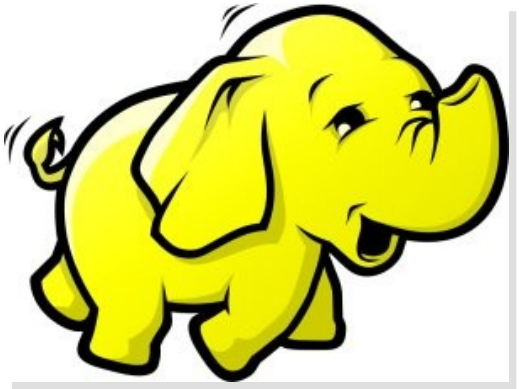
當企鵝龍遇上小飛象

DRBL-Hadoop

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw



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

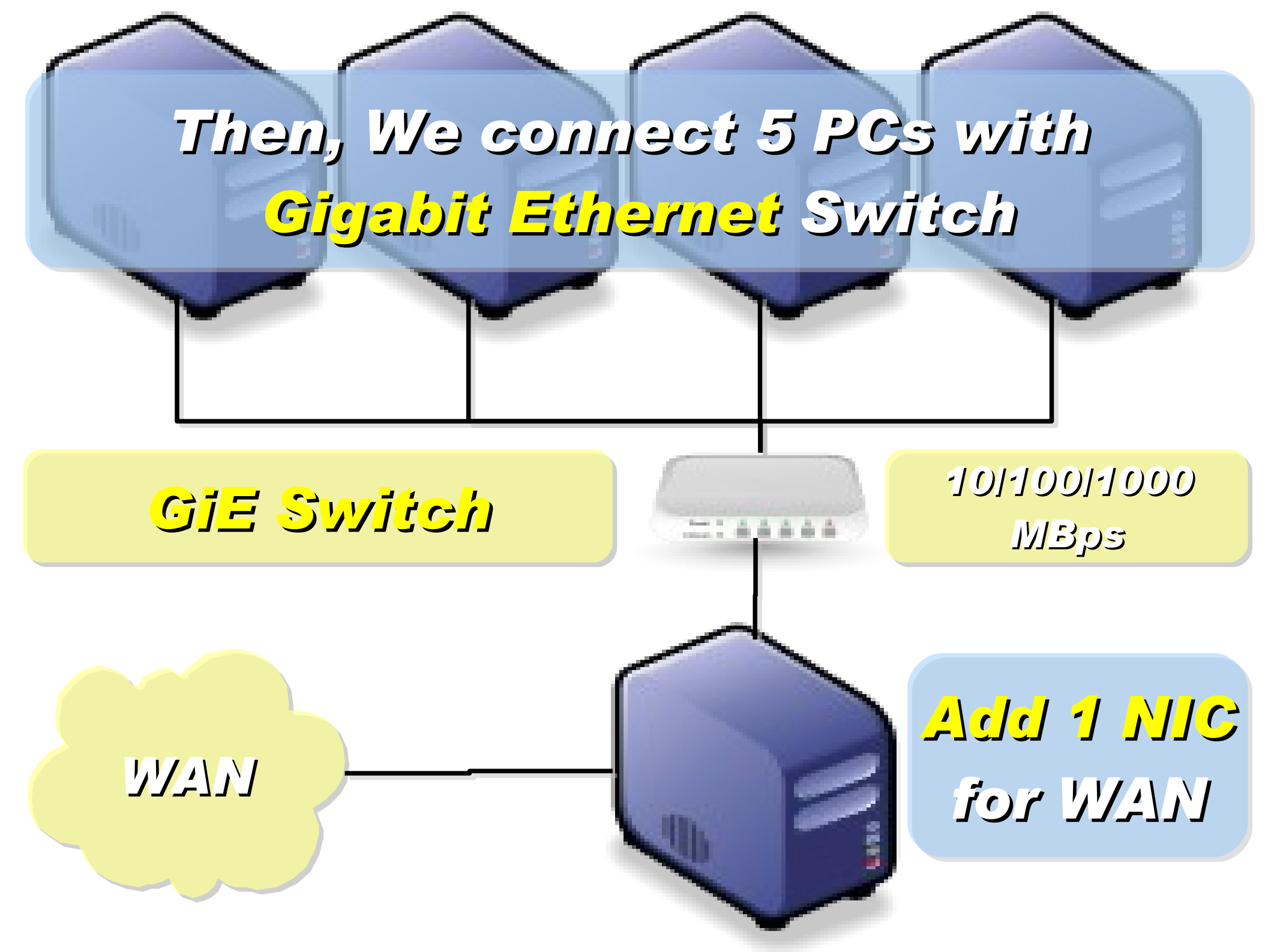
At First, We have “ 4 + 1 ” PC Cluster

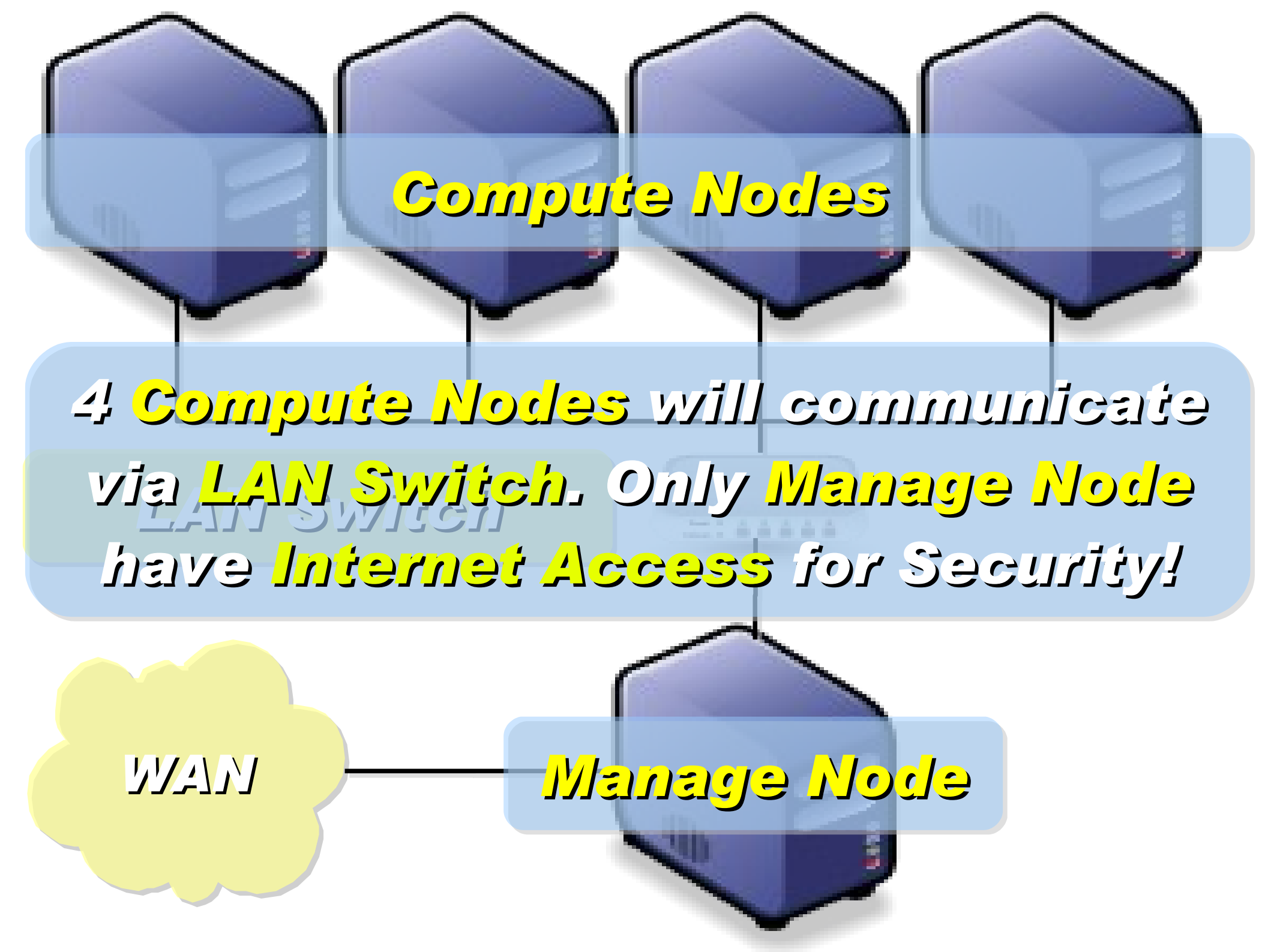
It'd better be

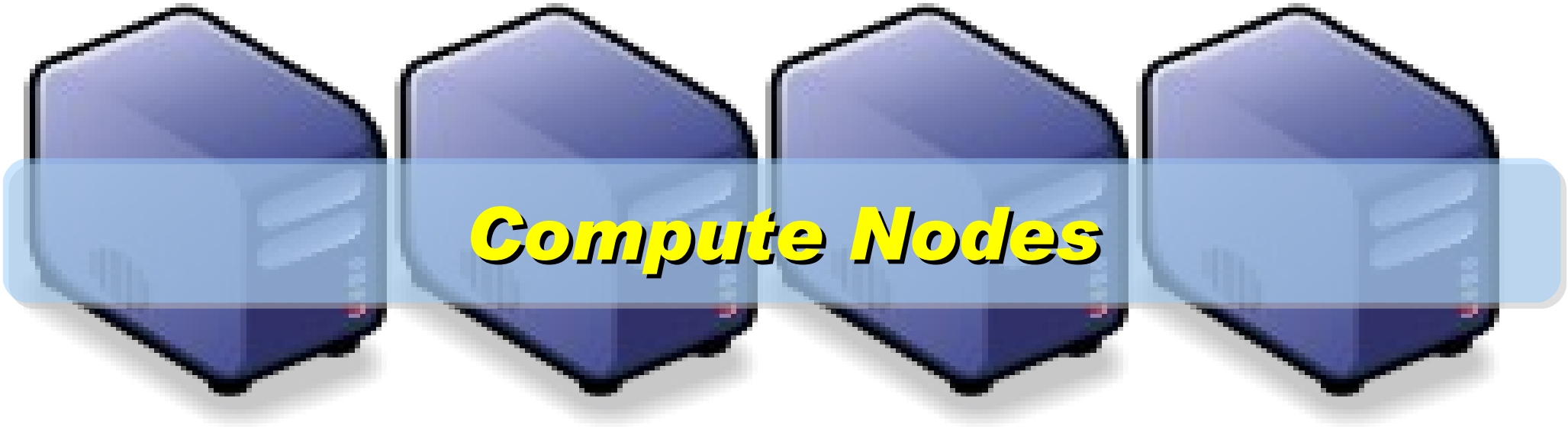
2^n

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



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.

OpenPBS

File Sharing

NFS

Extra

Messaging

Account Mgnt.

MPICH

SSHD

NIS

YP

GCC

GNU Libc

Bash

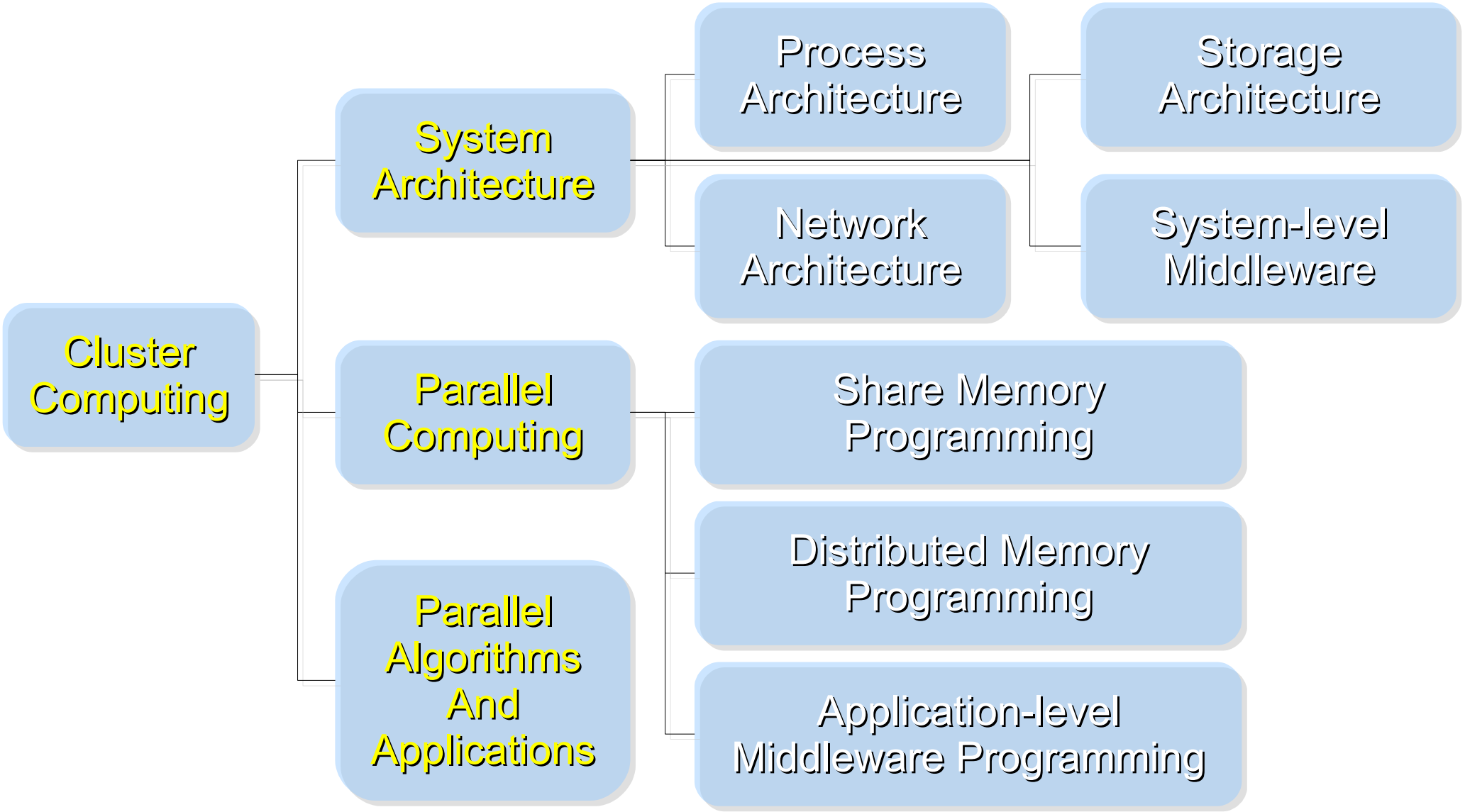
Perl

Kernel Module

Linux Kernel

Boot Loader

Research topics about PC Cluster



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
Kadeploy	ALL	No	Image	Yes	Yes	Yes
FAI	Debian-Based	Yes	Package	Yes	No	No

PART 2-1 :

