當企鵝龍遇上小飛象

快速佈建 Hadoop 叢集、效能瓶頸量測與生物資訊應用

When DRBL meet Hadoop: cluster deployment, performance measurement and its application in bioinformatics

Jazz Wang

Yao-Tsung Wang

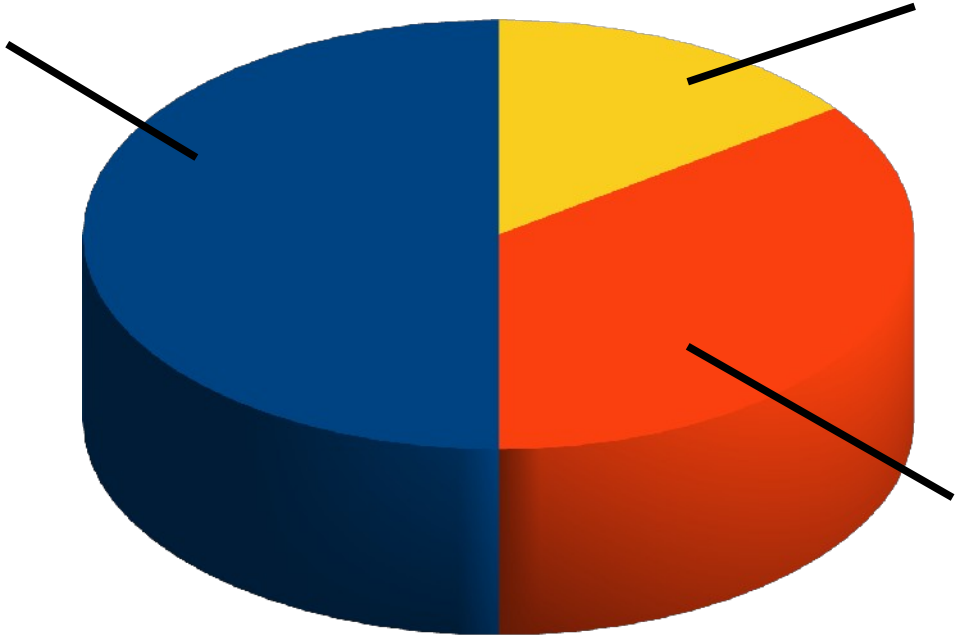
jazz@nchc.org.tw

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

當企鵝龍遇上小飛象

PART 1：

叢集佈署工具簡介：企鵝龍與聰明蛙

PART 2：

常用除錯、效能量測工具與關鍵指標

PART 3：

Hadoop 相關生物資訊應用

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
  - Puppet : configuration management tool

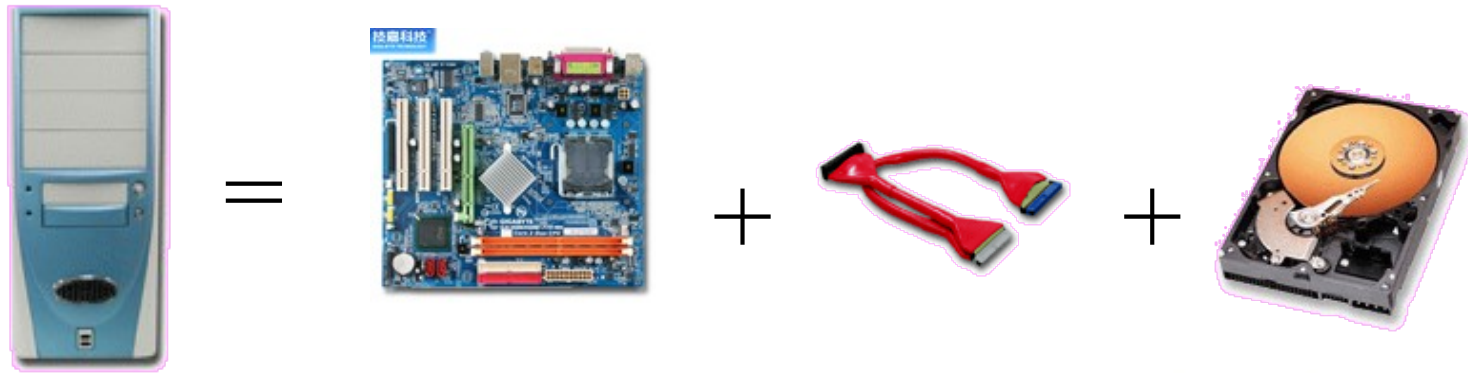
叢集佈署工具比較表

	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

2012 年重要的 IT 技術 Puppet

如果要在 Amazon EC2 上佈署 Hadoop 等軟體，可以考慮 Puppet
因爲作業系統已由虛擬機器的範本裝好了，只能用「有碟」的作法！

hstack

Blog

Hadoop/HBase automated deployment using Puppet

with 23 comments

Introduction

72 tweets

retweet

Deploying and configuring Hadoop and HBase across clusters is a complex task. In this article I will show what we do to make it easier, and share the deployment recipes that we use.

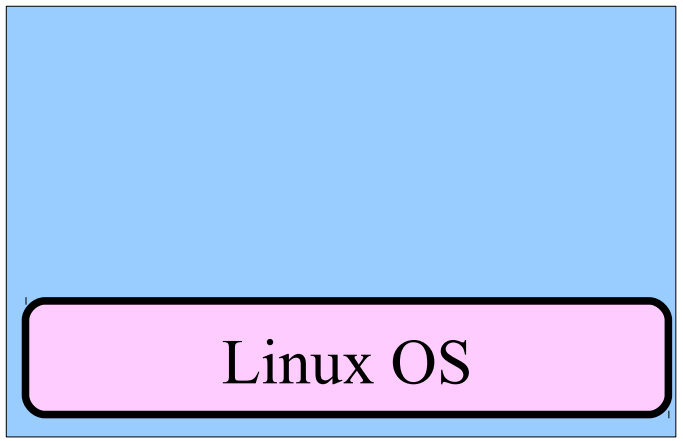
For the tl;dr crowd: go get the code here.

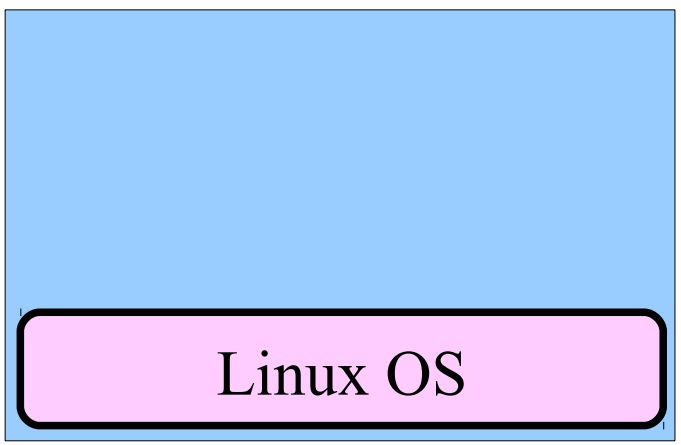
Cool tools

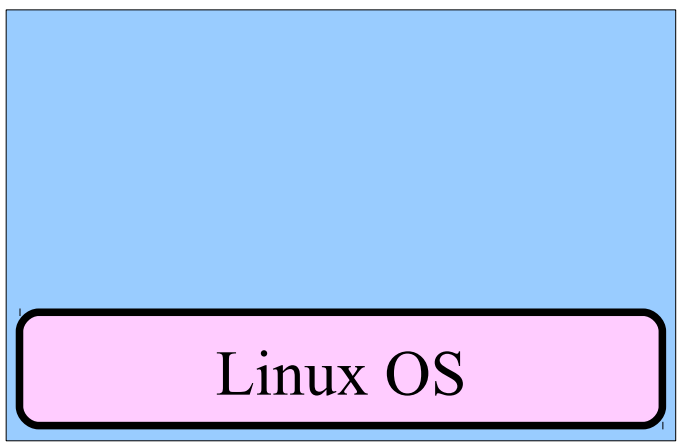
Before going into how we do things, here is the list of tools that we are using, and which I will mention in this article. I will try to put a link next to any tool-specific term, but you can always refer to its specific home-page for further reference.

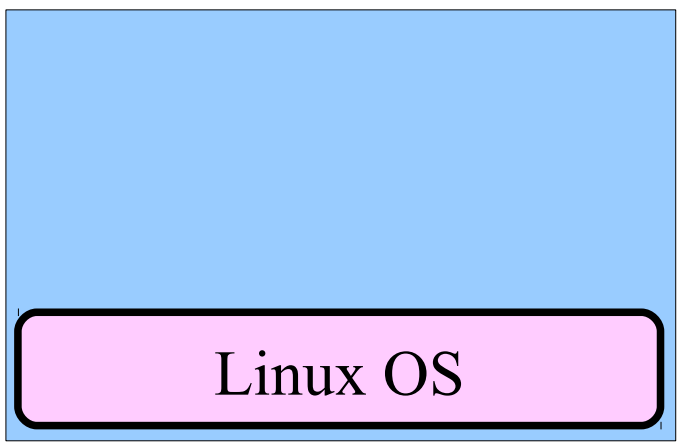
- Hudson – this is a great CI server, and we are using it to build Hadoop, HBase, Zookeeper and more
- The Hudson Promoted Builds Plug-in – allows defining operations that run after the build has finished, manually or automatically
- Puppet – configuration management tool We don't have a dedicated operations team to hand off a list of instructions on how we want our machines to look like. The operations team helping us just makes sure the servers are in the rack, networked and powered up, but once we have a set of IPs (usually from IPMI cards) we're good to go ourselves. We are our own devops team, and as such we try to automate as much as possible, where possible, and using the tools above helps a lot.

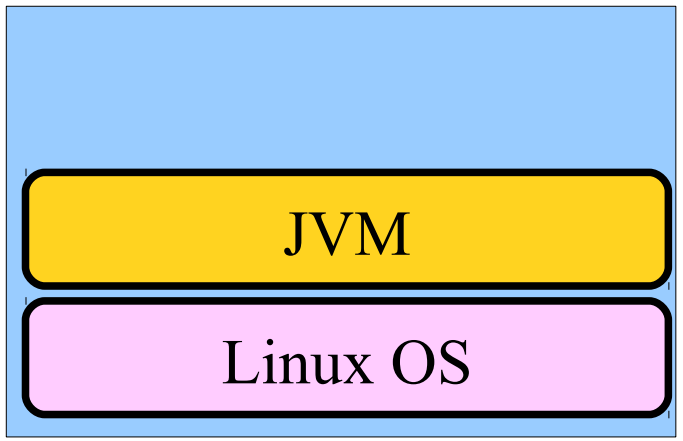
https://github.com/hstack/puppet
http://hstack.org/hstack-automated-deployment-using-puppet/
http://www.cioinsight.com/images/stories/slideshows/SS_142511_CIO_TechSkills/

