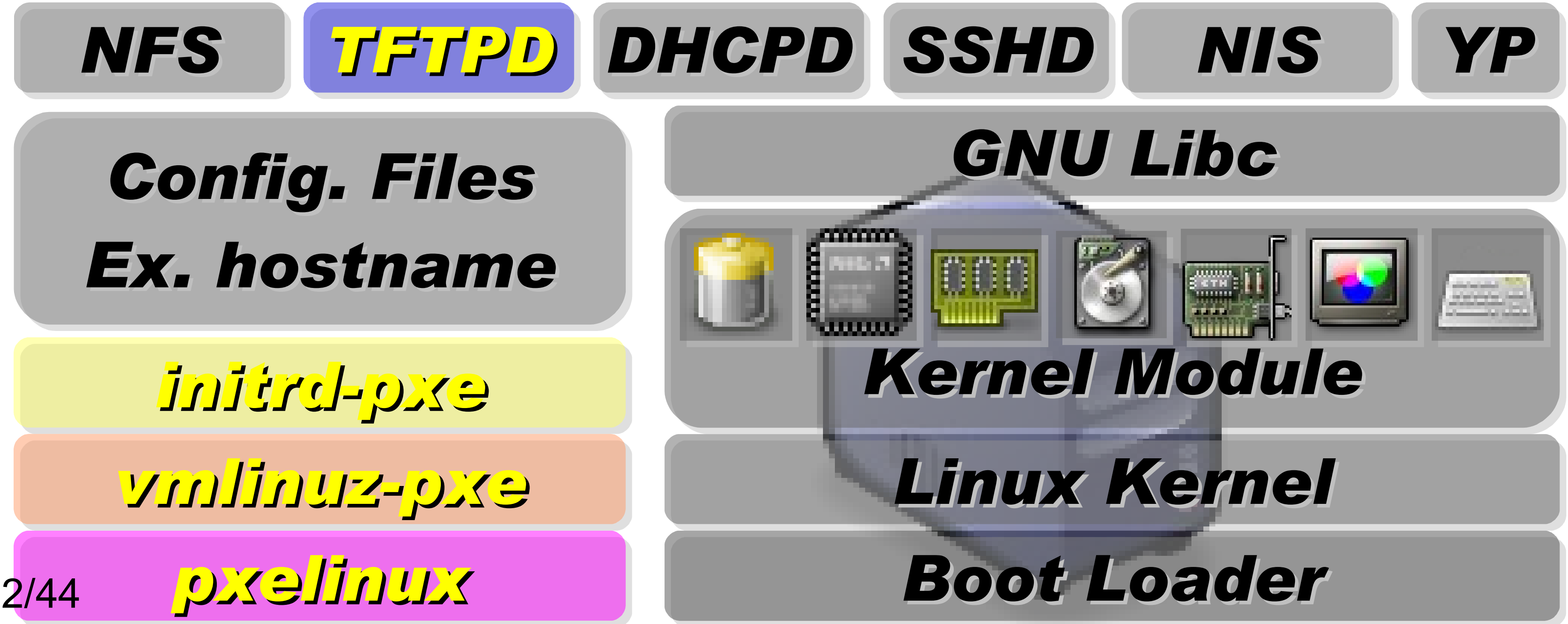


IP 1

IP 2

IP 3

IP 4



NFS

TFTPD

DHCPD

SSHD

NIS

YP

Config. Files
Ex. hostname

GNU Libc

Kernel Module

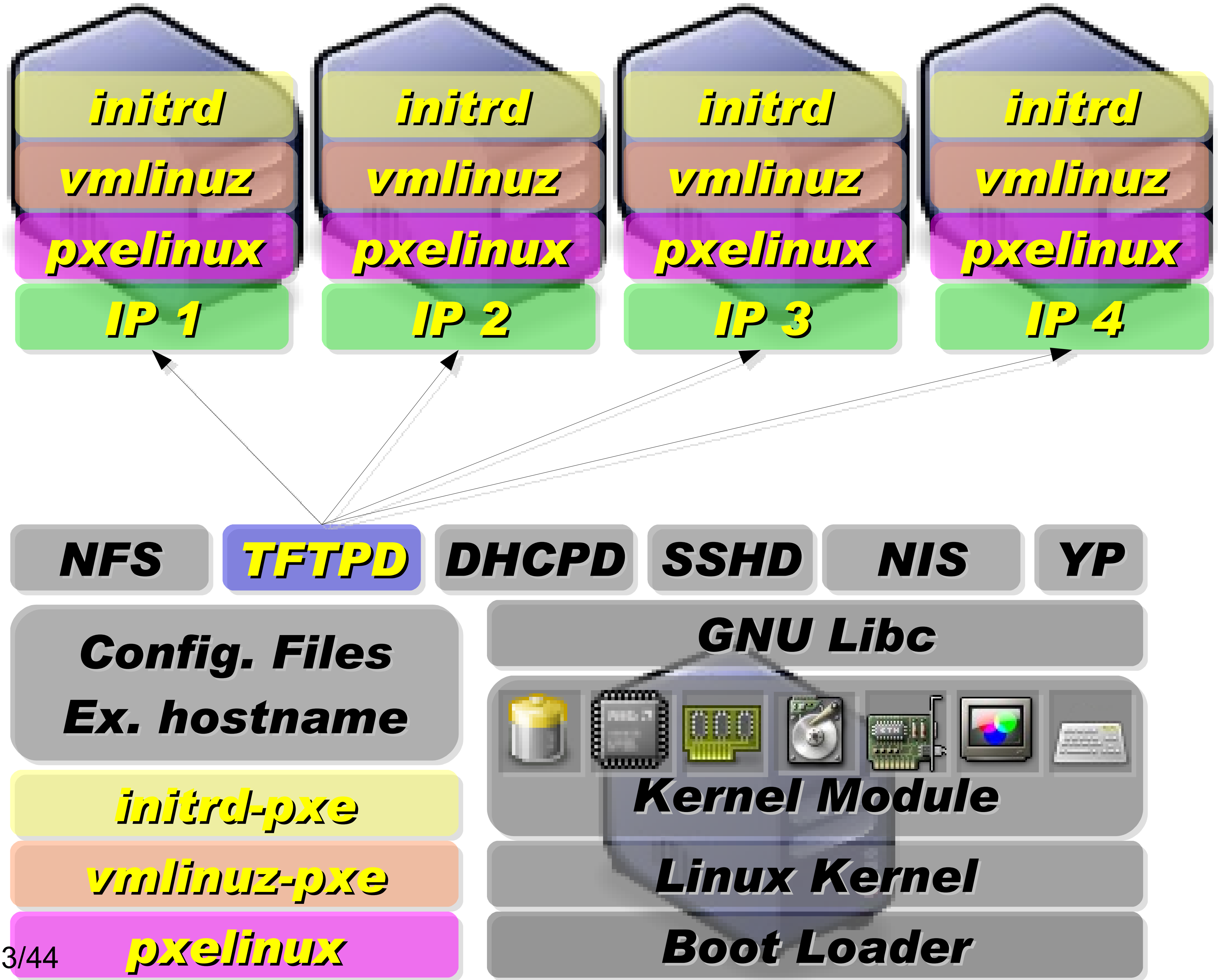
initrd-pxe

vmlinuz-pxe

Linux Kernel

pxelinux