Hadoop Deployment Tool

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

SmartFrog

- 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

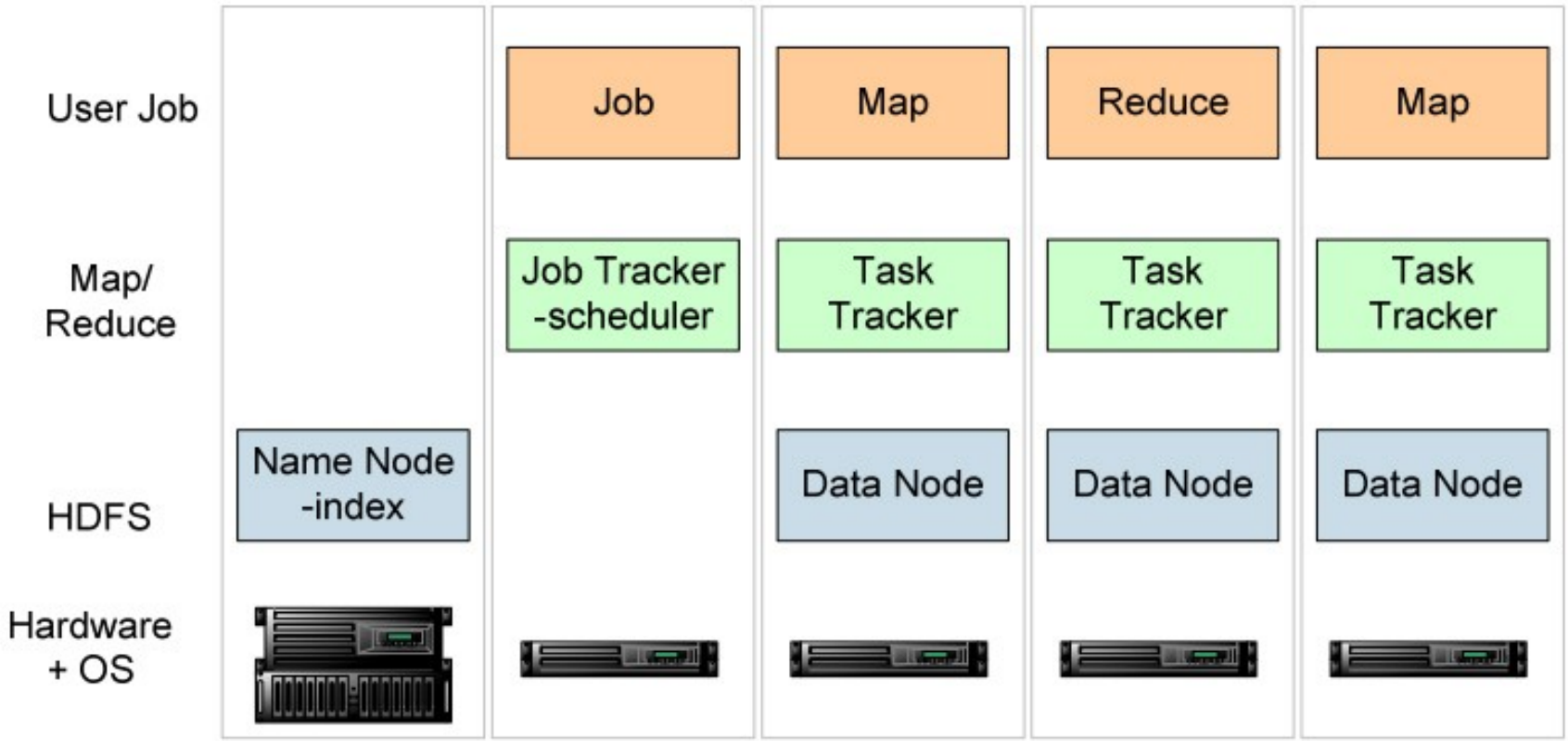
SmartFrog - HPLabs' CM tool

- 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



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

The hand-managed cluster

- Manual install onto machines
- SCP/FTP in Hadoop zip
- copy out hadoop-site.xml and other files
- edit /etc/hosts, /etc/rc5.d, SSH keys …
- Installation scales O(N)
- Maintenance, debugging scales worse

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

The locked-down cluster

- PXE Preboot of OS images
- RedHat Kickstart to serve up (see instalinux.com)
- Maybe: LDAP to manage state, or custom RPMs

Requires:

uniform images, central LDAP service, good ops team, stable configurations, home-rolled RPMs

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

CM-tool managed cluster

Configuration Management tools

- State Driven: observe system state, push it back into the desired state
- Workflow: apply a sequence of operations to change a machine's state
- Centralized: central DB in charge
- Decentralized: machines look after themselves

