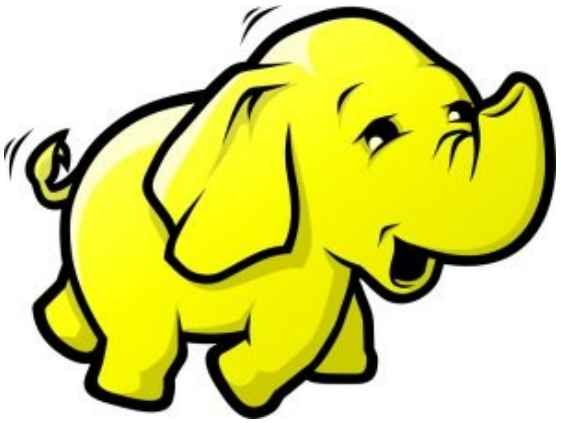
Hadoop 簡介：源起與術語

Introduction to Hadoop : History and Terminology

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw



What is Hadoop ?

用一句話解釋 *Hadoop* 是什麼 ??

Hadoop is a software platform that lets one easily write and run applications that process vast amounts of data.

Hadoop 是一個讓使用者簡易撰寫並執行處理海量資料應用程式的軟體平台。

亦可以想像成一個處理海量資料的生產線，只須學會定義 ***map*** 跟 ***reduce*** 工作站該做哪些事情。

Features of Hadoop …
Hadoop 這套軟體的特色是 …

- **海量 Vast Amounts of Data**
  - 擁有儲存與處理大量資料的能力
  - Capability to STORE and PROCESS vast amounts of data.
- **經濟 Cost Efficiency**
  - 可以用在由一般 PC 所架設的叢集環境內
  - Based on large clusters built of commodity hardware.
- **效率 Parallel Performance**
  - 透過分散式檔案系統的幫助，以致得到快速的回應
  - With the help of HDFS, Hadoop have better performance.
- **可靠 Robustness**
  - 當某節點發生錯誤，能即時自動取得備份資料及佈署運算資源
  - Robustness to add and remove computing and storage resource without shutdown entire system.

Founder of Hadoop – Doug Cutting

Hadoop 這套軟體的創辦人 Doug Cutting

Doug Cutting Talks About The Founding Of Hadoop

http://www.youtube.com/watch?v=qxC4urJOchs

History of Hadoop ... 2002~2004

Hadoop 這套軟體的歷史源起 ... 2002~2004

- Lucene
  - http://lucene.apache.org/
  - 用Java 設計的高效能文件索引引擎API
  - a high-performance, full-featured **text search engine library** written entirely in **Java**.
  - 索引文件中的每一字，讓搜尋的效率比傳統逐字比較還要高的多
  - Lucene create an inverse index of every word in different documents. It enhance performance of text searching.

History of Hadoop ... 2002~2004

Hadoop 這套軟體的歷史源起 ... 2002~2004

- Nutch
  - http://nutch.apache.org/
  - Nutch 是基於開放原始碼所開發的網站搜尋引擎
  - Nutch is open source web-search software.
  - 利用Lucene 函式庫開發
  - It builds on Lucene and Solr, adding web-specifics, such as a crawler, a link-graph database, parsers for HTML and other document formats, etc.

Three Gifts from Google ....

來自 Google 的三個禮物 ....

- Nutch 後來遇到儲存大量網站資料的瓶頸
- Nutch encounter storage issue
- Google 在一些會議分享他們的三大關鍵技術
- Google shared their design of web-search engine
  - SOSP 2003 : “The Google File System”
  - http://labs.google.com/papers/gfs.html
  - OSDI 2004 : “MapReduce : Simplifed Data Processing on Large Cluster”
  - http://labs.google.com/papers/mapreduce.html
  - OSDI 2006 : “Bigtable: A Distributed Storage System for Structured Data”
  - http://labs.google.com/papers/bigtable-osdi06.pdf

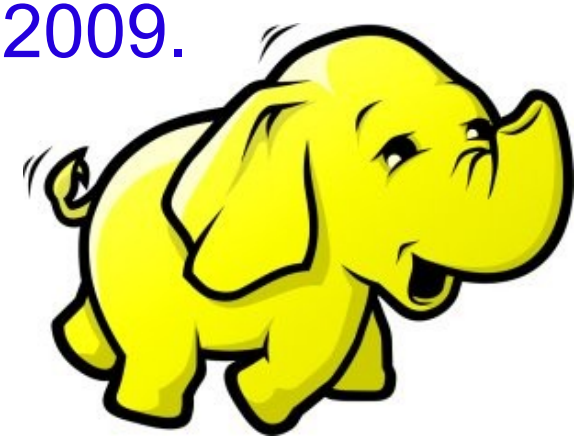
History of Hadoop … 2004 ~ Now

Hadoop 這套軟體的歷史源起 ... 2004 ~ Now

- Dong Cutting reference from Google's publication
- Added DFS & MapReduce implement to Nutch
- According to user feedback on the mail list of Nutch ....
- Hadoop became separated project since Nutch 0.8
- Nutch DFS → Hadoop Distributed File System (HDFS)
- Yahoo hire Dong Cutting to build a team of web search engine at year 2006.
  - Only 14 team members (engineers, clusters, users, etc.)
- Doung Cutting joined Cloudera at year 2009.

YAHOO!

cloudera



Who Use Hadoop ??

有哪些公司在用 Hadoop 這套軟體 ??

- Yahoo is the key contributor currently.
- IBM and Google teach Hadoop in universities …
- http://www.google.com/intl/en/press/pressrel/20071008_ibm_univ.html
- The New York Times used 100 Amazon EC2 instances and a Hadoop application to process 4TB of raw image TIFF data (stored in S3) into 11 million finished PDFs in the space of 24 hours at a computation cost of about $240 (not including bandwidth)

  – from http://en.wikipedia.org/wiki/Hadoop
- http://wiki.apache.org/hadoop/AmazonEC2
- http://wiki.apache.org/hadoop/PoweredBy

- A9.com
- ADSDAQ by Contextweb
- EHarmony
- Facebook
- Fox Interactive Media
- IBM
- ImageShack
- ISI
- Joost
- Last.fm
- Powerset
- The New York Times
- Rackspace
- Veoh
- Metaweb

Hadoop in production run ....
商業運轉中的 Hadoop 應用 ....

- February 19, 2008
- Yahoo! Launches World's Largest Hadoop Production Application
- http://developer.yahoo.net/blogs/hadoop/2008/02/yahoo-worlds-largest-production-hadoop.html

Number of links between pages in the index	roughly 1 trillion links
Size of output	over 300 TB, compressed!
Number of cores used to run single Map-Reduce job	over 10,000
Raw disk used in the production cluster	over 5 Petabytes

Hadoop in production run ....
商業運轉中的 Hadoop 應用 ....

- September 30, 2008
- Scaling Hadoop to 4000 nodes at Yahoo!
- http://developer.yahoo.net/blogs/hadoop/2008/09/scaling_hadoop_to_4000_nodes_a.html

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**

Comparison between Google and Hadoop
Google 與 Hadoop 的比較表

Develop Group	Google	Apache
Sponsor	Google	Yahoo, Amazon
Algorithm Method	MapReduce	MapReduce
Resource	open document	open source
File System (MapReduce)	GFS	HDFS
Storage System (for structure data)	big-table	HBase
Search Engine	Google	Nutch
OS	Linux	Linux / GPL

Why should we learn Hadoop ?
爲何需要學習 Hadoop ??

Search Jobs | Browse Jobs | Local Jobs | Salaries | Employment Trends

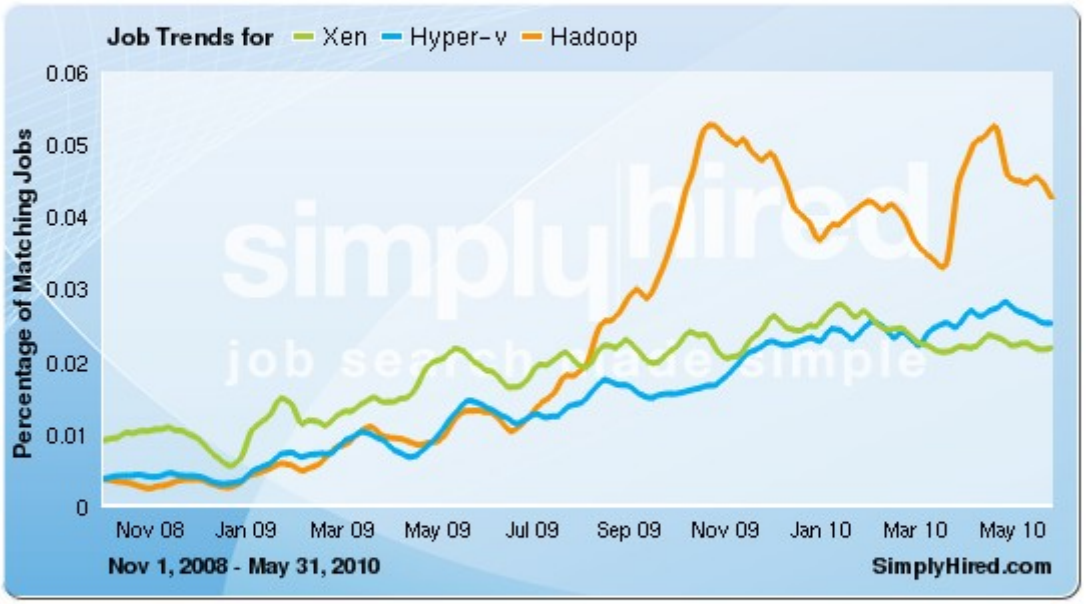
simplyhired®
job search made simple

Employment Trends

Xen, Hyper-V, Hadoop

Tip: You can compare trends by separating them with commas.

Xen, Hyper-v, Hadoop Trends



Xen, Hyper-v, Hadoop Job Trends

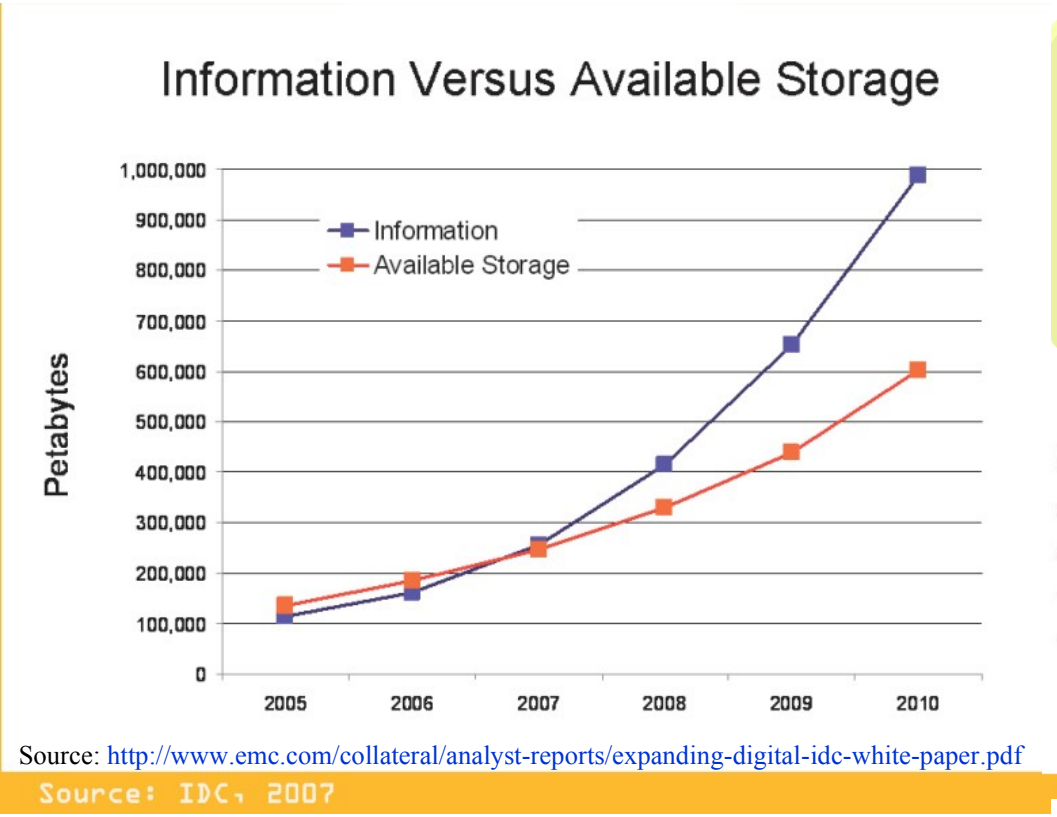
This graph displays the percentage of jobs with your search terms anywhere in the job listing. Since November 2008, the following has occurred:

- Xen jobs increased 141%
- Hyper-v jobs increased 551%
- Hadoop jobs did not change or there is no data available

1. Data Explore
資訊大爆炸

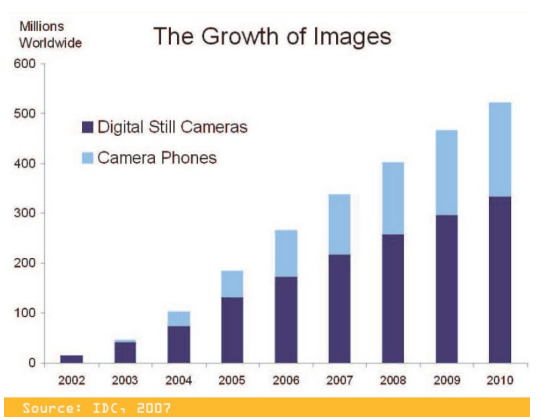
2. Data Mining Tool
方便作資料探勘的工作

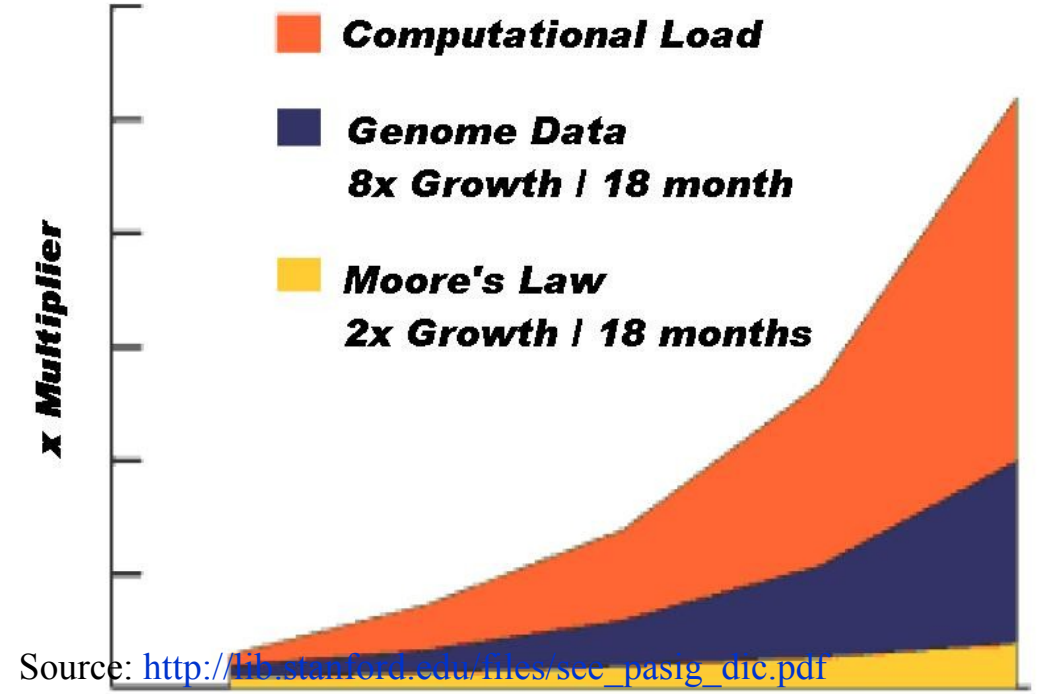
3. Looking for Jobs
好找工作 !!

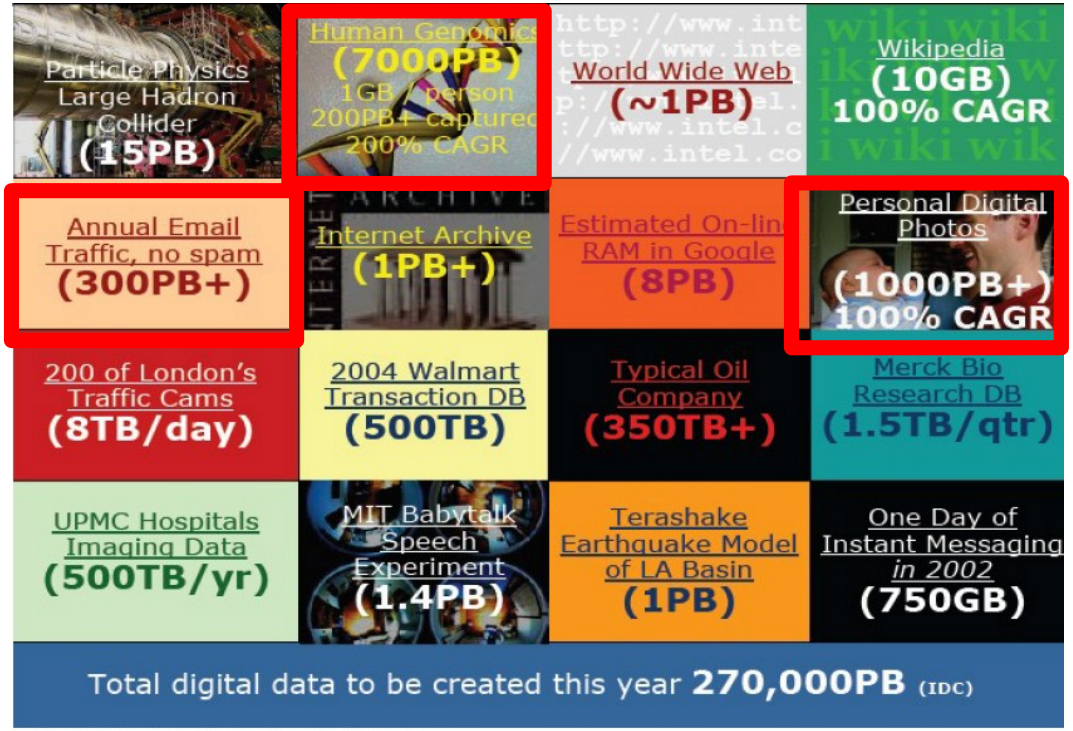


2007 Data Explore

Top 1 : Human Genomics – 7000 PB / Year
Top 2 : Digital Photos – 1000 PB+/ Year
Top 3 : E-mail (no Spam) – 300 PB+ / Year



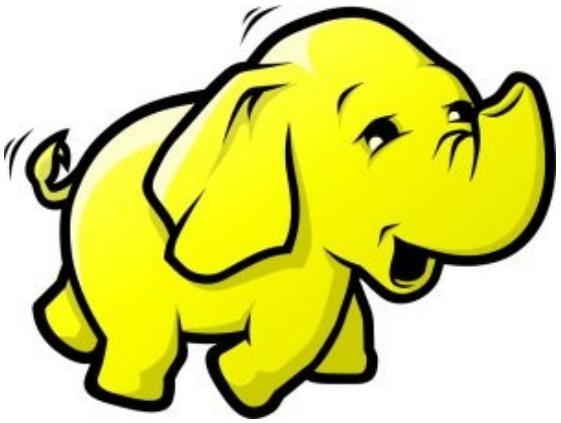




Hadoop 專業術語

Introduction to Hadoop Terminology

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



Two Key Elements of Operating System
作業系統兩大關鍵組成元素

Scheduler
程序排程

File System
檔案系統

Terminologies of Hadoop

Hadoop 文件中的專業術語

- Job
  - 任務
- Task
  - 小工作
- JobTracker
  - 任務分派者
- TaskTracker
  - 小工作的執行者
- Client
  - 發起任務的客戶端
- Map
  - 應對
- Reduce
  - 總和

- Namenode
  - 名稱節點
- Datanode
  - 資料節點
- Namespace
  - 名稱空間
- Replication
  - 副本
- Blocks
  - 檔案區塊 (64M)
- Metadata
  - 屬性資料

Two Key Roles of HDFS
HDFS 軟體架構的兩種關鍵角色

名稱節點 NameNode

- Master Node
- Manage NameSpace of HDFS
- Control Permission of Read and Write
- Define the policy of Replication
- Audit and Record the NameSpace
- Single Point of Failure

資料節點 DataNode

- Worker Nodes
- Perform operation of Read and Write
- Execute the request of Replication
- Multiple Nodes

Two Key Roles of Job Scheduler
程序排程的兩種關鍵角色

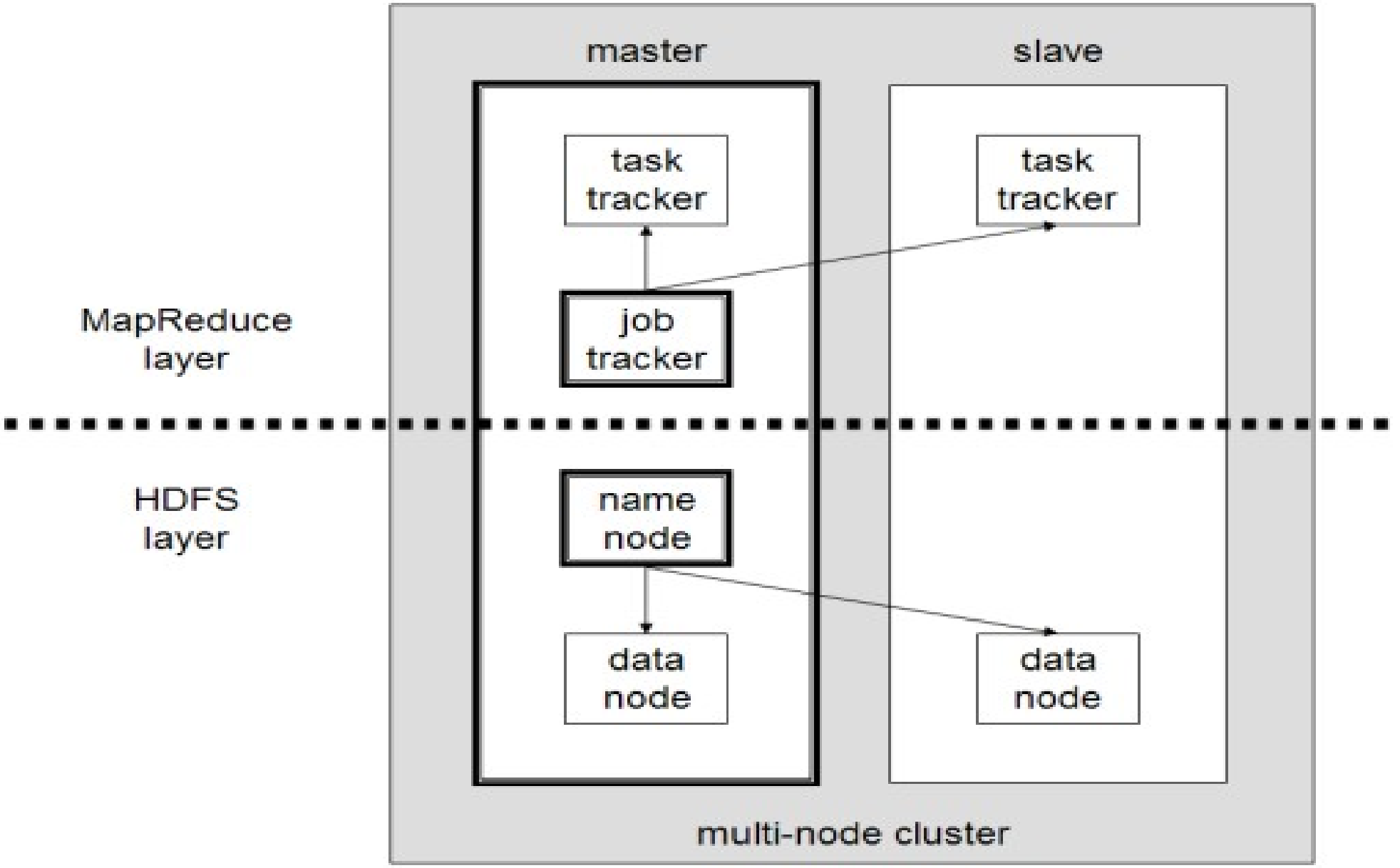
JobTracker

- **Master Node**
- **Receive Jobs from Hadoop Clients**
- **Assigned Tasks to TaskTrackers**
- **Define Job Queuing Policy, Priority and Error Handling**
- **Single Point of Failure**

