運用企鵝龍打造多人 Hadoop 叢集

hadoop.nchc.org.tw 營運經驗分享

Building Multi-user Hadoop Cluster using DRBL & Clonezilla

Jazz Wang

Yao-Tsung Wang

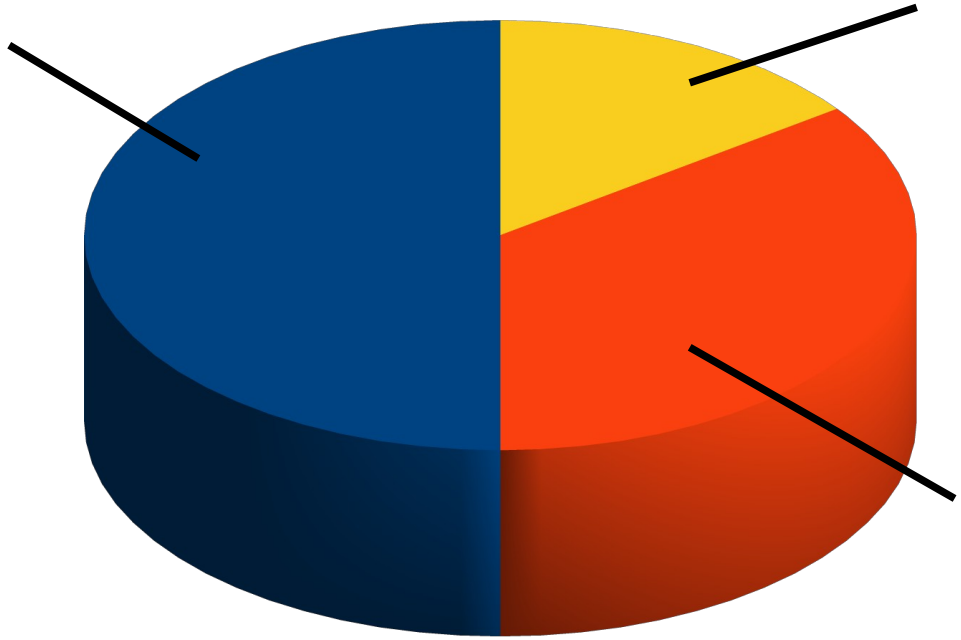
jazz@nchc.org.tw

Powered by DRBL

WHO AM I？這傢伙是誰啊？JAZZ？

- 講者介紹：
  - 國網中心　王耀聰　副研究員 / 交大電控碩士
  - jazz@nchc.org.tw
- 所有投影片、參考資料與操作步驟均在網路上
  - 由於雲端資訊變動太快，愛護地球，請減少不必要之列印。

FLOSS 使用者
Debian/Ubutnu
Access Grid
Motion/VLC
Red5
Debian Router
DRBL/Clonezilla
Hadoop



行動力薄弱的開發者
DRBL/Clonezilla
Hadoop Ecosystem

推廣者
DRBL/Clonezilla
Partclone/Tuxboot
Hadoop Ecosystem

運用企鵝龍打造多人 Hadoop 叢集

PART 1 :

叢集佈署工具簡介：企鵝龍與聰明蛙

PART 2 :

運用企鵝龍佈署資料探勘平台的經驗分享

- PaaS : Data Processing (DRBL-Hadoop)

PART 3 :

運用再生龍從小硬碟搬家到大硬碟

叢集佈署工具簡介：企鵝龍與聰明蛙

Introduction to SSI and CMT : DRBL & SmartFrog

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

傳統實驗室佈署電腦叢集的方法

1. Setup one **Template** machine

2. Cloning to multiple machine

3. Configure Settings

↓

4. Install **Job Scheduler**

↓

5. Running **Benchmark**

傳統方式容易面臨的叢集管理問題

Add New User Account ?

Upgrade Software ?

How to share user data ?

Configuration Syncronization

萬一您要佈署四千台以上的叢集呢??