CM tools are the only way to manage big clusters

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

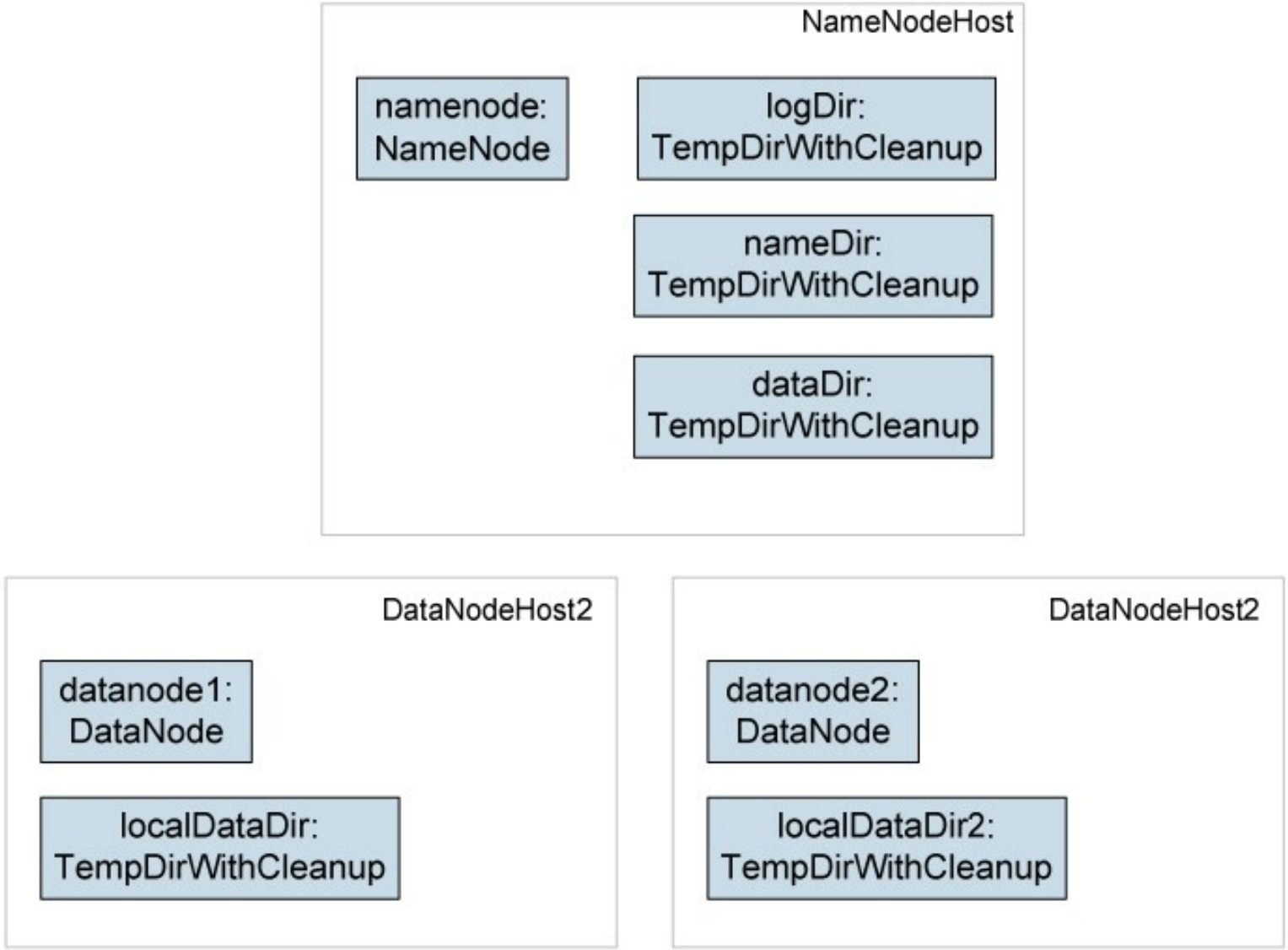
Model the system in the SmartFrog language

```
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

The runtime deploys the model



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

Steps to deployability

1. Configure Hadoop from an SmartFrog description
2. Write components for the Hadoop nodes
3. Write the functional tests
4. Add *workflow* components to work with the filesystem; submit jobs
5. Get the tests to pass

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

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

Powered by DRBL

何謂企鵝龍 DRBL ??

- ***Diskless Remote Boot in Linux***
- 網路是便宜的，人的時間才是昂貴的。
- 企鵝龍簡單來說就是.....
  - 用網路線取代硬碟排線
  - 所有學生的電腦都透過網路連接到一台伺服器主機

Powered by DRBL

Diskfull PC

=

+

+

Diskless PC

Server

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