TaskTracker

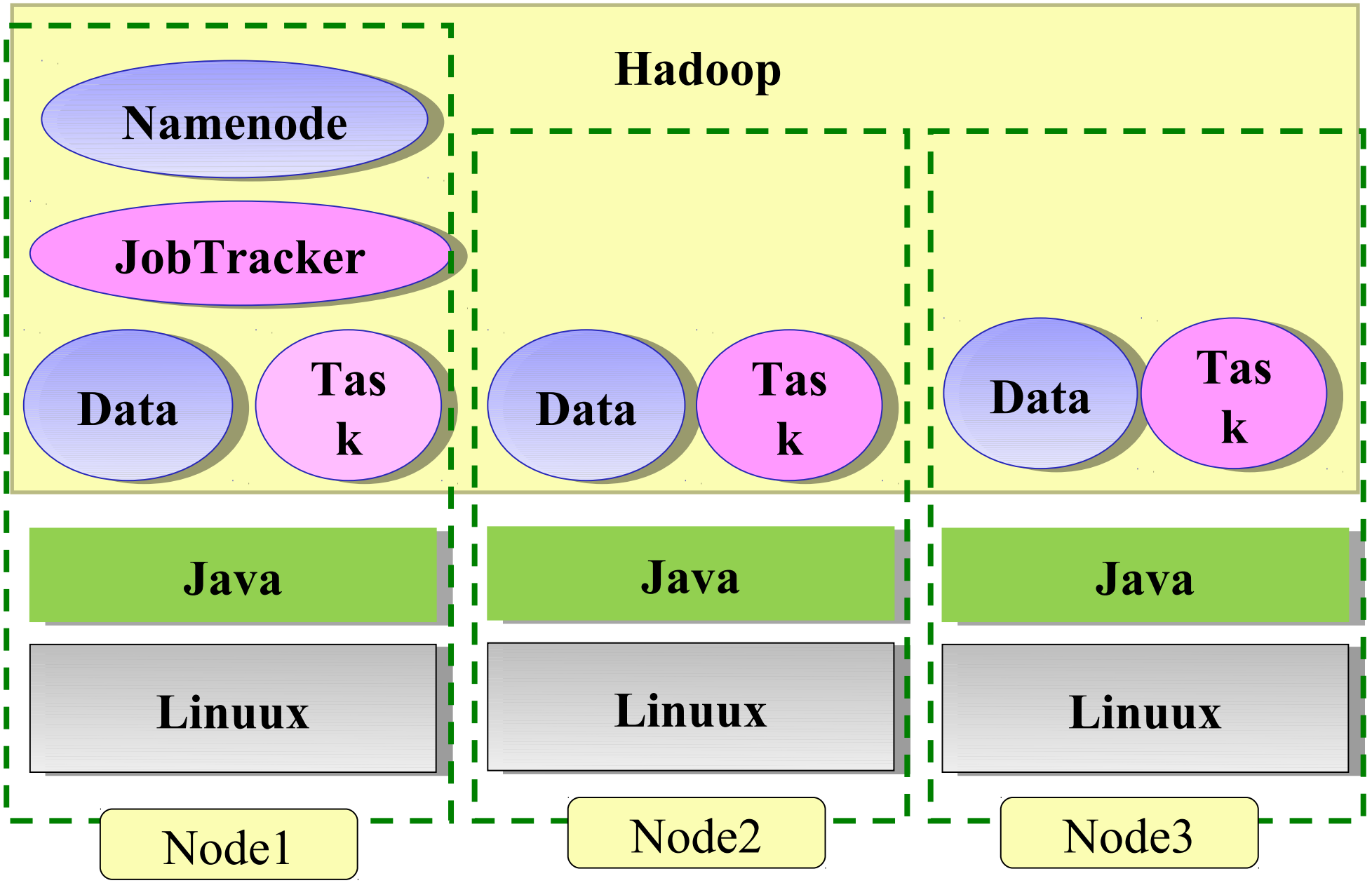
- **Worker Nodes**
- **Excute Mapper and Reducer Tasks**
- **Save Results and report task status**
- **Multiple Nodes**

Different Roles of Hadoop Architecture
Hadoop 軟體架構中的不同角色

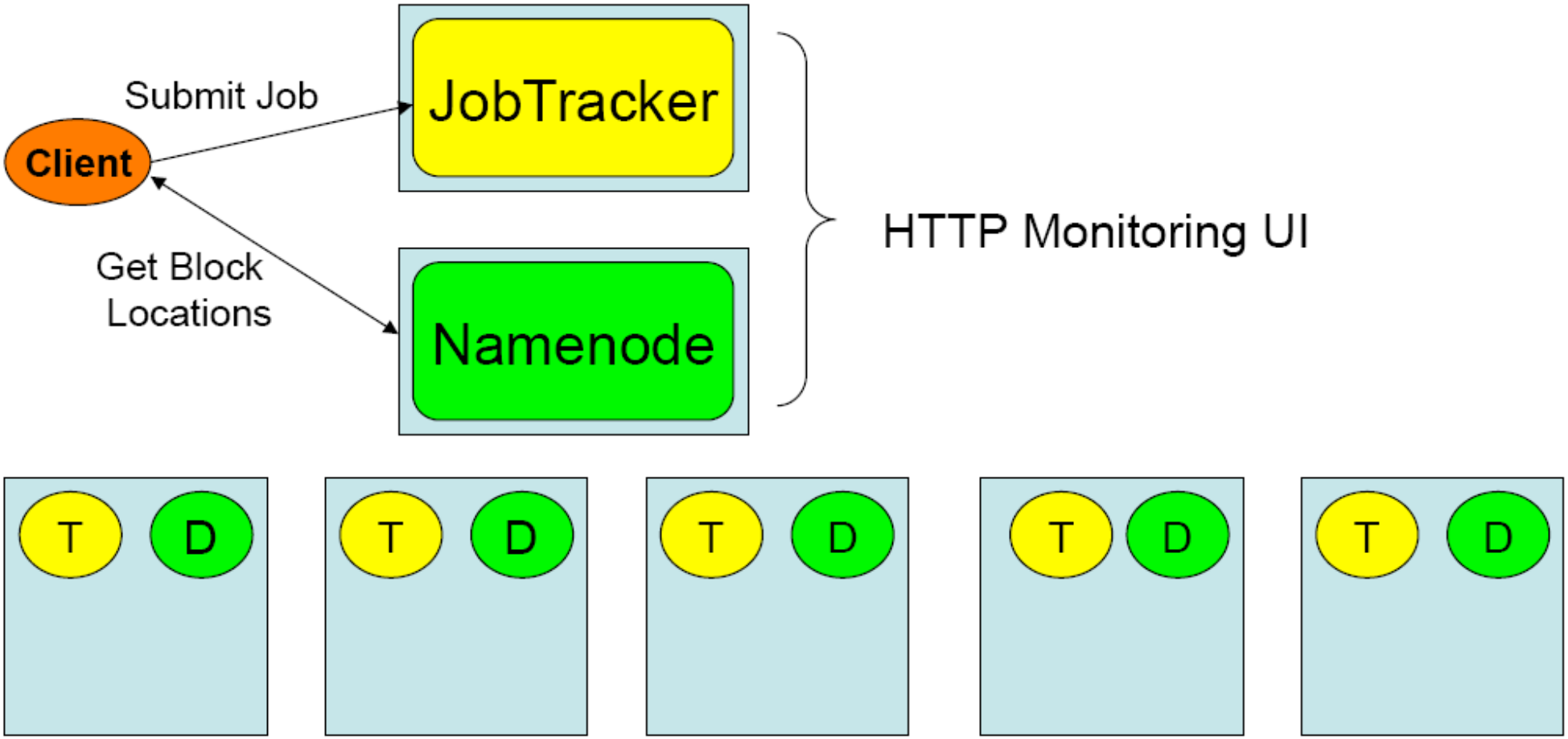


Distributed Operating System of Hadoop

Hadoop 建構成一個分散式作業系統



About Hadoop Client ...
不在雲裡的 Hadoop Client



What we learn today ?

WHAT	Hadoop 是運算海量資料的軟體平台 !! hadoop is a software platform to process vast amount of data!!
WHO	始祖是 Doug Cutting ， Apache 社群支持， Yahoo 贊助 From Doug Cutting to Apache Community, Yahoo and more !
WHEN	Hadoop 是 2004 年從 Nutch 分裂出來的專案 !! Hadoop became separate project since year 2004 !!
WHY	資料大爆炸、資料探勘、找工作 Data Explore, Data Mining, Jobs !!
HOW	建構在大型的個人電腦叢集之上 Install on large clusters built of commodity hardware !!

Questions?

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

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

Powered by DRBL

Lab 0 : 準備實驗環境

- 下載 Ubuntu 9.10 安裝光碟 ISO 檔
- 下載 WinSCP 與 PieTTY
- 下載 VirtualBox
- 安裝 VirtualBox

Ubuntu 9.10 (Karmic Koala)

This directory contains the most frequently downloaded Ubuntu images. Other images, including DVDs and source CDs, may be available on the cdimage server. See also the list of download mirrors.

Select an image

Ubuntu is distributed on five types of images described below.

Desktop CD

The desktop CD allows you to try Ubuntu without changing your computer at all, and at your option to install it permanently later. This type of CD is what most people will want to use. You will need at least 256MB of RAM to install from this CD.

There are two images available, each for a different type of computer:

PC (Intel x86) desktop CD

For almost all PCs. This includes most machines with Intel/AMD/etc type processors and almost all computers that run Microsoft Windows, as well as newer Apple Macintosh systems based on Intel processors. Choose this if you are at all unsure.

64-bit PC (AMD64) desktop CD

Choose this to take full advantage of computers based on the AMD64 or EM64T architecture (e.g., Athlon64, Opteron, EM64T Xeon, Core 2). If you have a non-64-bit processor made by AMD, or if you need full support for 32-bit code, use the Intel x86 images instead.

連線至 http://free.nchc.org.tw/ubuntu-cd/karmic/

下載 Ubuntu 9.10PC (Intel x86) desktop CD

連線至 http://ntu.csie.org/~piaip/pietty/

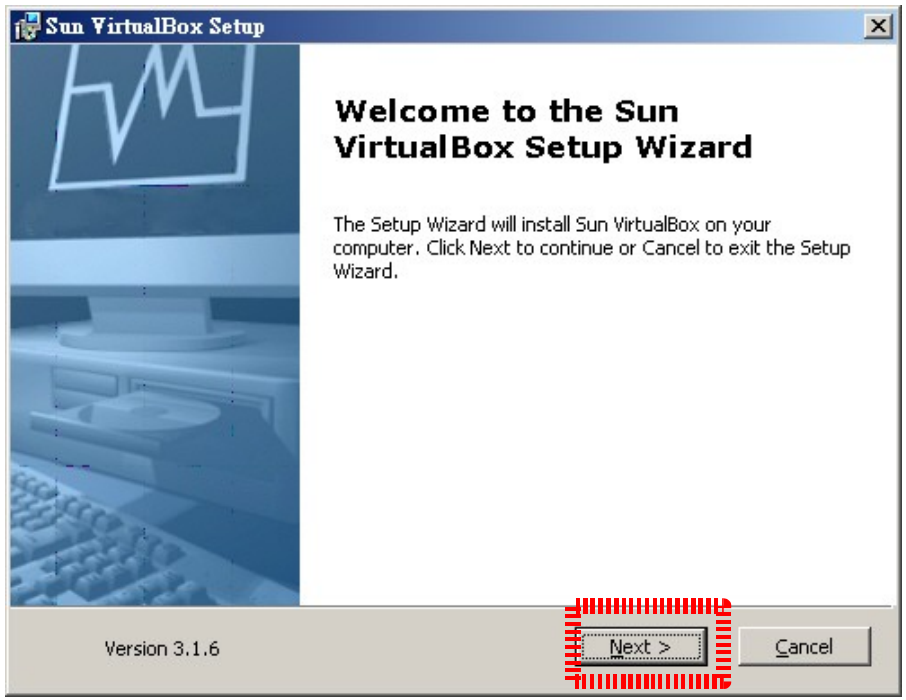
下載 PieTTY (SSH Client)

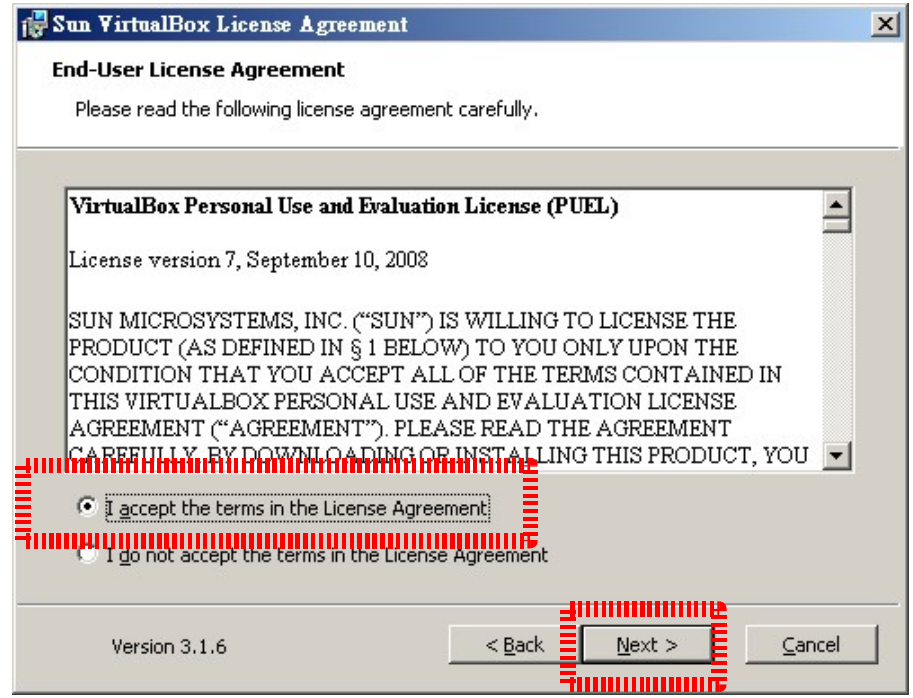
連線至 http://winscp.net

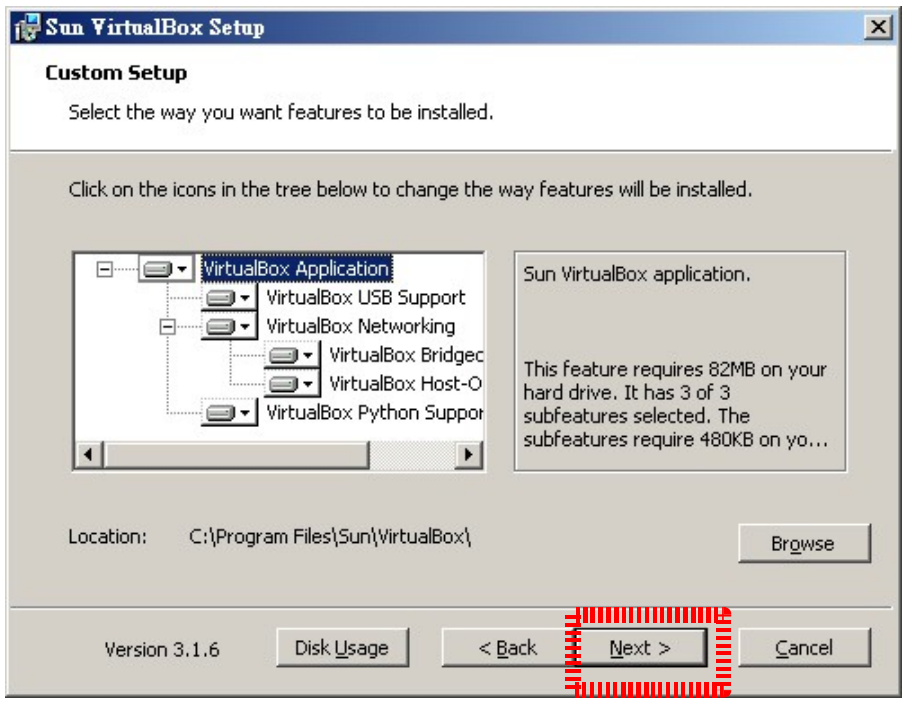
下載 WinSCP (SFTP Client)

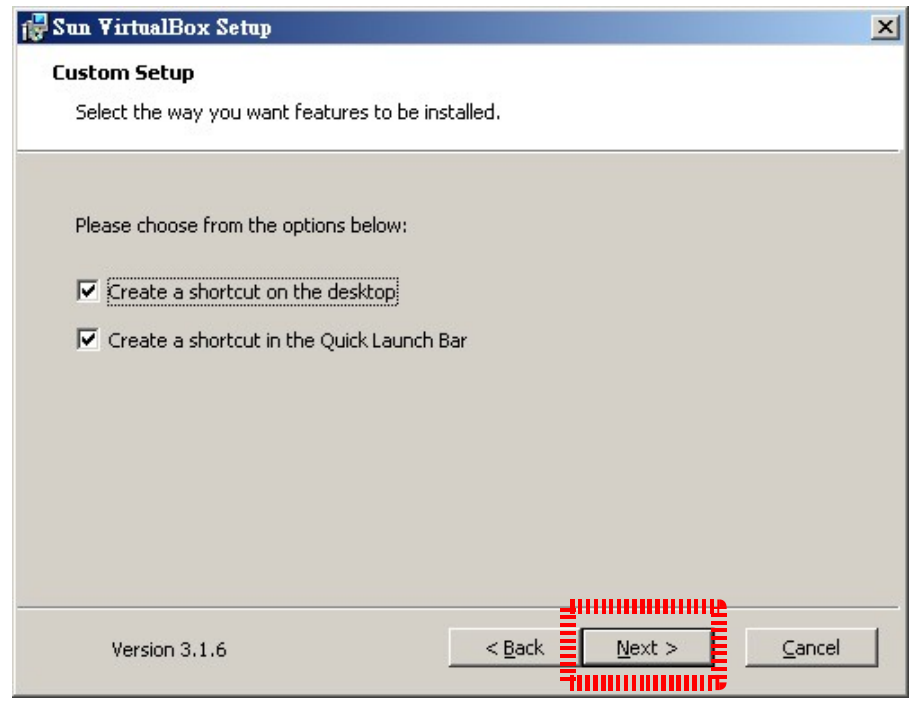
連線至 http://www.virtualbox.org/wiki/Downloads
下載 VirtualBox 虛擬化工具