Boot Loader



initrd

vmlinuz

pxelinux

IP 1

initrd

vmlinuz

pxelinux

IP 2

initrd

vmlinuz

pxelinux

IP 3

initrd

vmlinuz

pxelinux

IP 4

NFS

TFTPD

DHCPD

SSHD

NIS

YP

Config. Files
Ex. hostname

GNU Libc

Kernel Module

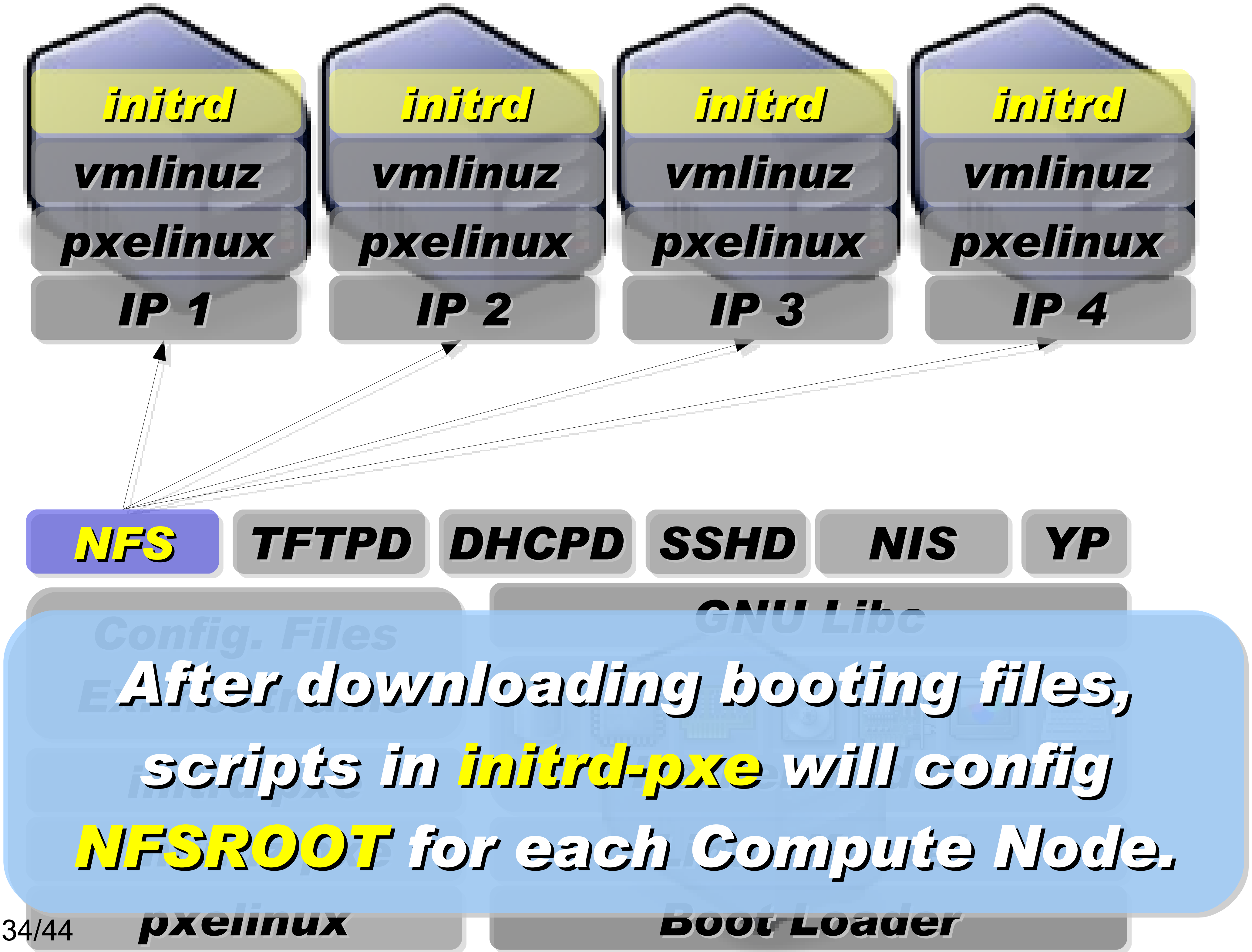
initrd-pxe

vmlinuz-pxe

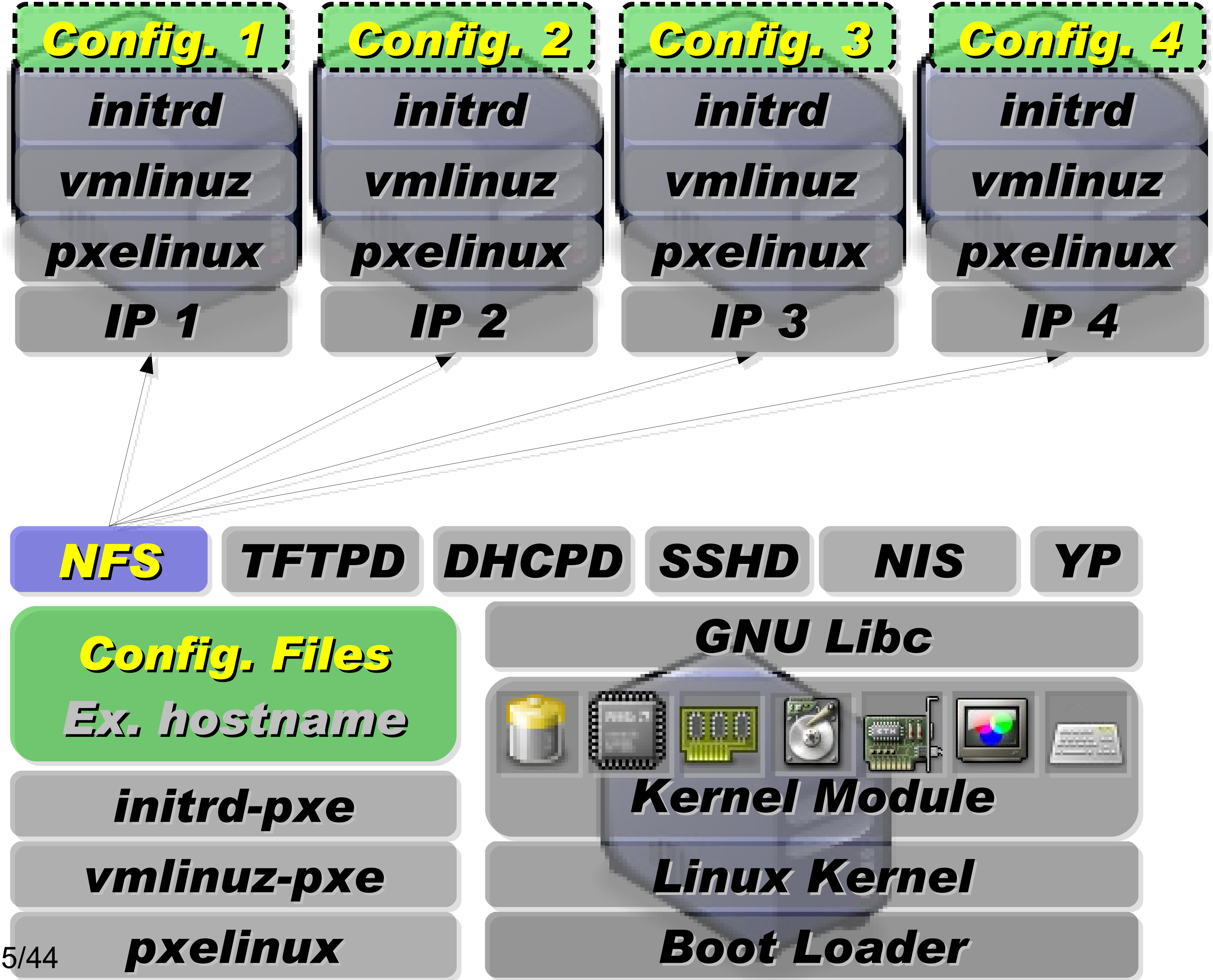
Linux Kernel

pxelinux

Boot Loader



After downloading booting files, scripts in **initrd-pxe** will config **NFSROOT** for each Compute Node.