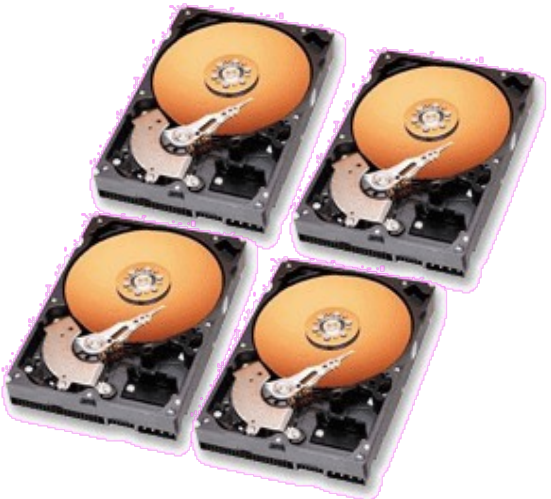
何謂再生龍 Clonezilla ??

- **Clone (複製) + zilla = Clonezilla (再生龍)**
- 裸機備分還原工具
- **Norton Ghost** 的自由軟體版替代方案

Disk to Disk

Disk to Image

Image to N Disks



降低資訊教育管理成本

需要「化繁為簡」的解決方案！

一般國內小學的電腦教室

☑人力、時間成本高

教師 1 人維護管理多組設備

教學同時分派或收集作業

☑設備維護成本高

需分別處理設定 (每班約 40 台)

如：電腦中毒、環境設定

系統操作問題、開關機、

備份還原等

平衡商業軟體與知識教育

知識和軟體都需要讓孩子「帶著走」！

☑商業軟體授權高成本

在校學習，也需回家複習

學校每台(平均)2萬

學生家用(平均)4萬

☑知識與法治的學習

教育知識，也需教育尊重

尊重智財權觀念

國網中心自由軟體開發

多元化資訊教學的新選擇！

以個人叢集電腦 (PC Cluster) 經驗發展 DRBL&Clonezilla

企鵝龍 DRBL

(Diskless Remote Boot in Linux)

適合將整個電腦教室轉換
成純自由軟體環境

再生龍 Clonezilla

適用完整系統備份、裸機
還原或災難復原

是自由！不是免費…

分送、修改、存取、使用軟體的自由。免費是附加價值。

企鵝龍 DRBL 與再生龍 Clonezilla

電腦教室管理的新利器！

■ 以每班 40 台電腦為估算單位

DRBL&Clonezilla	未使用	使　用
管理簡化	分別管理40台	管理 1台 伺服器
硬體設備成本	每台都需配備周邊硬體	伺服器控制，節約每台學生機之周邊硬體
軟體授權成本	40台:3000*40=120,000 (MS Windows授權1台電腦之授權費NT$3,000)	軟體授權 NT$0
合法複製、分享	需負擔授權費	複製合法 NT$0
多元化電腦教學	不同系統無法並存	Linux 與M$ Windows可並存