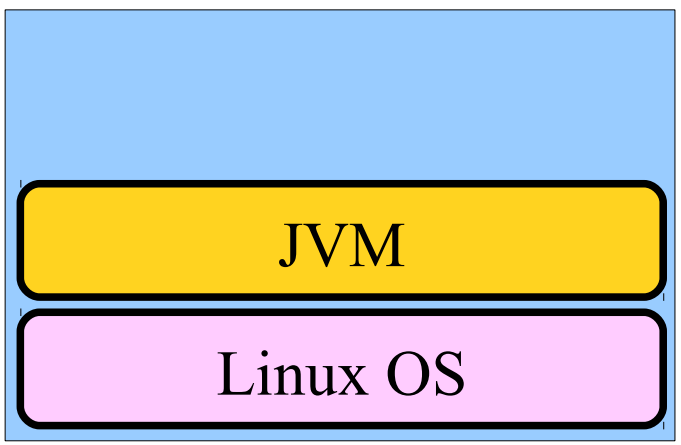

Linux OS

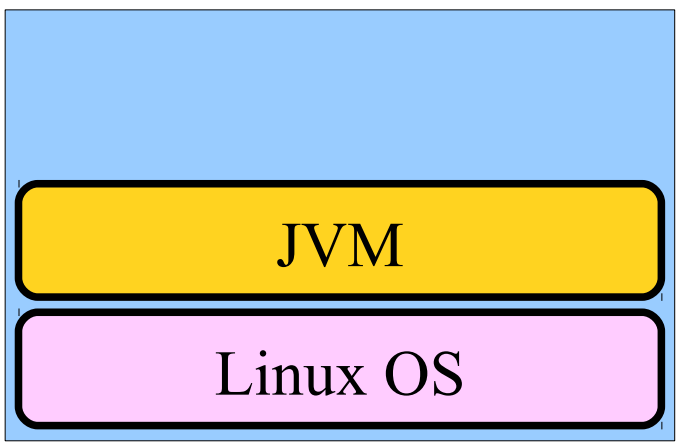

Linux OS

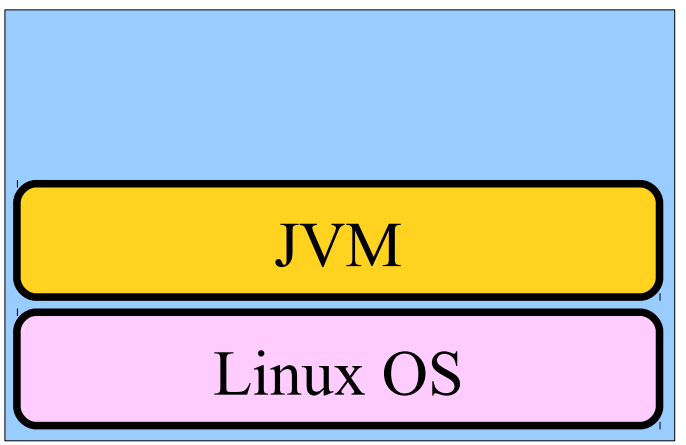

Linux OS

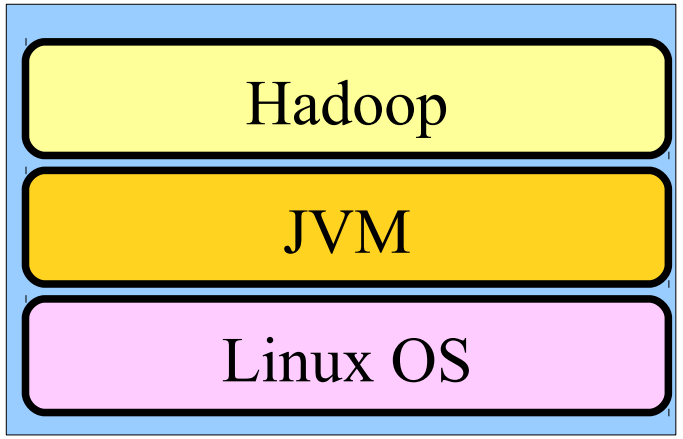

Linux OS

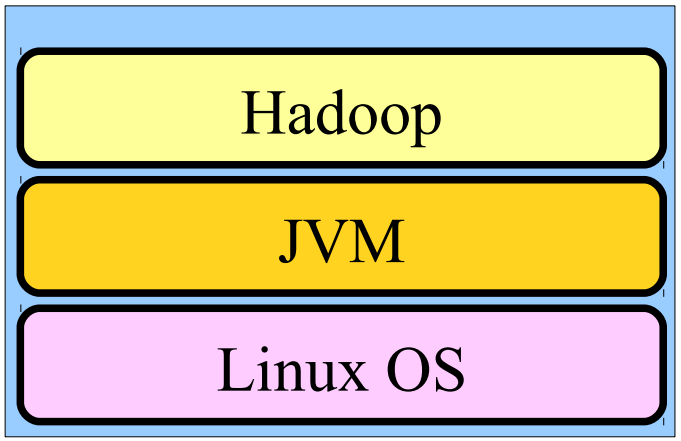

JVM
Linux OS

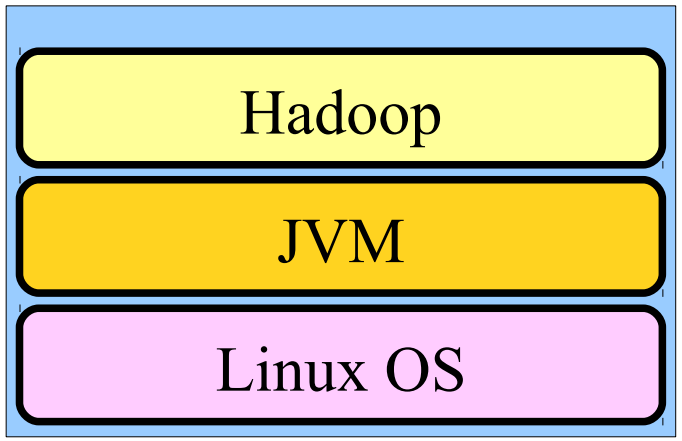

JVM
Linux OS

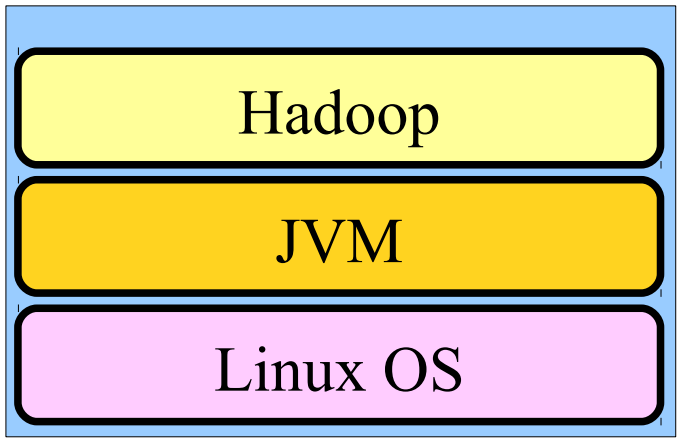

JVM
Linux OS

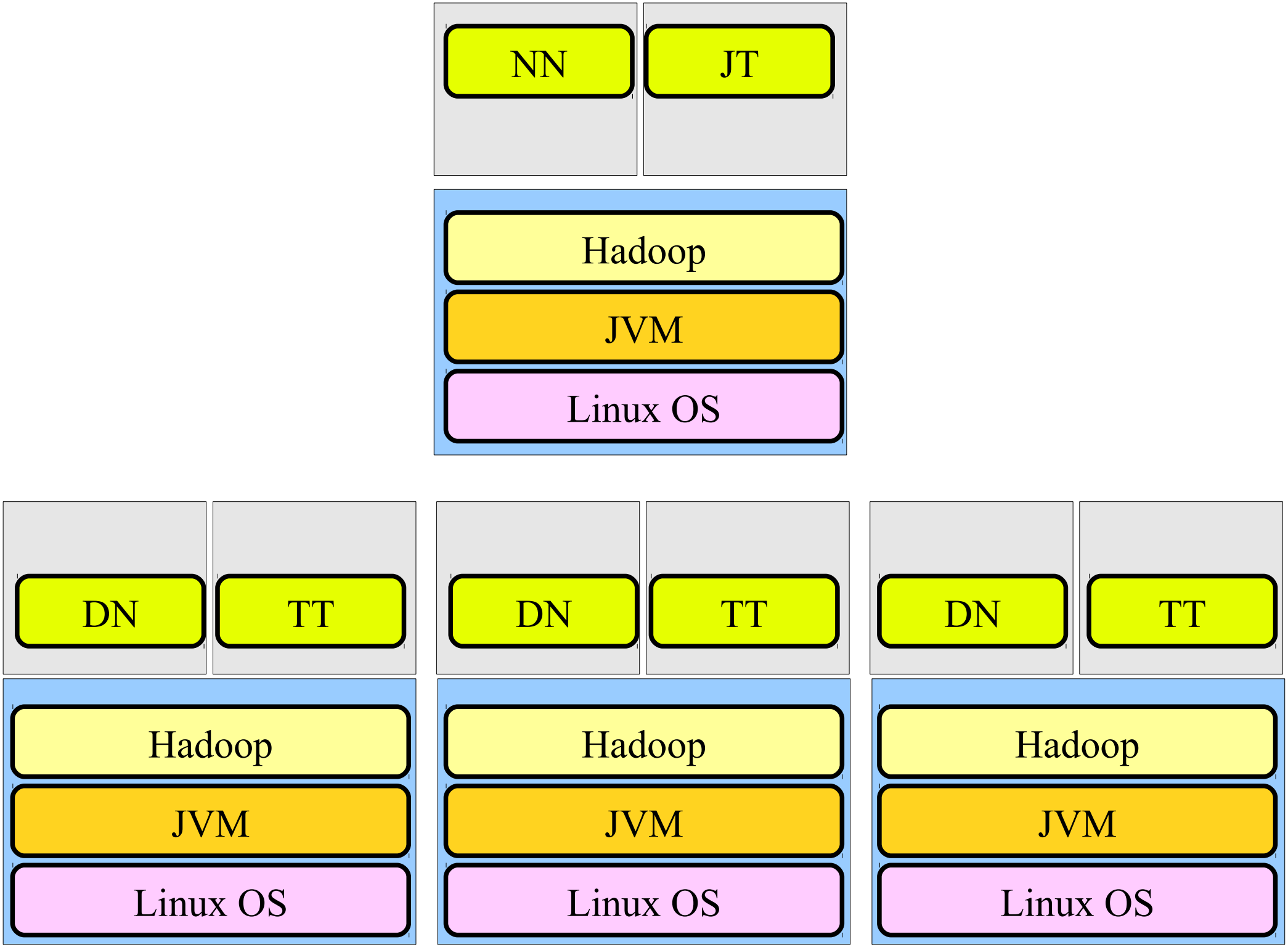

JVM
Linux OS