Config. 1
initrd
vmlinuz
pxelinux
IP 1

Config. 2
initrd
vmlinuz
pxelinux
IP 2

Config. 3
initrd
vmlinuz
pxelinux
IP 3

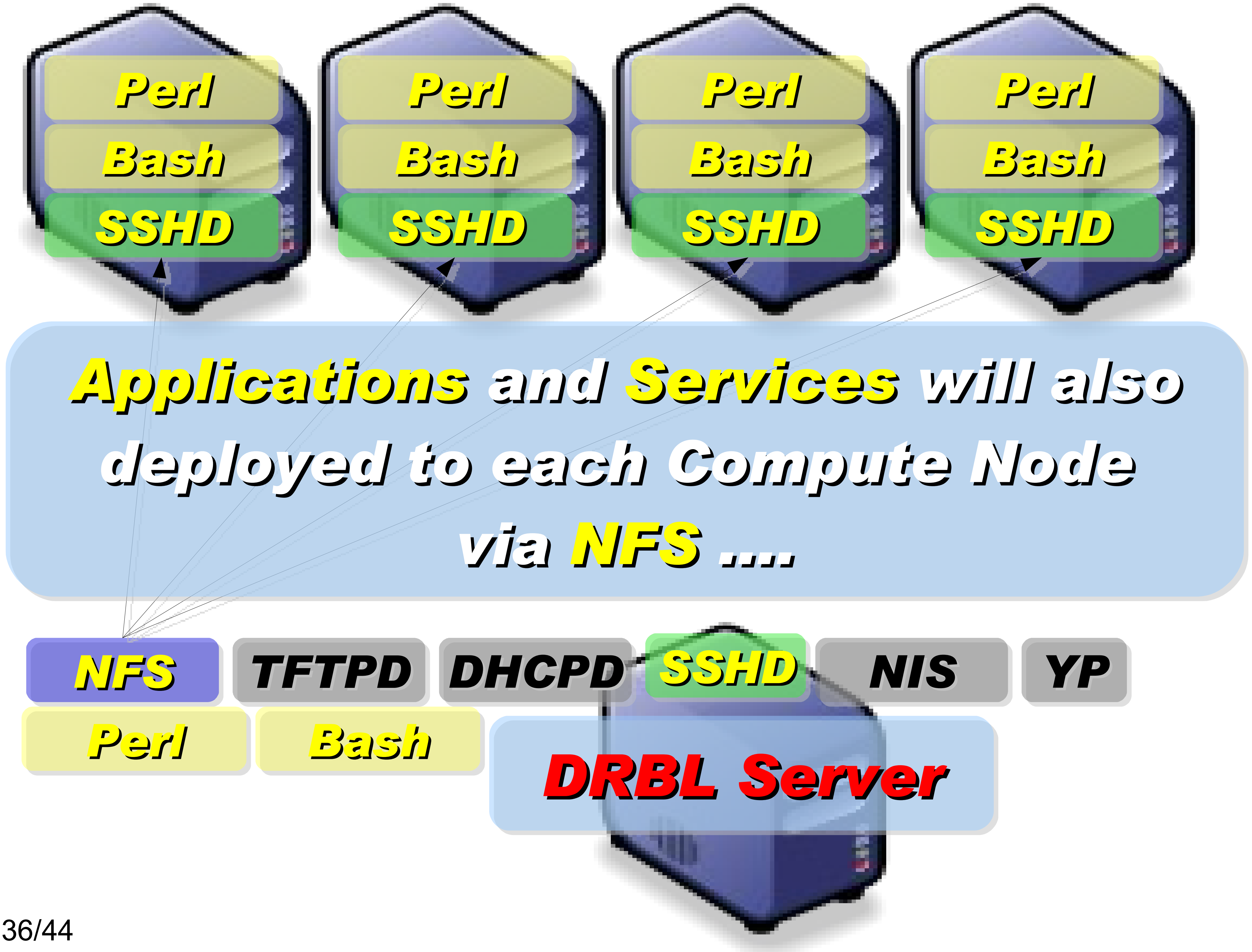
Config. 4
initrd
vmlinuz
pxelinux
IP 4

NFS
TFTPD
DHCPD
SSHD
NIS
YP

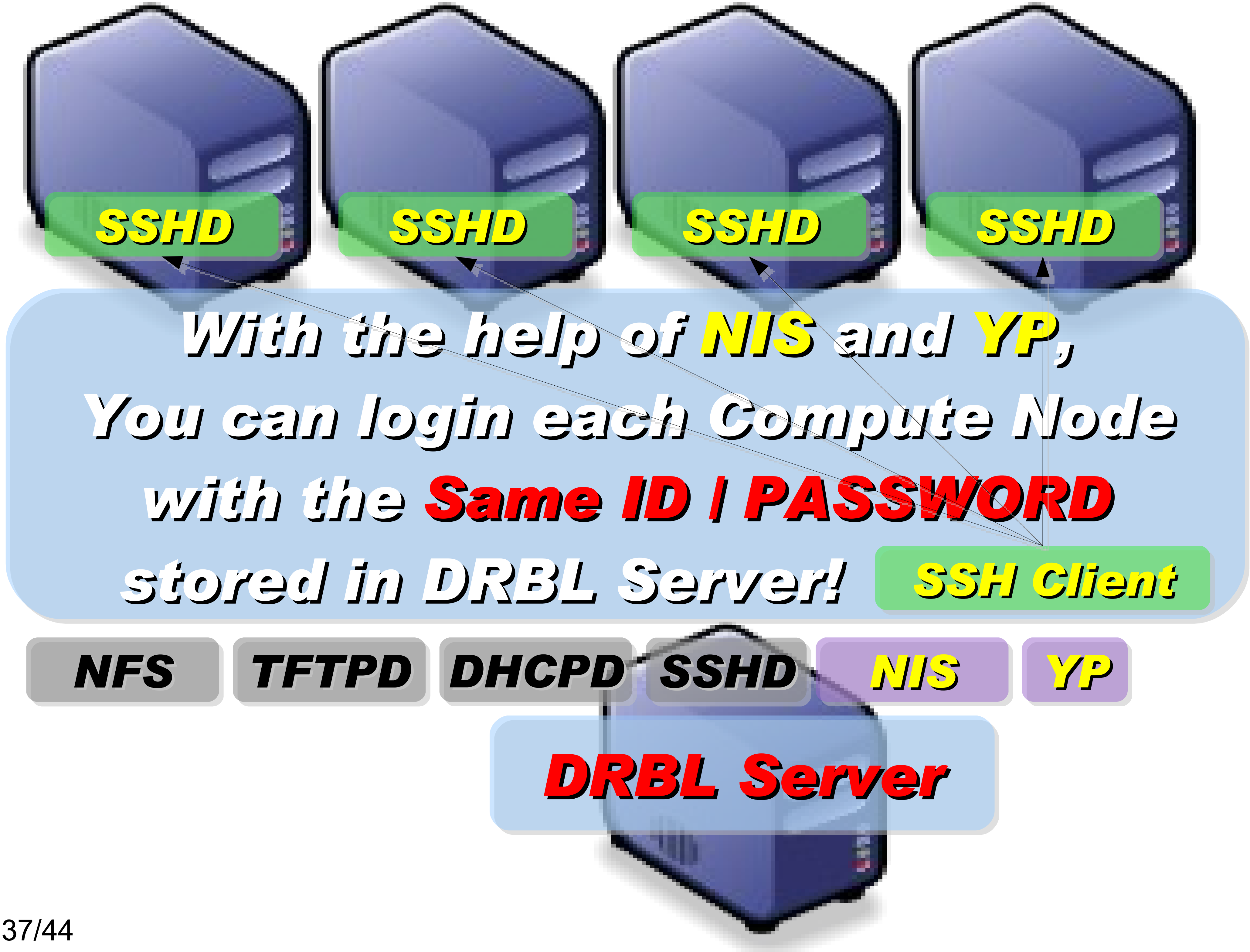
Config. Files
Ex. hostname

initrd-pxe
vmlinuz-pxe
pxelinux

GNU Libc
Kernel Module
Linux Kernel
Boot Loader



Applications and Services will also deployed to each Compute Node via NFS ....



SSHD SSHD SSHD SSHD

With the help of NIS and YP,