教育單位採用 DRBL

高速計算研究
資料儲存備援

降低管理維護成本
帶動自由軟體使用
節樽軟體授權成本（估計）

NT. 98,595,000 元

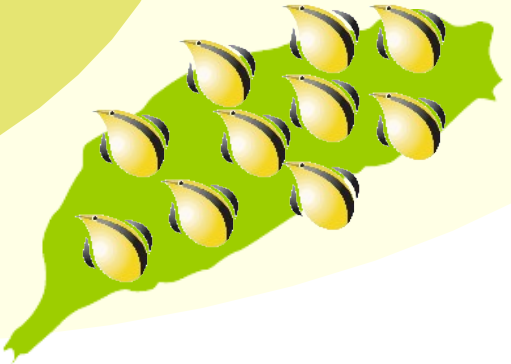
以某商業獨家軟體每機 3000 元授權費計，
每班 35 台電腦 (3000*35*939)

擴至全國各單位

節省龐大軟體授權費

降低台灣盜版率

提升台灣形象



PART 1-3 :

企鵝龍的開機原理

Jazz Wang

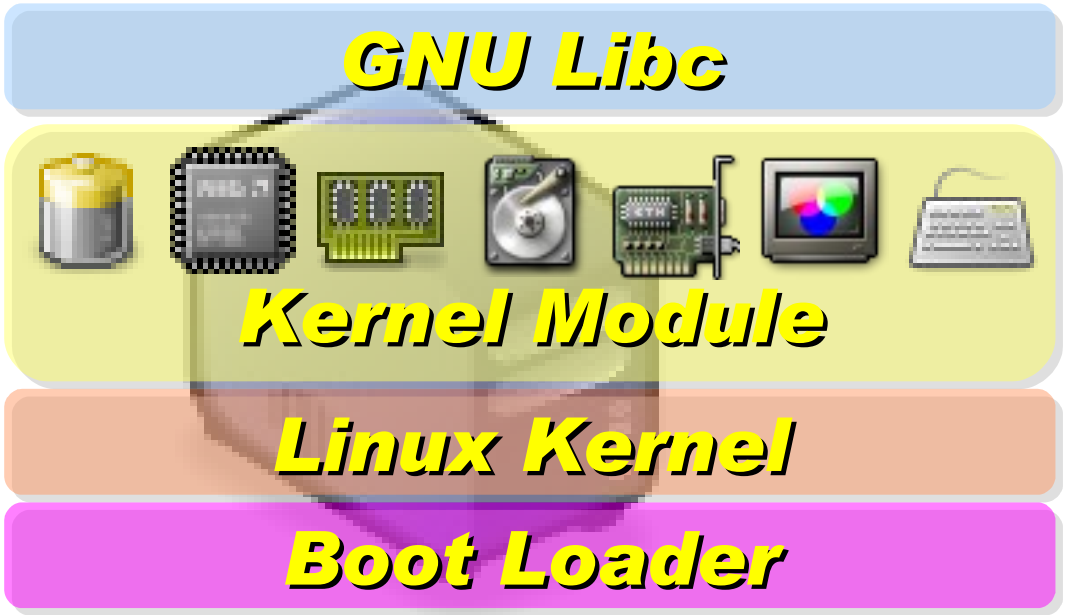
Yao-Tsung Wang

jazz@nchc.org.tw

Powered by DRBL

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

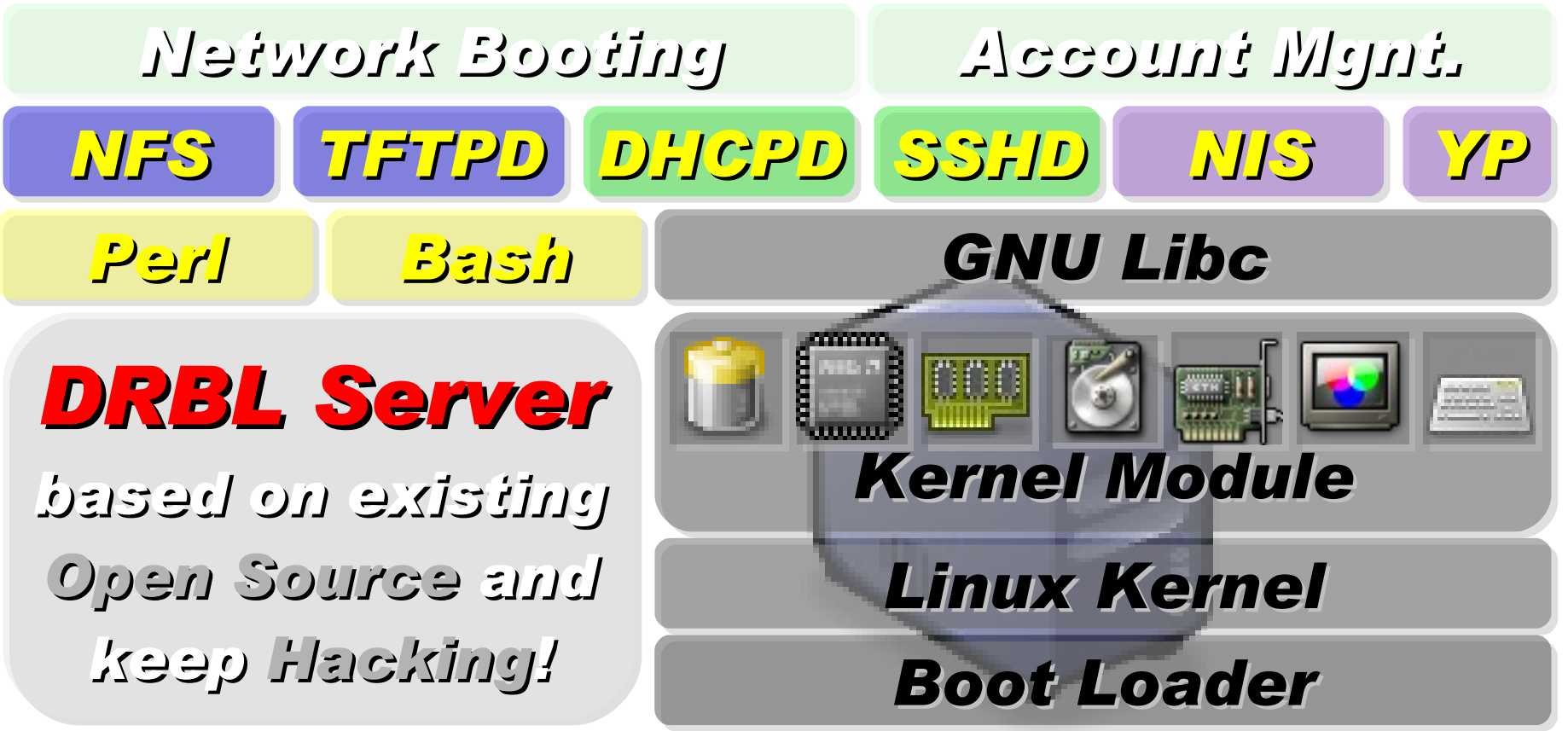
Redhat, Fedora, CentOS, Mandriva, Ubuntu, Debian, ...



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

There are lots of service needed:

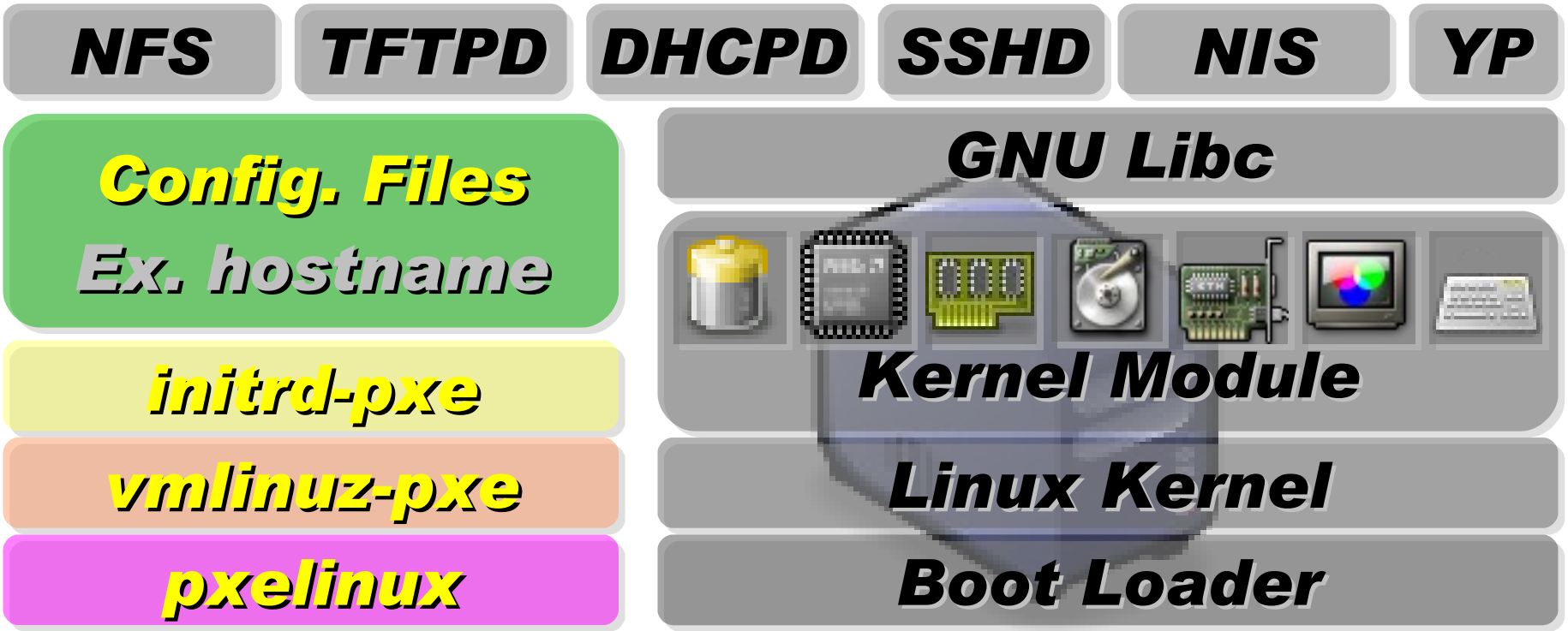
SSHD, DHCPD, TFTPD, NFS Server,
NIS Server, YP Server ...



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



3nd, We enable PXE function in BIOS configuration.

BIOS PXE | BIOS PXE | BIOS PXE | BIOS PXE

NFS | TFTPD | DHCPD | SSHD | NIS | YP

Config. Files
Ex. hostname

initrd-pxe

vmlinuz-pxe

pxelinux