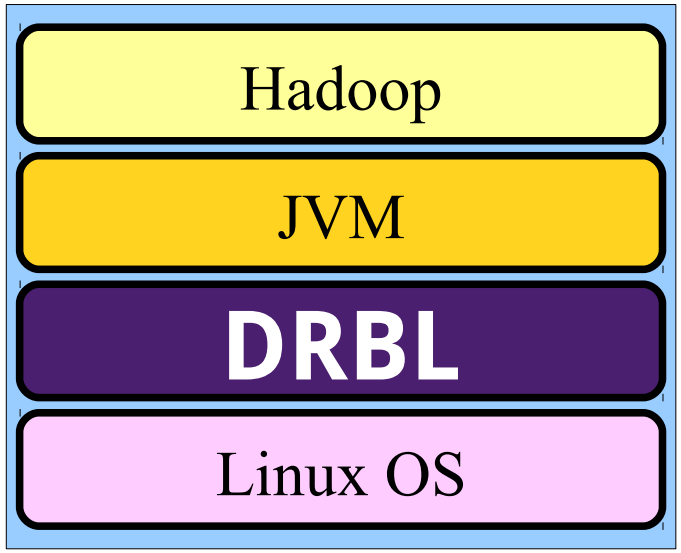

Hadoop
JVM
Linux OS

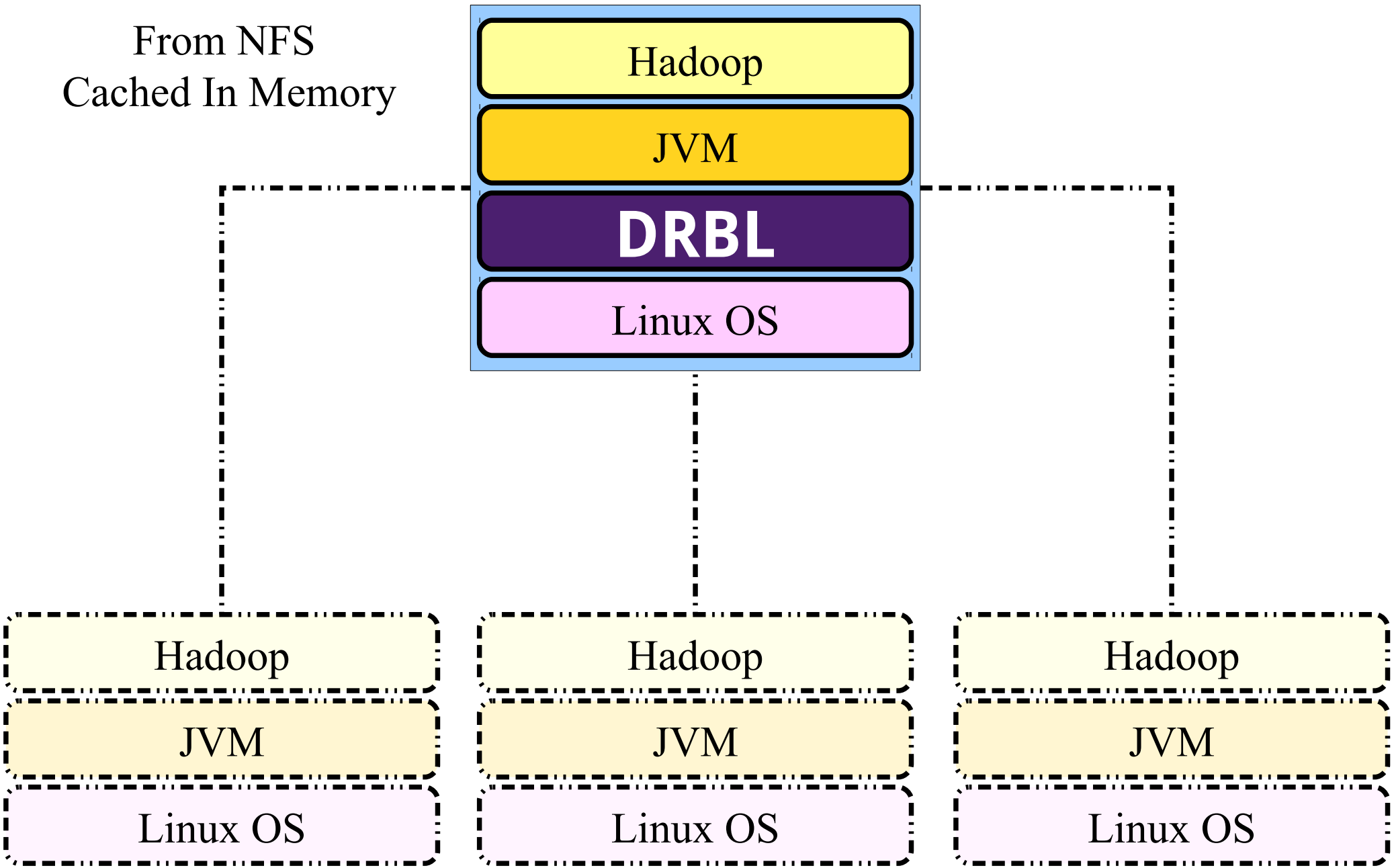

Hadoop
JVM
Linux OS

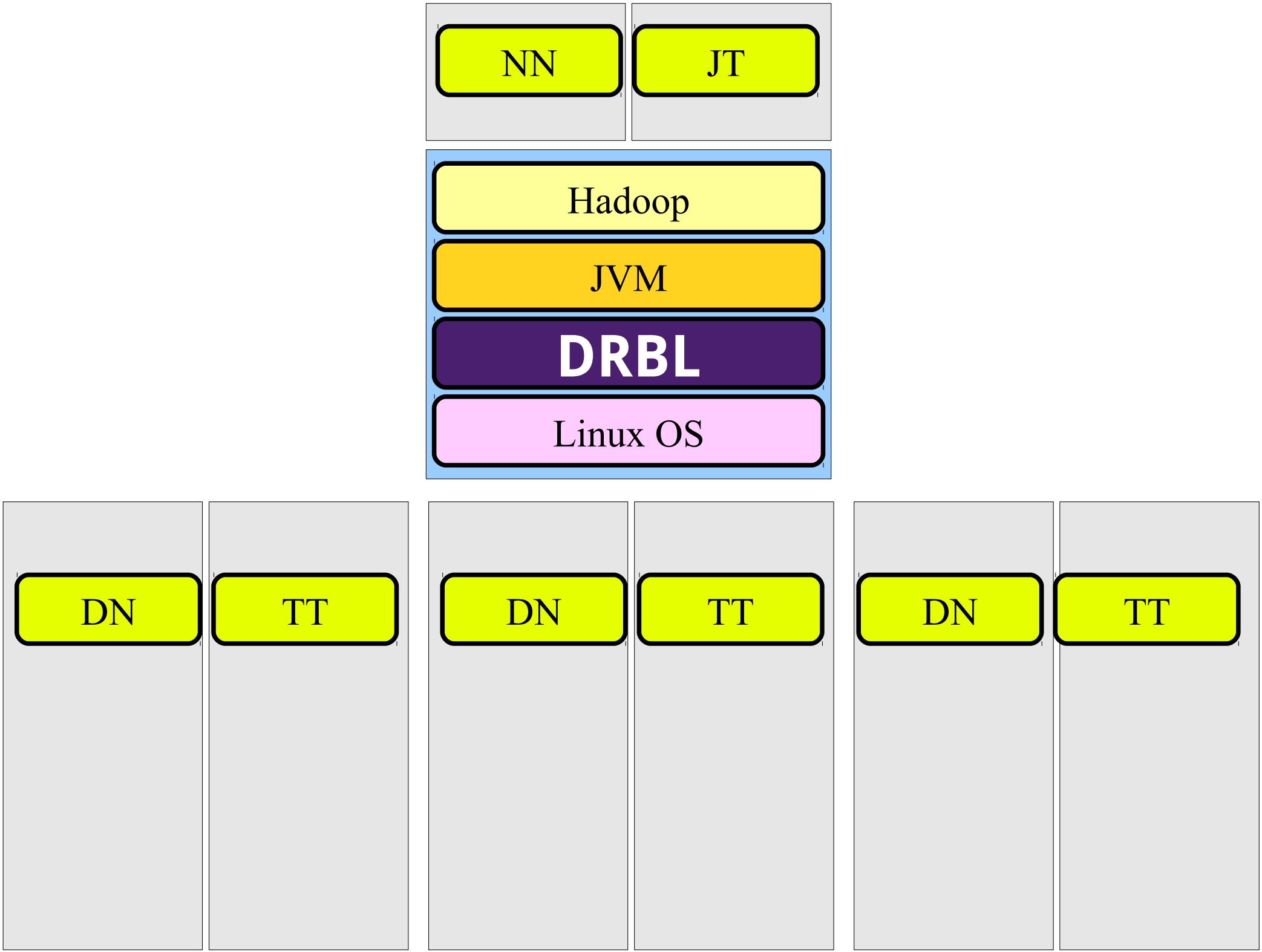

Hadoop
JVM
Linux OS


Hadoop
JVM
Linux OS


NN
JT
Hadoop
JVM
Linux OS
DN
TT
DN
TT
DN
TT
Hadoop
JVM
Linux OS
Hadoop
JVM
Linux OS
Hadoop
JVM
Linux OS


Hadoop
JVM
DRBL
Linux OS


From NFS
Cached In Memory
Hadoop
JVM
DRBL
Linux OS
Hadoop
JVM
Linux OS
Hadoop
JVM
Linux OS
Hadoop
JVM
Linux OS