You can login each Compute Node
with the Same ID / PASSWORD
stored in DRBL Server!

SSH Client

NFS TFTPD DHCPD SSHD NIS YP

DRBL Server

PART 2 -1:

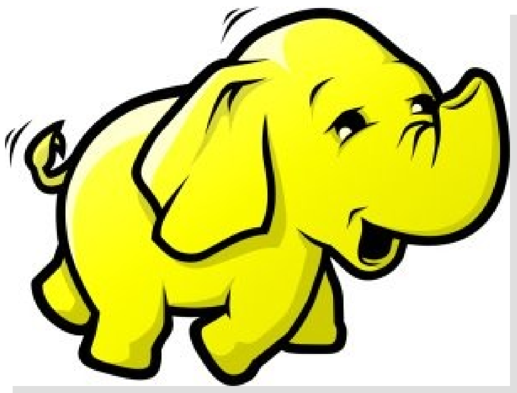
當企鵝龍遇上小飛象

When DRBL meet Hadoop

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw



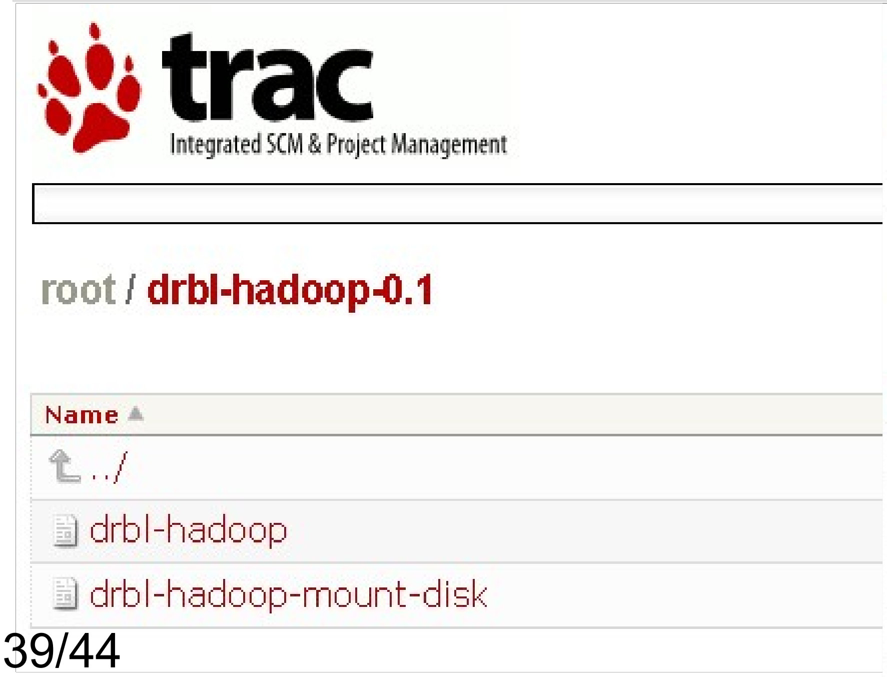
Deploy Hadoop Using DRBL

- **Under development. Need packaging.**
- **drbl-hadoop – mounting local hard disk for HDFS**

```
svn co http://trac.nchc.org.tw/pub/grid/drbl-hadoop
```