GNU Libc

Kernel Module

Linux Kernel

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

initrd-pxe

vmlinuz-pxe

pxelinux

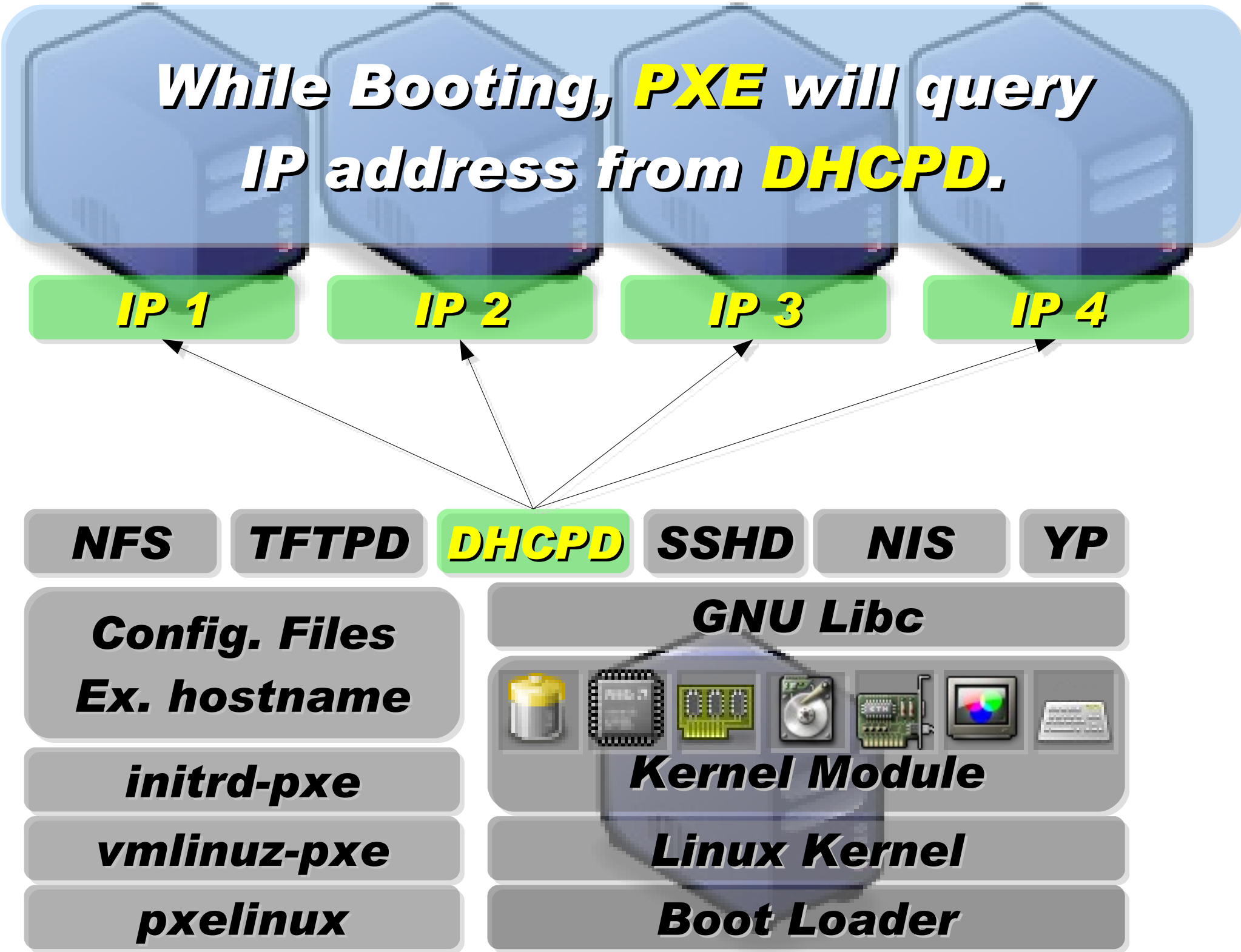
GNU Libc

Kernel Module

Linux Kernel

Boot Loader

While Booting, PXE will query IP address from DHCPD.



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

initrd-pxe

vmlinuz-pxe

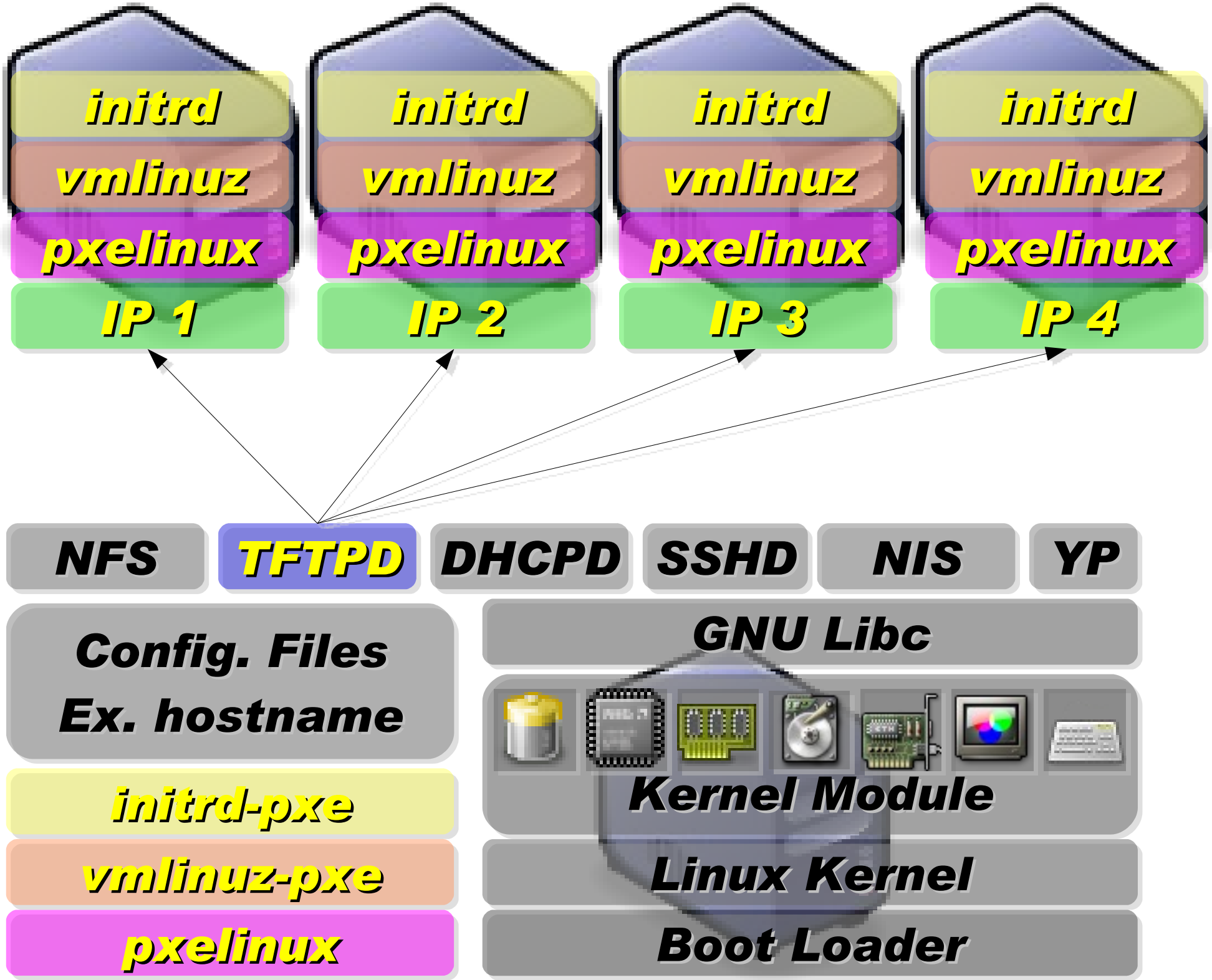
pxelinux

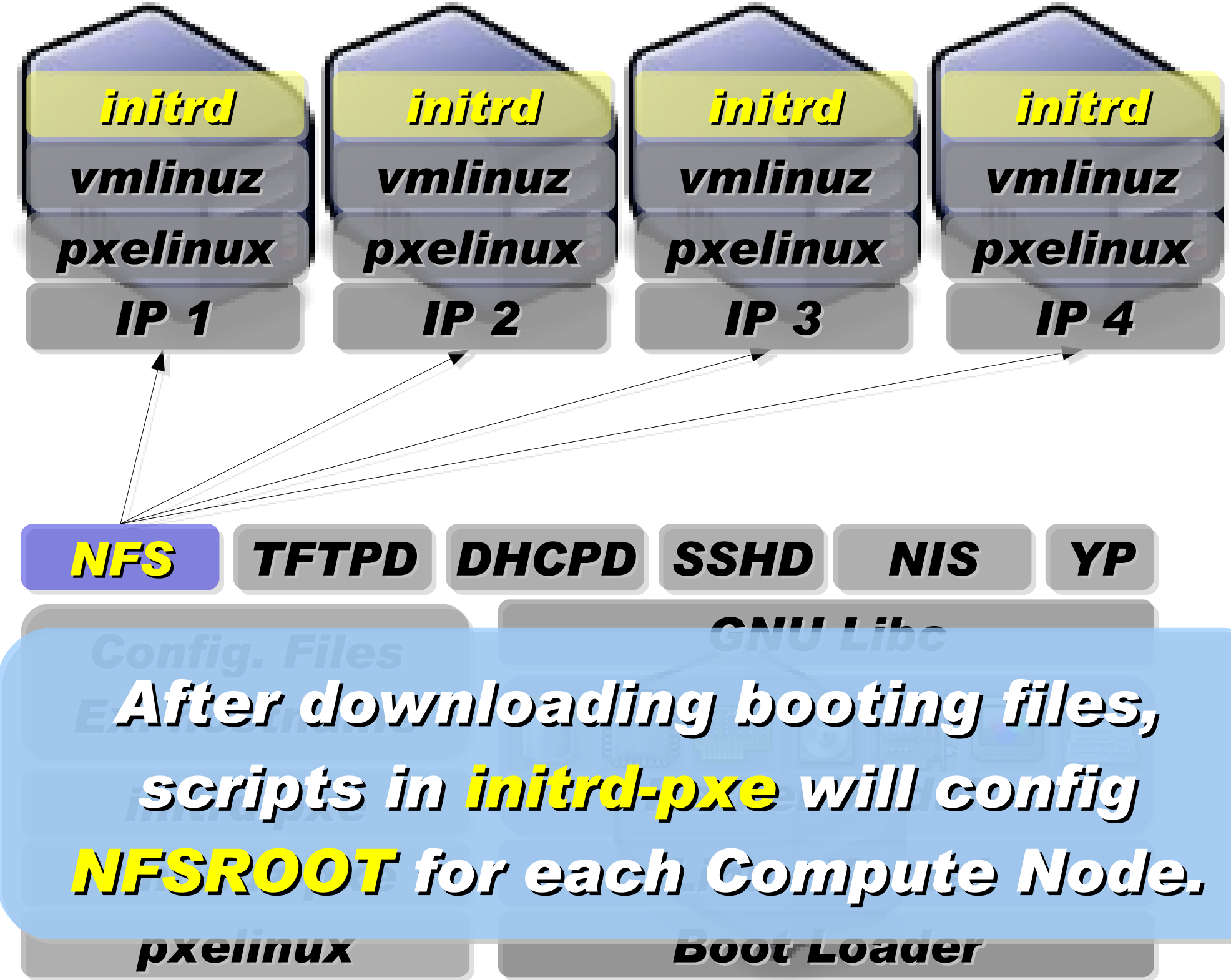
GNU Libc

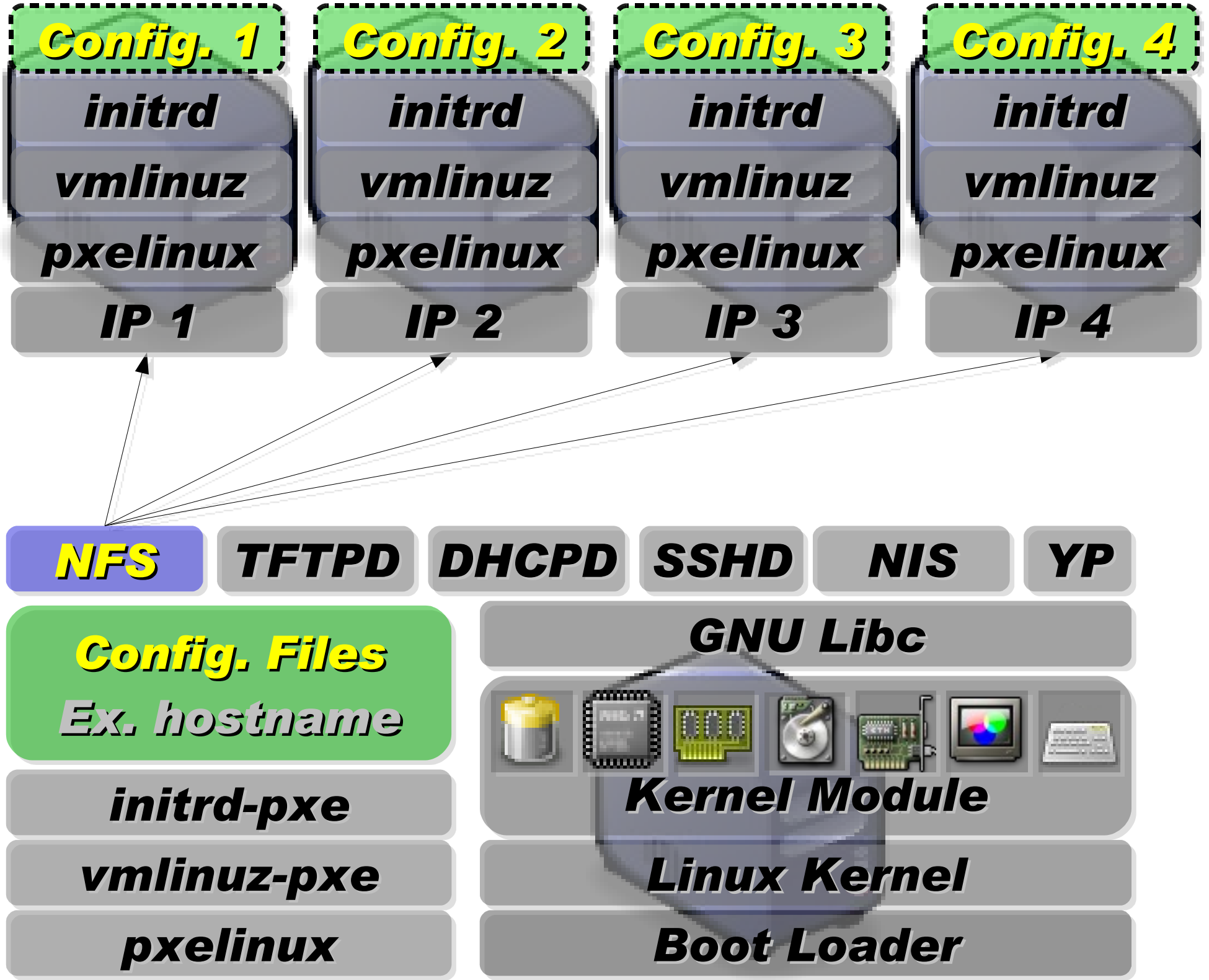
Kernel Module

Linux Kernel

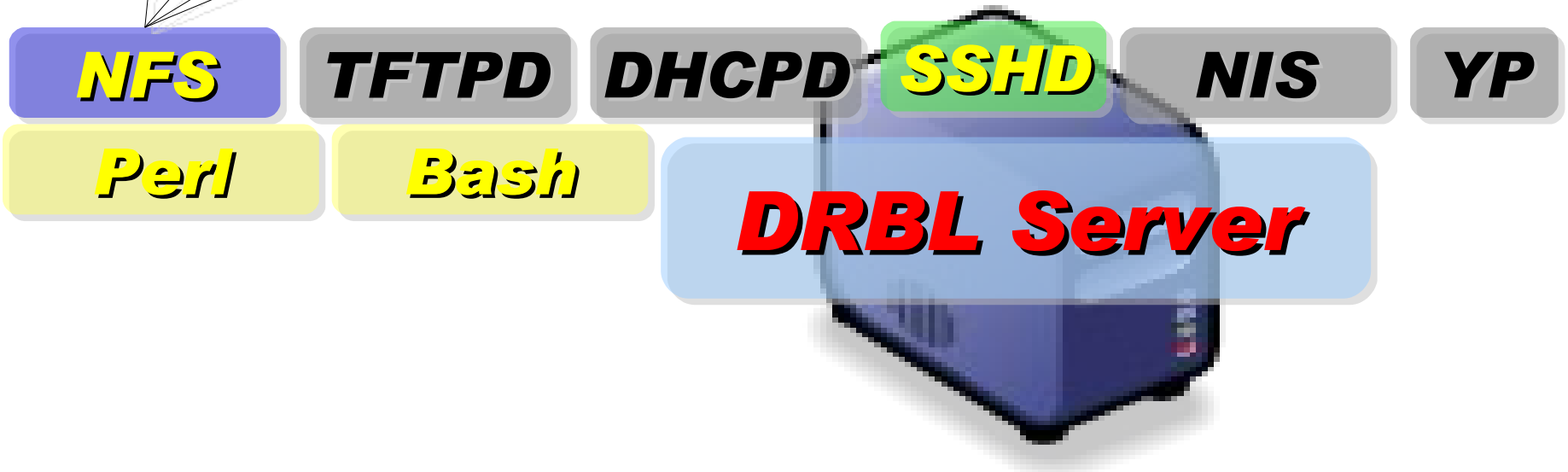
Boot Loader

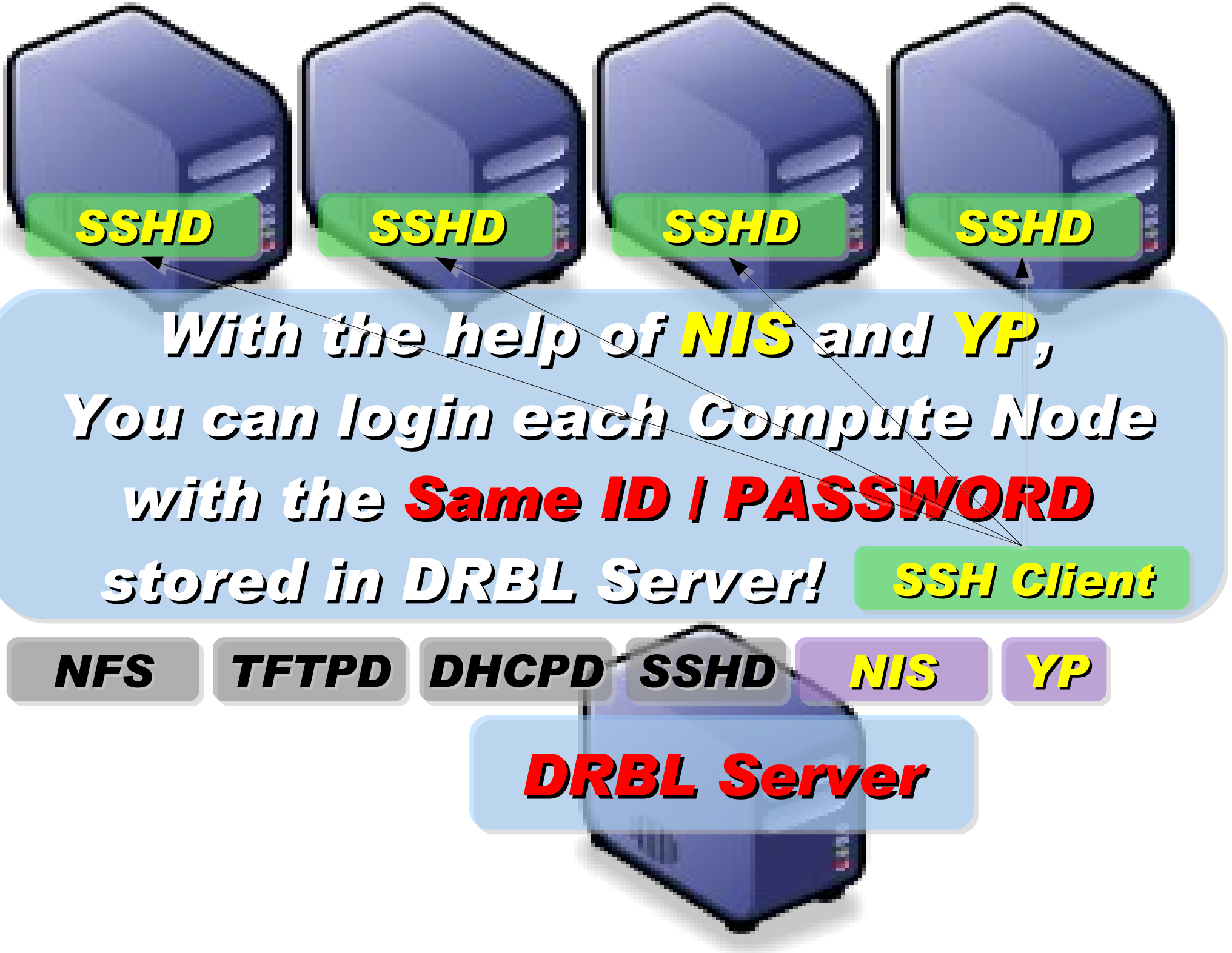

initrd
vmlinuz
pxelinux
IP 1
initrd
vmlinuz
pxelinux
IP 2
initrd
vmlinuz
pxelinux
IP 3
initrd
vmlinuz
pxelinux
IP 4
NFS
TFTPD
DHCPD
SSHD
NIS
YP
Config. Files
Ex. hostname
initrd-pxe
vmlinuz-pxe
pxelinux
GNU Libc
Kernel Module
Linux Kernel
Boot Loader


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