NN
JT
Hadoop
JVM
DRBL
Linux OS
DN
TT
DN
TT
DN
TT

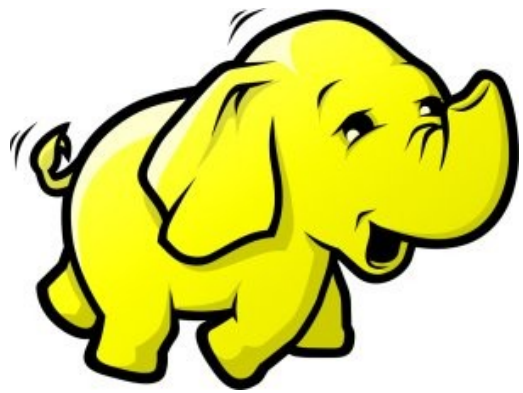
運用企鵝龍佈署資料探勘平台的經驗分享

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

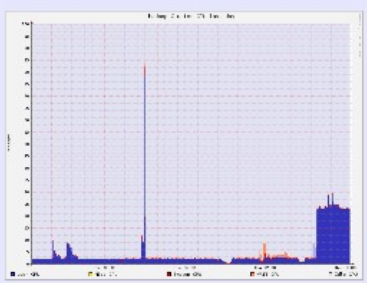
歡迎加入討論群組，以利接收即時公告事宜

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/

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

Hadoop 除錯、效能監控與調校指標

How to debug, measure the performance and key index of performance

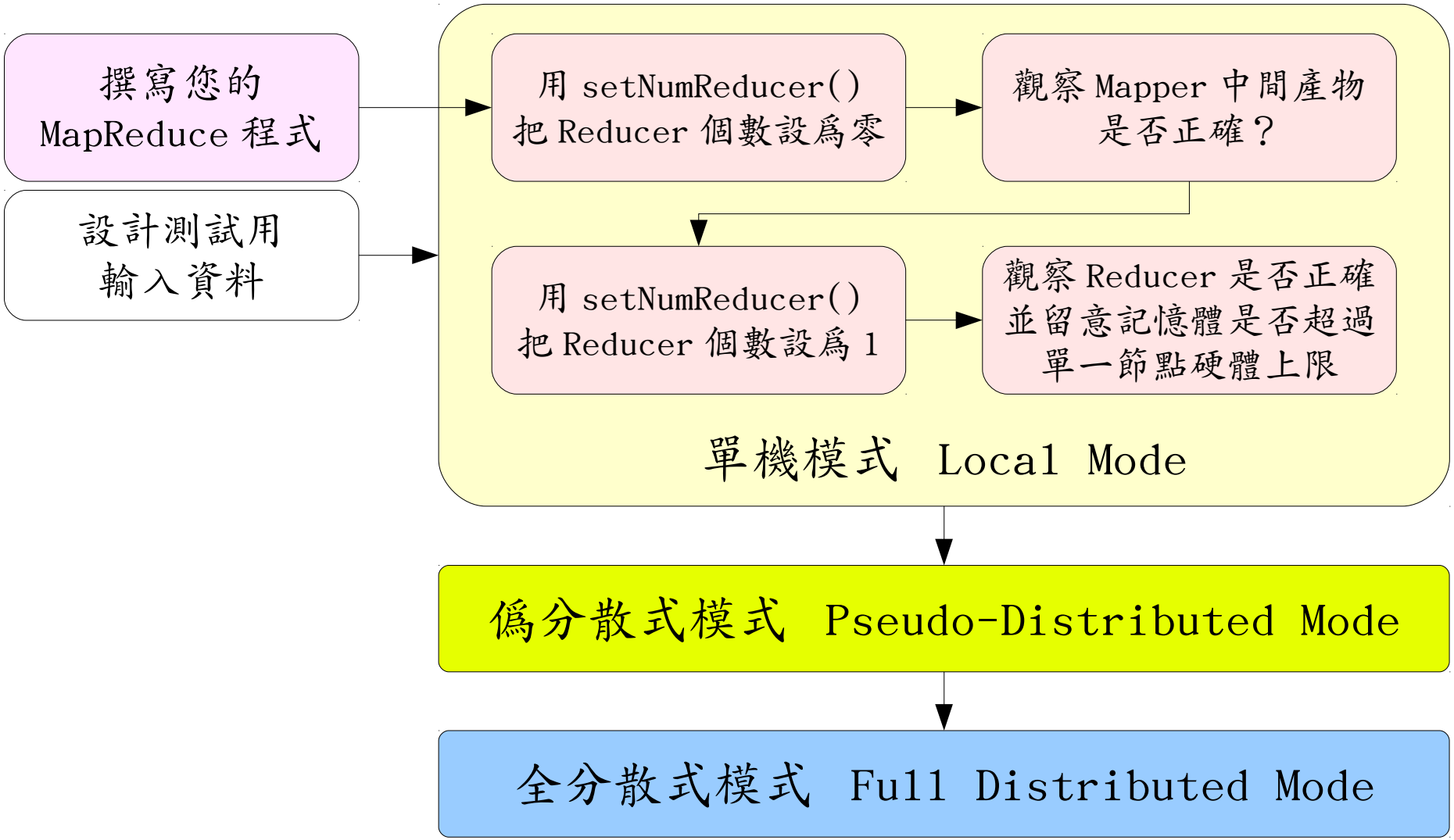
Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

Powered by DRBL

Hadoop Debug Process 標準除錯程序



http://wiki.apache.org/hadoop/HowToDebugMapReducePrograms

Java Remote Debug

- 有時靠 System.out.println() 是不夠的，有人想要 Step Trace 怎麼辦？
- 先在 Local Mode 執行，啓動 Java Remote Debug 的參數，然後用 Eclipse 的 Step Trace 功能來觀察程式的行爲。
- export HADOOP_OPTS="-agentlib:jdwp=transport=dt_socket,server=y,suspend=y,address=5000"
- http://javarevisited.blogspot.com/2011/02/how-to-setup-remote-debugging-in.html
- http://code.google.com/p/hadoop-clusternet/wiki/DebuggingJobsUsingEclipse

How to setup java remote debugging in eclipse

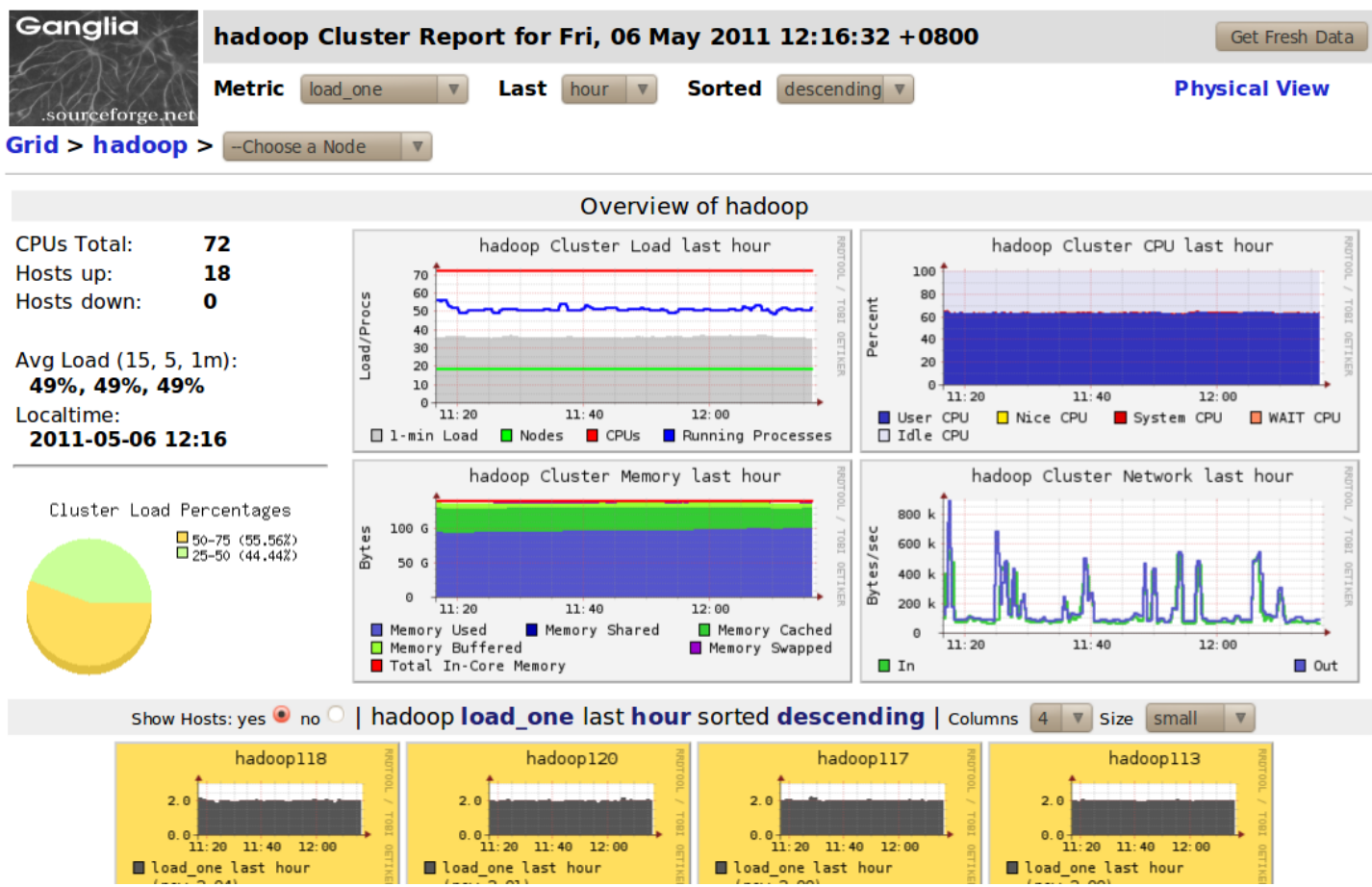
by JAVIN PAUL • FEB. 26, 2011

READ LATER

Remote debugging is not a new concept and many of you are aware of this just for who don't know what is remote debugging? It's a way of debugging any process could be **Java** or C++ running on some other location from your development machine. Since debugging is essential part of development and ability to debug your application not only

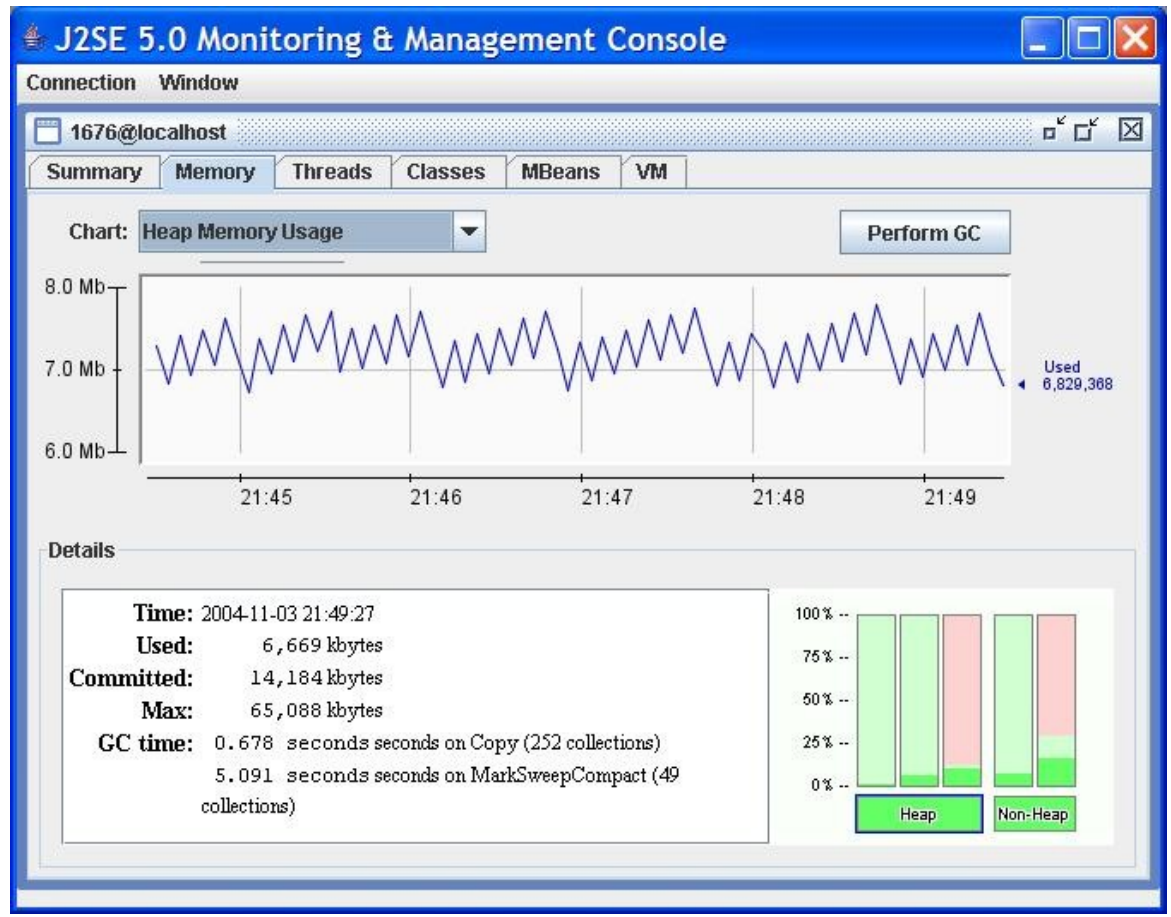
系統狀態監控 Ganglia

- Hadoop 預設可以產生效能數據（ Metrics ）給 Ganglia
- 請根據您的 Ganglia 安裝情形設定 conf/hadoop-metrics.properties
- http://ganglia.sourceforge.net/