安装 VirtalBox（1）

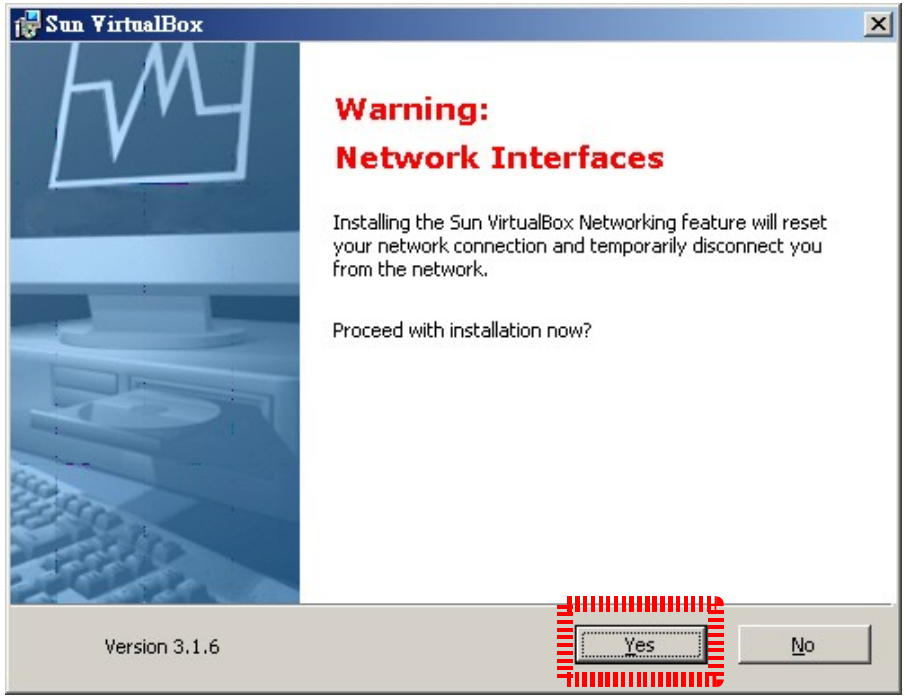


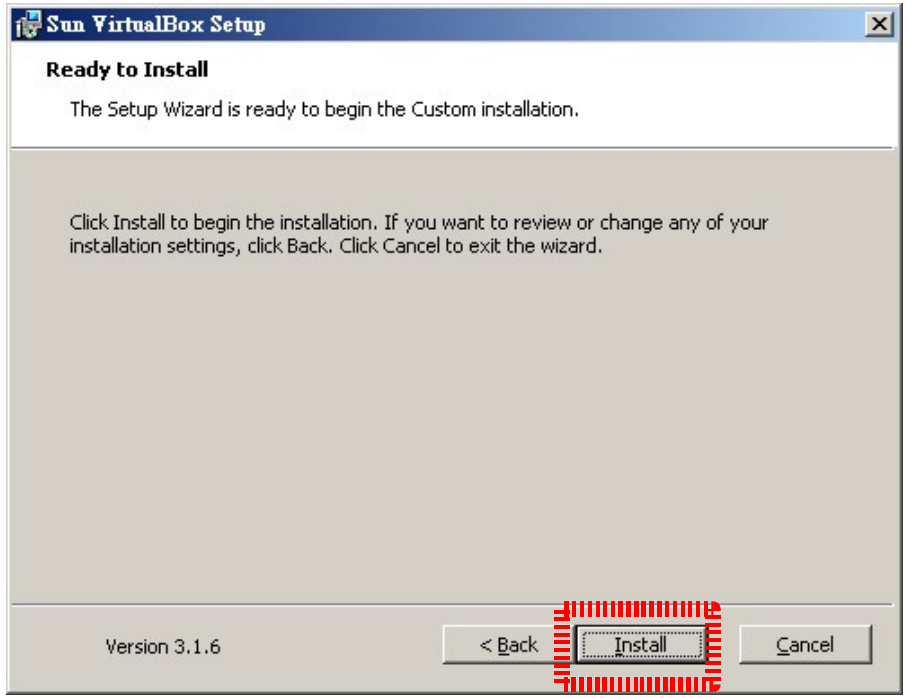


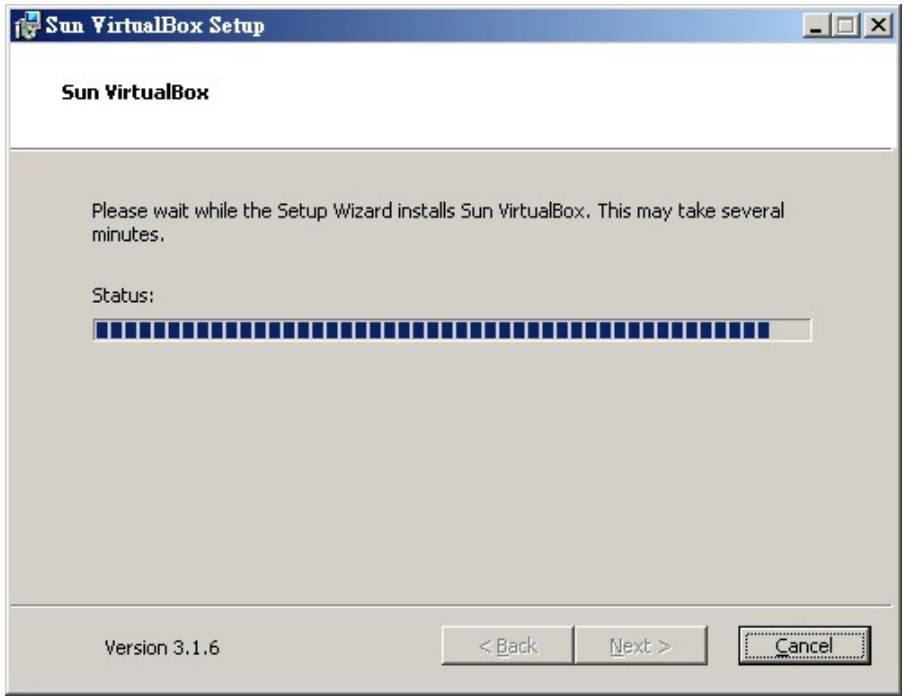


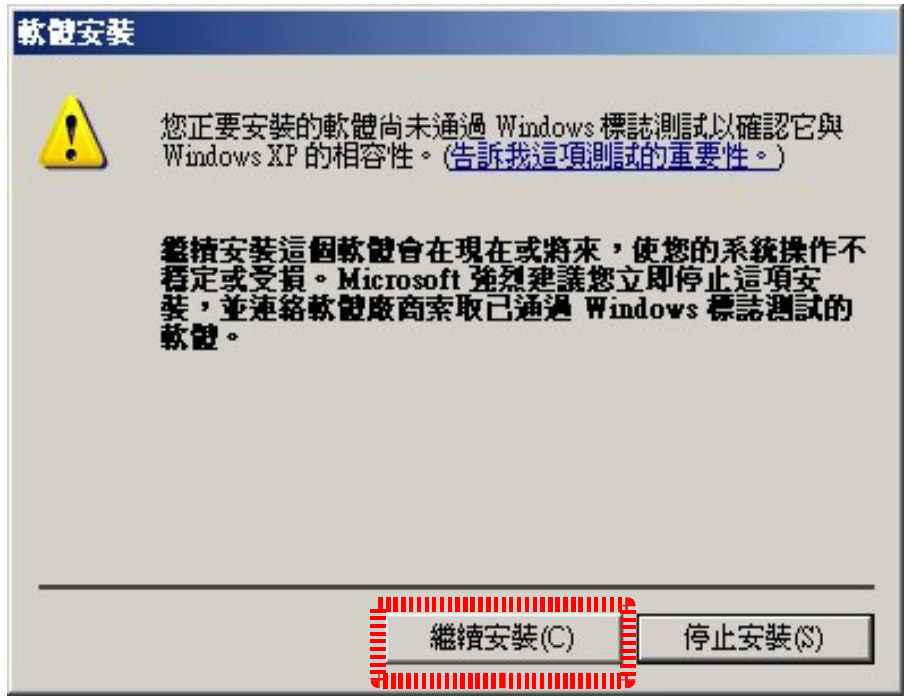


安裝 VirtalBox（2）

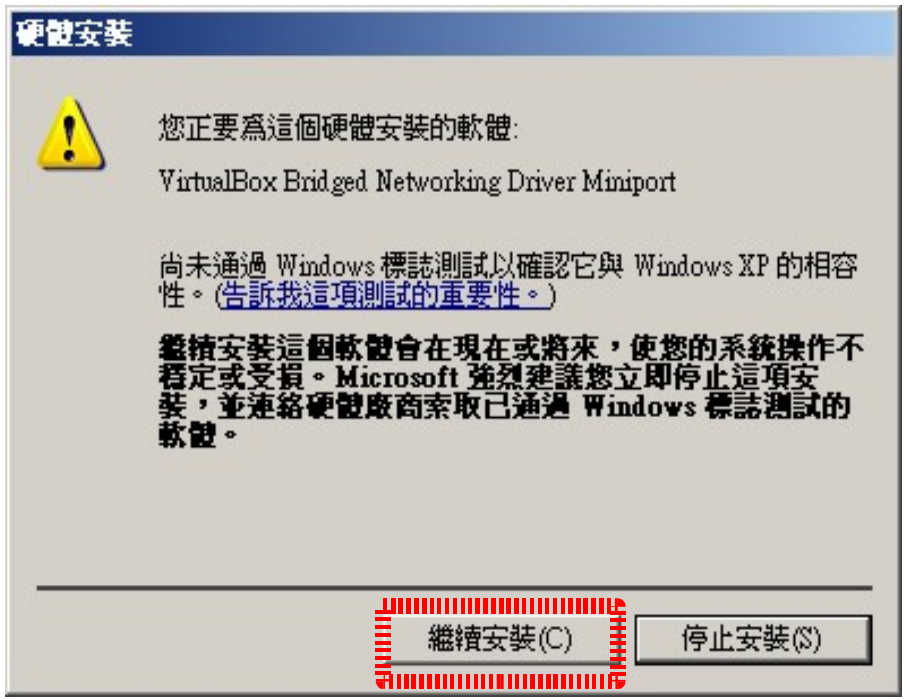


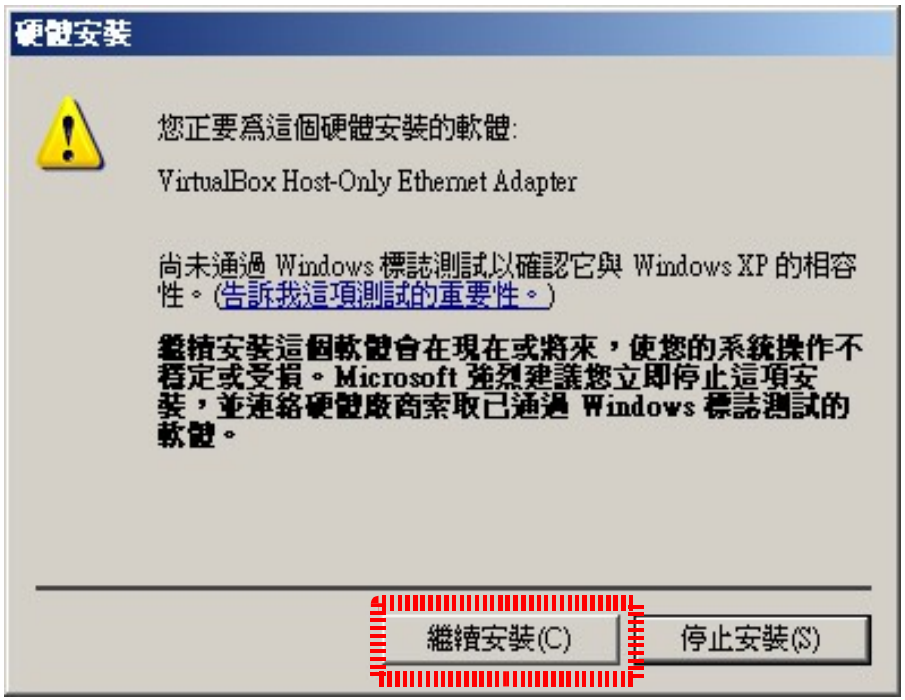


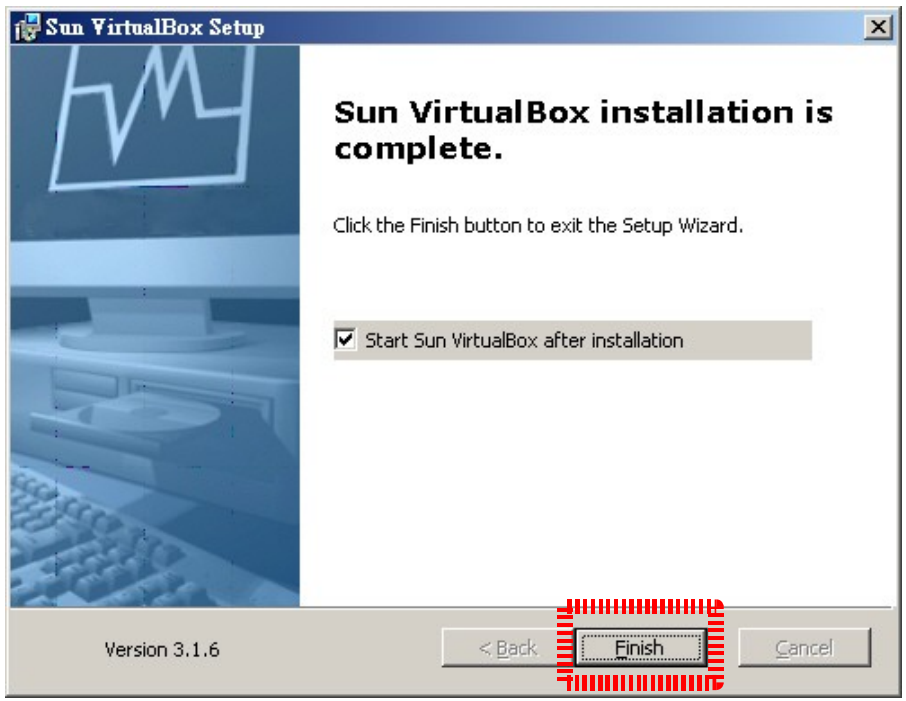


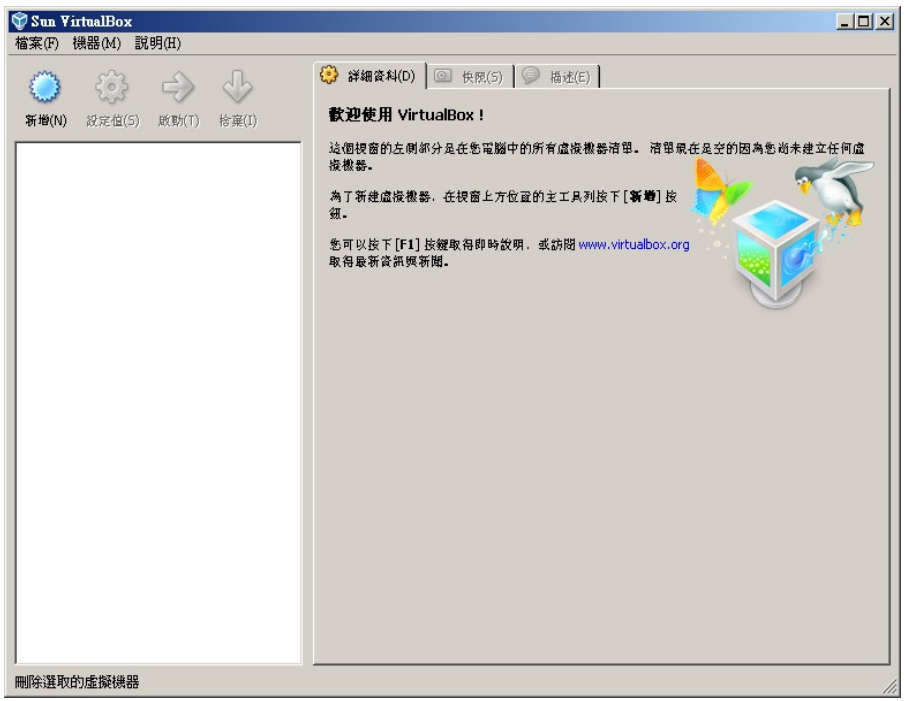


安裝 VirtalBox（3）



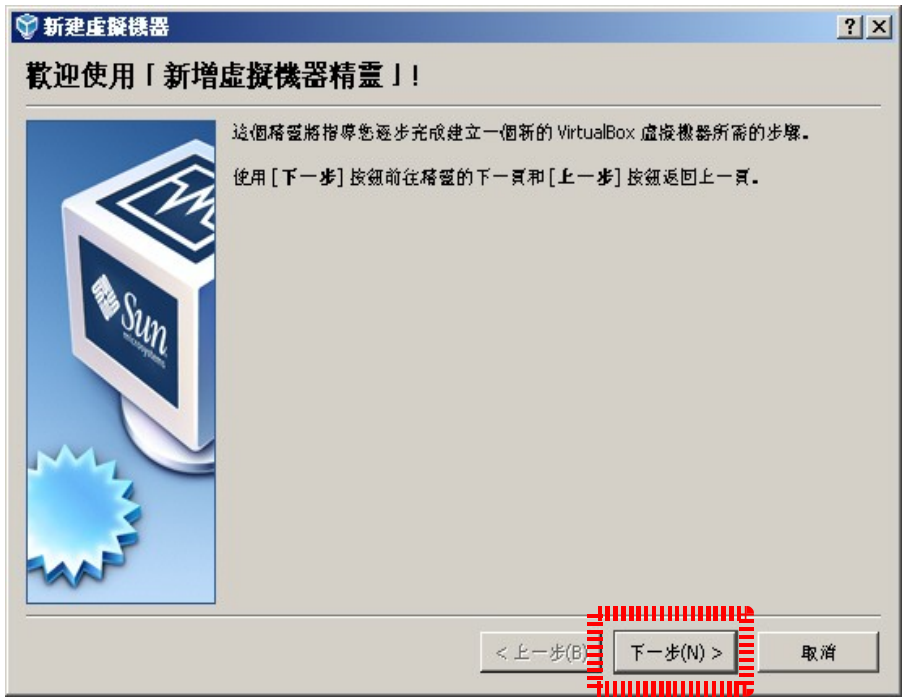


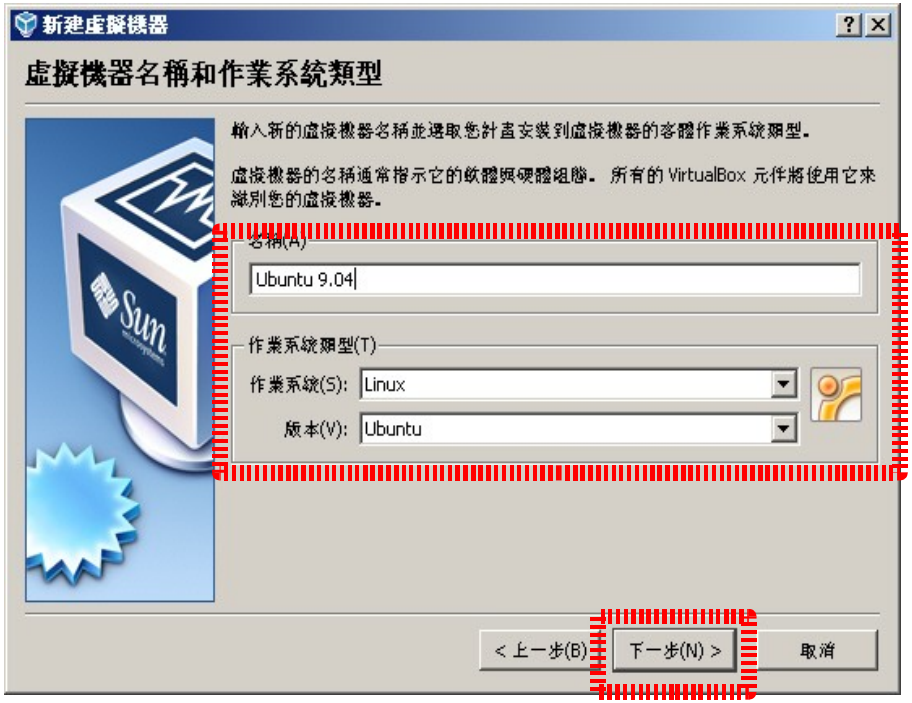


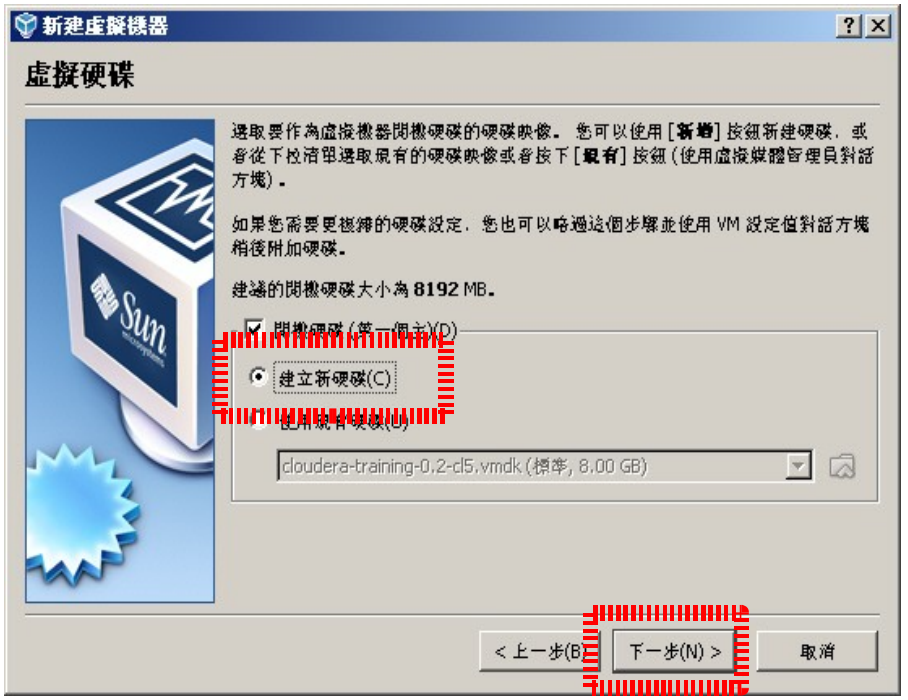


Lab 1 : 安裝 Ubuntu 9.10 GNU/Linux

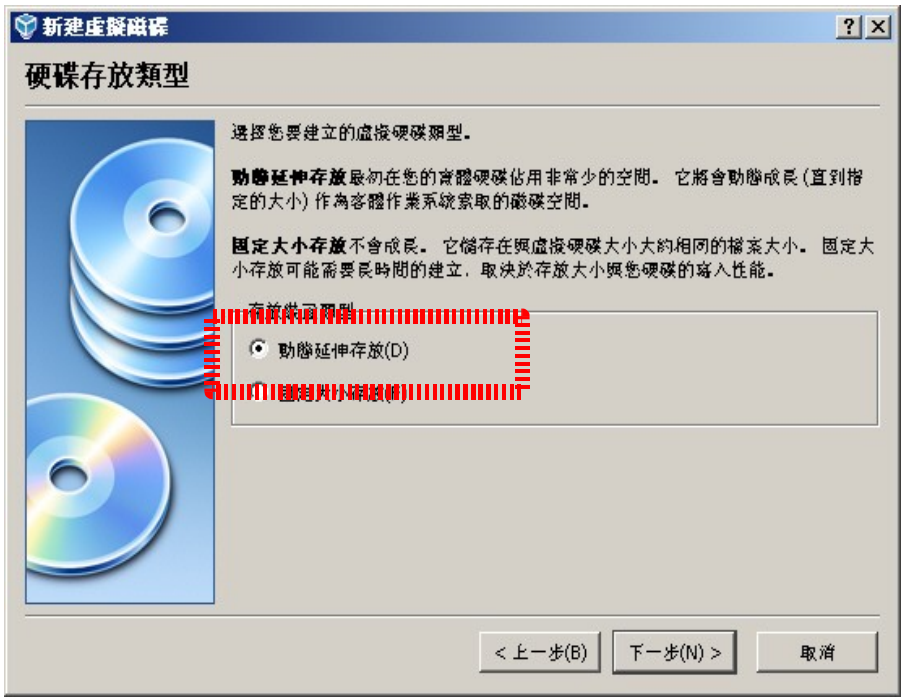
安裝 Ubuntu 9.10 (1)



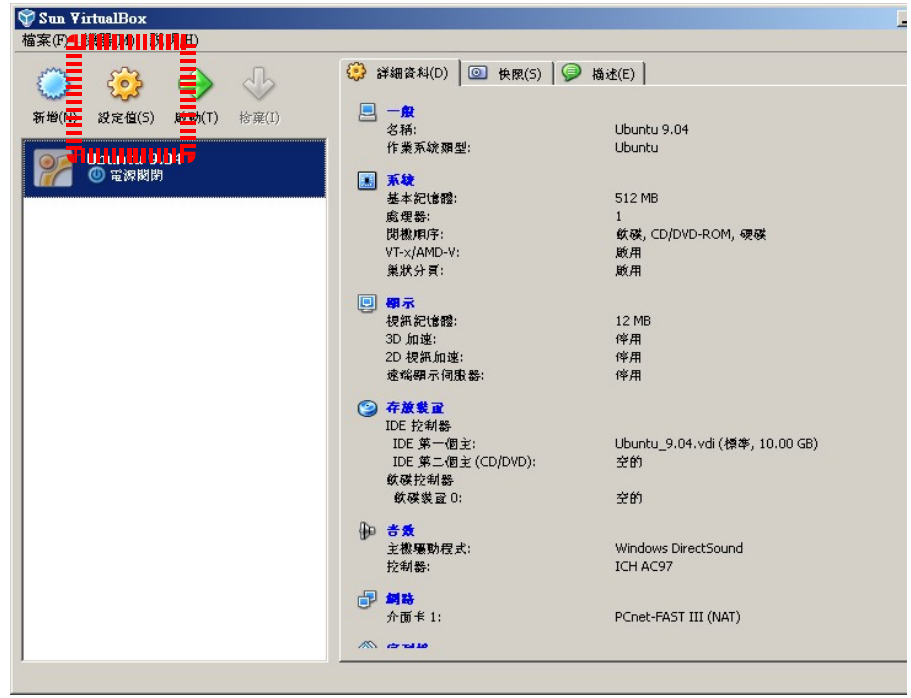


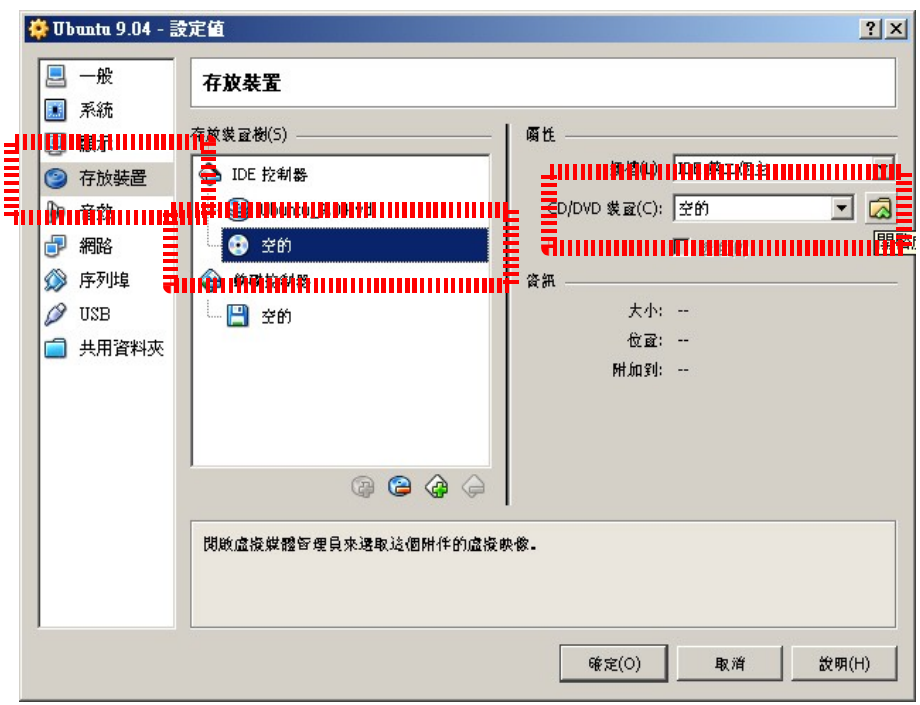


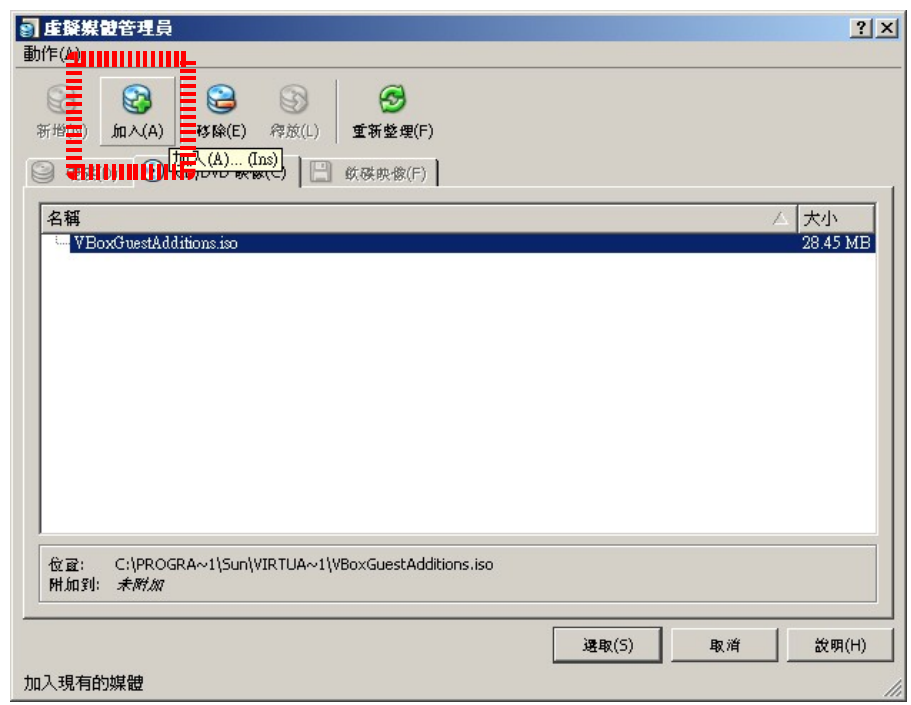
安裝 Ubuntu 9.10 (2)



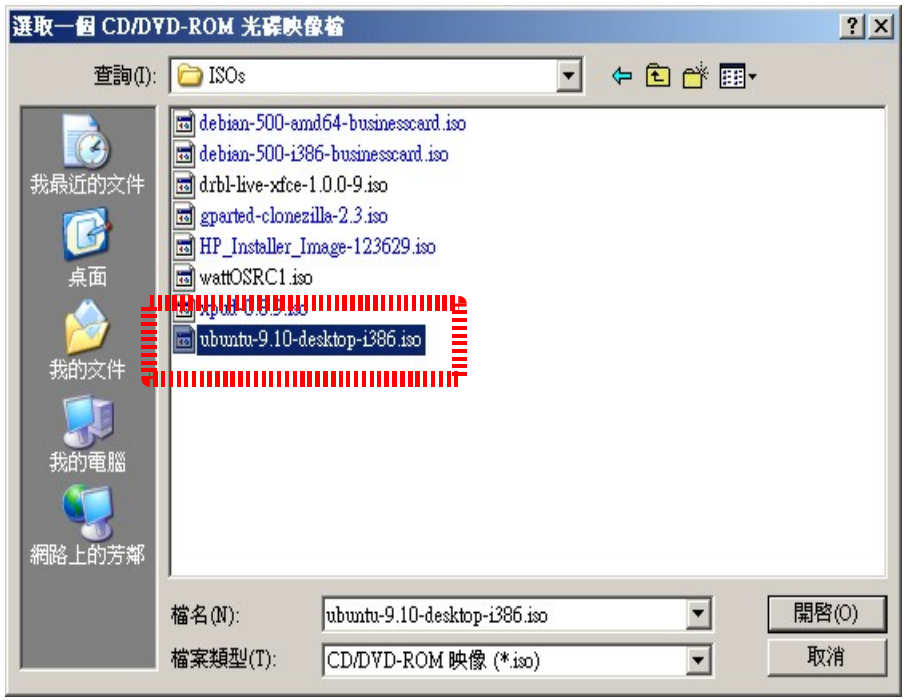
安裝 Ubuntu 9.10 (3)