Config. 1
Config. 2
Config. 3
Config. 4
initrd
vmlinuz
pxelinux
IP 1
IP 2
IP 3
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

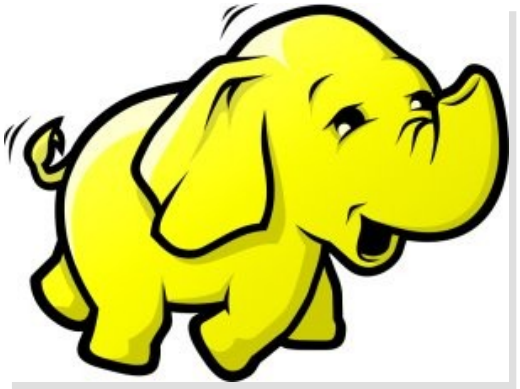
PART 2 -1:

當企鵝龍遇上小飛象

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw



使用 DRBL 佈署 Hadoop

- 仍在開發中，待整理套件
- **drbl-hadoop –** 掛載本機硬碟給 **HDFS** 用

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

- **hadoop-register –** 註冊網站與 **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

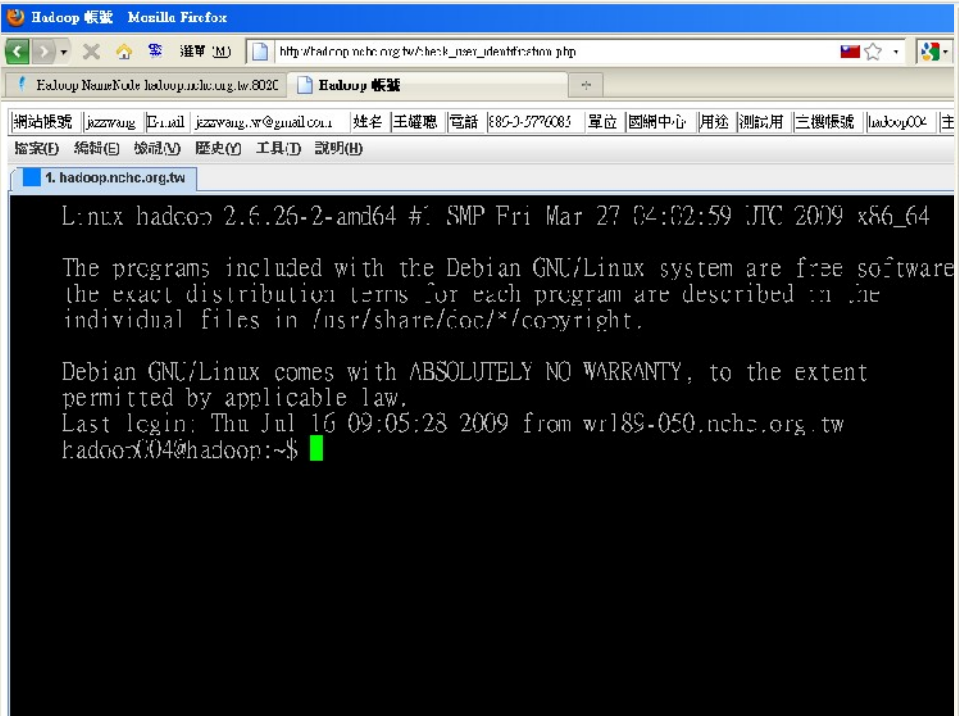
trac
Integrated SCM & Project Management

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

關於 hadoop.nchc.org.tw

- **DRBL Server - 1** 台 **(hadoop)** ，加大 **/home** 與 **/tftpboot** 空間。
- **DRBL Client - 19** 台 **(hadoop101~hadoop119)**
- 使用 **Cloudera** 的 **Debian** 套件
- 使用 **drbl-hadoop** 的設定跟 **init.d script** 來協助部署
- 使用 **hadoop-register** 來提供使用者註冊與 **ssh applet** 介面



Lesson Learn

- **Cloudera** 套件的好處：使用 **init.d script** 來啟動關閉