Hadoop 系統狀態監控 JMX

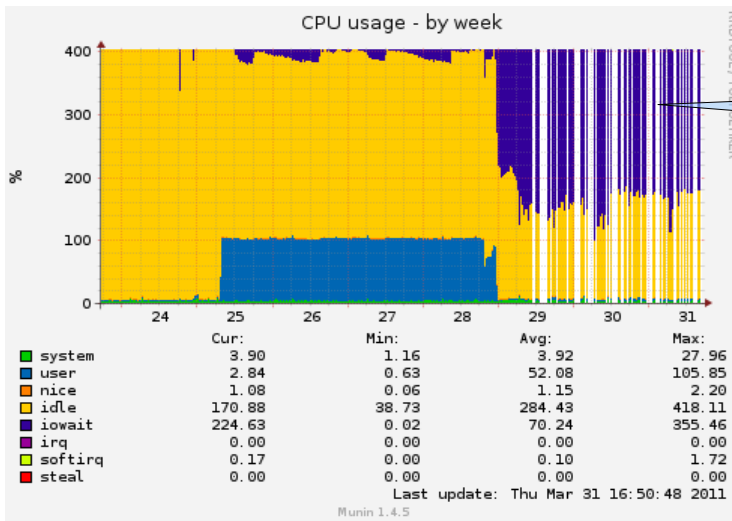
- 由於 Ganglia 的取樣頻率一般是 10 秒一次到一分鐘一次，若是需要更即時的狀態資料，可以使用 JMX Client 來讀取 Hadoop 送出的 Metrics
- 像是 jconsole、 Hyperic或 Nagios等。

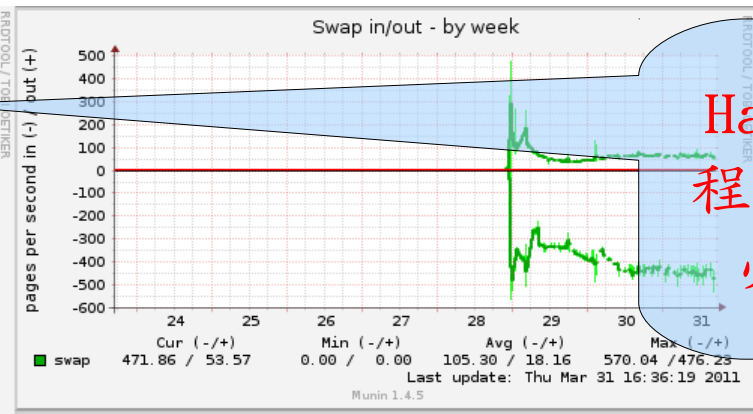


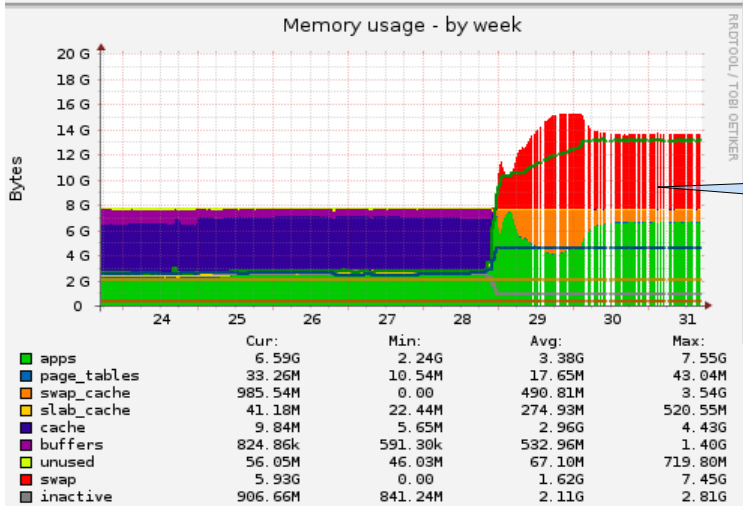
http://www.cloudera.com/blog/2009/03/hadoop-metrics/

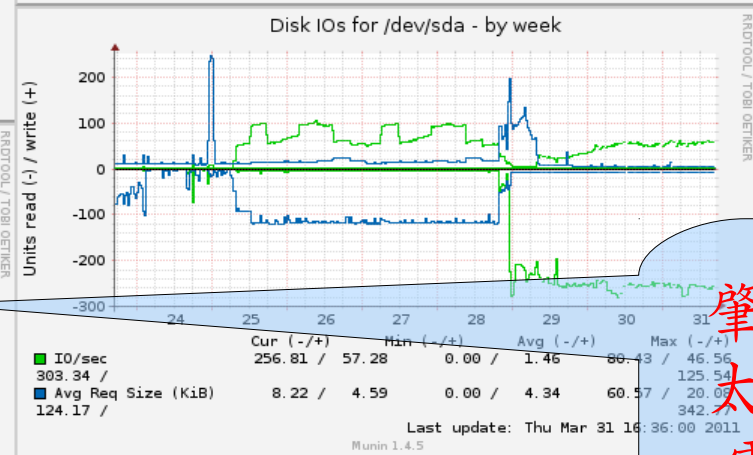
系統 I/O 狀態監控 Munin

- 由於 Ganglia 所蒐集的資料並沒有每顆硬碟的 I/O 數據，有時會使用 Munin 這套軟體來了解每顆硬碟的 I/O 情形，進而分析讀寫效能。
- 當讀寫 I/O 遠慢於 CPU 運算時，會發生 IOWAIT







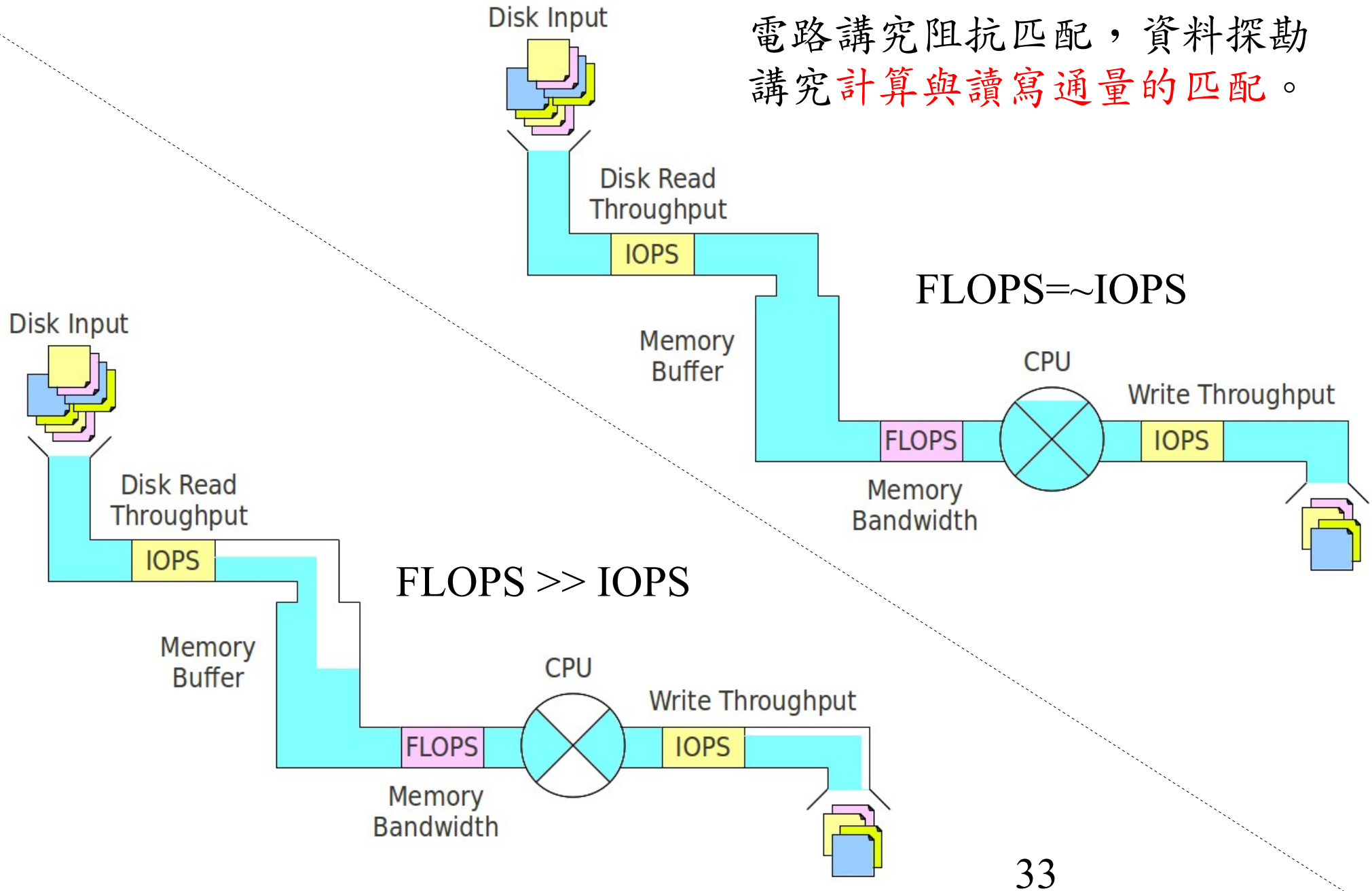


Hadoop MapReduce 程式若出現 IOWait 必然效能不彰！

肇事原因常是用了太多記憶體，結果需要改用 SWAP～

Optimization = I/O Impedance ? 資料通量達成匹配？

電路講究阻抗匹配，資料探勘講究計算與讀寫通量的匹配。



Hadoop 於生物資訊領域的應用

The Application of Hadoop in Bioinformatics Field

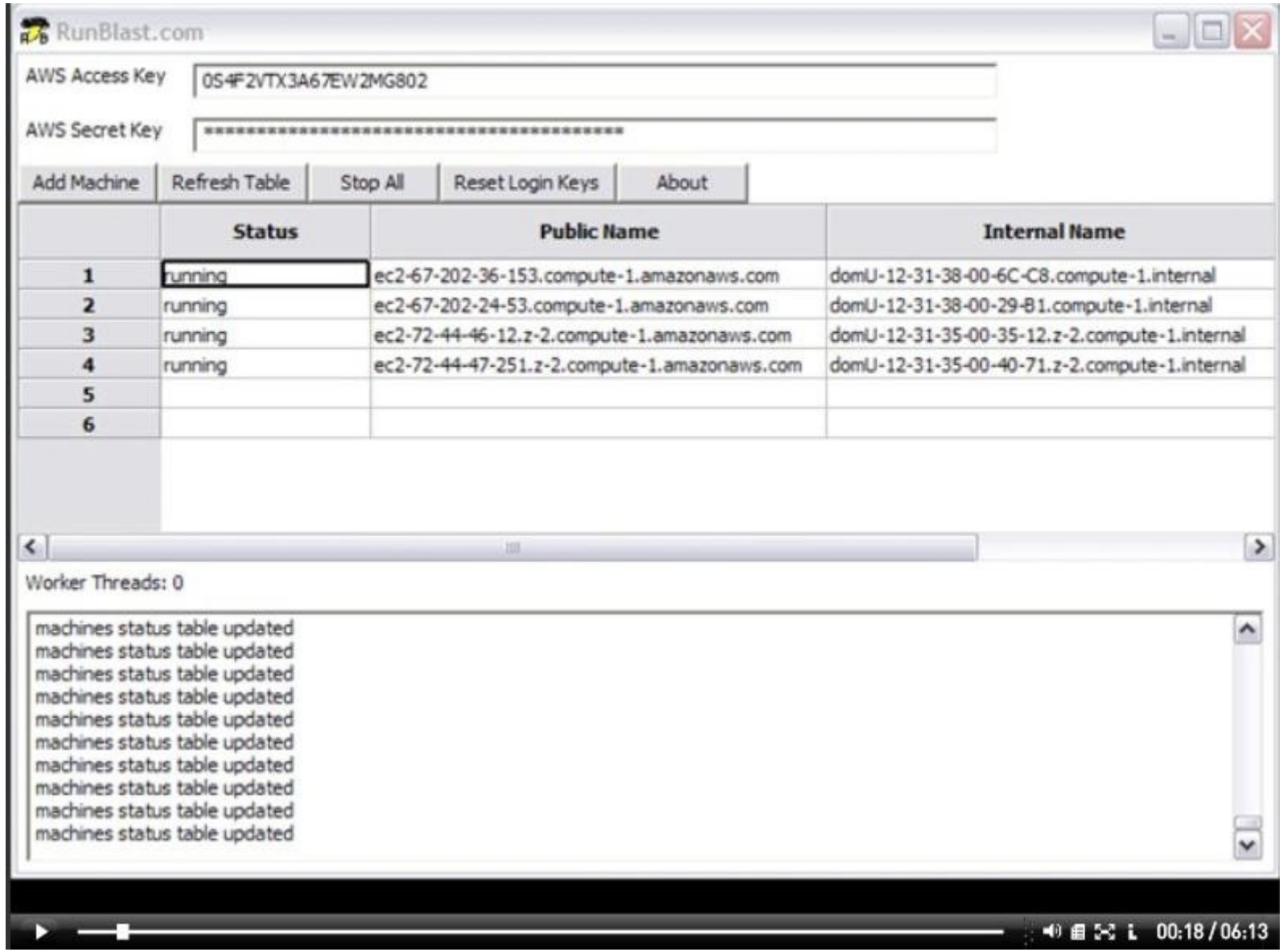
Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