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**

進階叢集佈署工具

- 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

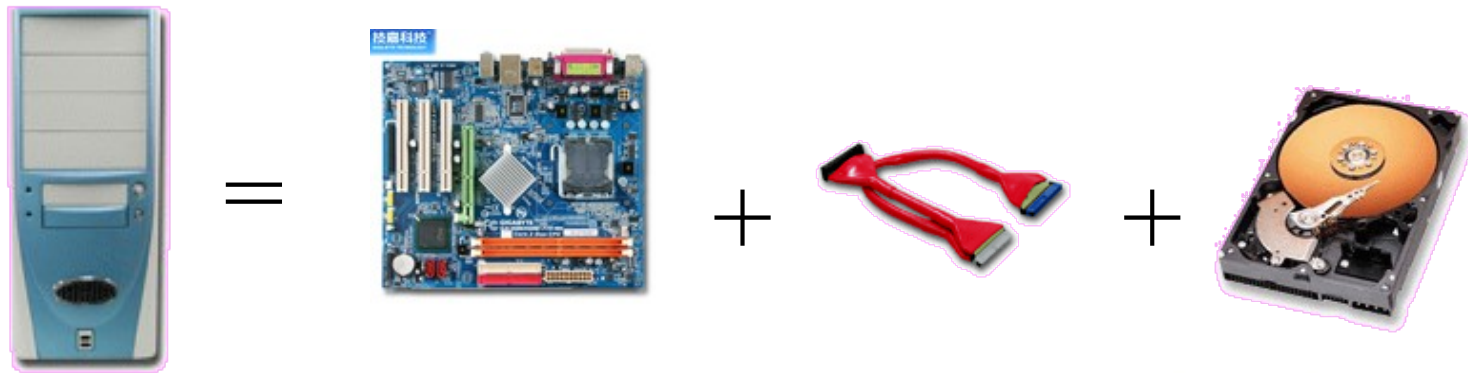
叢集佈署工具比較表

	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

國網中心企鵝龍（DRBL）簡介

- **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

Diskfull PC



Diskless PC

Server

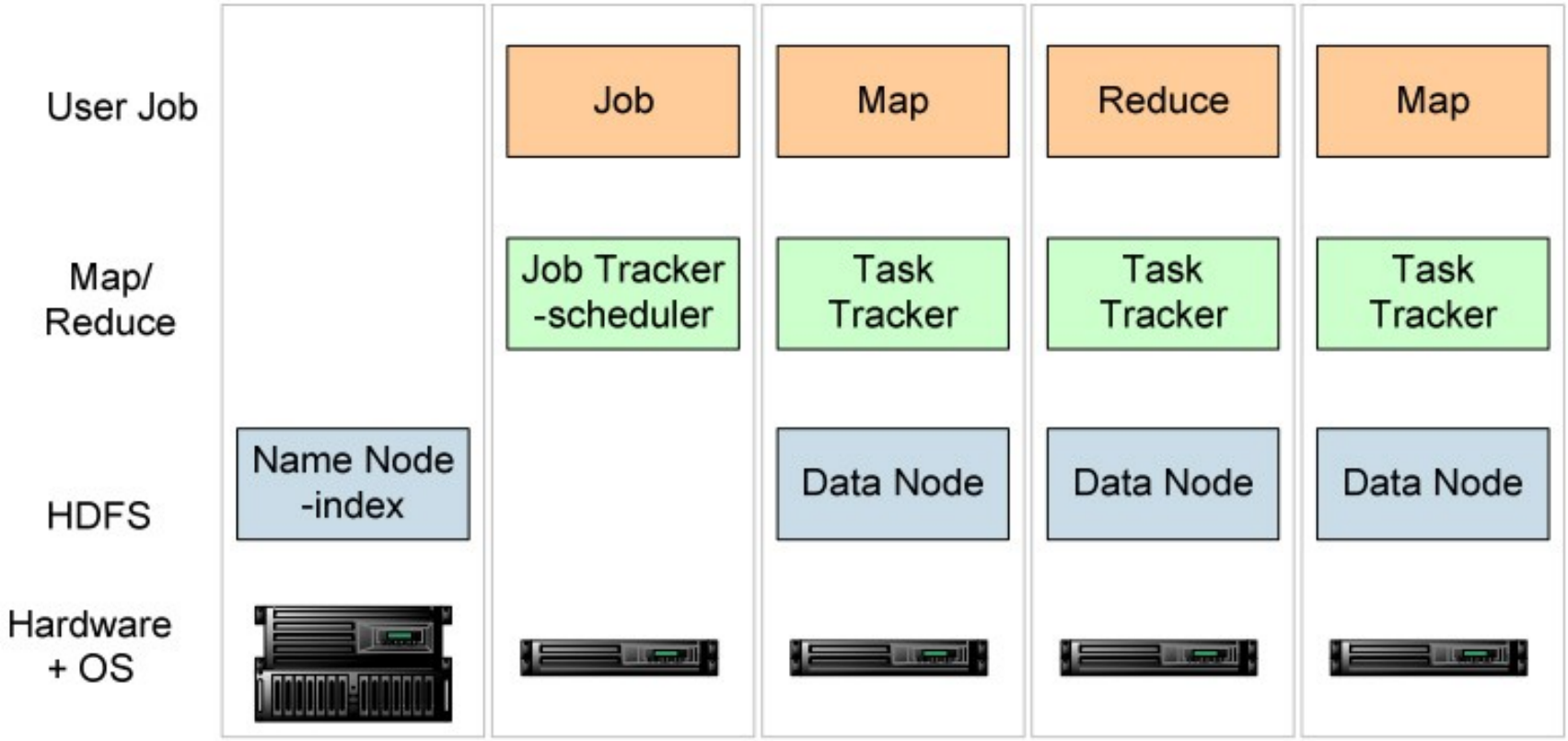
source: http://www.mren.com.tw

惠普實驗室的聰明蛙（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

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

企鵝龍的開機原理

Installation and Booting Procedure of DRBL

Jazz Wang

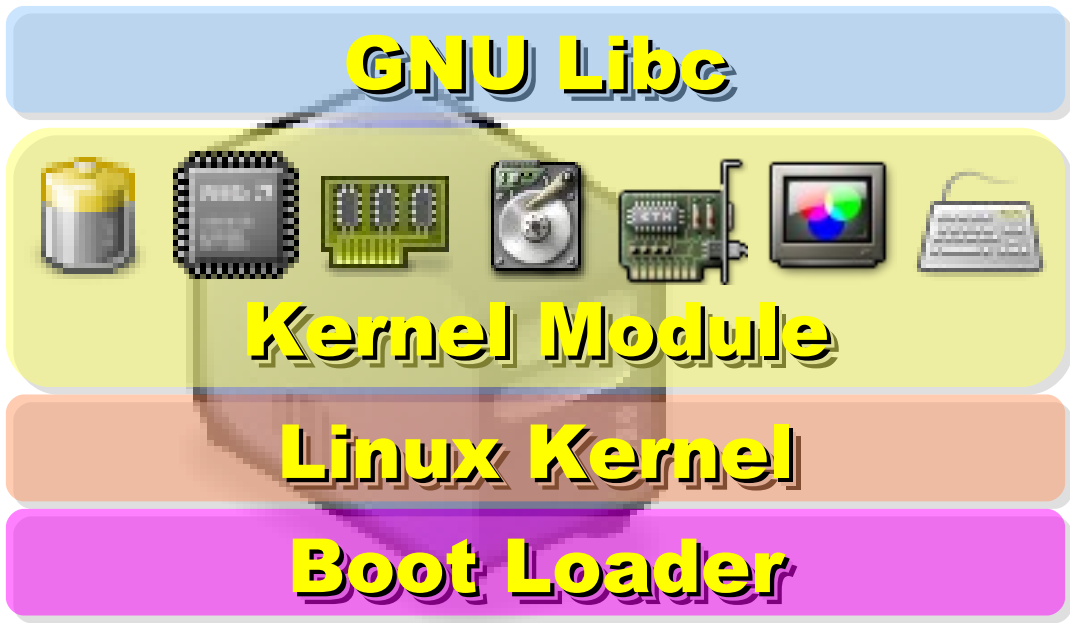
Yao-Tsung Wang

jazz@nchc.org.tw

1st, We install Base System of GNU/Linux on Management Node.

You can choose:

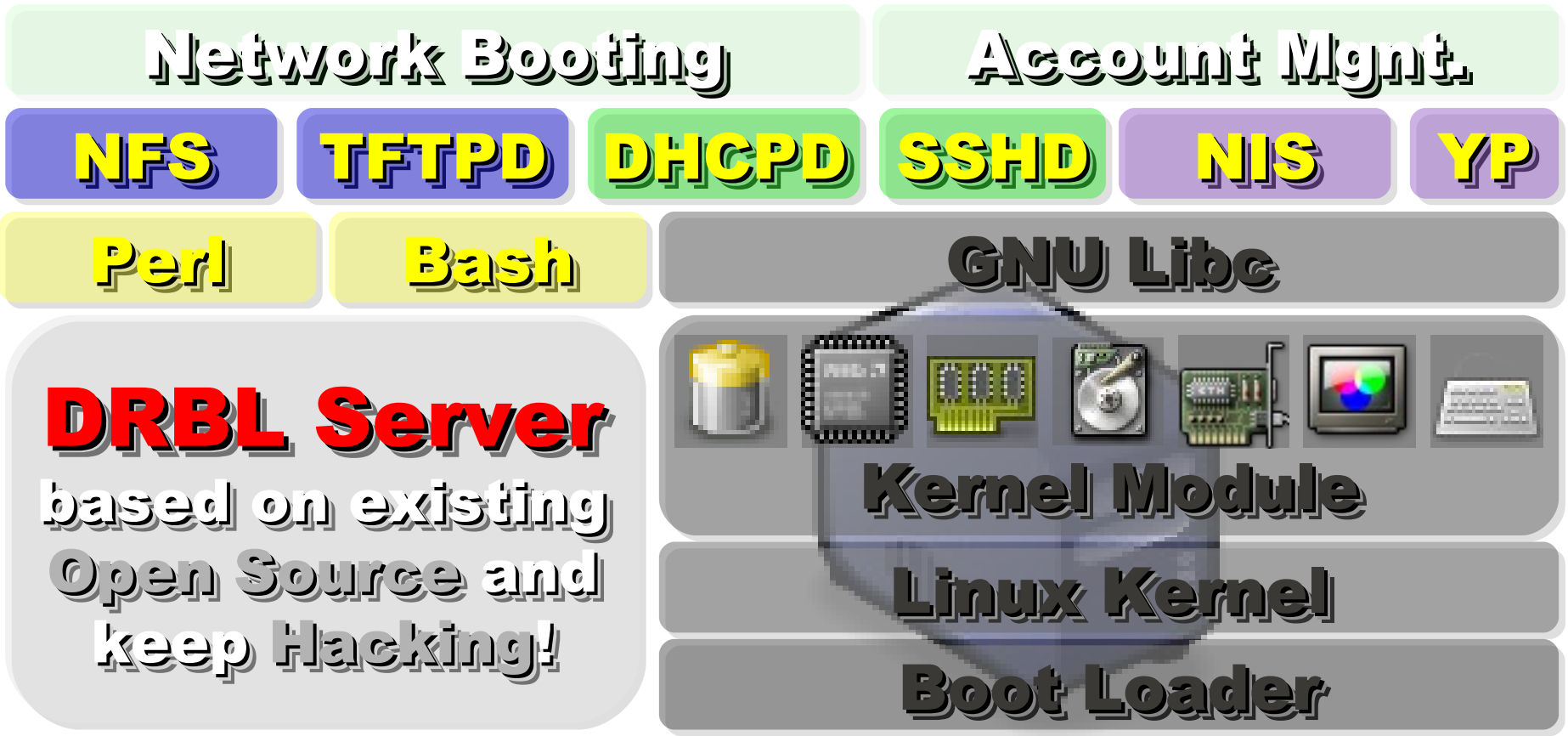
Redhat, Fedora, CentOS, Mandriva, Ubuntu, Debian, ...



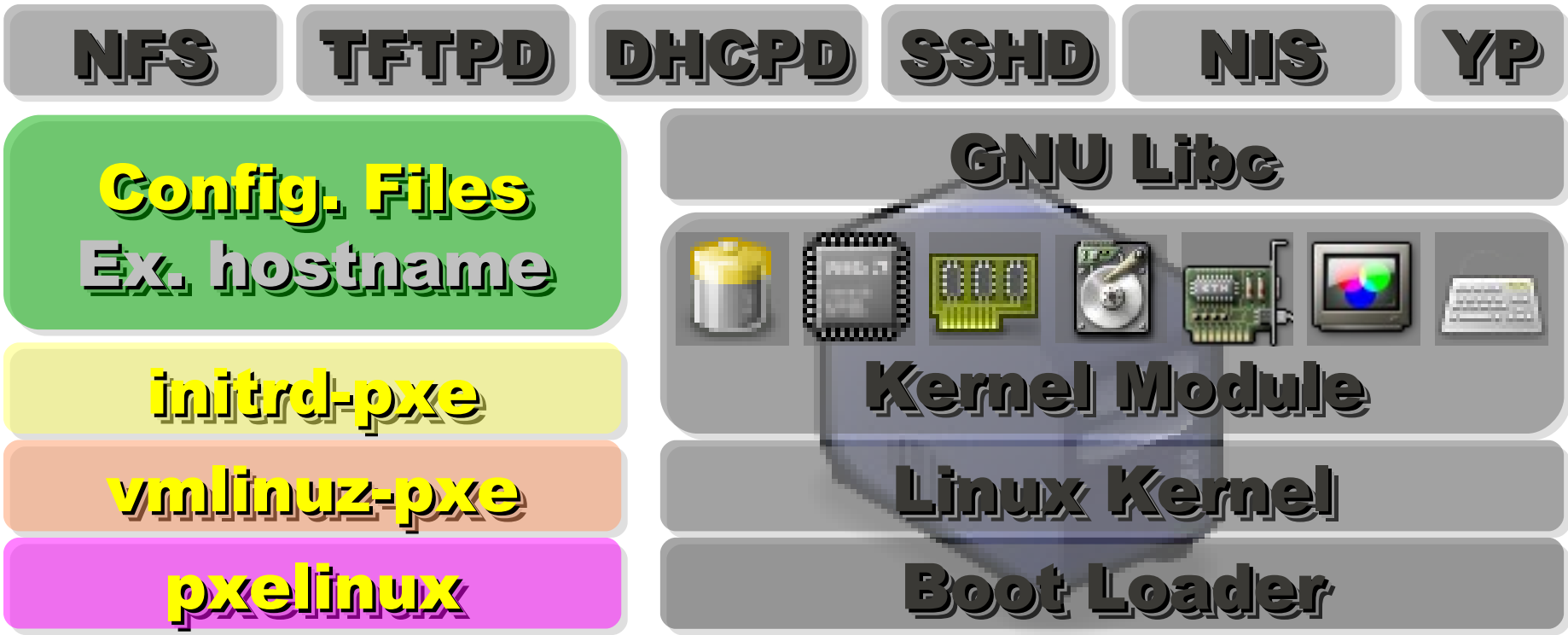
2nd, We install DRBL package and configure it as DRBL Server.

There are lots of service needed:

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



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

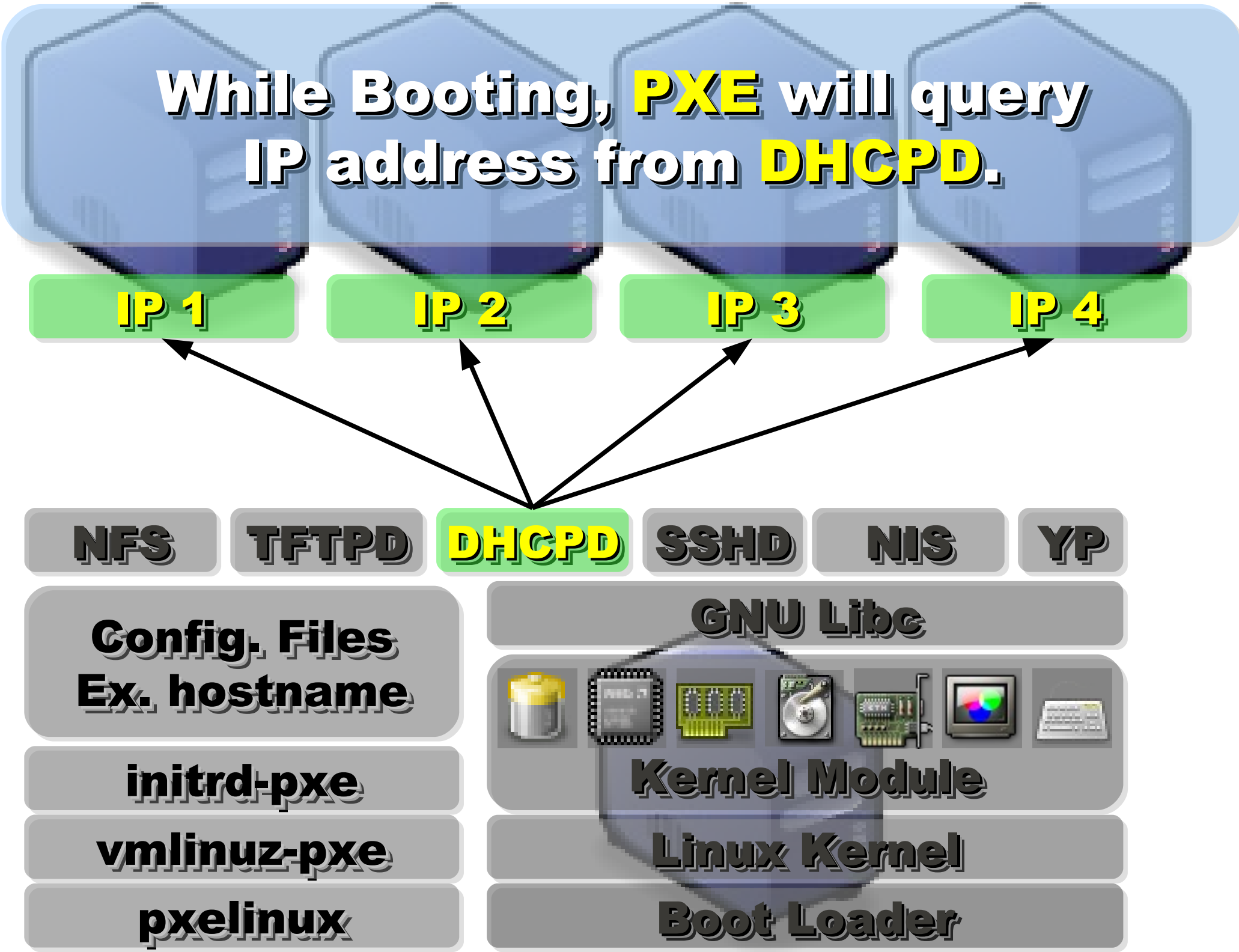
pxelinux

GNU Libc

Kernel Module

Linux Kernel

Boot Loader


While Booting, PXE will query IP address from DHCPD.
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