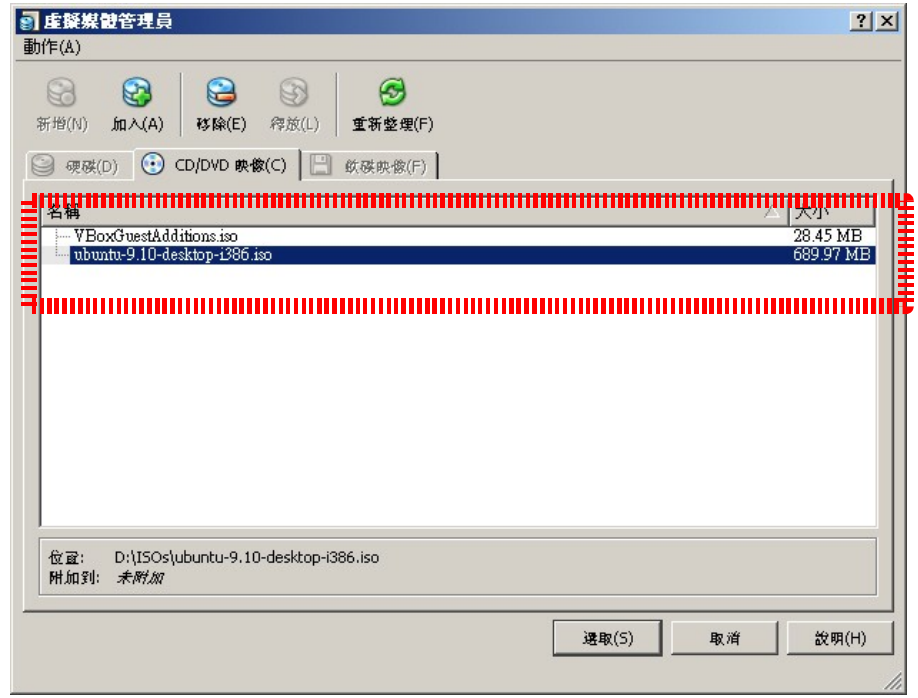


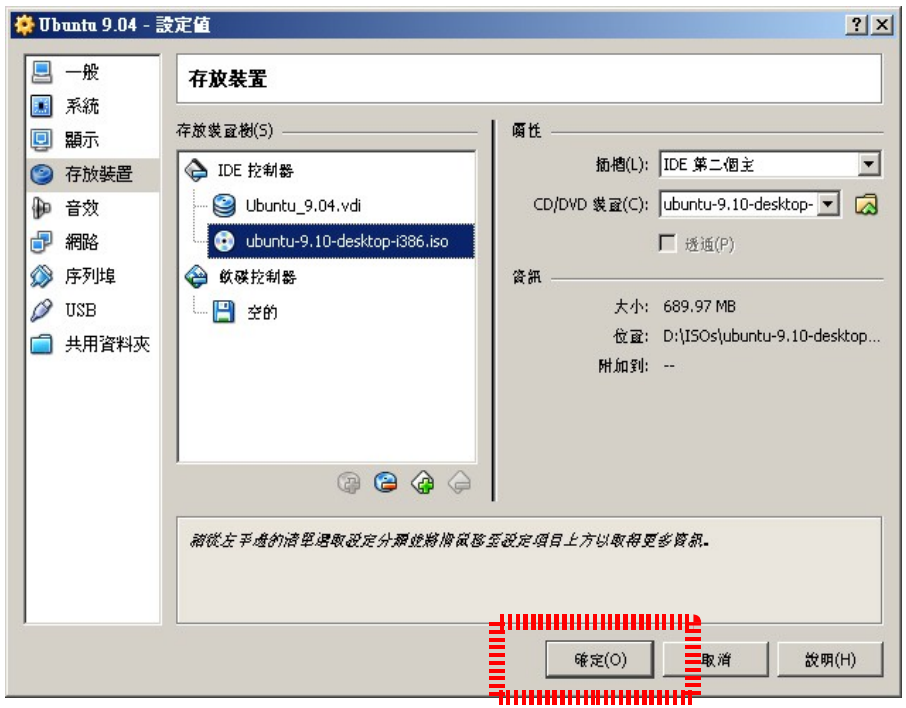


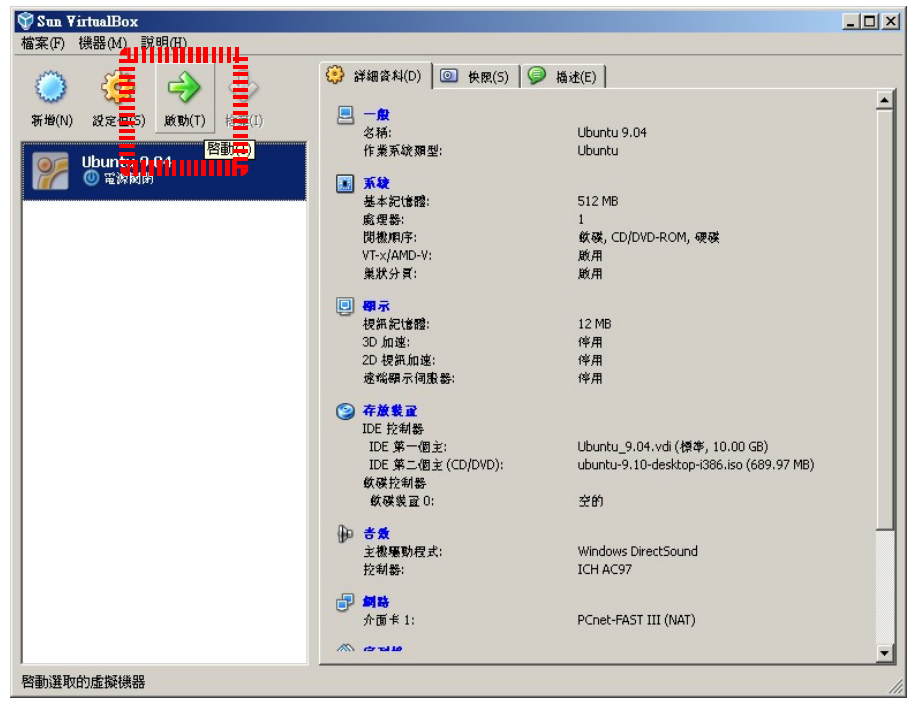


安裝 Ubuntu 9.10 (4)

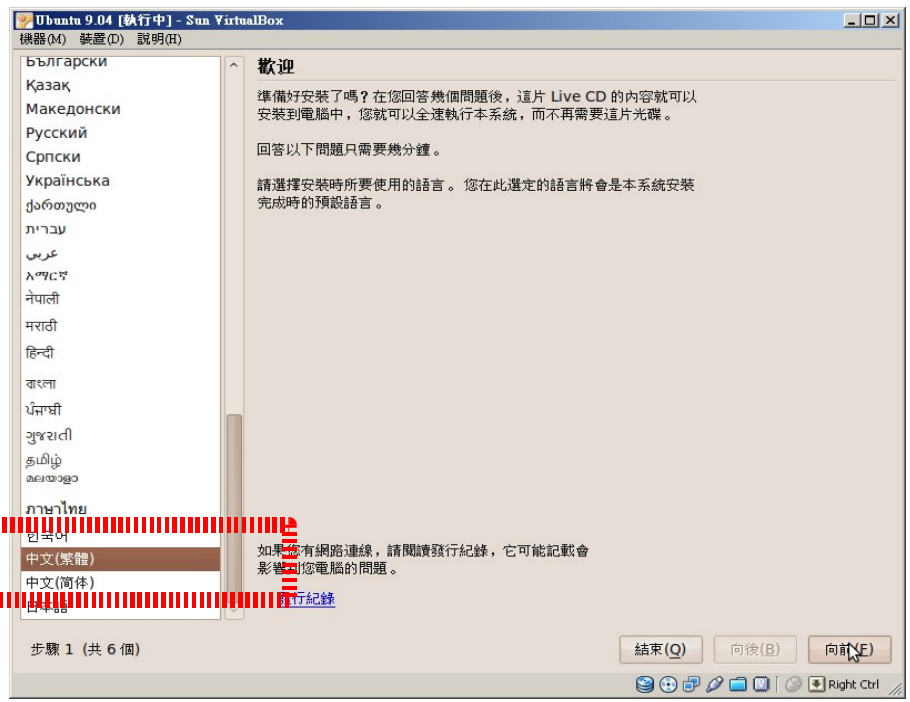


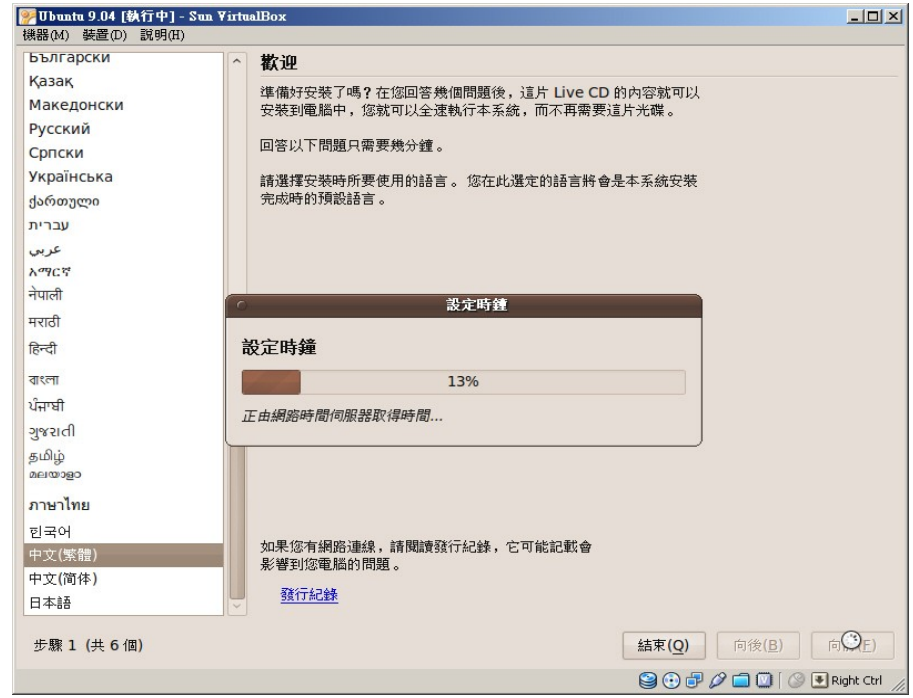


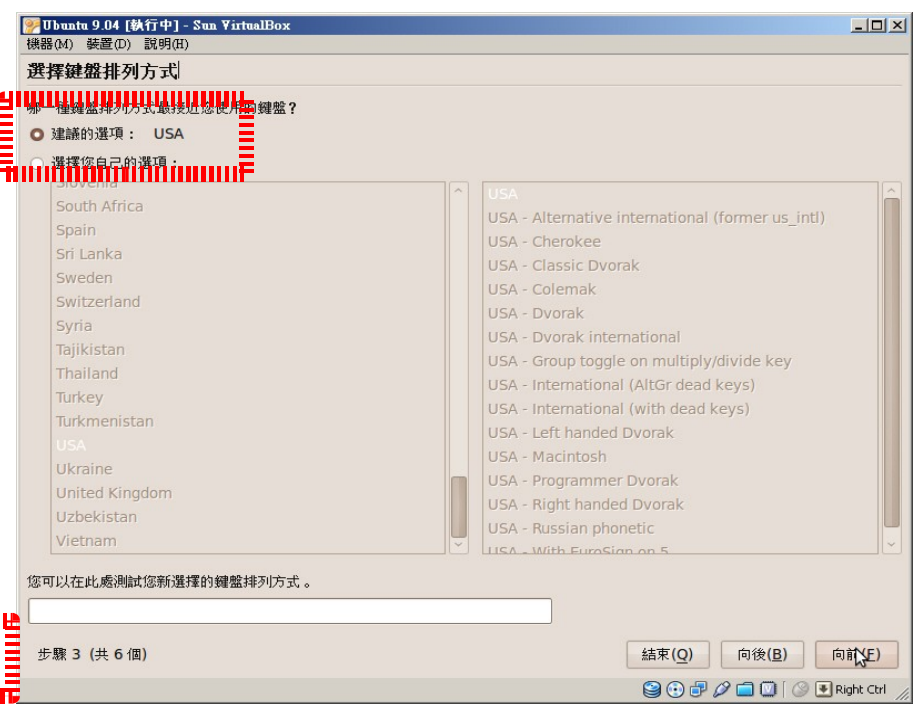




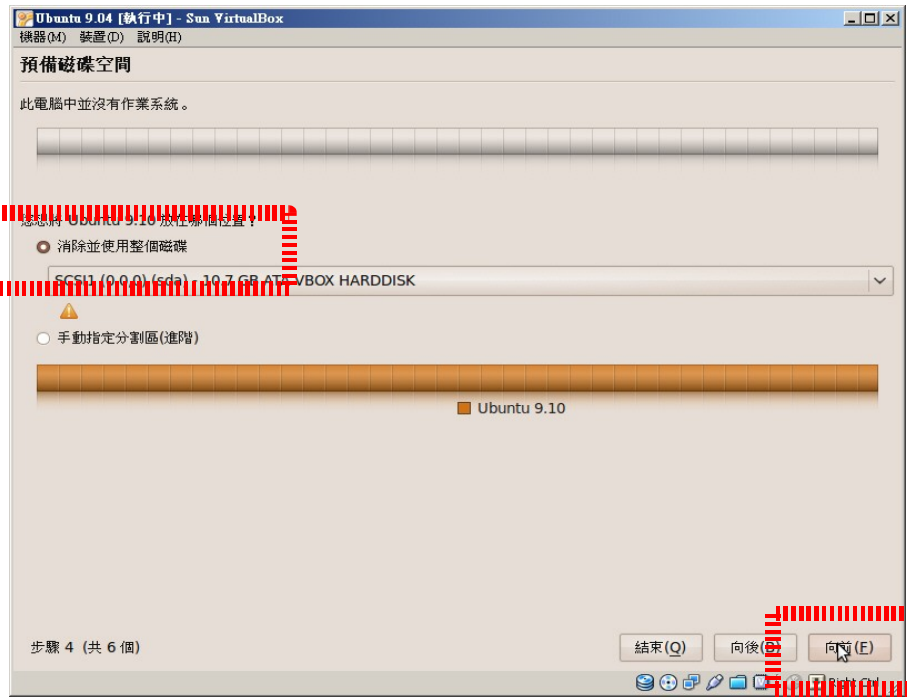
安裝 Ubuntu 9.10 (5)

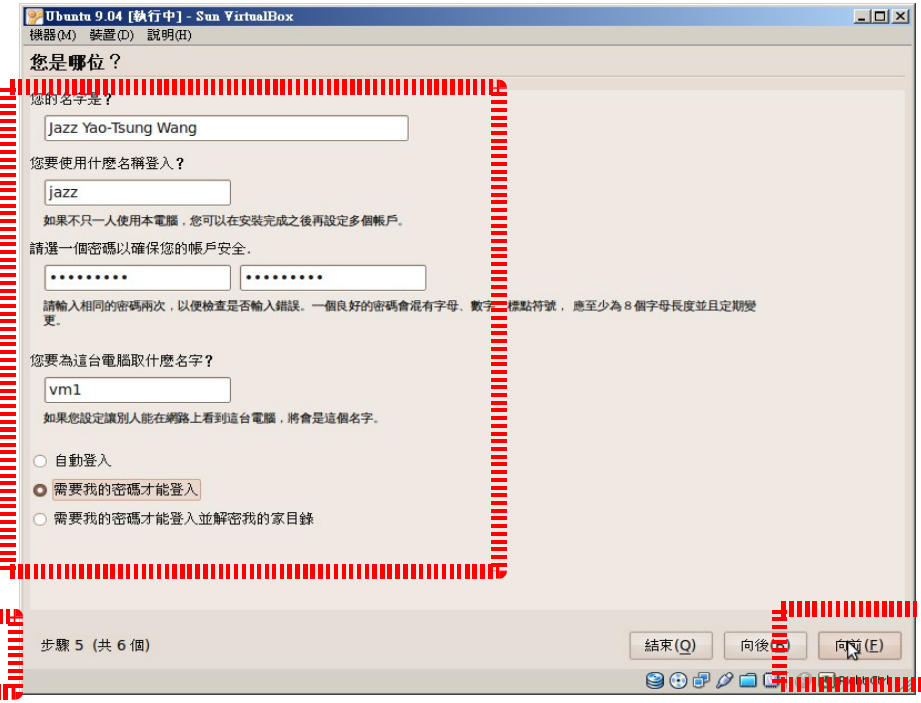


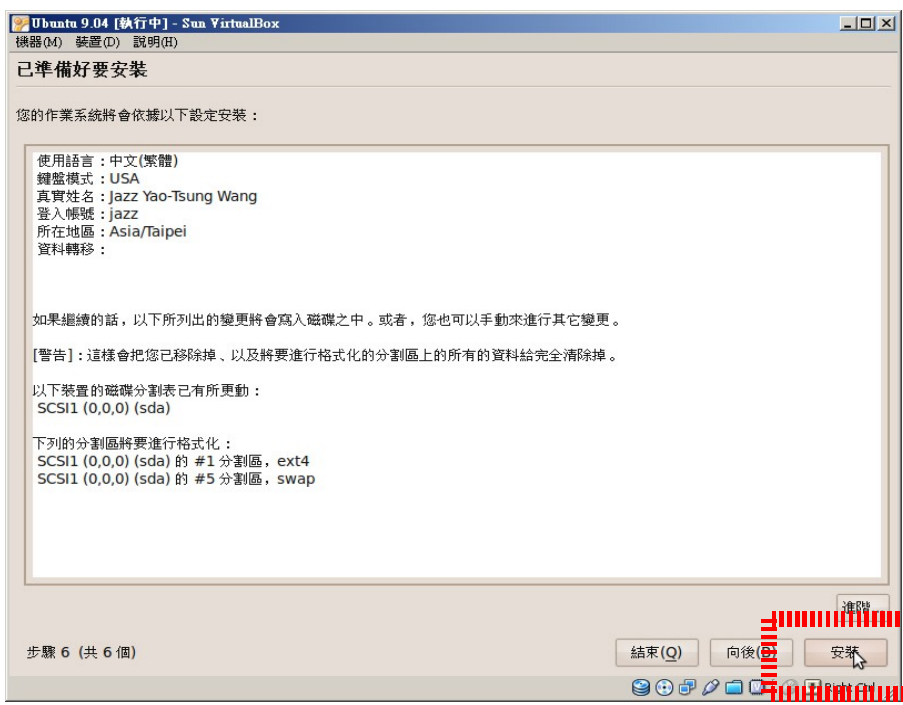




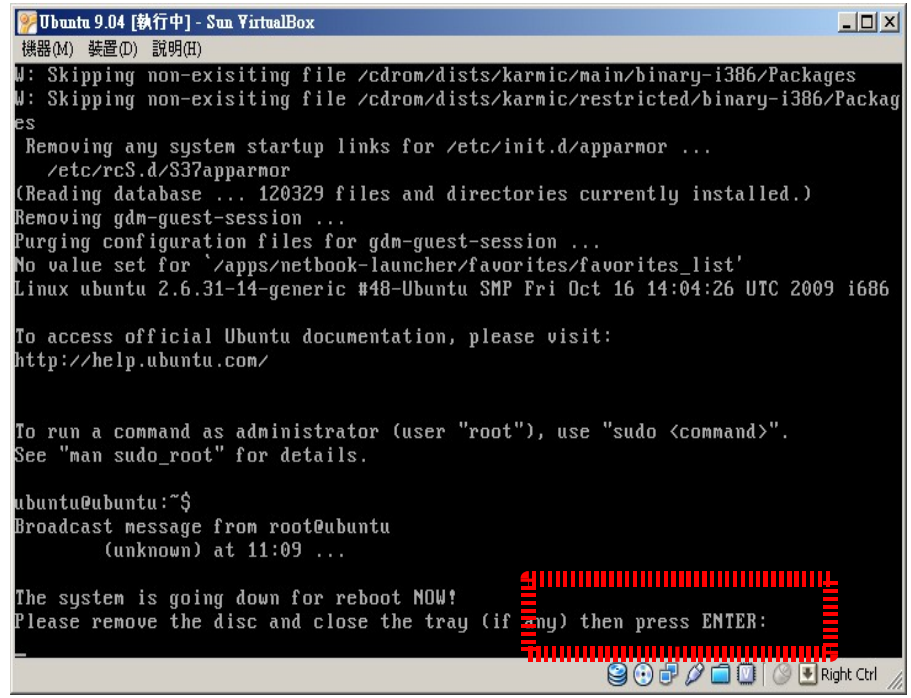
安裝 Ubuntu 9.10 (6)

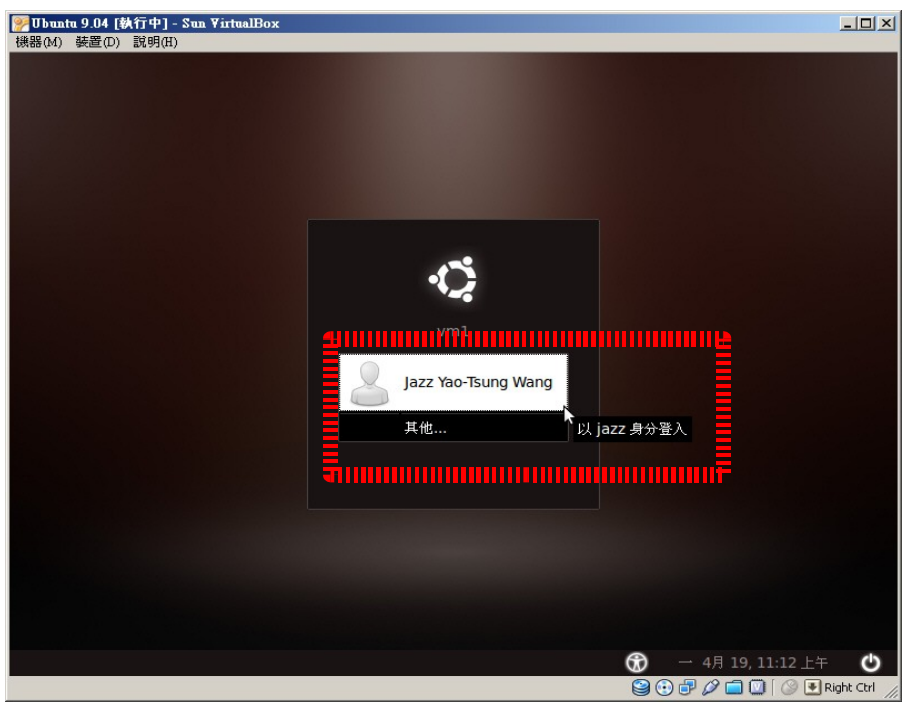


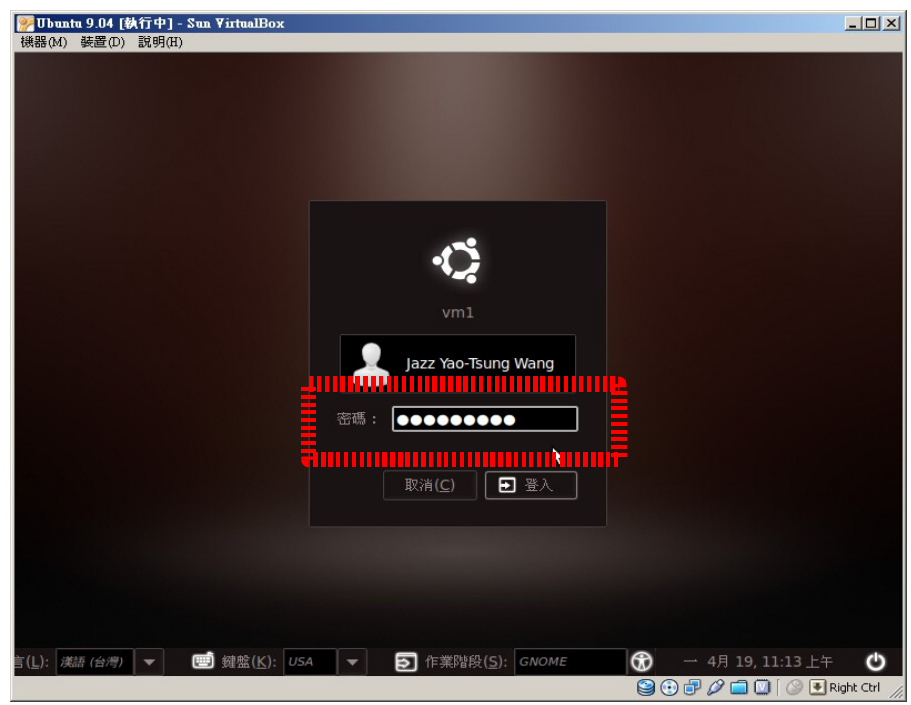




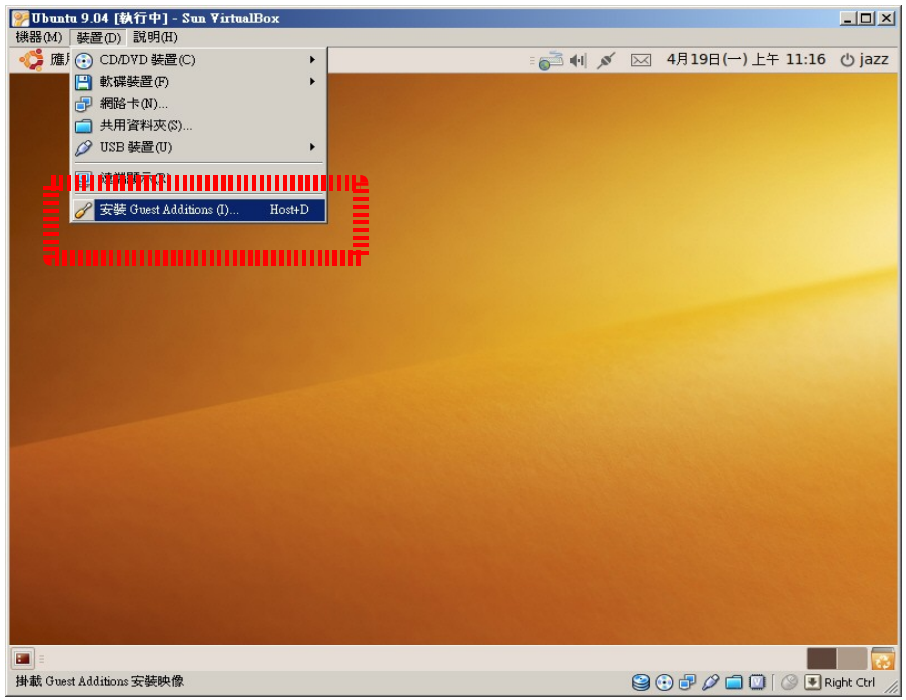
安裝 Ubuntu 9.10 (7)