Powered by DRBL

RunBLAST : mpiBLAST in Amazon EC2



Video: http://www.runblast.com/videos/runblast-blastwizard.swf

CloudBLAST

- "CloudBLAST: Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications", eScience 2008
- 特點：採用 MapReduce 演算法進行 BLAST 運算

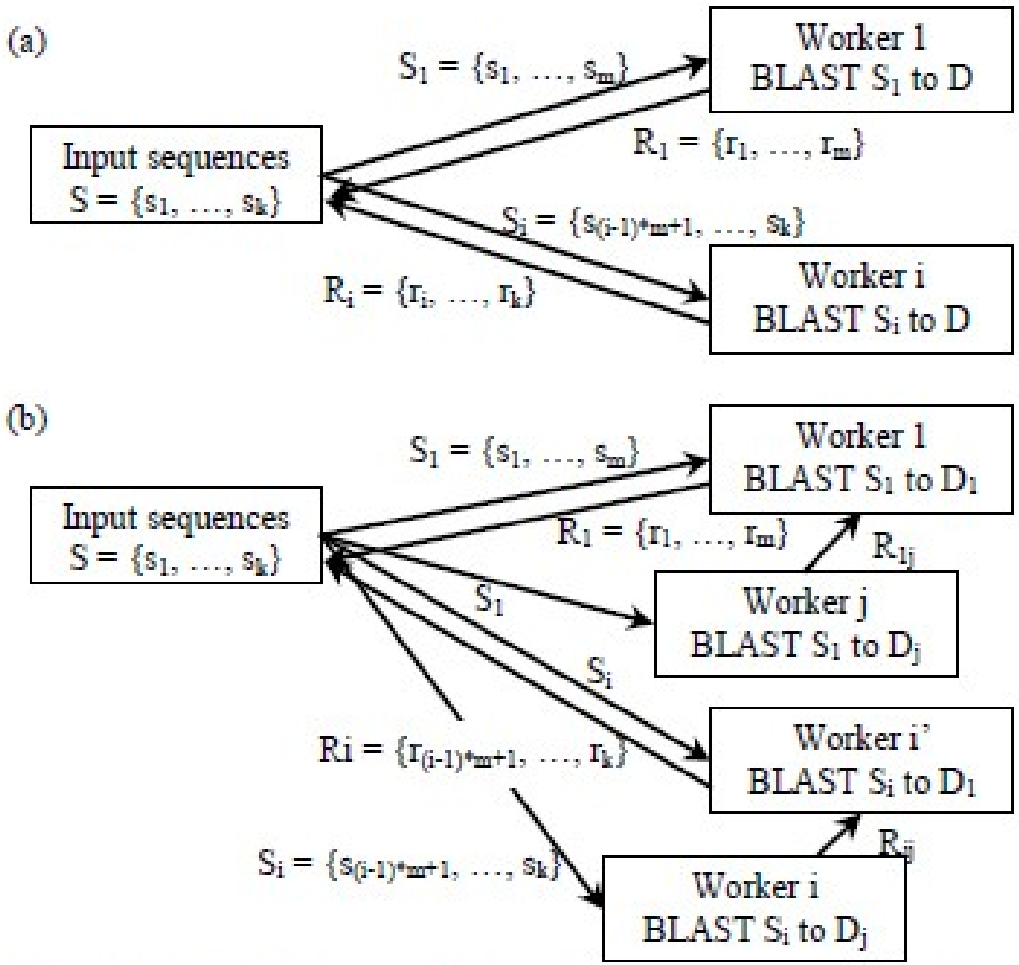


Figure 1. Parallelization of BLAST: (a) input sequences are partitioned into i subsets, each of which is processed by a worker, and results are

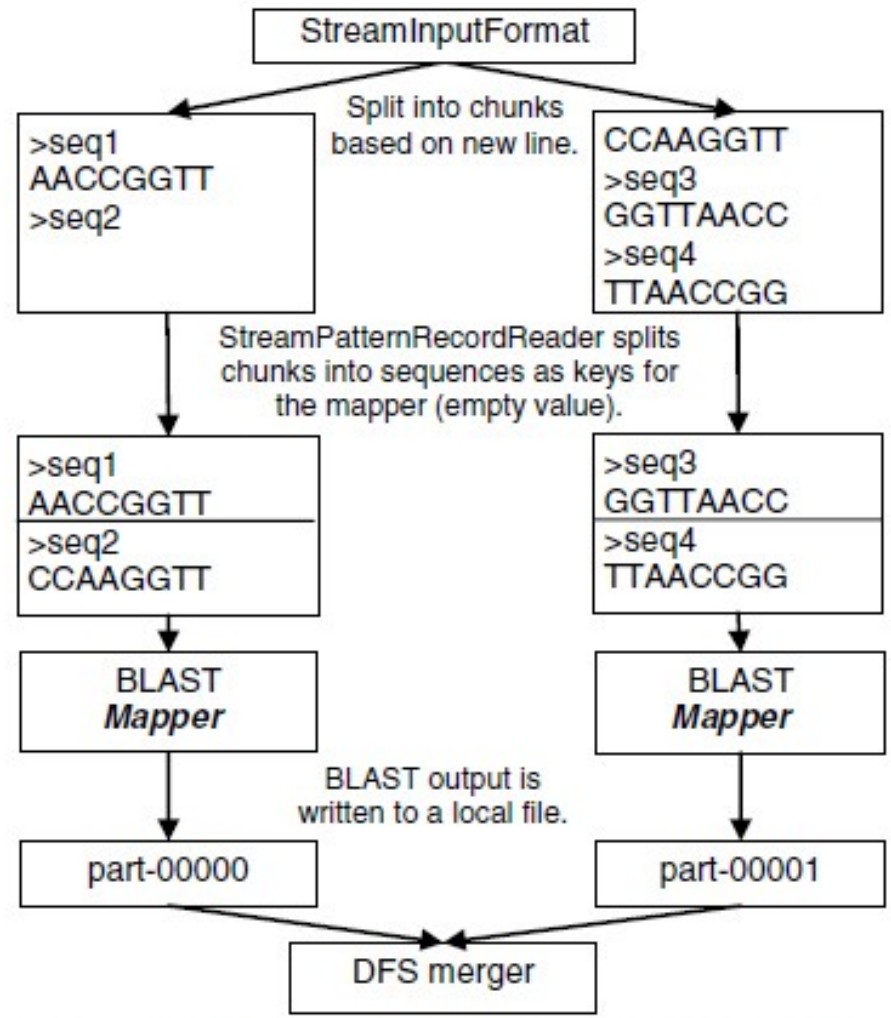


Figure 2. BLASTing with MapReduce. Given a set of input sequences

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

Powered by DRBL

企鵝龍的開機原理

Installation and Booting Procedure of DRBL

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

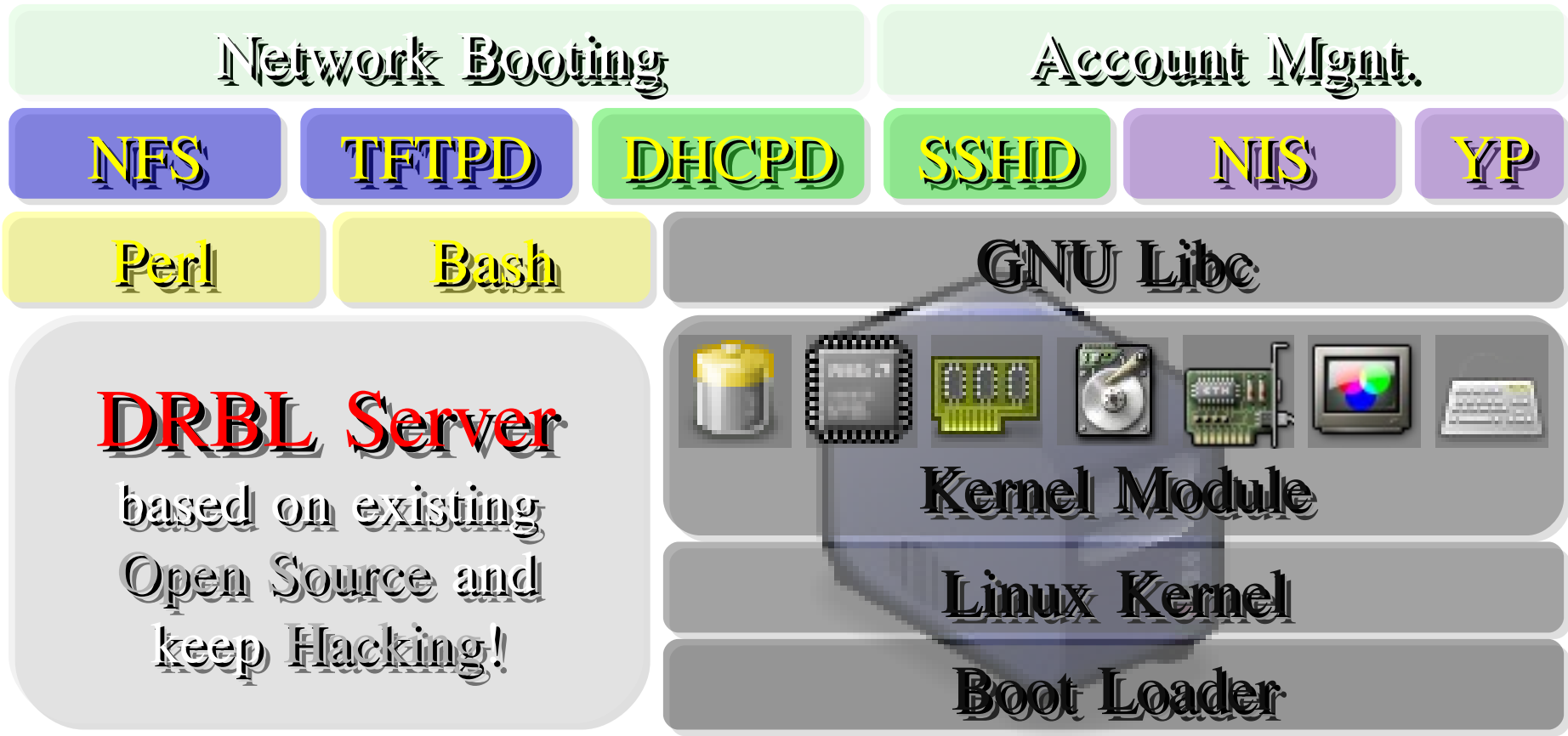
1st, We install Base System of GNU/Linux on Management Node. You can choose:

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

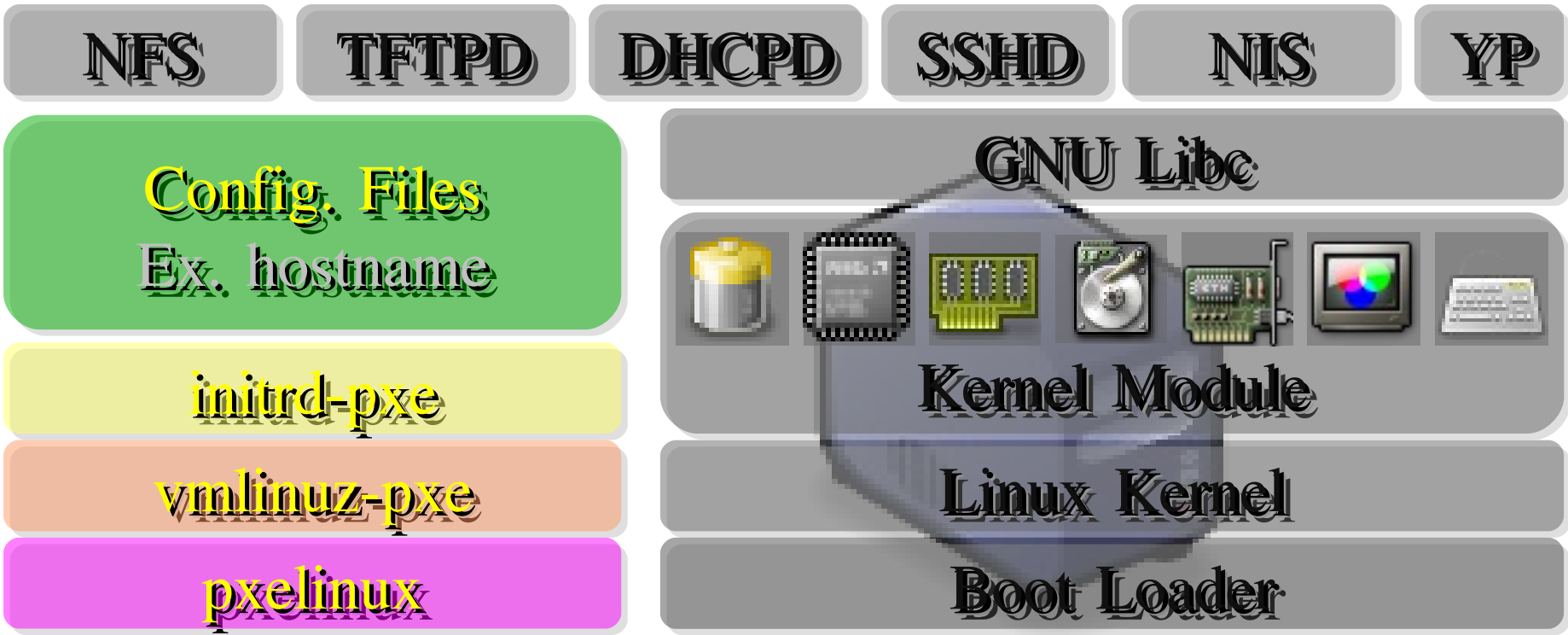
2nd, We install DRBL package and configure it as DRBL Server.