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

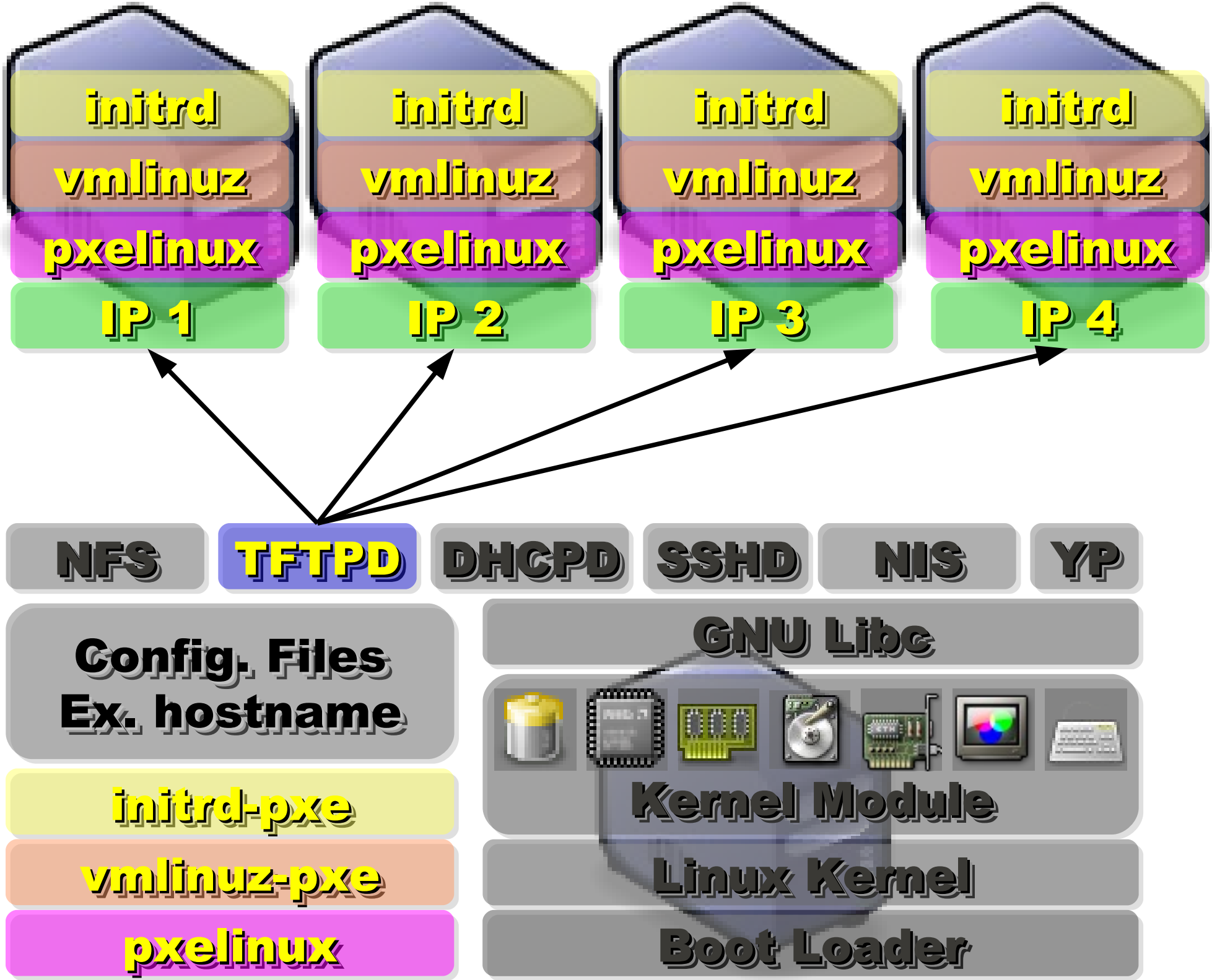
pxelinux

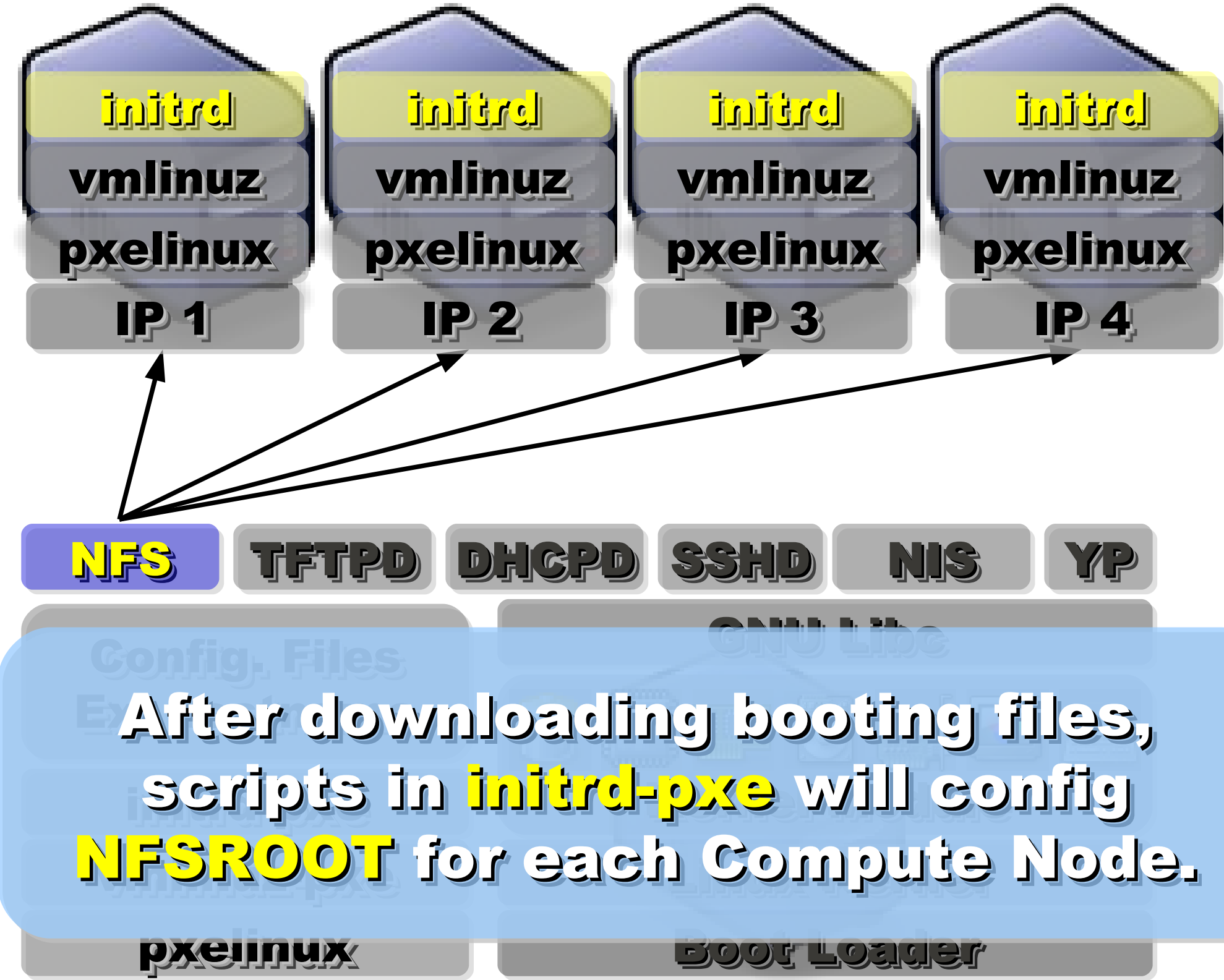
GNU Libc

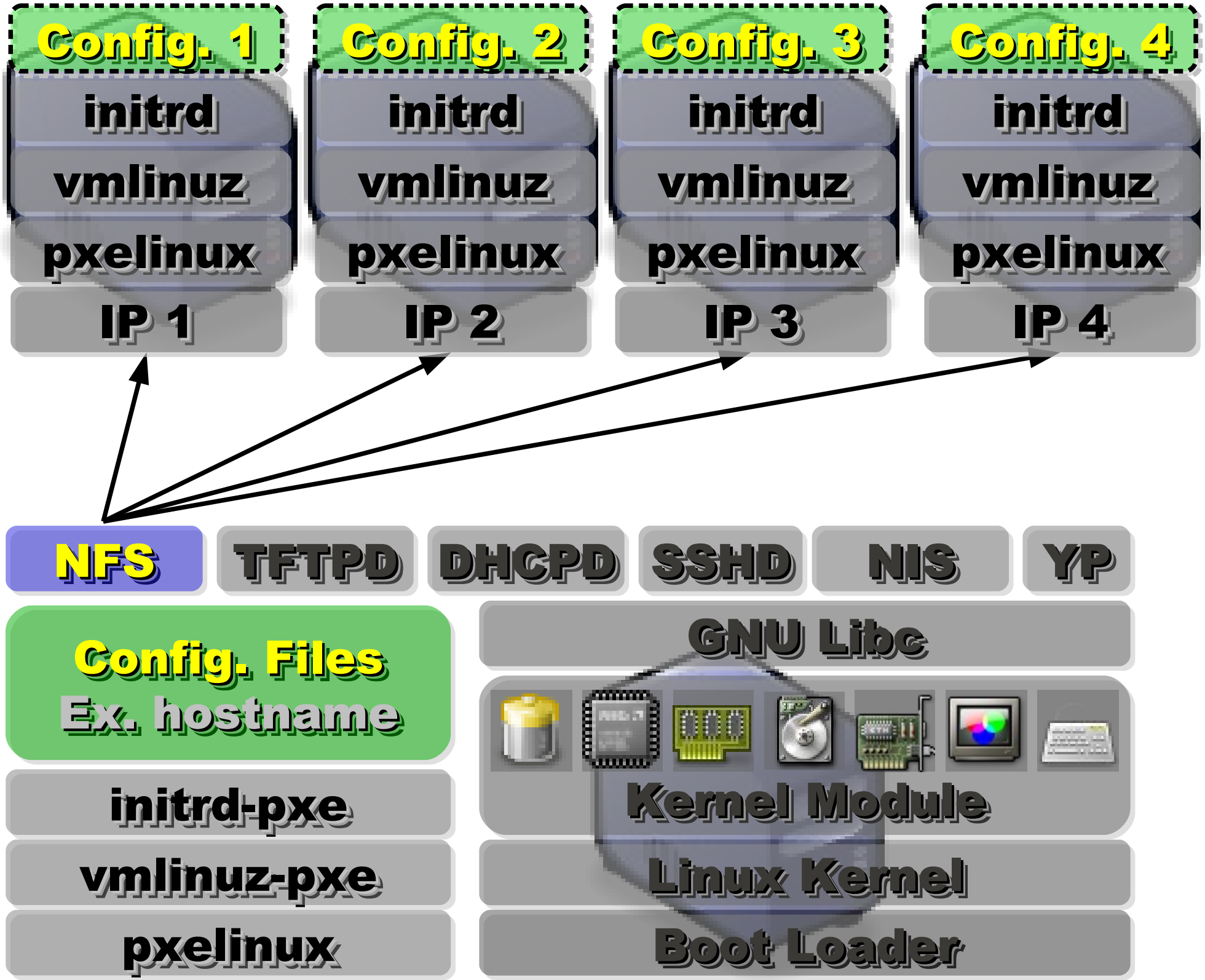
Kernel Module

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
Config. Files
pxelinux
Boot Loader
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


Perl
Bash
SSHD
Perl
Bash
SSHD
Perl
Bash
SSHD
Perl
Bash
SSHD
Applications and Services will also deployed to each Compute Node via NFS ....
NFS
TFTPD
DHCPD
SSHD
NIS
YP
Perl
Bash
DRBL Server