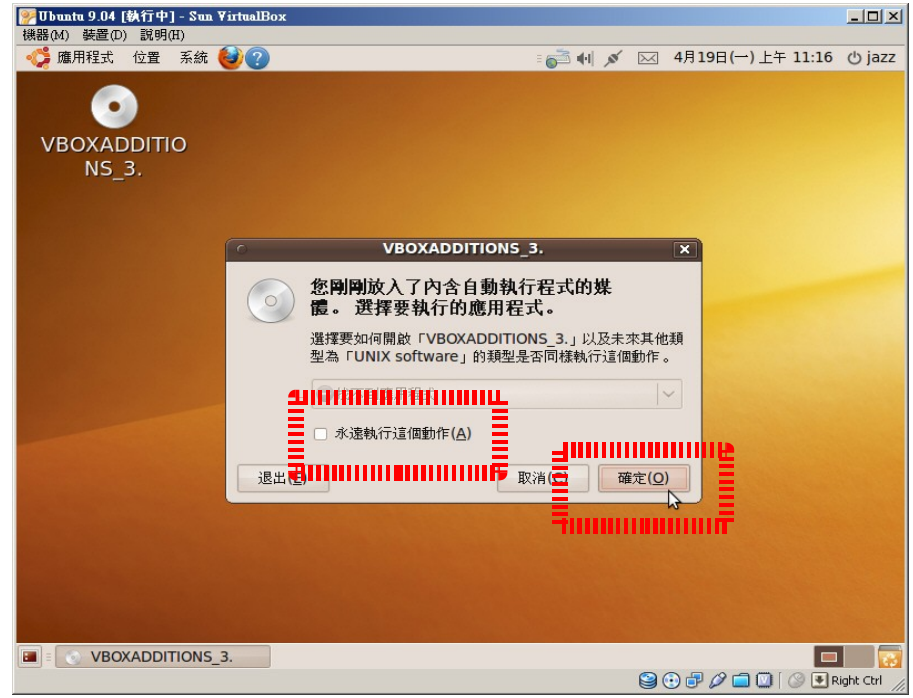


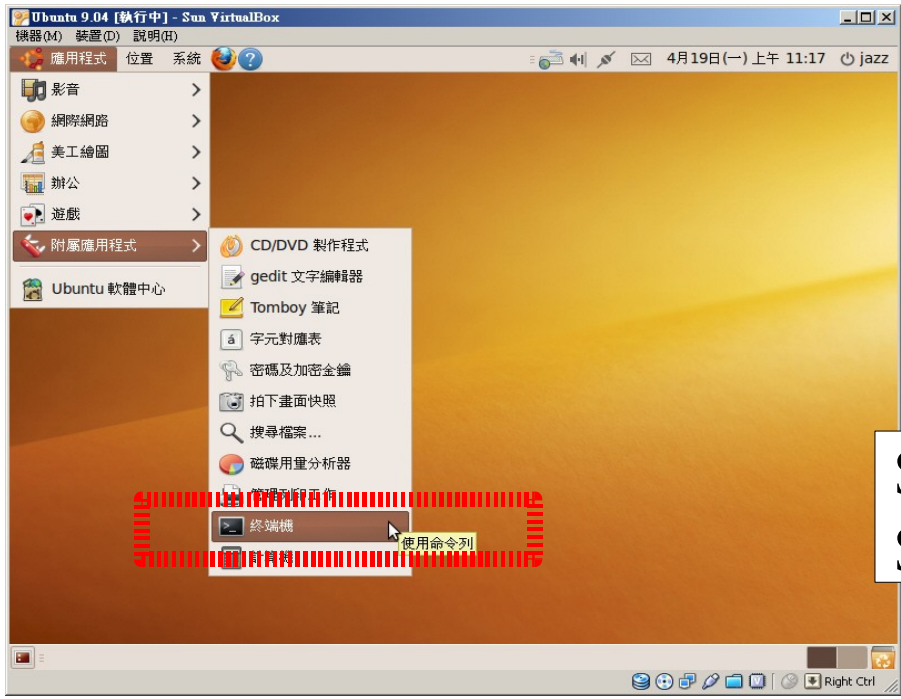


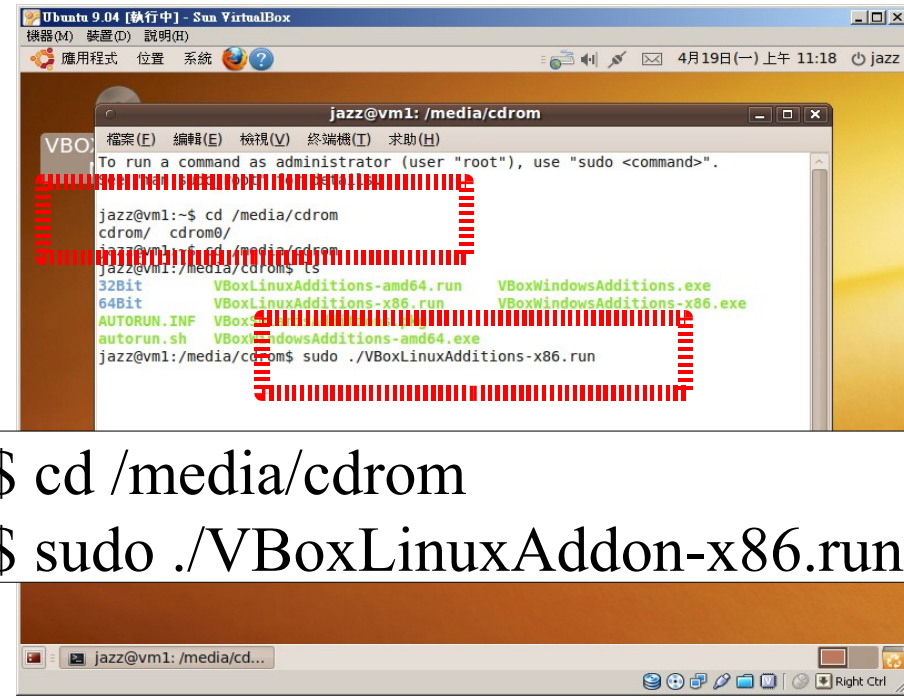


安裝 Ubuntu 9.10 (8)



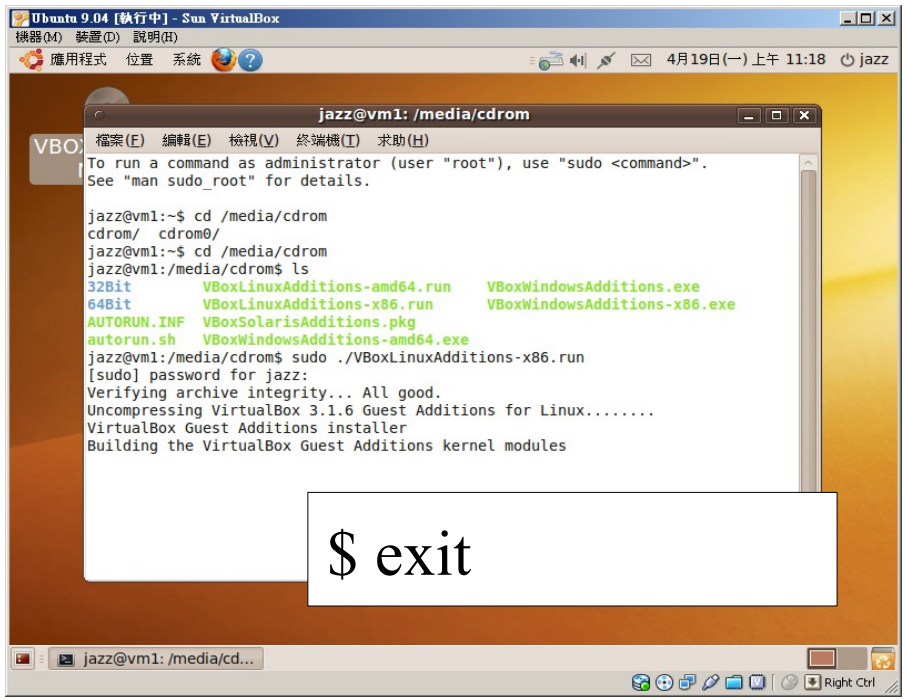


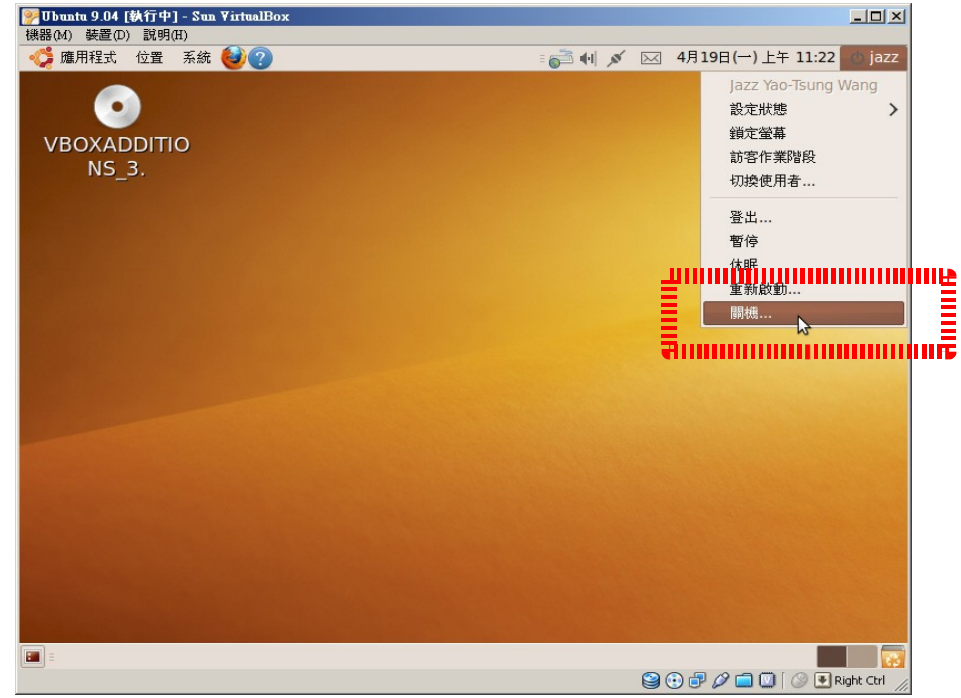





```
$ cd /media/cdrom
$ sudo ./VBoxLinuxAddon-x86.run
```

安裝 Ubuntu 9.10 (9)





一個令人哭笑不得的說法

學習使用免費的 Linux，

除非你的時間也是免費的！！

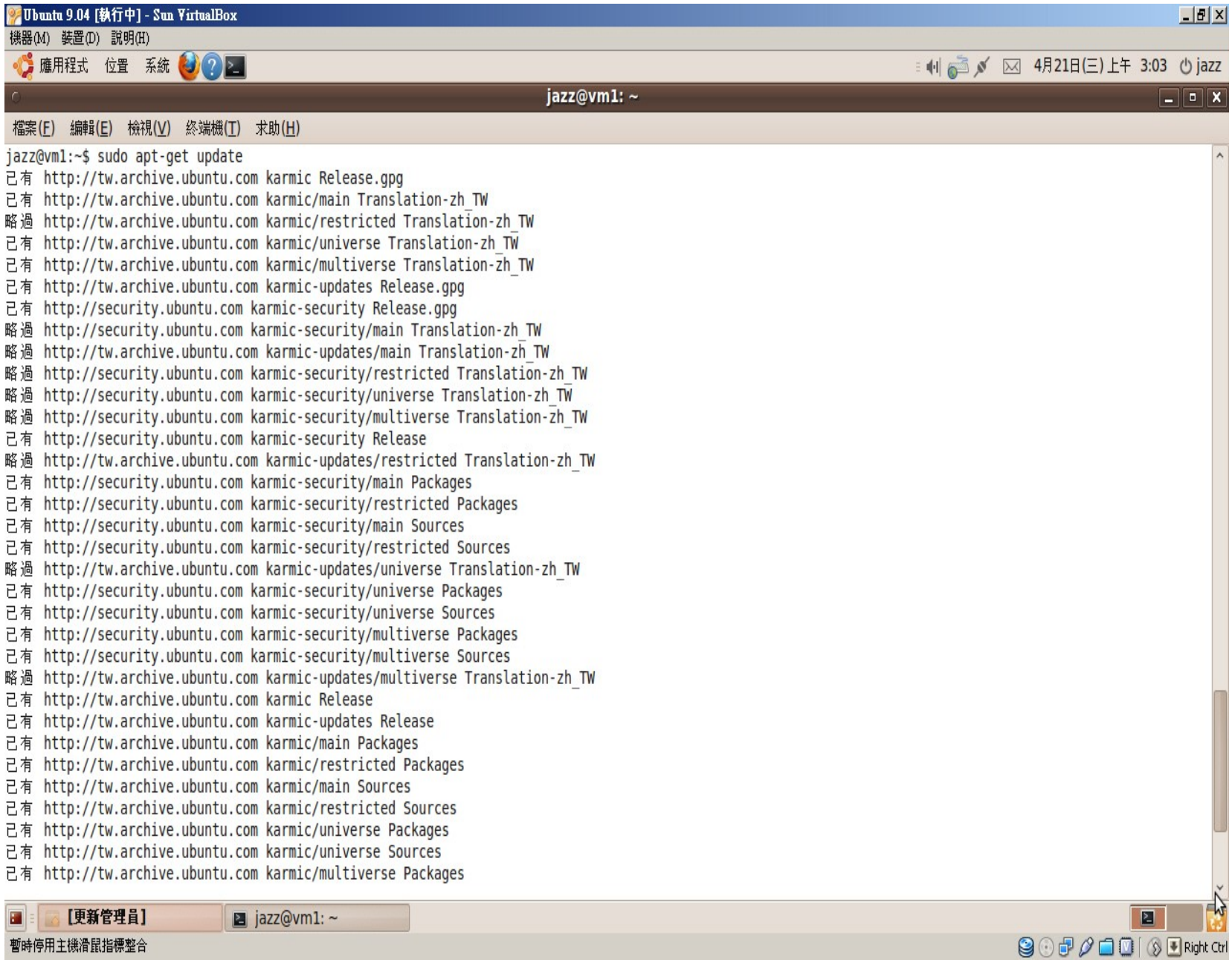
Linux 入門資源

時間可以存起來嘛？把作過的事情紀錄下來吧！
這部分的介紹，我 2001 年作過就請大家自己看了

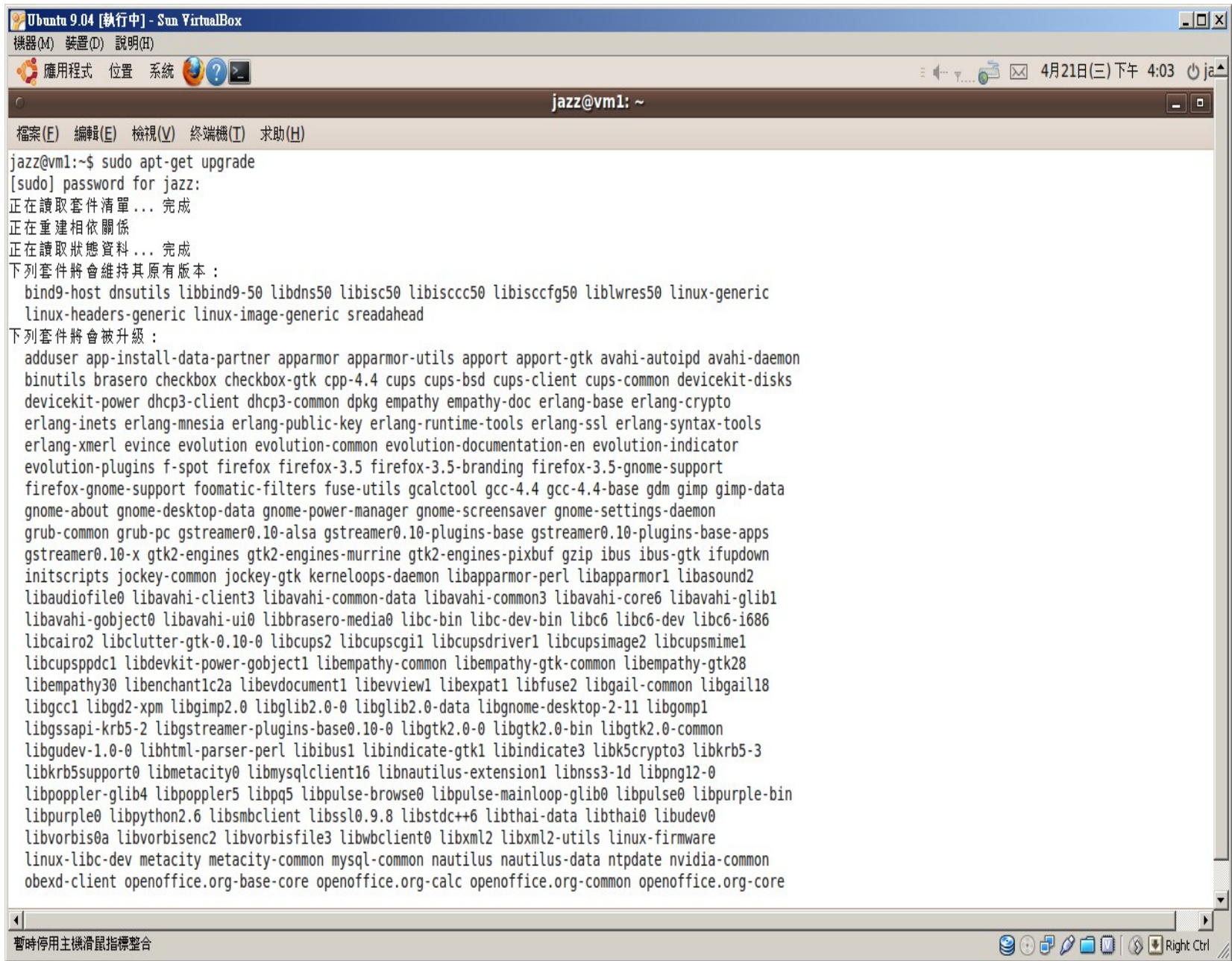
http://angkor.jazzbear.idv.tw/Linux_DivX/02.avi