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

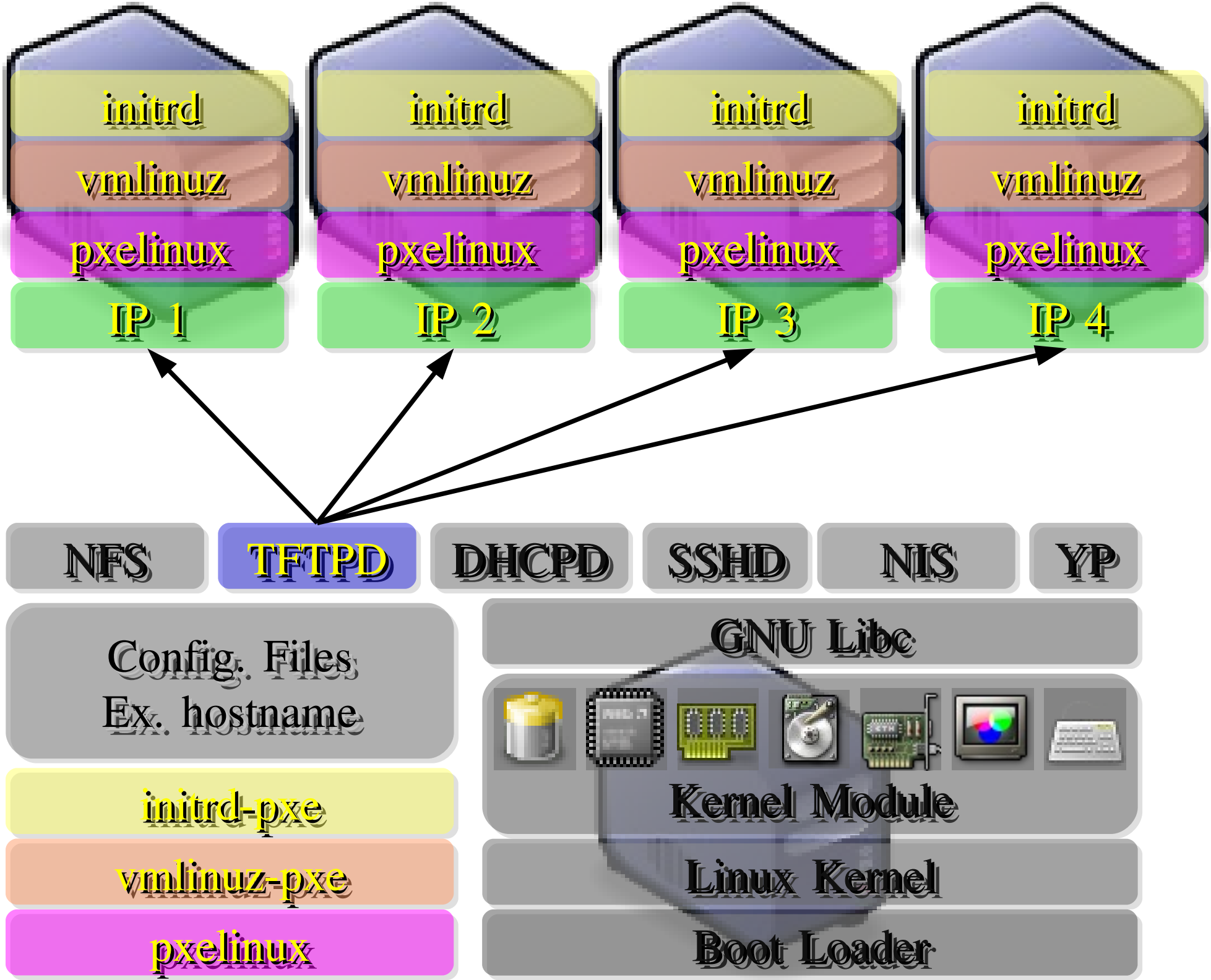
pxelinux

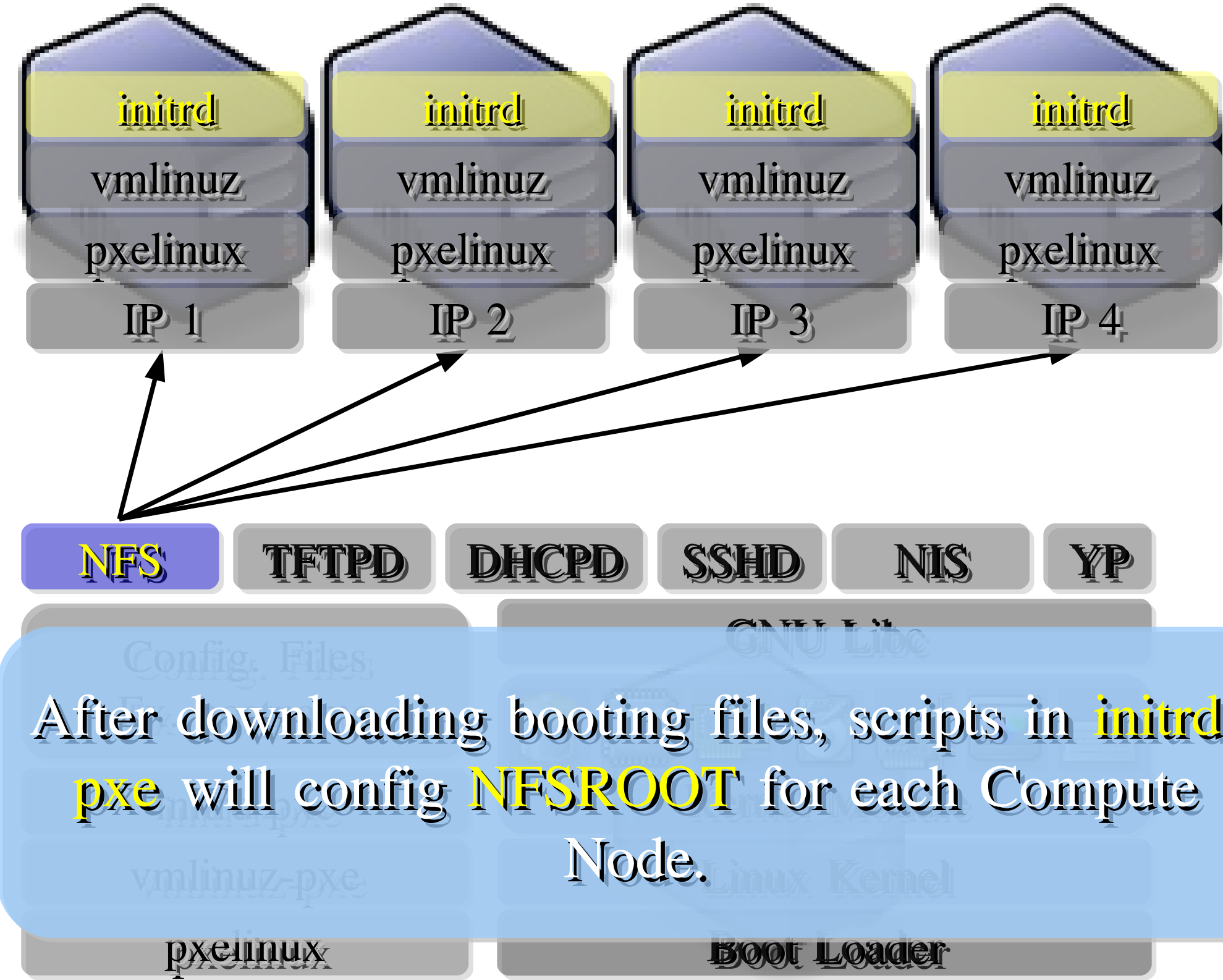
GNU Libc

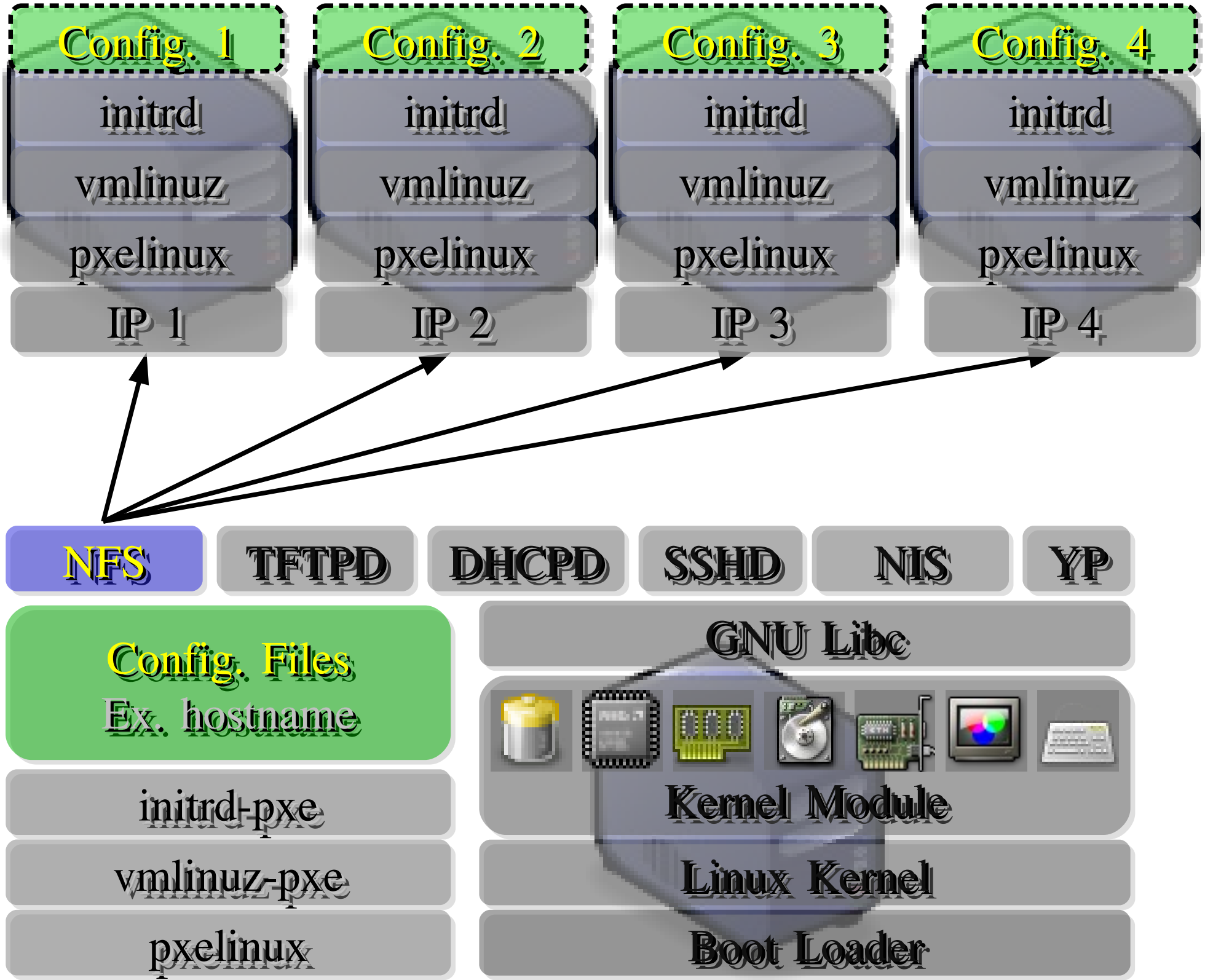
Kernel Module

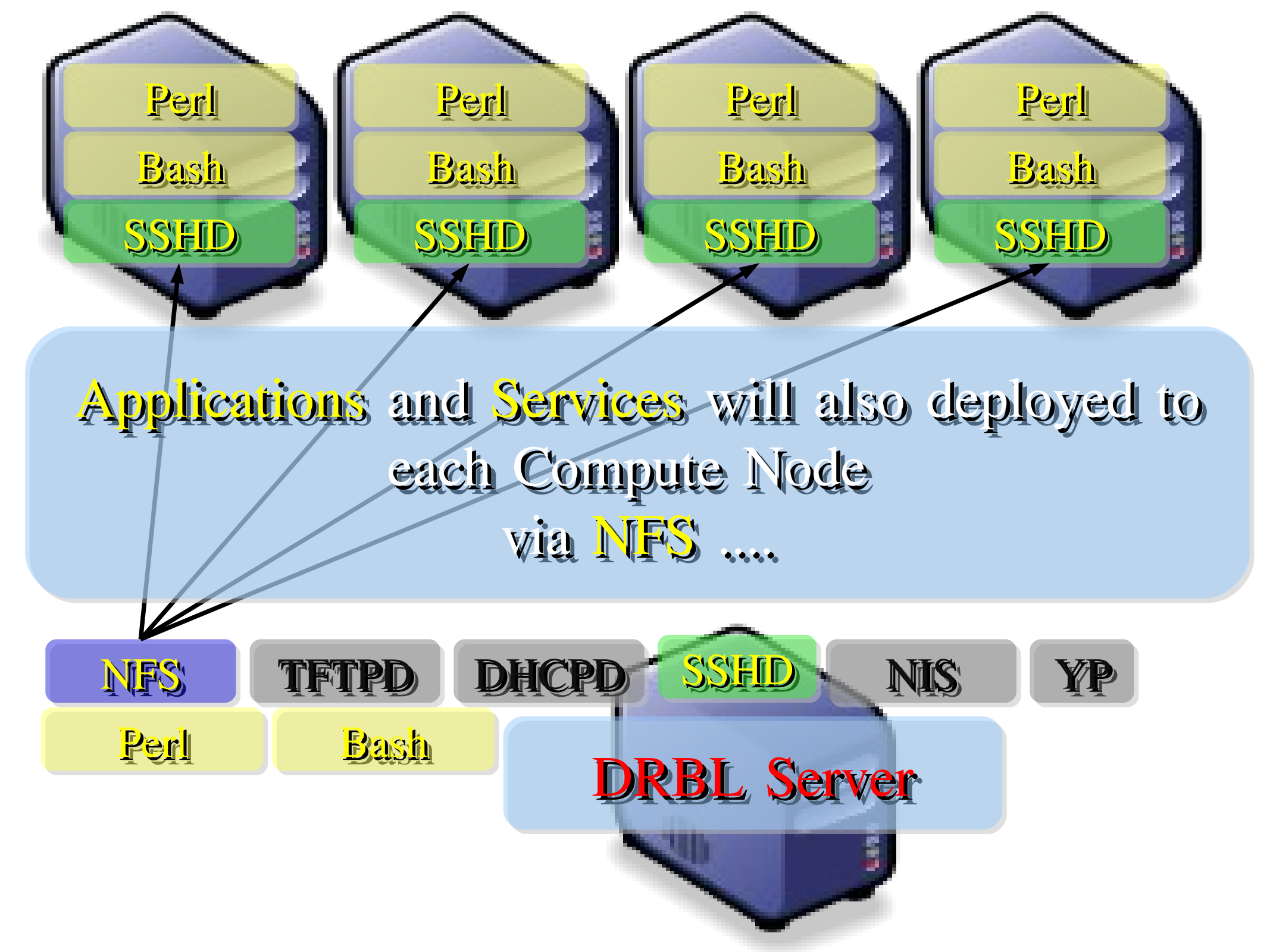
Linux Kernel

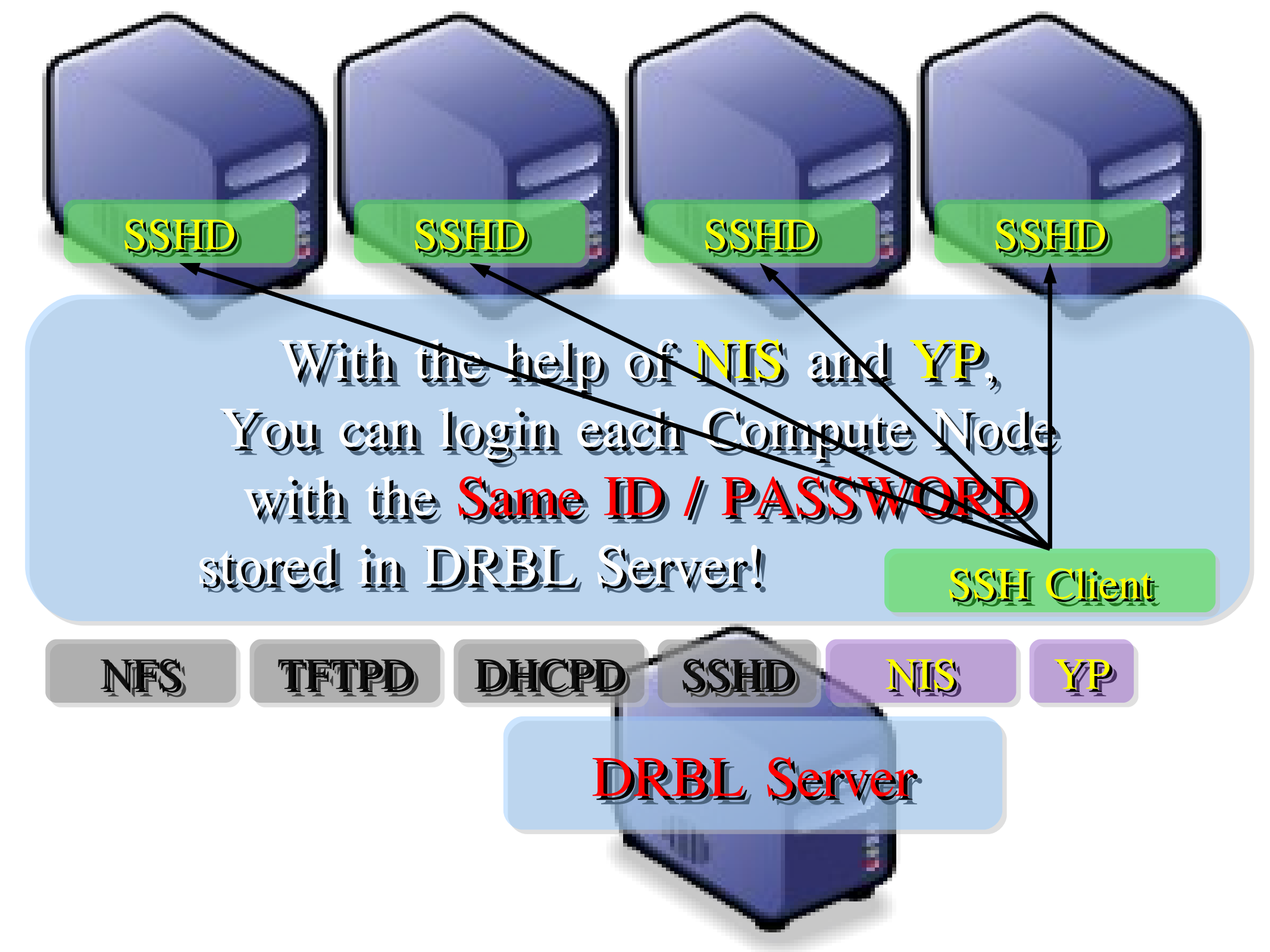
Boot Loader


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
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