- **hadoop-register – Website and ssh applet**

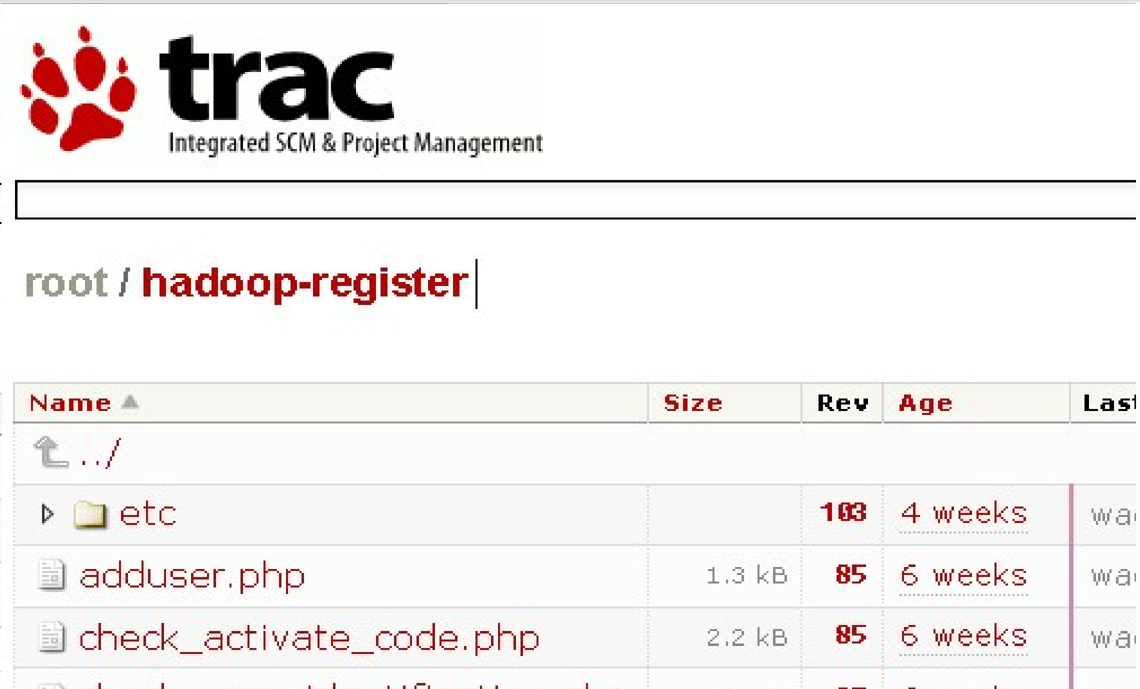
```
svn co http://trac.nchc.org.tw/pub/cloud/hadoop-register
```



trac
Integrated SCM & Project Management

root / drbl-hadoop-0.1

Name ▲
../
drbl-hadoop
drbl-hadoop-mount-disk



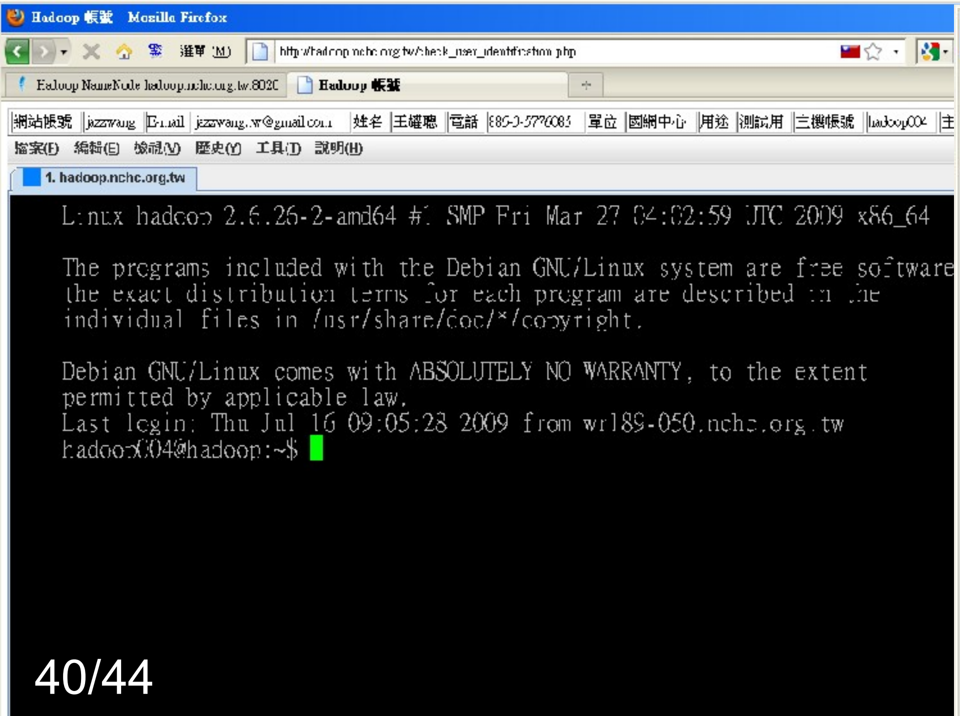
trac
Integrated SCM & Project Management

root / hadoop-register

Name ▲	Size	Rev	Age	Last
../				
▷ etc		103	4 weeks	wa
adduser.php	1.3 kB	85	6 weeks	wa
check_activate_code.php	2.2 kB	85	6 weeks	wa