- 如何找尋協助？
- Linux 基本指令操作
- Linux 管理者指令
- Linux 檔案系統簡介

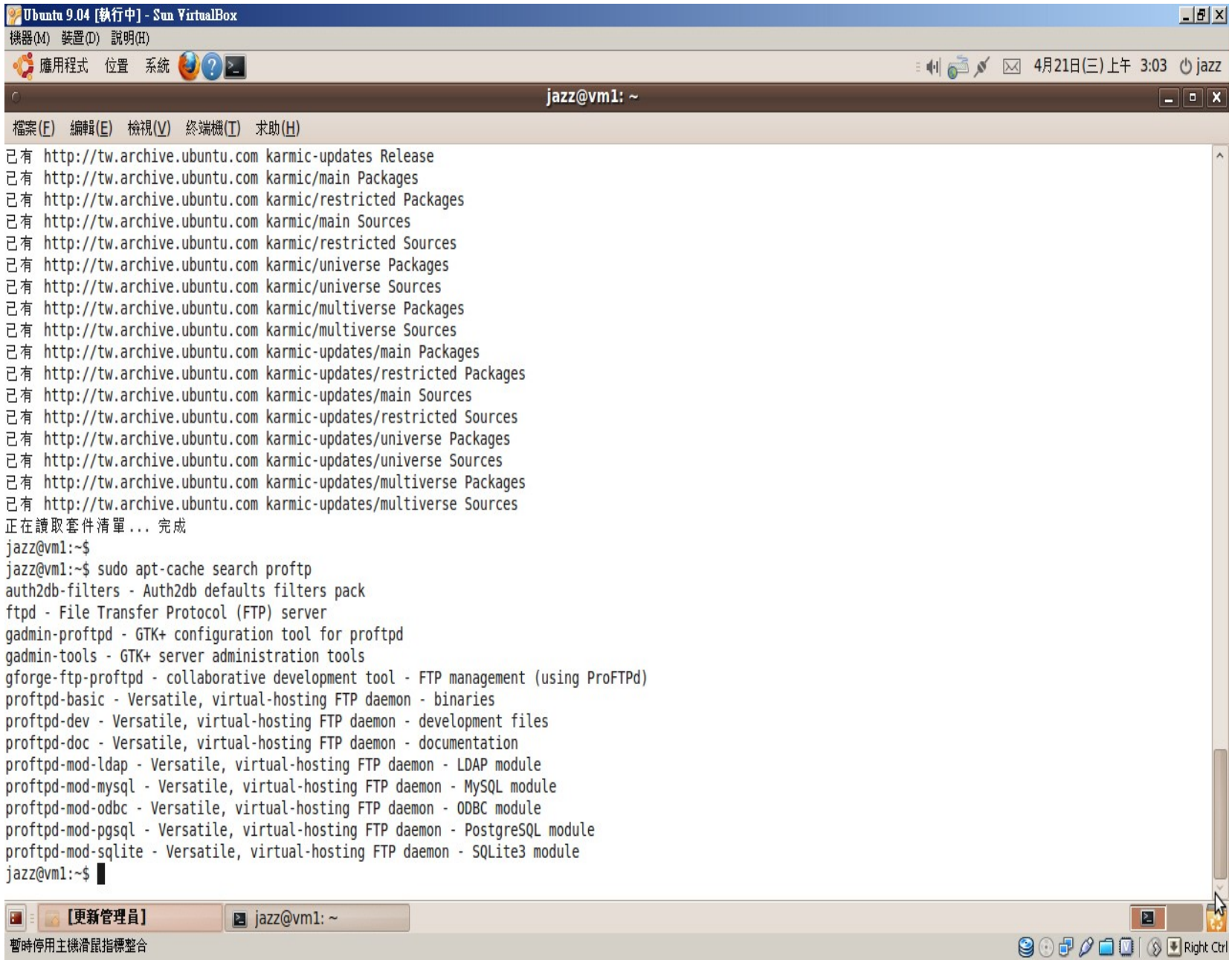
基本管理篇 [1] 用 apt-get update 更新套件



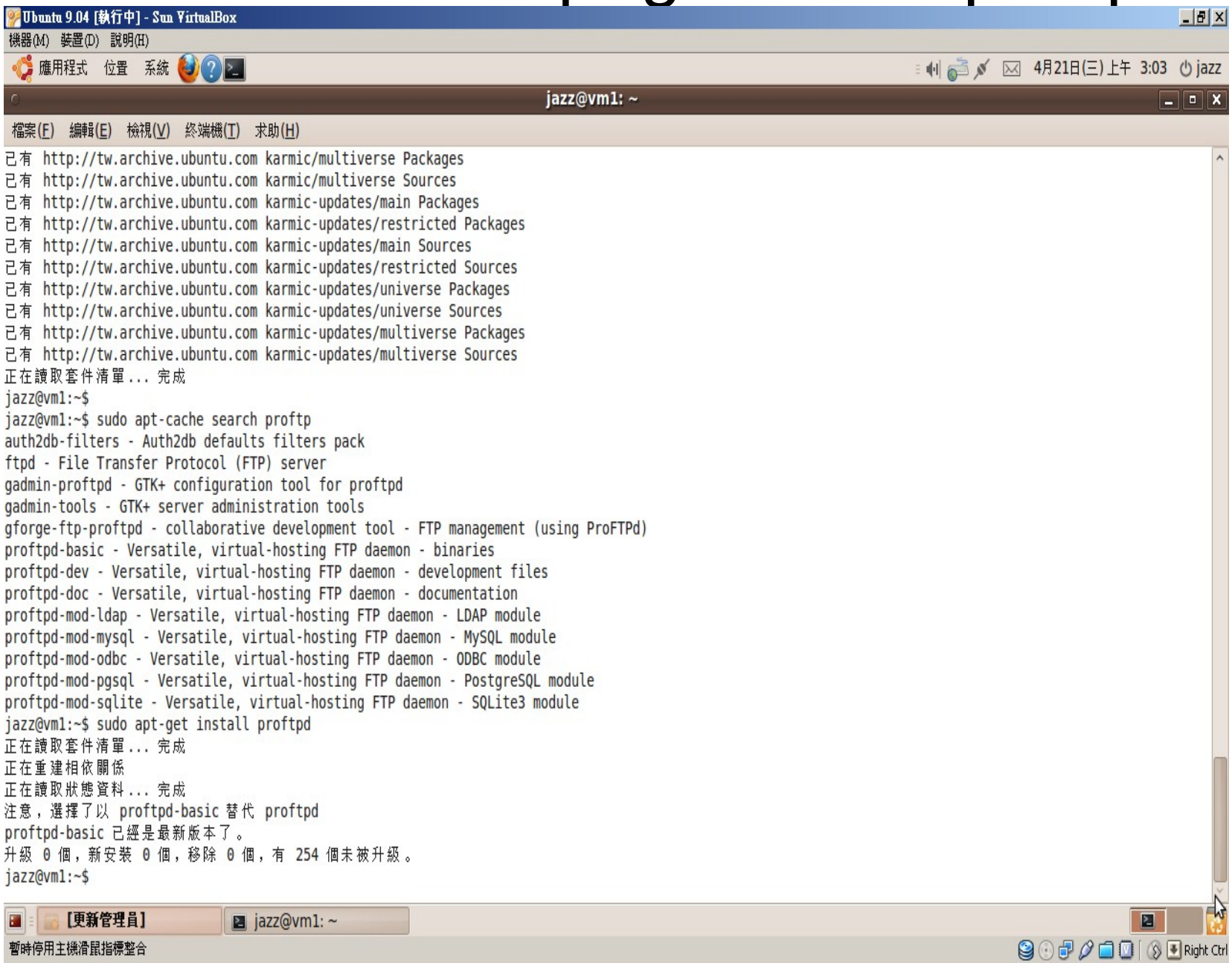
基本管理篇 [2] 用 apt-get upgrade 升級系統



基本管理篇 [3] 用 apt-cache search 搜尋



基本管理篇 [4] 用 apt-get install 安裝套件
Ex. apt-get install proftp



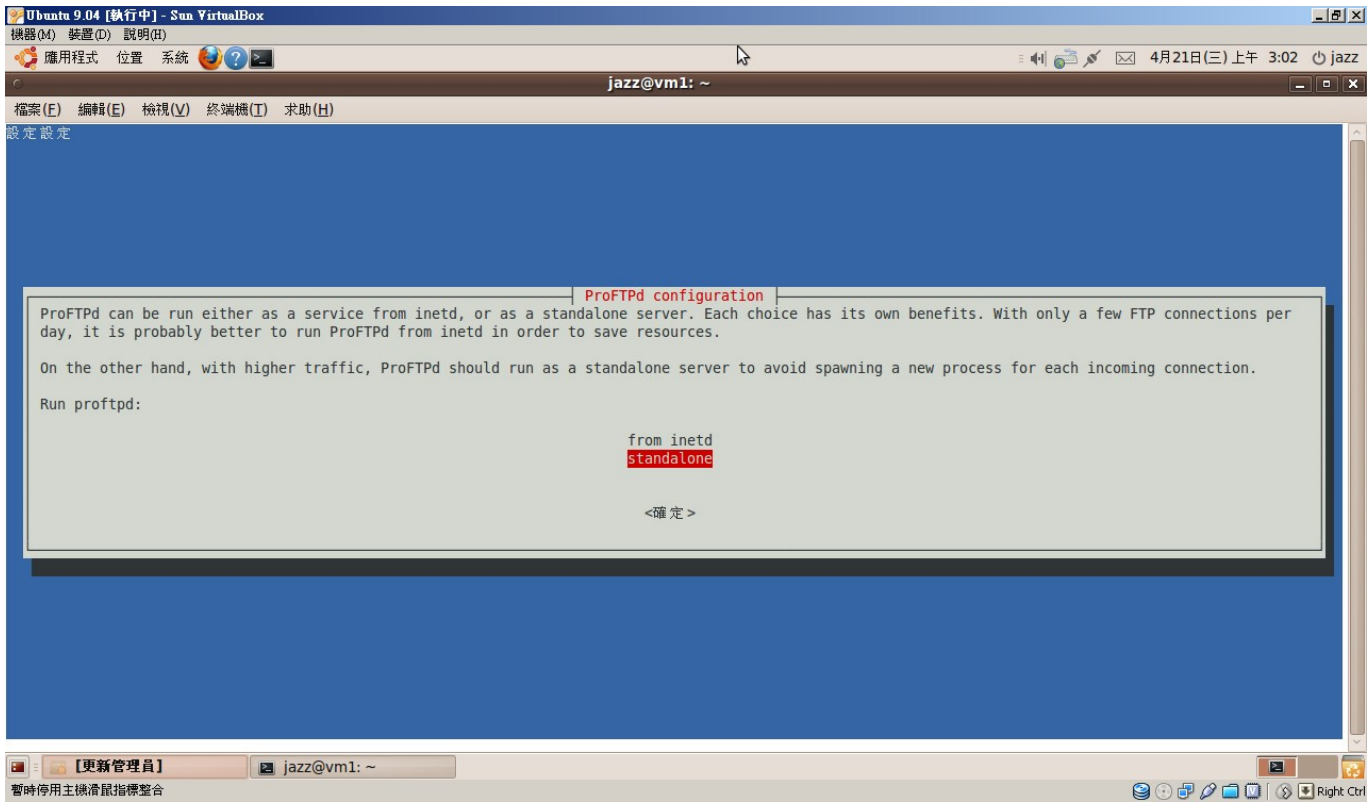
安裝 ProFTP

$ sudo apt-get install proftpd

$ sudo apt-get install ftp

- 測試

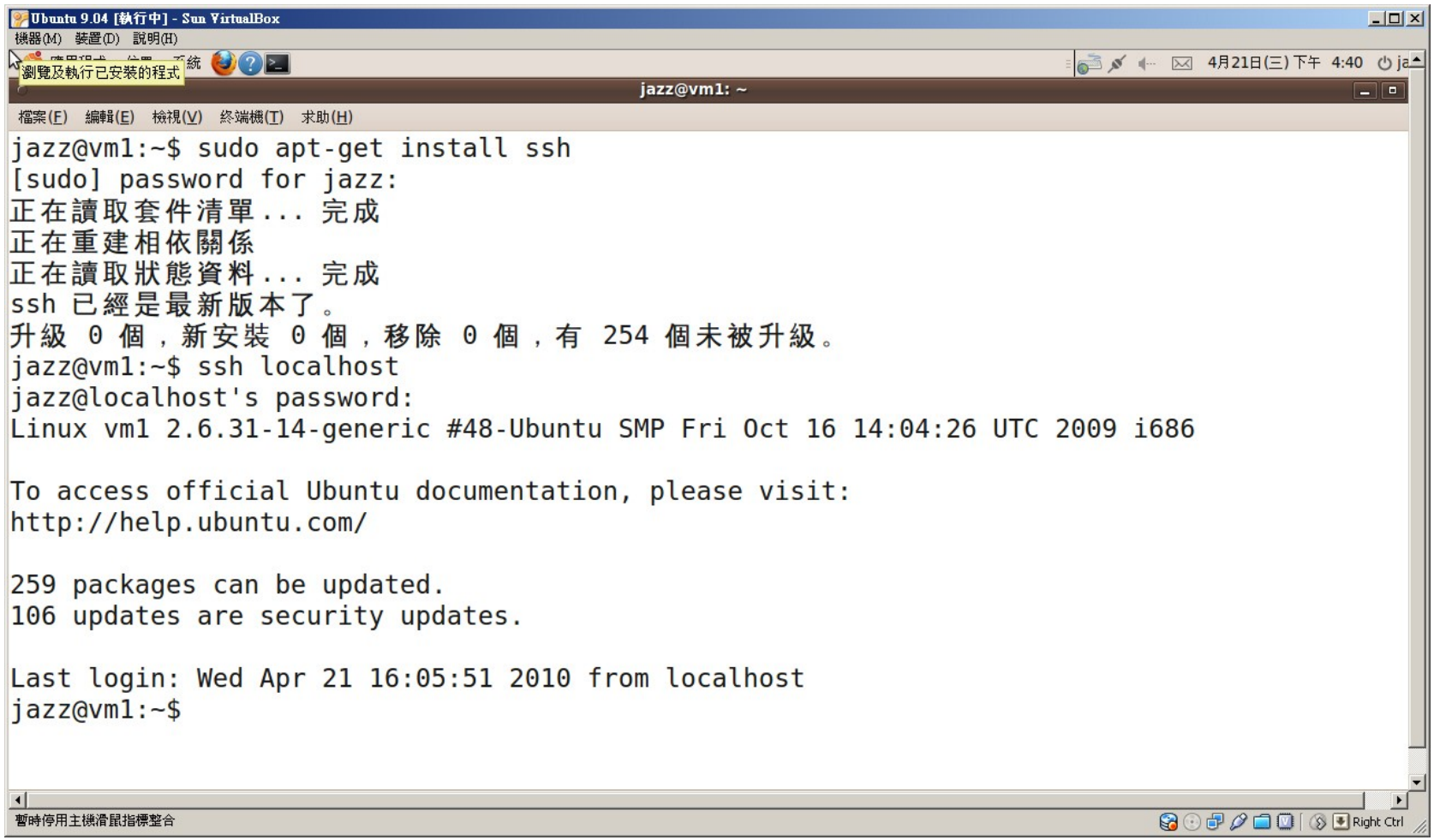
$ ftp localhost 21



安裝 SSH

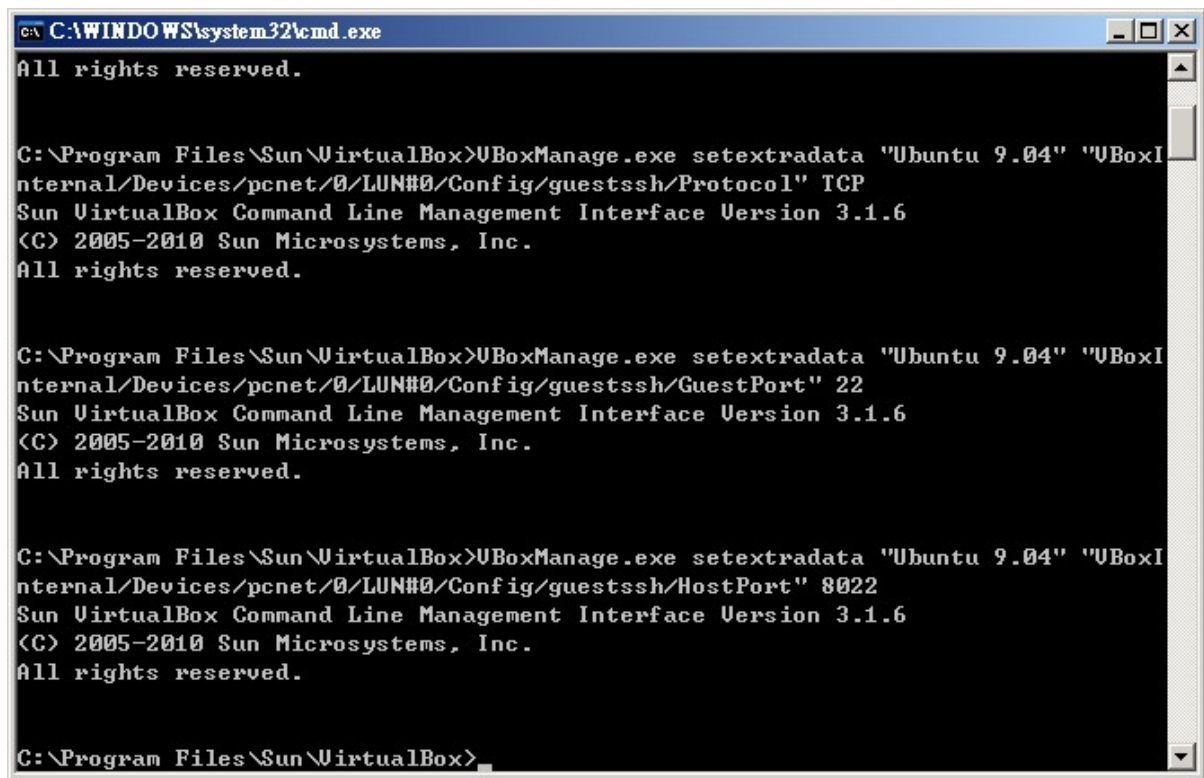
$ sudo apt-get install ssh

$ ssh localhost

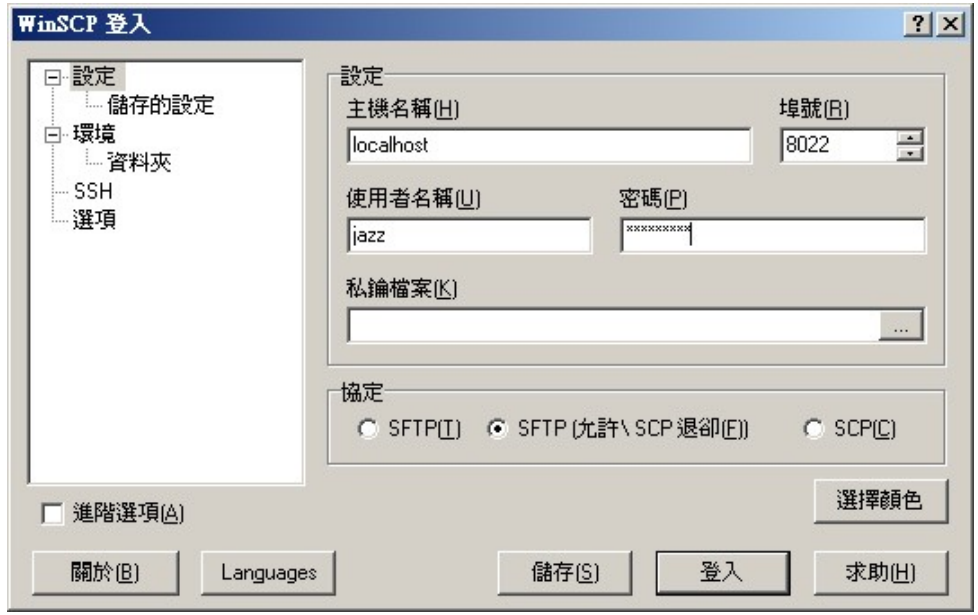


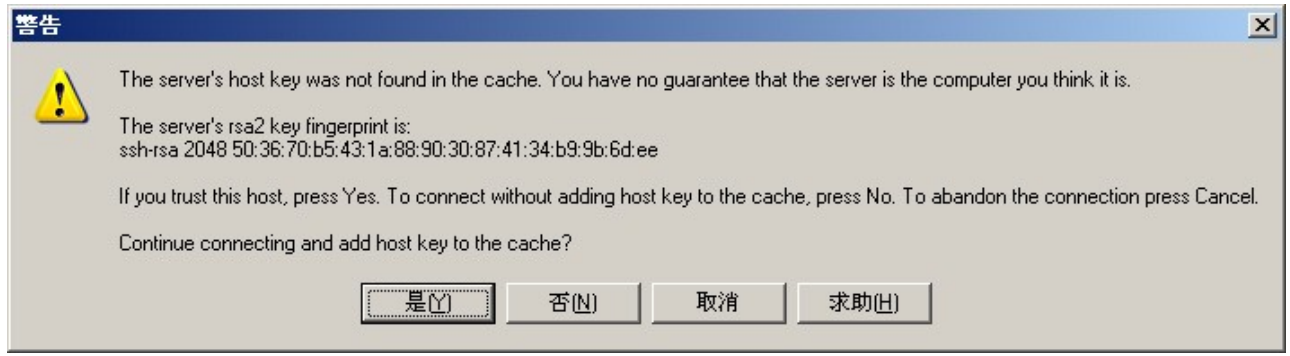
讓 Windows 可以連到虛擬機器

- VBoxManage.exe setextradata "Ubuntu 9.04" "VBoxInternal/Devices/pcnet/0/LUN#0/Config/guestssh/Protocol" TCP
- VBoxManage.exe setextradata "Ubuntu 9.04" "VBoxInternal/Devices/pcnet/0/LUN#0/Config/guestssh/GuestPort" 22
- VBoxManage.exe setextradata "Ubuntu 9.04" "VBoxInternal/Devices/pcnet/0/LUN#0/Config/guestssh/HostPort" 8022

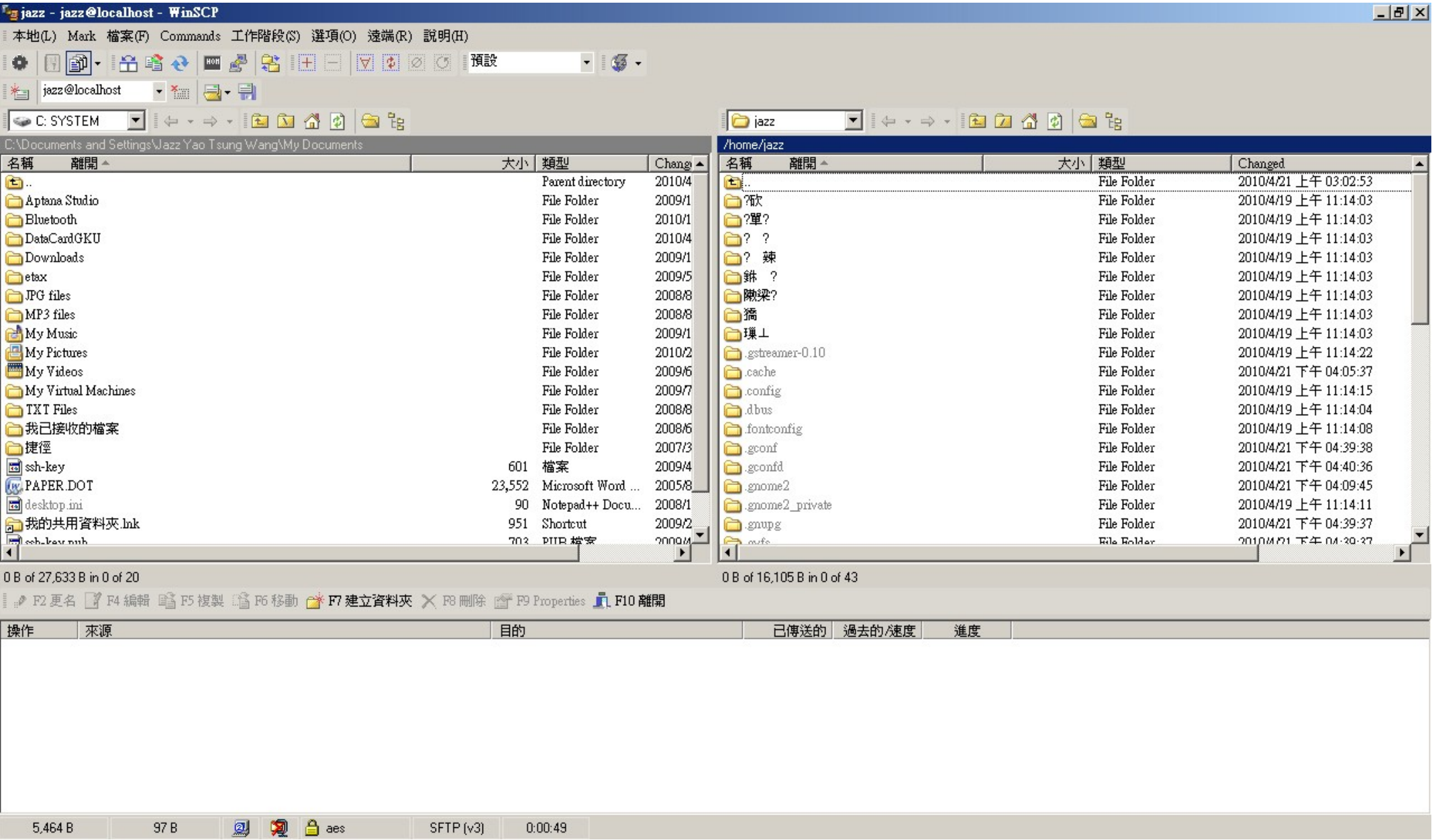


使用 WinSCP 上下傳檔案 (1)



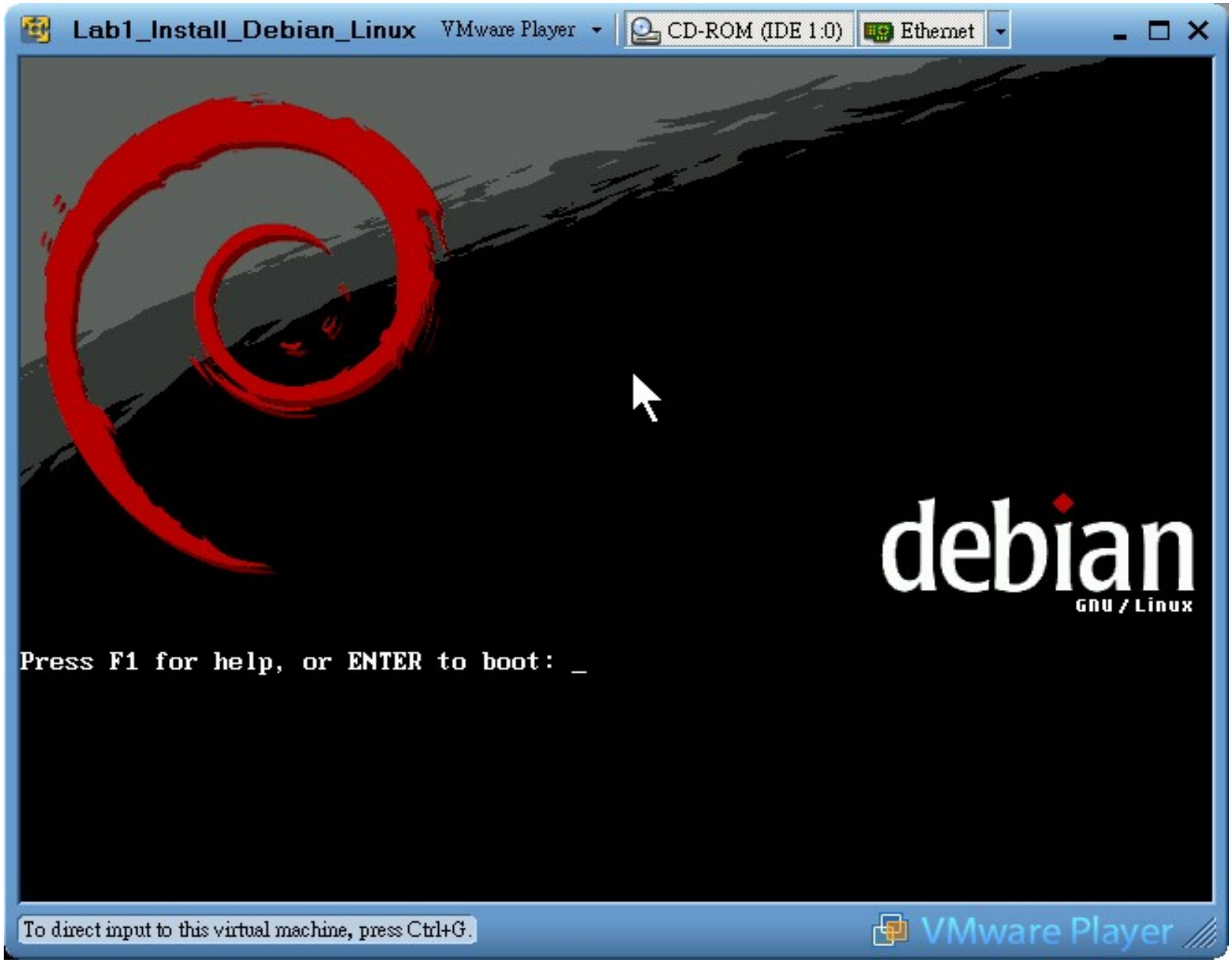


使用 WinSCP 上下傳檔案 (2)