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

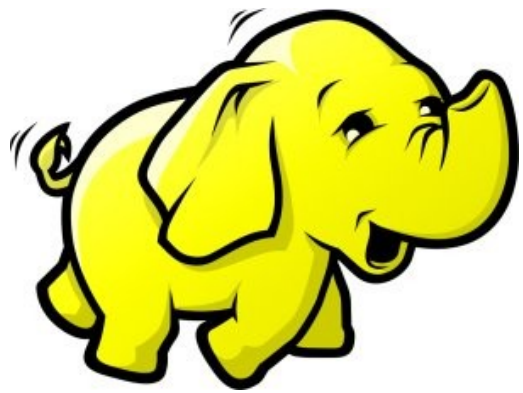
運用企鵝龍佈署資料探勘平台的經驗分享

Building Multi-user Hadoop Cluster using DRBL

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw



關於 hadoop.nchc.org.tw

- **DRBL Server – 1 台 (hadoop)**，加大 **/home** 與 **/tftpboot** 空間。
- **DRBL Client – 20 台 (hadoop101~hadoop120)**
- 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

DRBL+Hadoop=Haduzilla 黑肚龍系統架構

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-0.1/
```

- **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
check_user_identification.php	2.9 kB	85	6 weeks	wa

使用者註冊頁面 Hadoop-Register

Hadoop 帳號申請

帳號:

密碼:

登入 重填

新增帳號 忘記密碼 操作問題回報

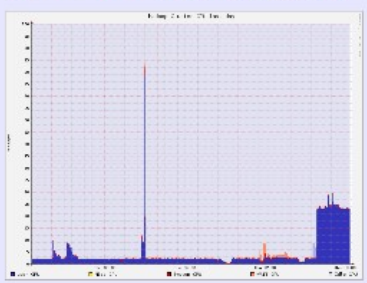
歡迎加入討論群組，以利接收即時公告事宜

家目錄空間吃緊中，請盡量上傳至HDFS後，清除家目錄檔案，謝謝！

註冊人數: 1460 / 1999 人

MapReduce 狀態 | HDFS 狀態

過去 24 小時 CPU 負載 - 查詢完整系統負載:



Running Jobs Quick Links

Jobid	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete
job_201104290234_0905	NORMAL	h1196	PA: Local Apriori over input: n/1mpy54 /input, with minSup: 15000, ep: 0.5	100.00%	10	10	100.00%
			PA: Local Apriori over				

NameNod

Started:	F
Version:	0
Compiled:	S
Upgrades:	T

Browse the filesyst
Namenode Logs

Cluster Summary

2079646 files and

WARNING : There a

Configured Cap
DFS Used
Non DFS Used

網站帳號 jazzwang E-mail [redacted] 姓名 王耀聰 電話 0 單位 0 用途 0 主機帳號 h998 主機密碼 [redacted] 登出

檔案(F) 編輯(E) 檢視(V) 歷史(Y) 工具(T) 說明(H)

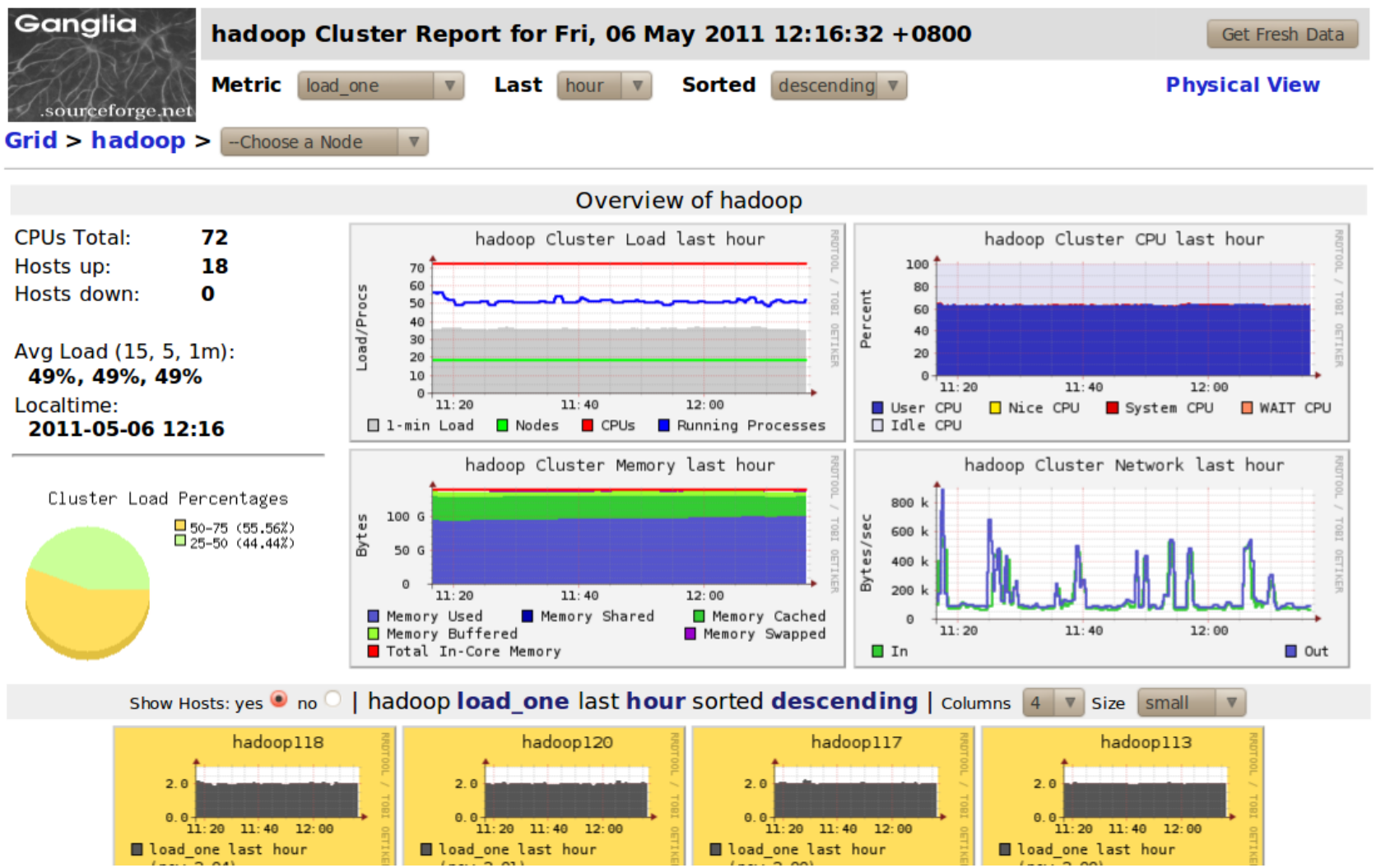
1. hadoop.nchc.org.tw