About hadoop.nchc.org.tw

- **DRBL Server x 1 (hadoop) with more space for /home and /tftpboot**
- **DRBL Client x 19 (hadoop101~hadoop119)**
- **Using Cloudera Hadoop Debian Packages**
- **Use drbl-hadoop and cloudera's init.d script to deploy hadoop**
- **Use hadoop-register to host web service and ssh applet**

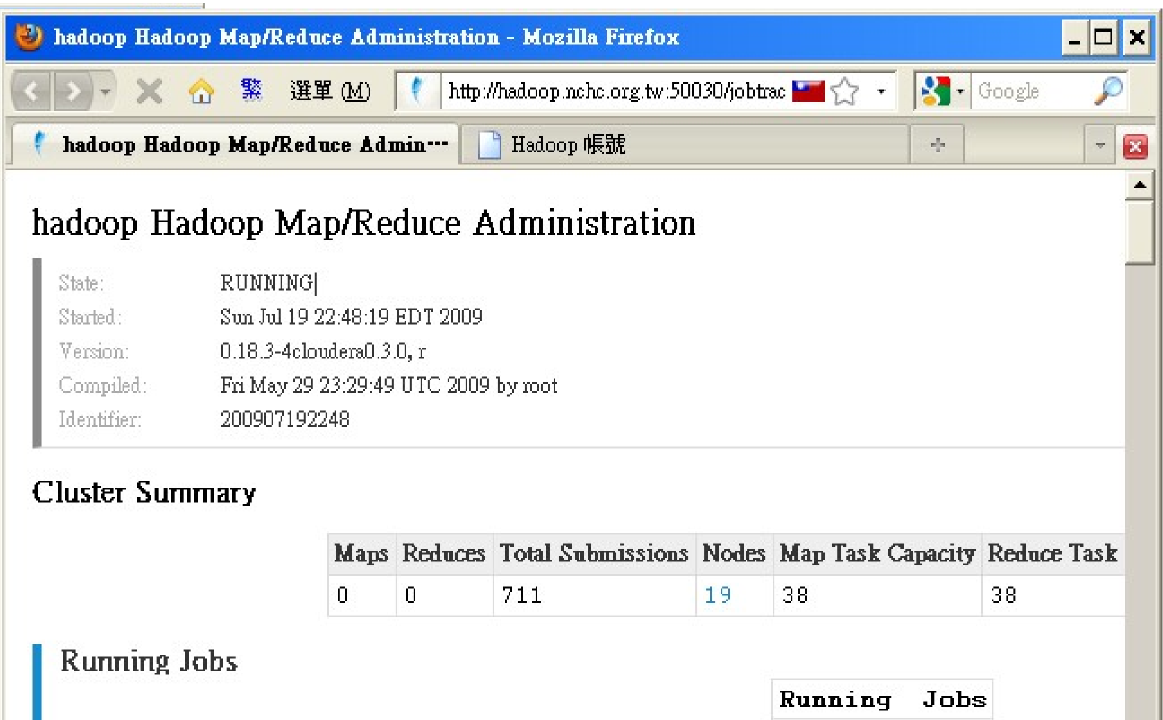

Hadoop 帳號 Mozilla Firefox
選單 (M)
http://hadoop.nchc.org.tw/check_user_identification.php
Hadoop NameNode hadoop.nchc.org.tw:8020
Hadoop 帳號
檔案(F) 編輯(E) 檢視(V) 歷史(Y) 工具(T) 說明(H)
1. hadoop.nchc.org.tw

網站帳號 jazzwang E-mail jazzwang.tw@gmail.com 姓名 王耀聰 電話 886-3-5776085 單位 國網中心 用途 測試用 主機帳號 hadoop004 主

```
Linux hadoop 2.6.26-2-amd64 #1 SMP Fri Mar 27 04:02:59 UTC 2009 x86_64

The programs included with the Debian GNU/Linux system are free software
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Thu Jul 16 09:05:28 2009 from wr189-050.nchc.org.tw
hadoop004@hadoop:~$
```


hadoop Hadoop Map/Reduce Administration - Mozilla Firefox
選單 (M)
http://hadoop.nchc.org.tw:50030/jobtrac
Google
hadoop Hadoop Map/Reduce Admin···
Hadoop 帳號

hadoop Hadoop Map/Reduce Administration

State: RUNNING
Started: Sun Jul 19 22:48:19 EDT 2009
Version: 0.18.3-4cloudera0.3.0, r
Compiled: Fri May 29 23:29:49 UTC 2009 by root
Identifier: 200907192248