  - **name node, data node, job tracker, task tracker**
- 建立大量帳號：
  - 可透過 **DRBL** 內建指令完成 **/opt/drbl/sbin/drbl-useradd**
- 使用者預設 **HDFS** 家目錄
  - 跑迴圈切換使用者，下 **hadoop fs -mkdir tmp**
- 設定使用者 **HDFS** 權限
  - 跑迴圈切換使用者，下 **hadoop dfs -chown $(id) /usr/$(id)**
- **HDFS** 會使用 **/var/lib/hadoop/cache/hadoop/dfs**
- **MapReduce** 會使用 **/var/lib/hadoop/cache/hadoop/mapred**

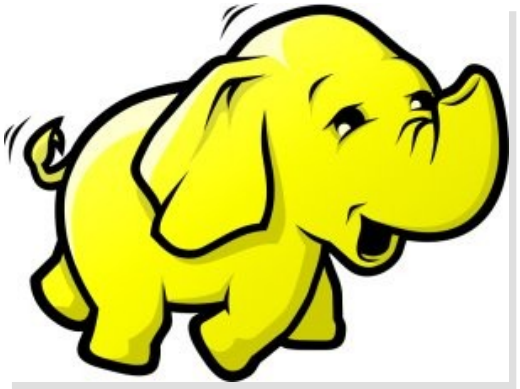
PART 2 -2:

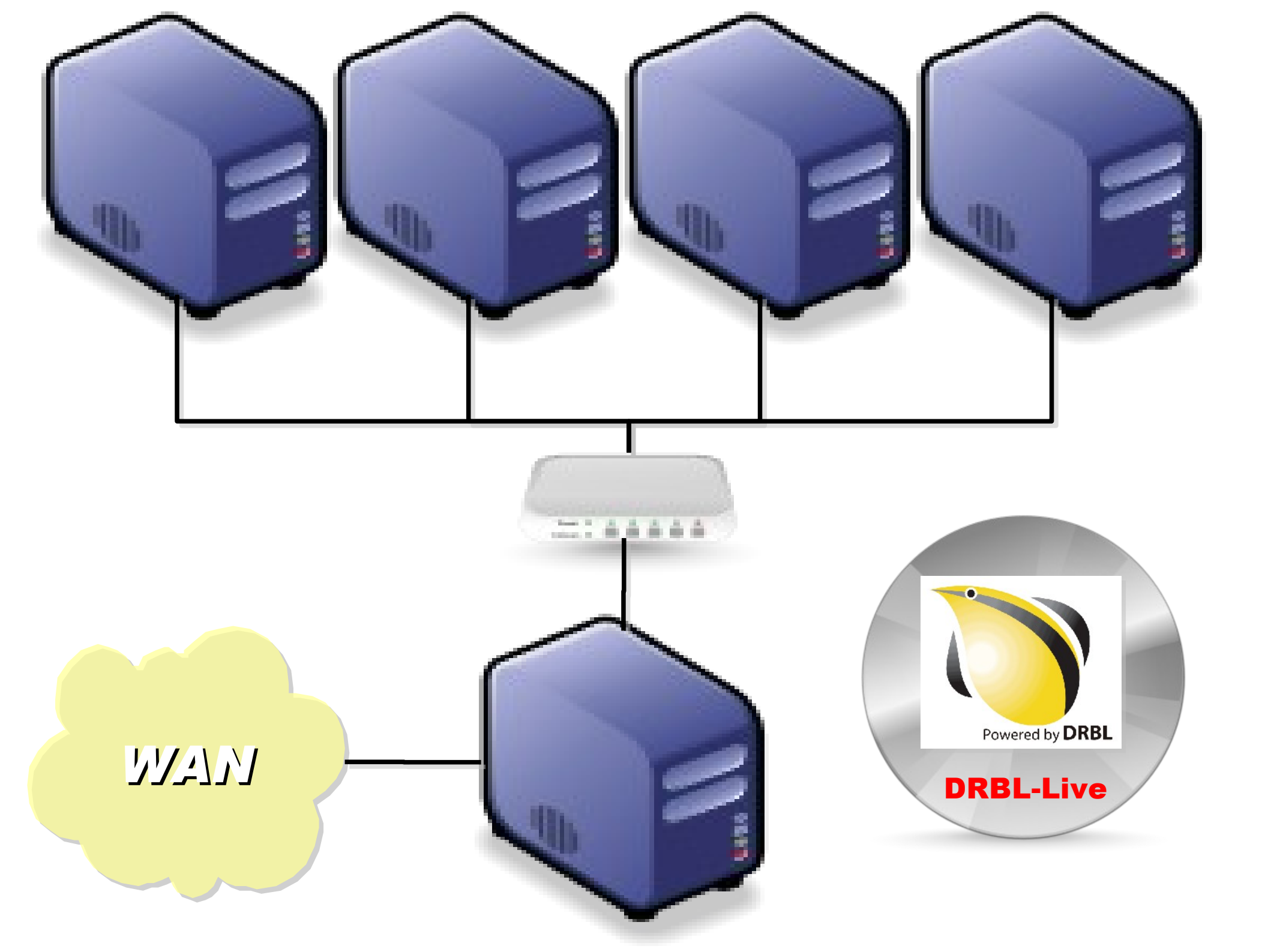
Live Demo

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw




WAN
Powered by DRBL
DRBL-Live

Demo with DRBL-Live CD

1. Boot Server with DRBL-Live CD

http://free.nchc.org.tw/drbl-live/stable/

2. Download DRBL-Hadoop Script

http://classcloud.org/drbl-hadoop-live.sh

http://classcloud.org/drbl-hadoop-live-run.sh

3. Follow the steps

http://classcloud.org/drbl-hadoop

Questions?

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

Powered by DRBL