```
Linux hadoop 2.6.32-5-amd64 #1 SMP Wed Jan 12 03:40:32 UTC 2011 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: Tue Apr 26 15:45:44 2011 from nat235.dynamic.cs.nctu.edu.tw
h998@hadoop:~$
```

Powered by Zterm
http://zhouer.org/ZTerm/

系統狀態監控 Ganglia

- 採用自由軟體 **Ganglia** 來蒐集電腦叢集的負載狀態
- **http://ganglia.sourceforge.net/**



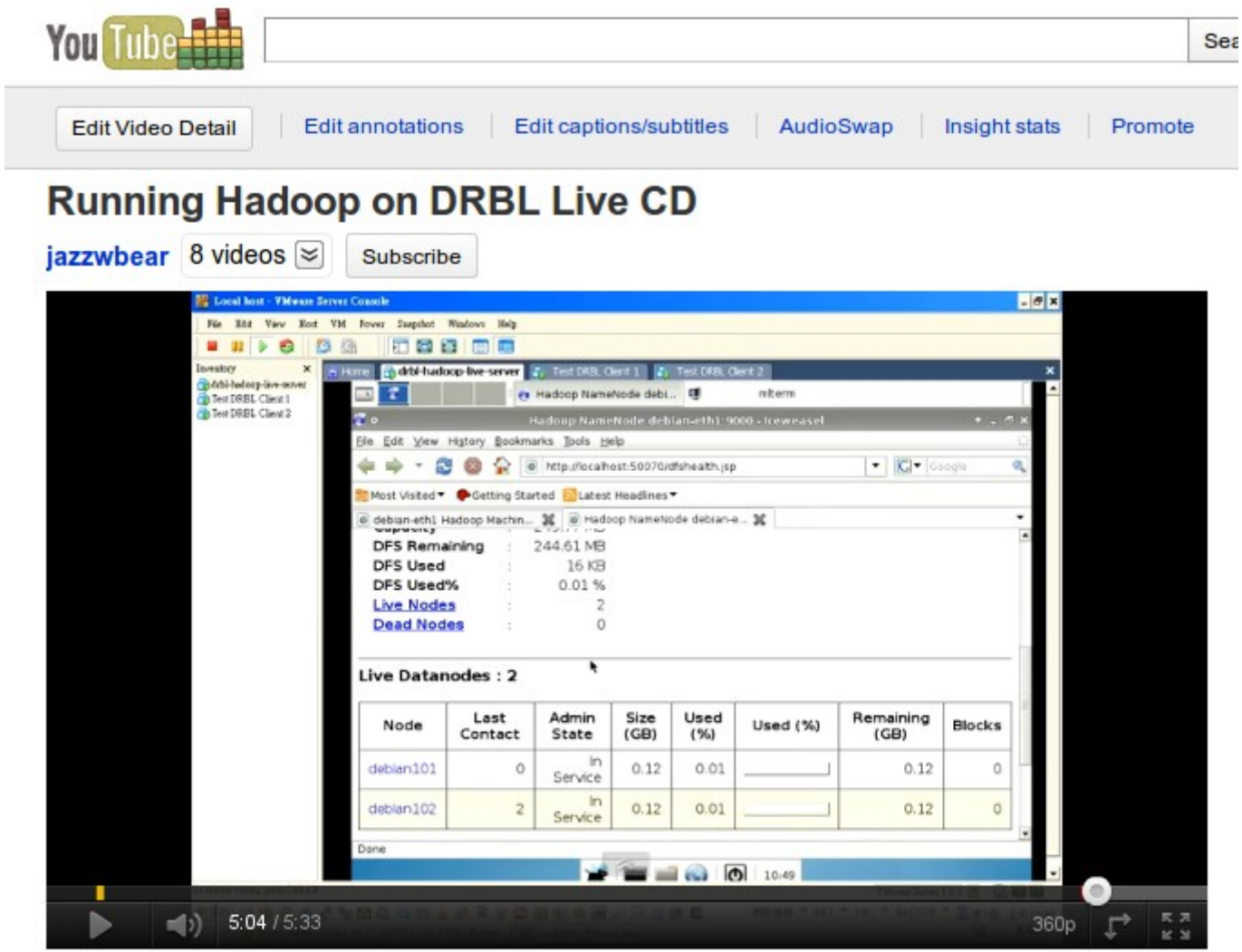
經驗分享（Lesson Learn）

- **Cloudera** 套件的好處：使用 **init.d script** 來啓動關閉
  - **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**

雛型開機光碟 DRBL-Hadoop Live CD

舊影片： http://www.youtube.com/watch?hl=en&v=Ix4WigGvE_A

下載點： http://drbl-hadoop.sf.net



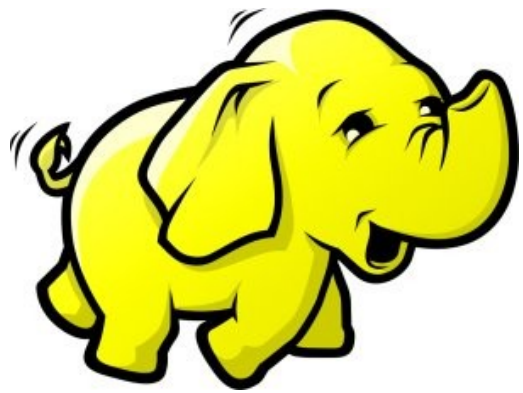
運用再生龍從小硬碟搬家到大硬碟

Hadoop Cluster disk migration using Clonezilla

Jazz Wang

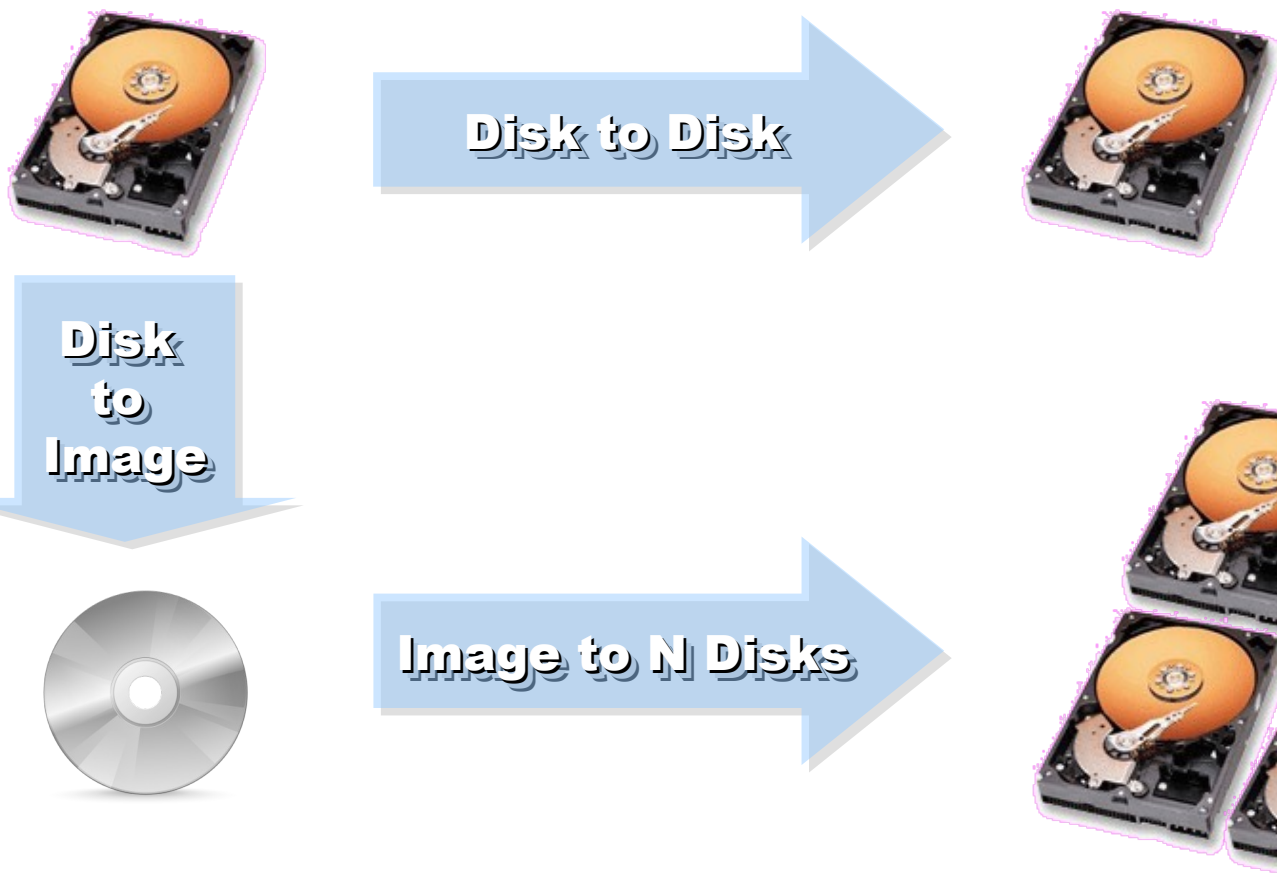
Yao-Tsung Wang

jazz@nchc.org.tw