Cluster Summary

Maps	Reduces	Total Submissions	Nodes	Map Task Capacity	Reduce Task
0	0	711	19	38	38

Running Jobs

Running Jobs

Lesson Learn

- **Cloudera Hadoop Package use init.d script to start/stop ...**
  - **name node, data node, job tracker, task tracker**
- **Creat 500 users in advanced** ：
  - **Use DRBL build-in command /opt/drbl/sbin/drbl-useradd**
- **Setup default HDFS home directory**
  - **Use for loop to run “hadoop fs -mkdir tmp“ for each user**
- **Setup permission of user HDFS folders**
  - **Use for loop to run “hadoop dfs -chown $(id) /usr/$(id)“**
- **HDFS use the space of /var/lib/hadoop/cache/hadoop/dfs**
- **MapReduce use the space of /var/lib/hadoop/cache/hadoop/mapred**

結論　Conclusion

- **Thanks to Cloudera to provide Hadoop related packages.**
- **Benefits**
  - **DRBL save your time and money. It make your life easier.**
  - **It's developed by Taiwan developers. Easy to communicate.**
  - **Using Network Booting could save power consumption ....**
- **Weakness**
  - **DRBL-Hadoop currently is only good for building experimental Hadoop Cluster.**
  - **If you are looking for operational product, maybe you can try SmartFrog.**

PART 2 -2:

Live Demo

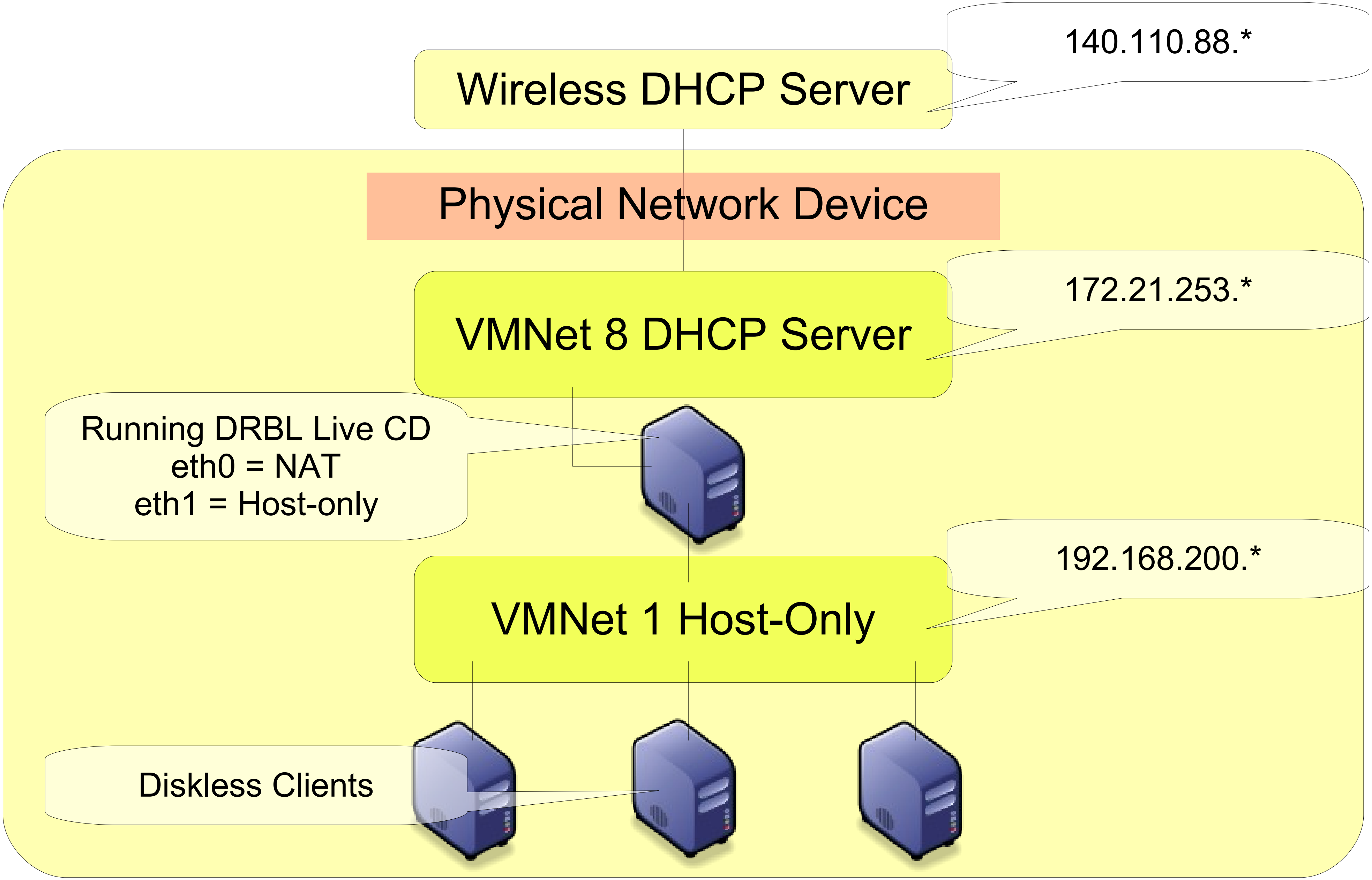
Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

Powered by DRBL

Demo Network Topology



Questions?

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

Powered by DRBL