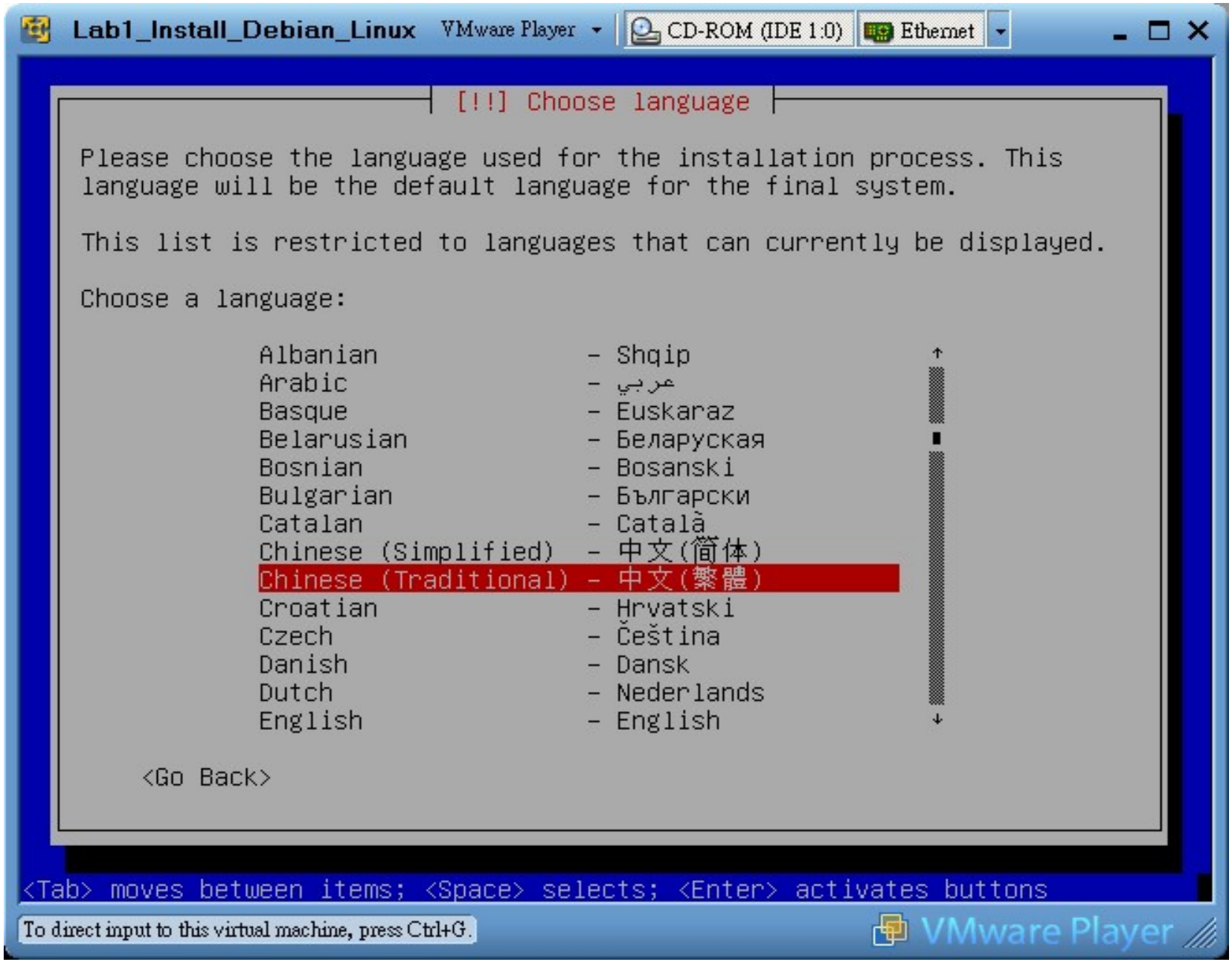


Appendex A :
安裝 Debian 4.0r3 GNU/Linux

開始模擬從光碟安裝


Lab1_Install_Debian_Linux
VMware Player
CD-ROM (IDE 1:0)
Ethernet
debian
GNU / Linux
Press F1 for help, or ENTER to boot: _
To direct input to this virtual machine, press Ctrl+G.
VMware Player

選擇中文(繁體)語系


Lab1_Install_Debian_Linux
VMware Player
CD-ROM (IDE 1:0)
Ethernet
[!!] Choose language
Please choose the language used for the installation process. This language will be the default language for the final system.
This list is restricted to languages that can currently be displayed.
Choose a language:
Albanian - Shqip
Arabic - عربي
Basque - Euskaraz
Belarusian - Беларуская
Bosnian - Bosanski
Bulgarian - Български
Catalan - Català
Chinese (Simplified) - 中文(简体)
Chinese (Traditional) - 中文(繁體)
Croatian - Hrvatski
Czech - Čeština
Danish - Dansk
Dutch - Nederlands
English - English
<Go Back>
<Tab> moves between items; <Space> selects; <Enter> activates buttons
To direct input to this virtual machine, press Ctrl+G.
VMware Player

選擇『台灣』

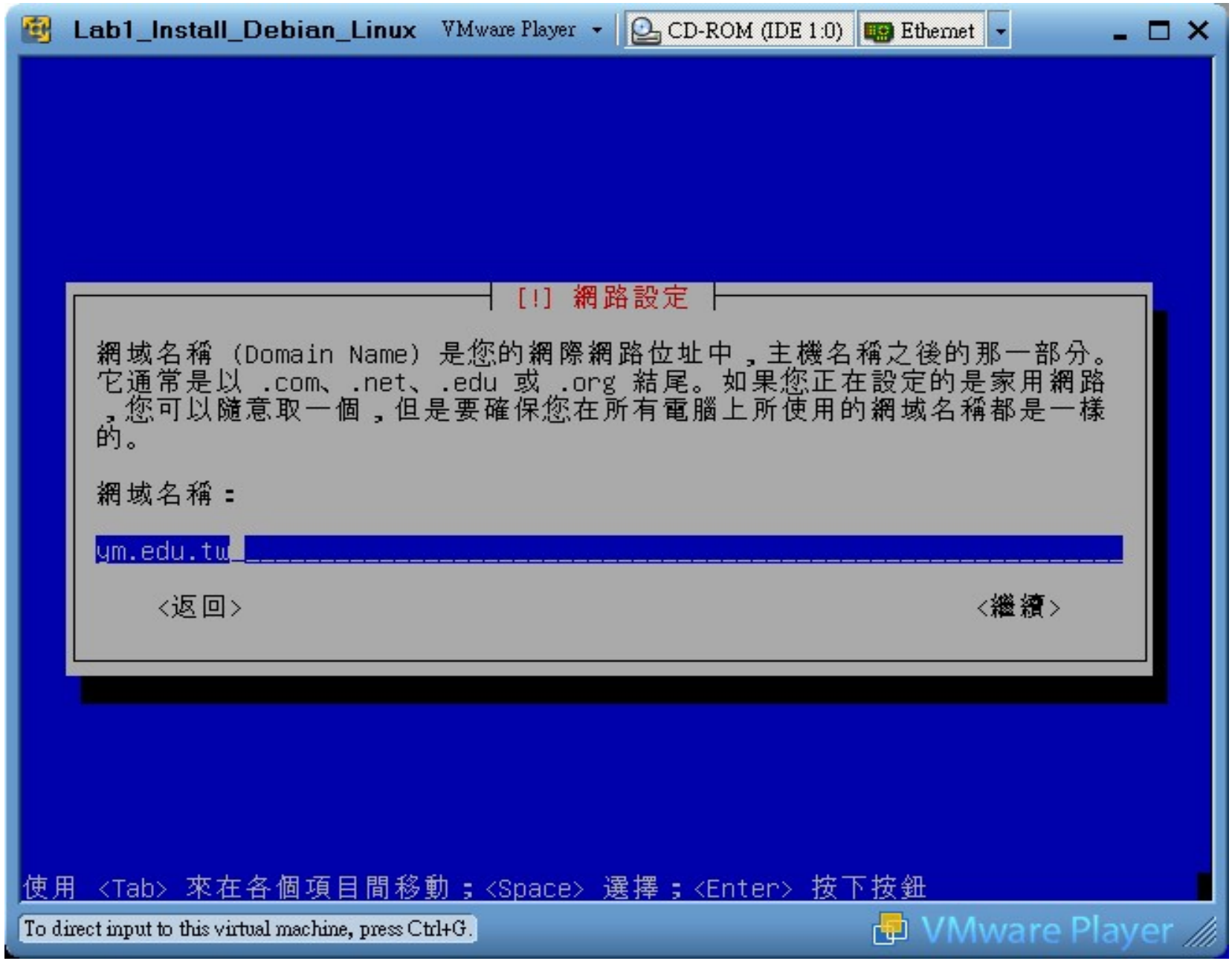

Lab1_Install_Debian_Linux
VMware Player
CD-ROM (IDE 1:0)
Ethernet
[!!] Choose language
根據您所使用的語言，您可能位於以下的國家或地區之一。
請選擇一個國家、領域或地區：
中國
台灣
新加坡
香港
其它
<返回>
使用 <Tab> 來在各個項目間移動；<Space> 選擇；<Enter> 按下按鈕
To direct input to this virtual machine, press Ctrl+G.
VMware Player

選擇『美語鍵盤』


Lab1_Install_Debian_Linux
VMware Player
CD-ROM (IDE 1:0)
Ethernet
[!] 選擇鍵盤排列方式
所使用的鍵盤格式：
愛沙尼亞語
拉丁美洲語
拉脫維亞語 拉丁(4)
挪威語
捷克語
斯洛伐克語
斯洛凡尼亞語
日語 (106 鍵)
比利時語
法語 (有 euro)
波蘭語
烏克蘭語
瑞典語
瑞士德語
瑞士法語
白俄羅斯語
立陶宛語
羅馬尼亞語
美語
<返回>
使用 <Tab> 來在各個項目間移動；<Space> 選擇；<Enter> 按下按鈕
To direct input to this virtual machine, press Ctrl+G.
VMware Player

輸入『主機名稱』 = drbl-srv

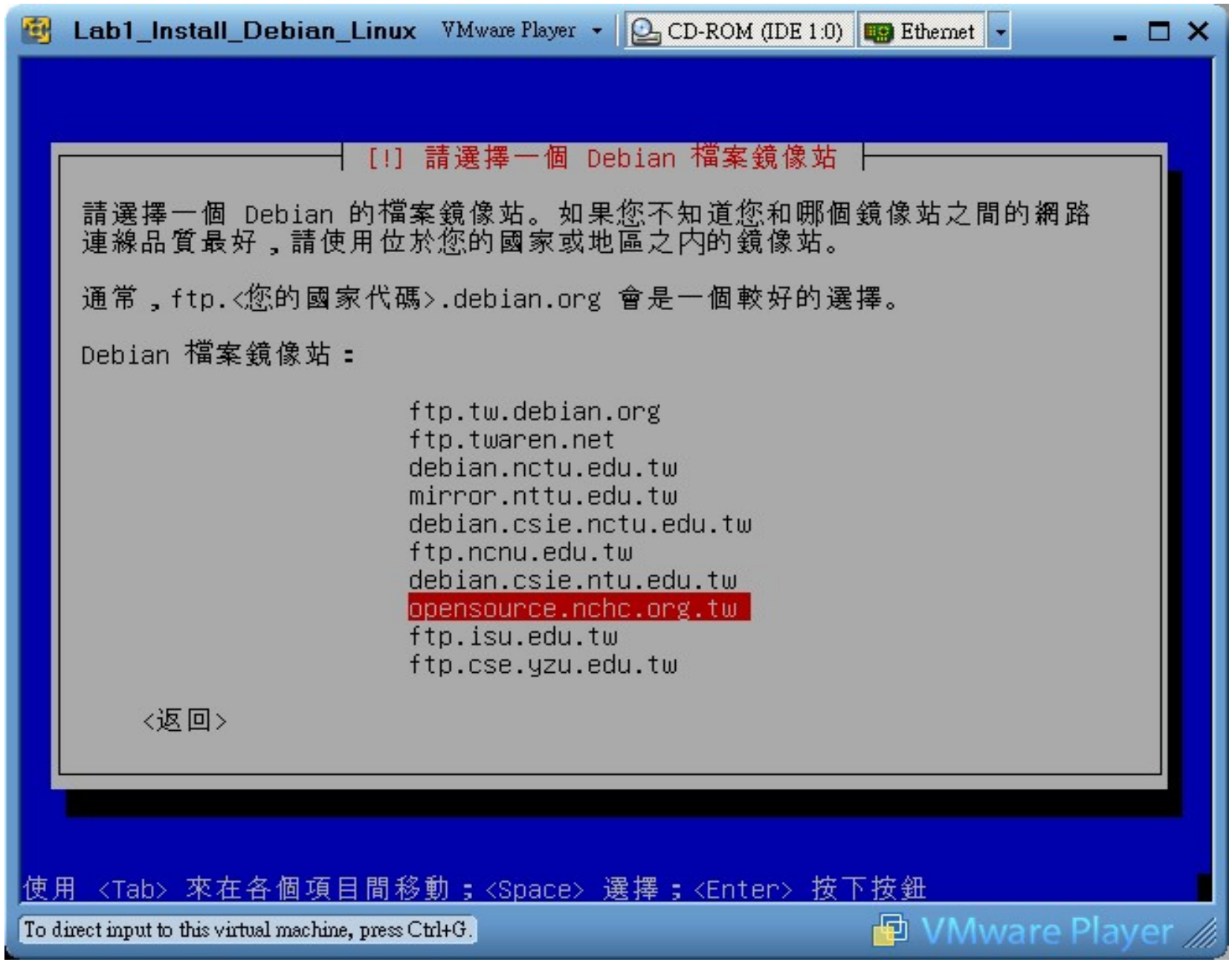
輸入『網域名稱』 = ym.edu.tw



選擇『 Debian 檔案映象站』(1) 台灣區

選擇『 Debian 檔案映象站』(2)

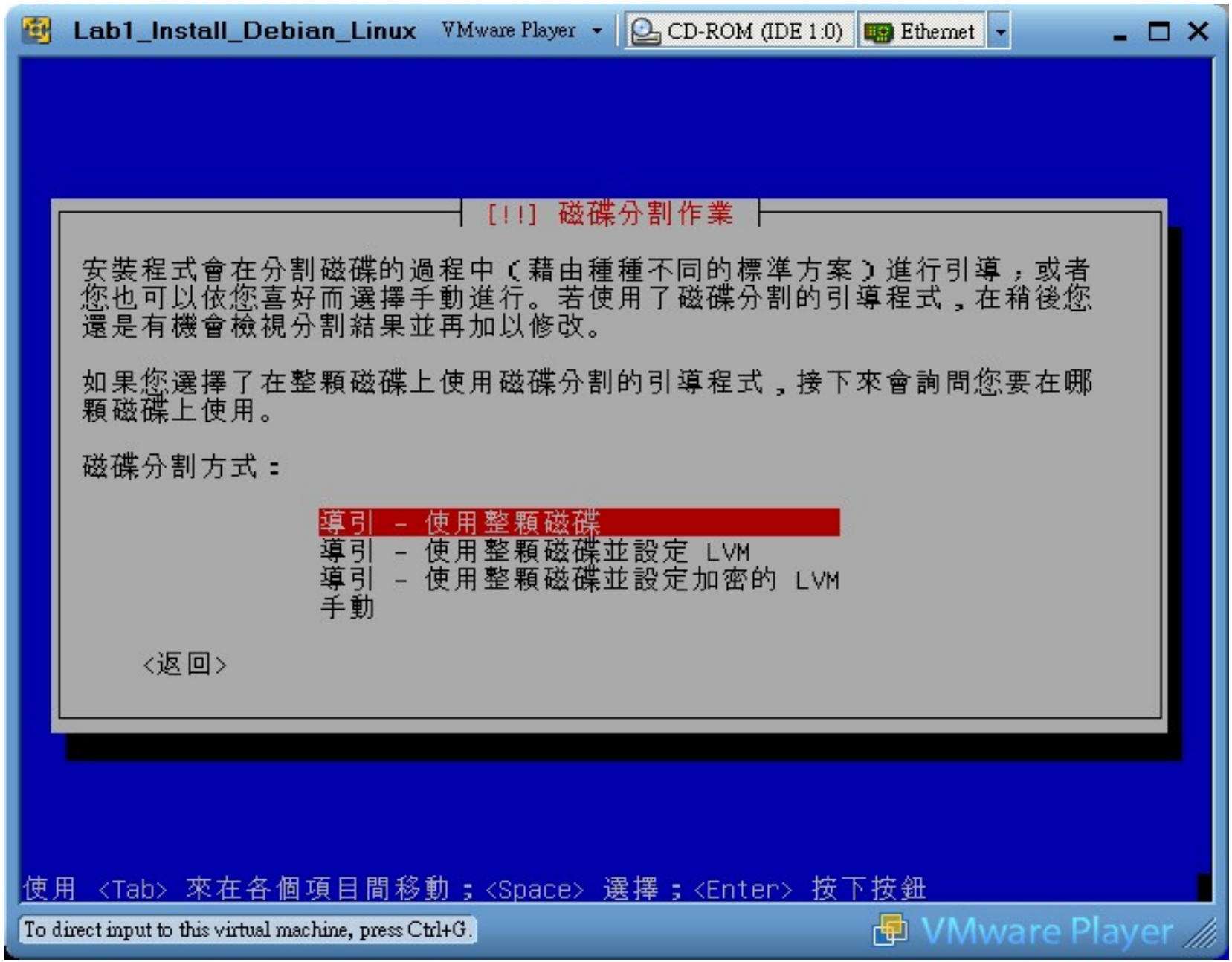
推薦 ftp.twaren.net 或 free.nchc.org.tw



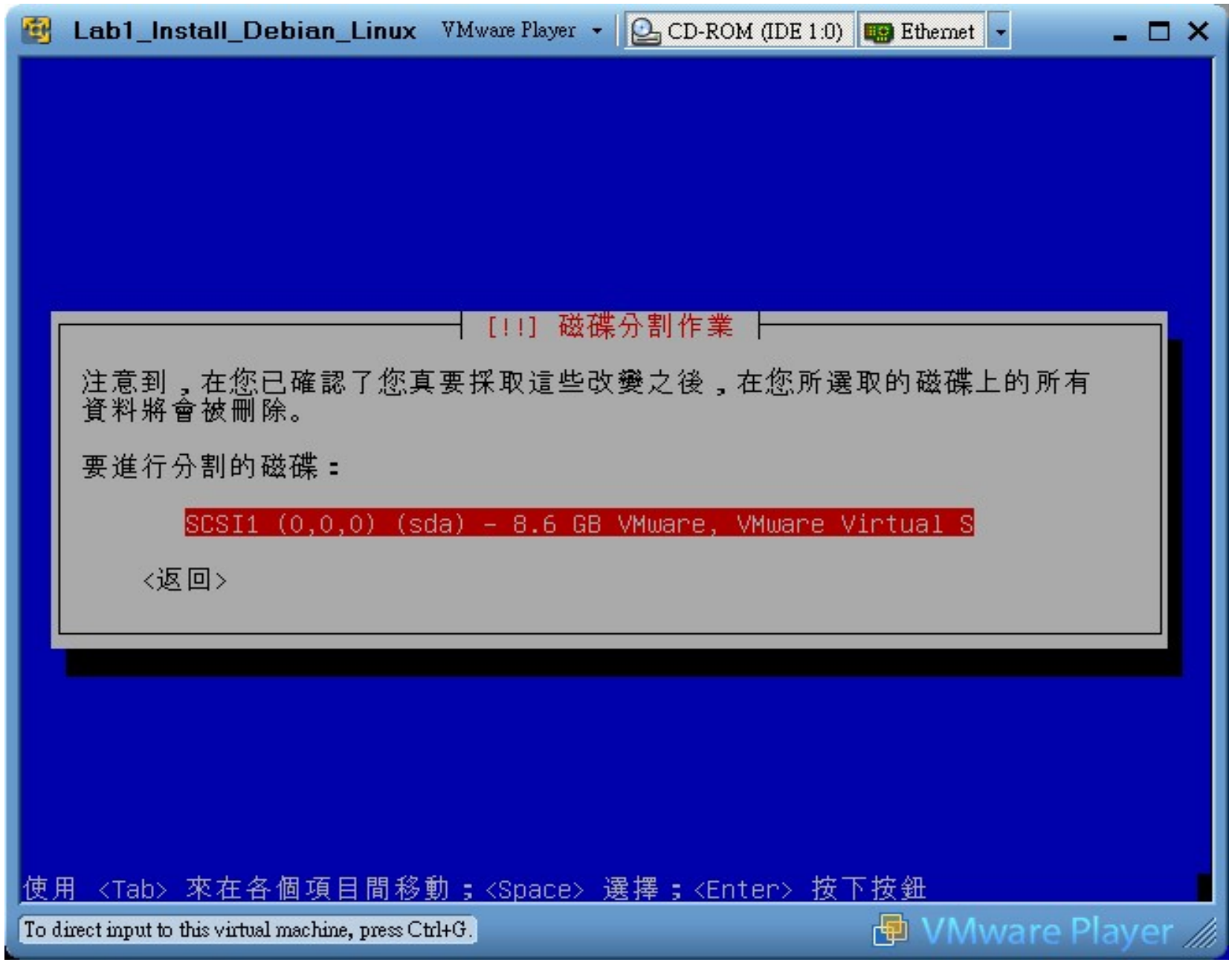
選擇『 Debian 檔案映象站』(3)