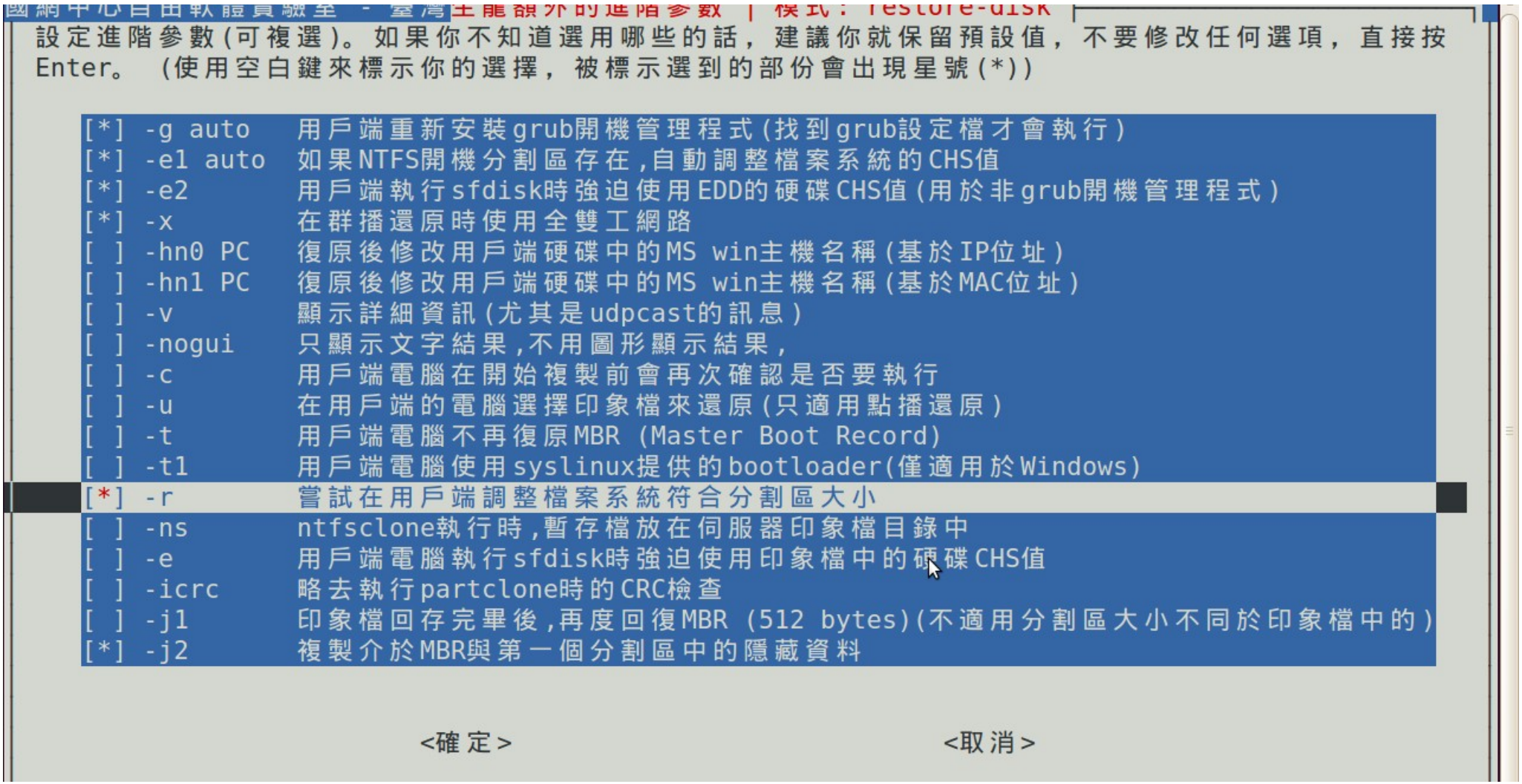
何謂再生龍 Clonezilla ??

- **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
- **http://clonezilla.nchc.org.tw , http://clonezilla.org**



您也用得上的再生龍功能！！

我要怎樣才能把小一點的硬碟複製到大一點的硬碟上？
http://drbl.nchc.org.tw/fine-print.php?path=./faq/1_DRBL_common/34_resize.faq#34_resize.faq



Attribution-Noncommercial-Share Alike 3.0 Taiwan

creative commons

姓名標示-非商業性-相同方式分享 3.0 台灣

您可自由：

分享 — 重製、散布及傳輸本著作

重混 — 修改本著作

惟需遵照下列條件：

姓名標示 — 您必須按照著作人或授權人所指定的方式，表彰其姓名（但不得以任何方式暗示其爲您或您使用本著作的方式背書）。

非商業性 — 您不得爲商業目的而使用本著作。

相同方式分享 — 若您變更、變形或修改本著作，您僅得依本授權條款或與本授權條款類似者來散布該衍生作品。

http://creativecommons.org/licenses/by-nc-sa/3.0/tw/

These slides could be distributed by Creative Commons License.

Questions?

Slides - http://trac.nchc.org.tw/cloud

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