通常不需要輸入 Proxy ，除非網路環境有限制

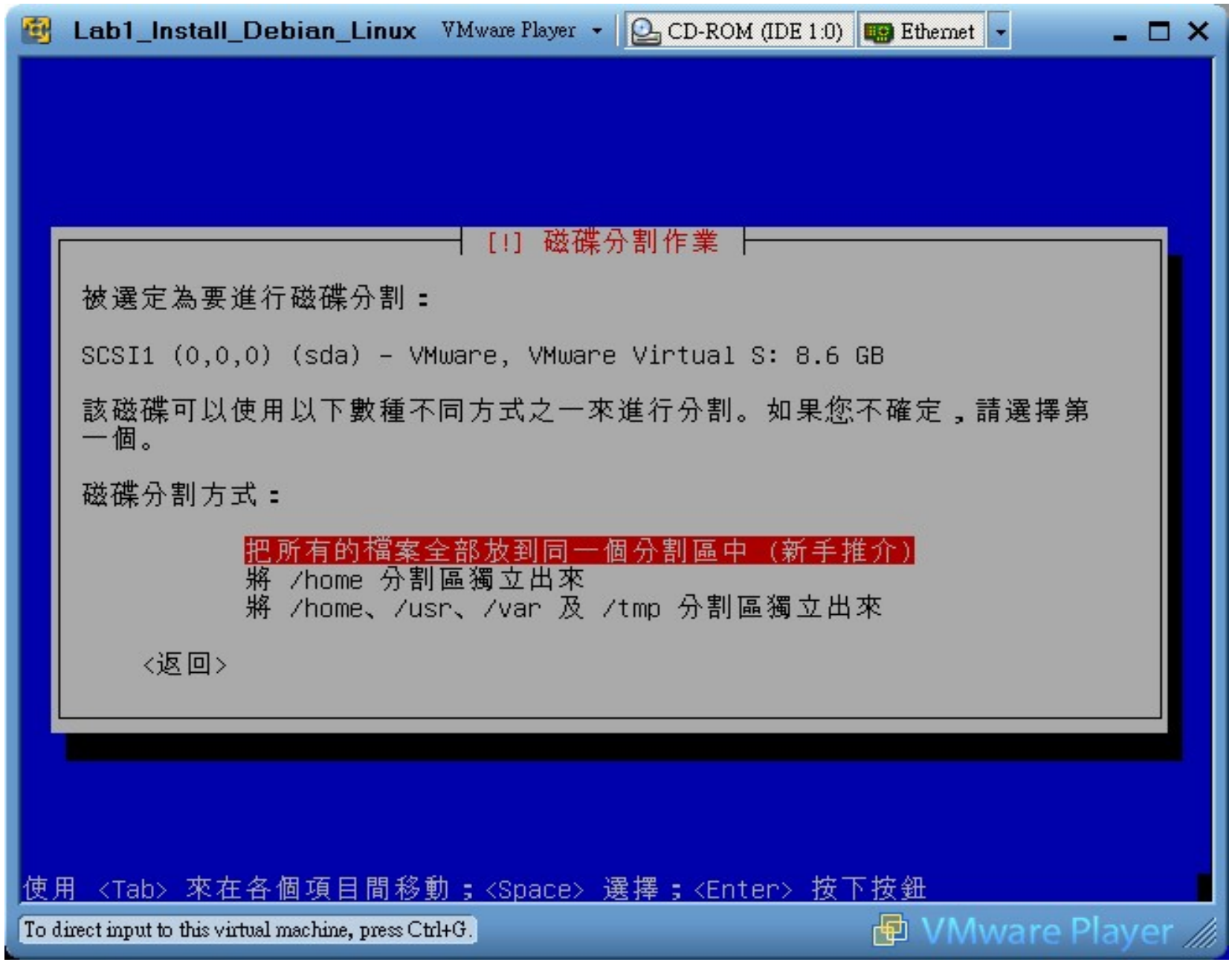
選擇『磁碟分割方式』
初學者建議採用第一種『引導 - 整顆硬碟』



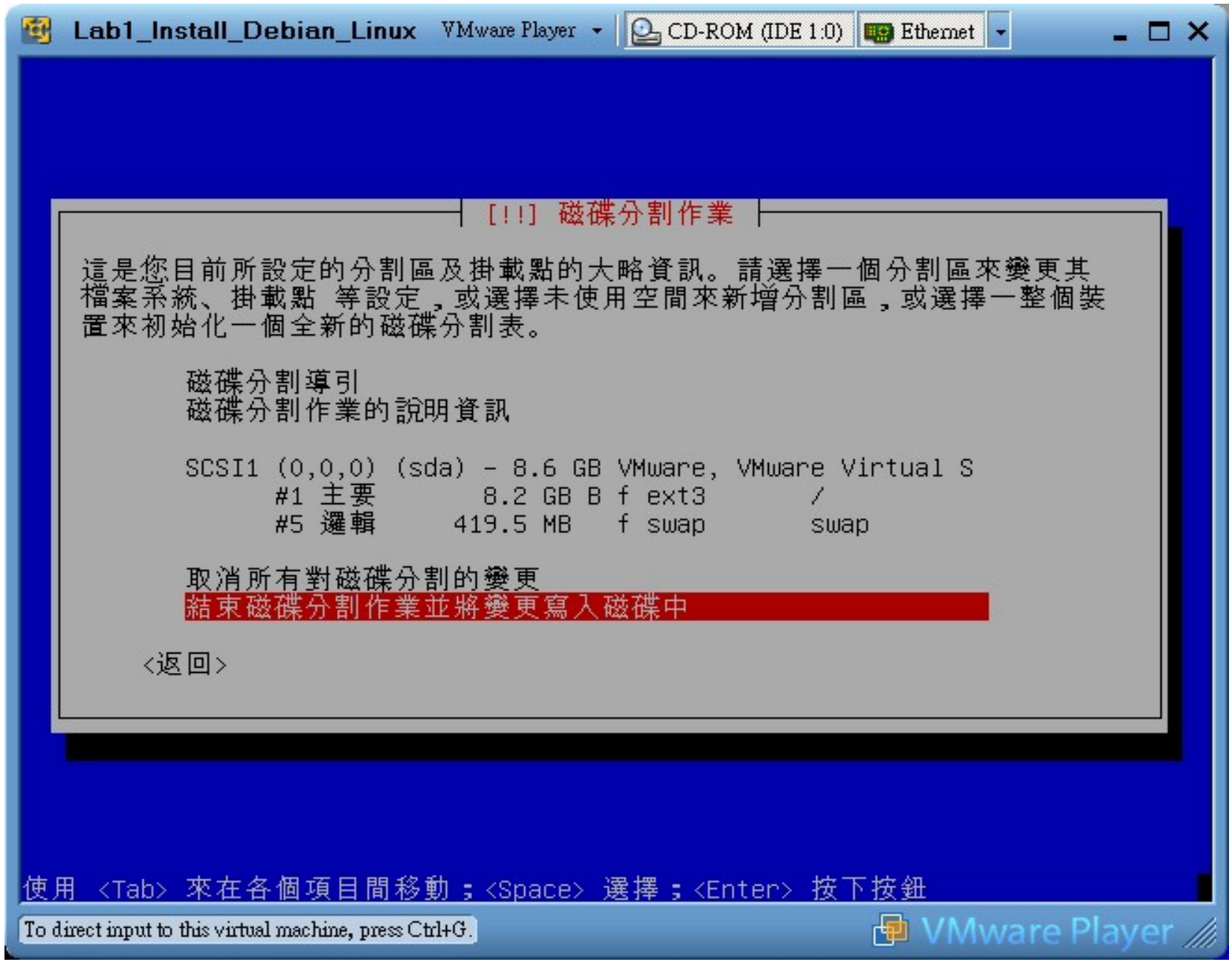
選擇『欲分割磁碟』
目前只有虛擬一顆硬碟，若實機有兩顆請小心！



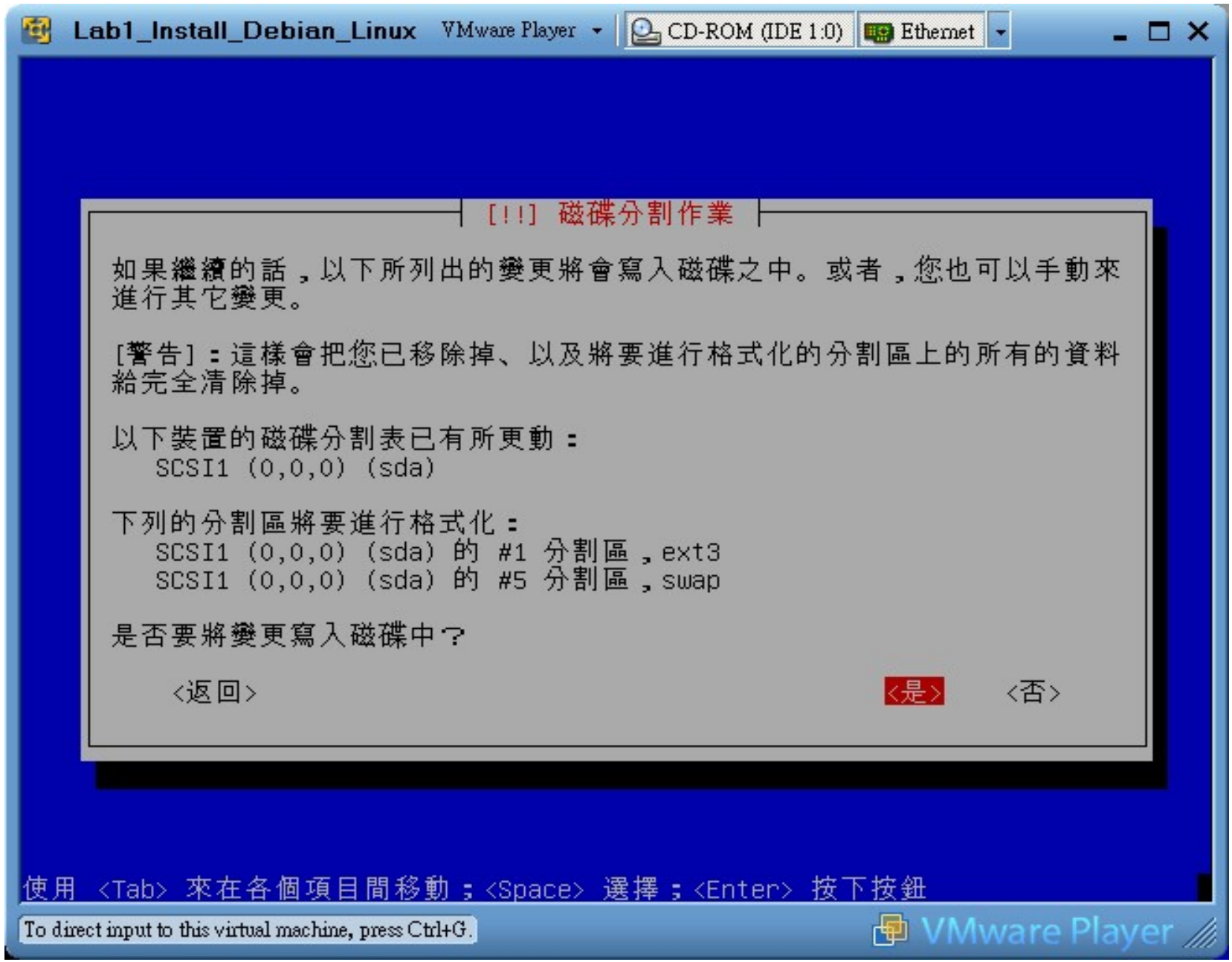
練習時可用同一個分割區，
服務用系統請至少獨立出 /home 分割區



導引模式會切割兩個磁區，一個根目錄 (root)
一個 swap (拿硬碟當記憶體空間用)



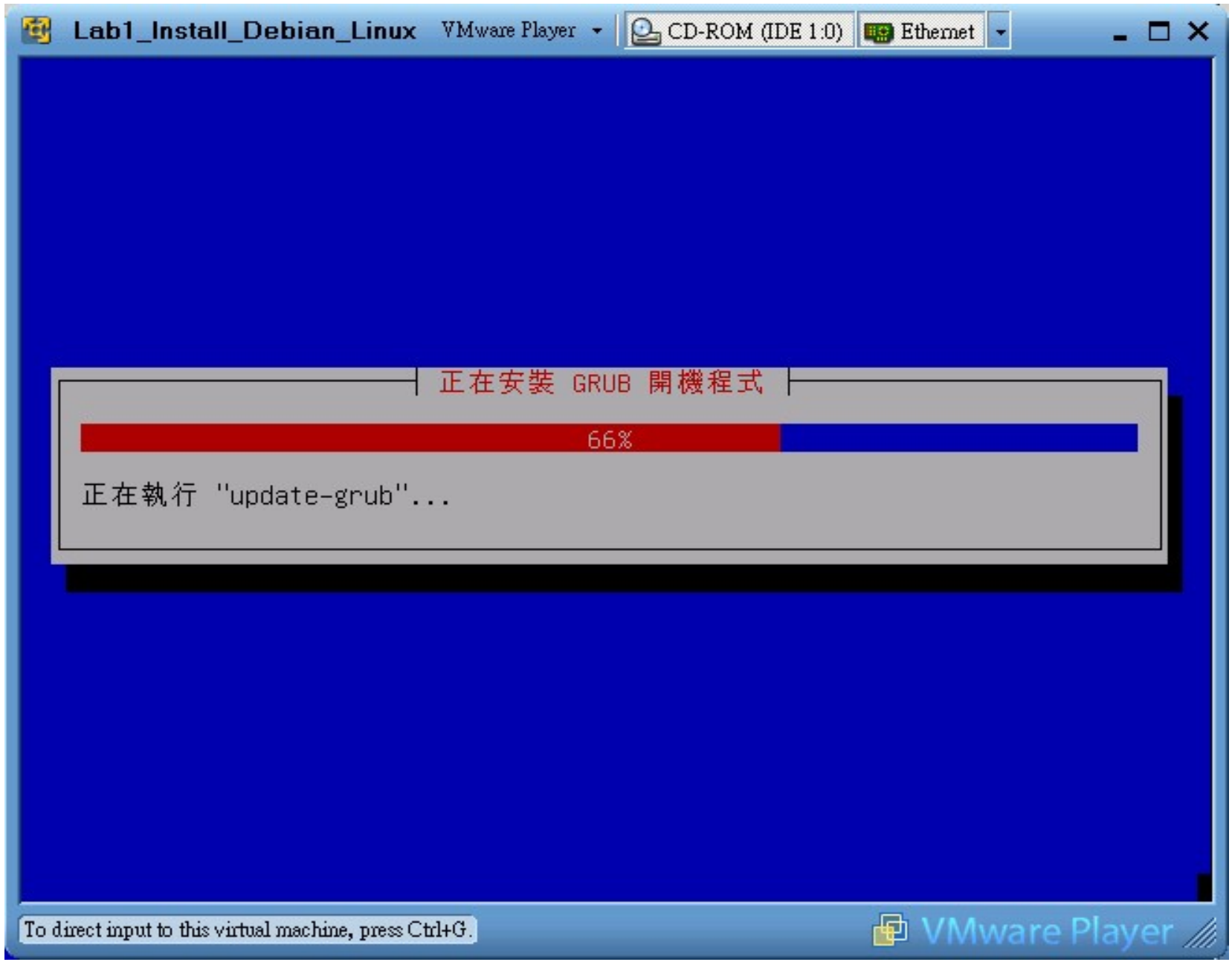
確認將目前的磁區分割表寫入硬碟
請用 TAB 選擇 < 是 >



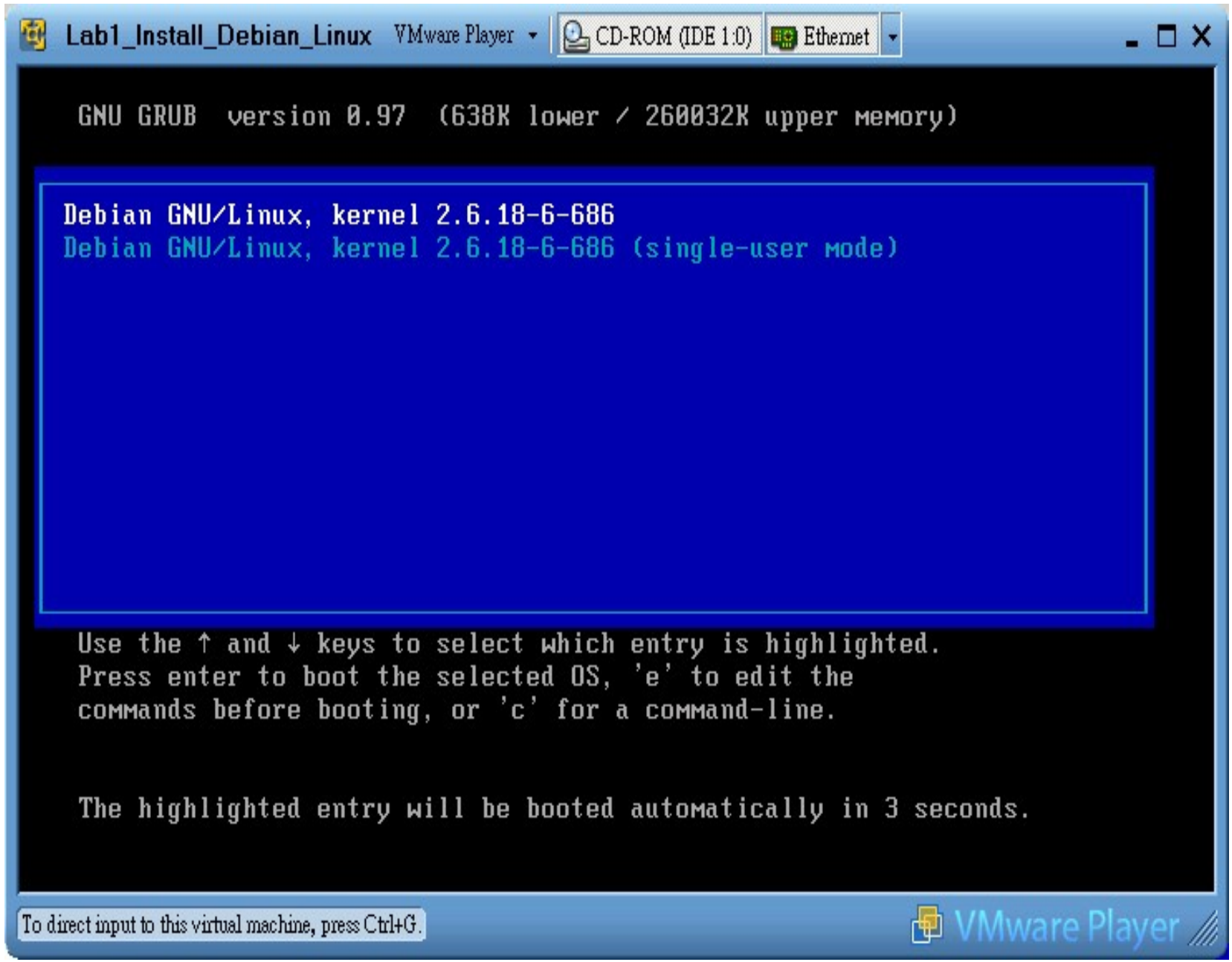
設定最高權限 root 管理者密碼

建立自己的帳號，習慣上不用 root 登入

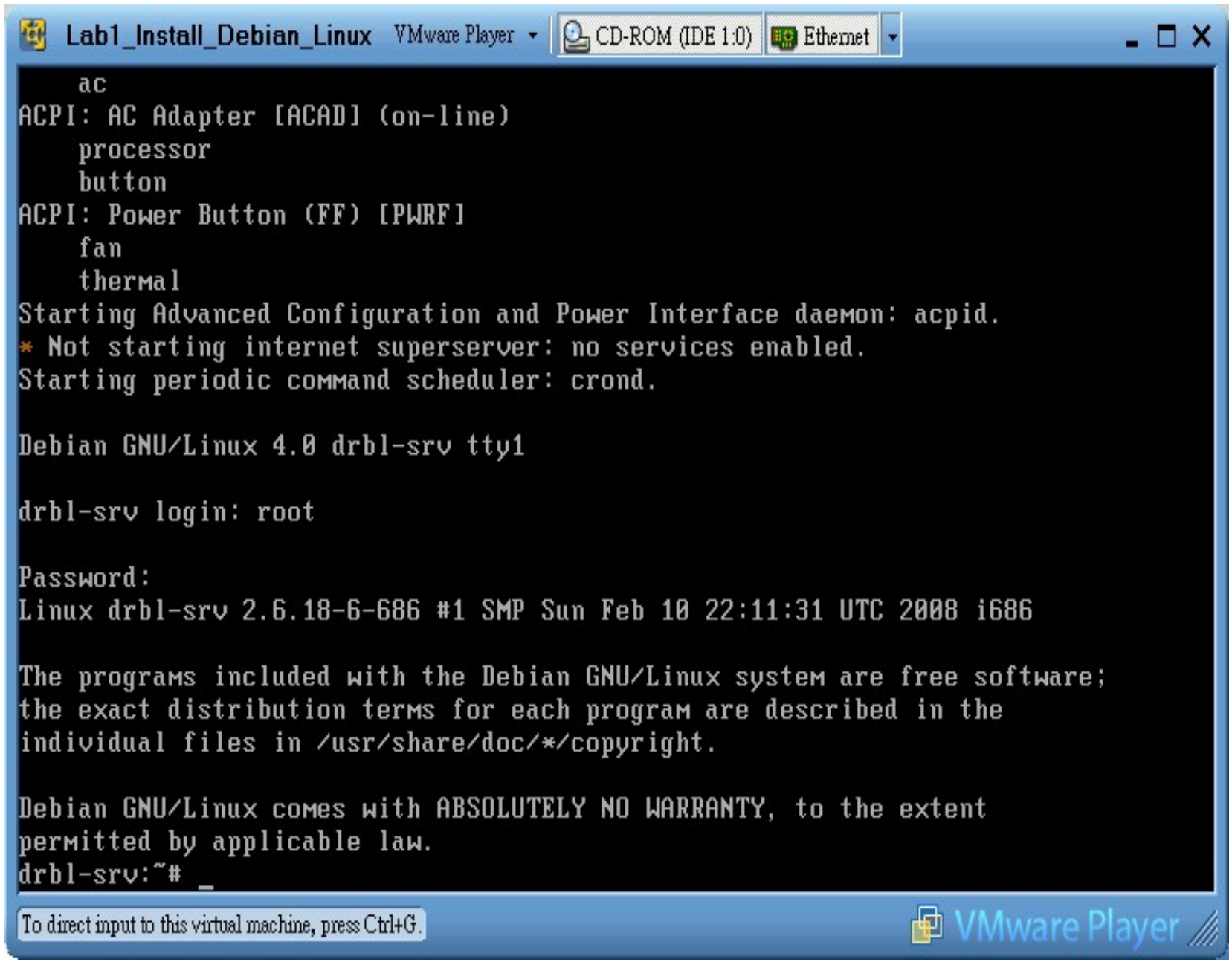
確認安裝 Grub Boot Loader



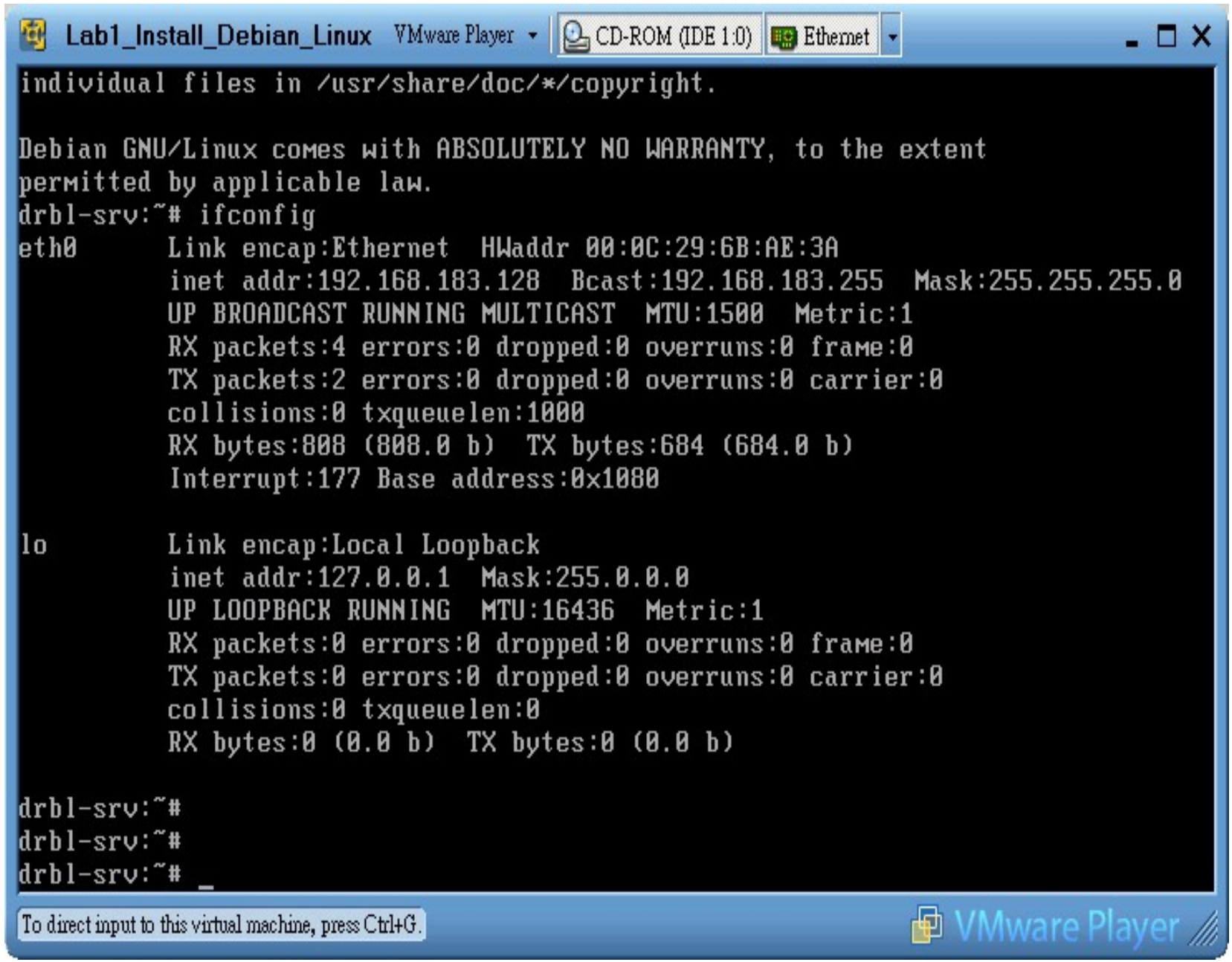
第一次系統開機


Lab1_Install_Debian_Linux
VMware Player
CD-ROM (IDE 1:0)
Ethernet
GNU GRUB version 0.97 (638K lower / 260032K upper memory)
Debian GNU/Linux, kernel 2.6.18-6-686
Debian GNU/Linux, kernel 2.6.18-6-686 (single-user mode)
Use the ↑ and ↓ keys to select which entry is highlighted.
Press enter to boot the selected OS, 'e' to edit the
commands before booting, or 'c' for a command-line.
The highlighted entry will be booted automatically in 3 seconds.
To direct input to this virtual machine, press Ctrl+G.
VMware Player

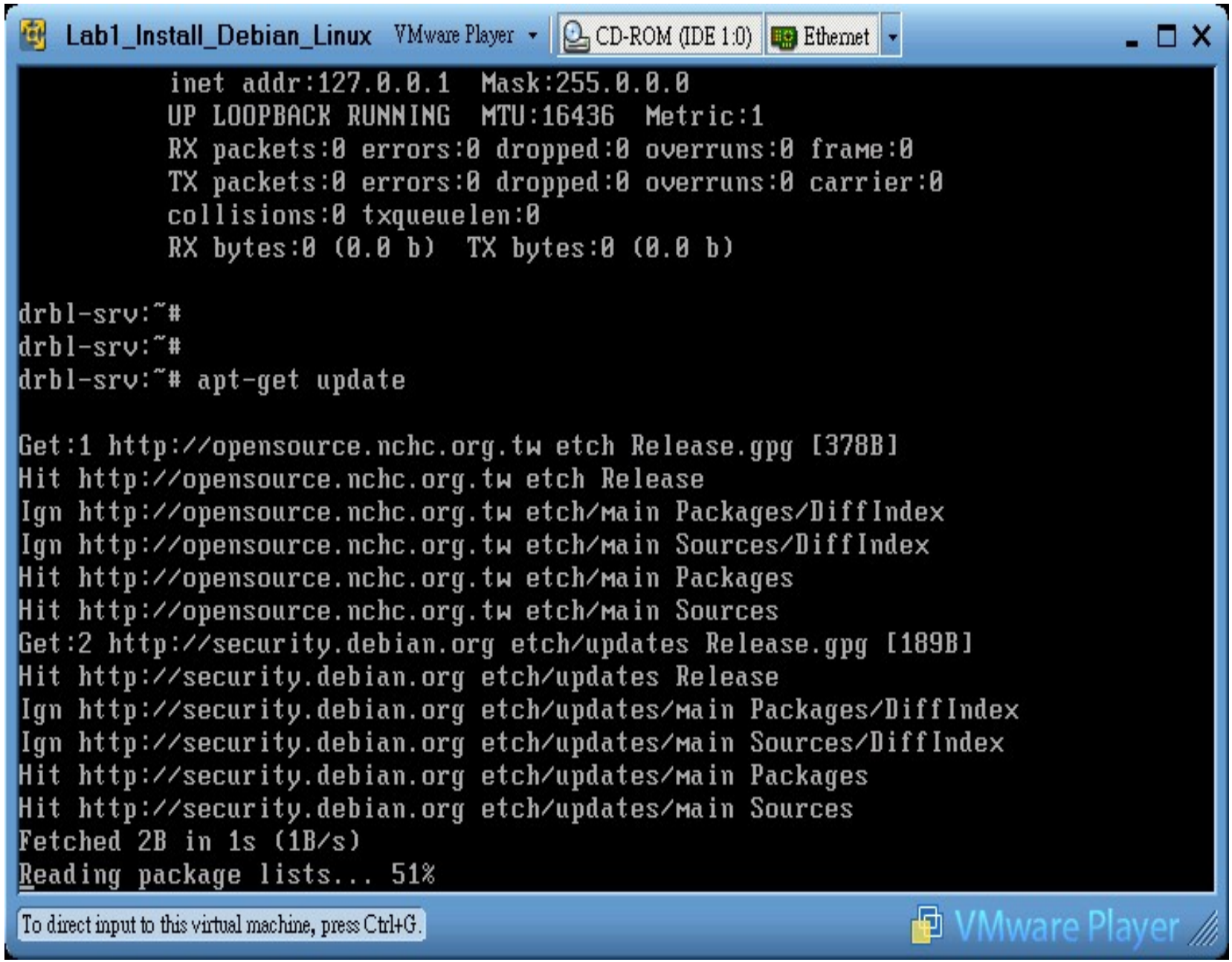
首次可使用 root 登入，
日後建議用一般身分登入



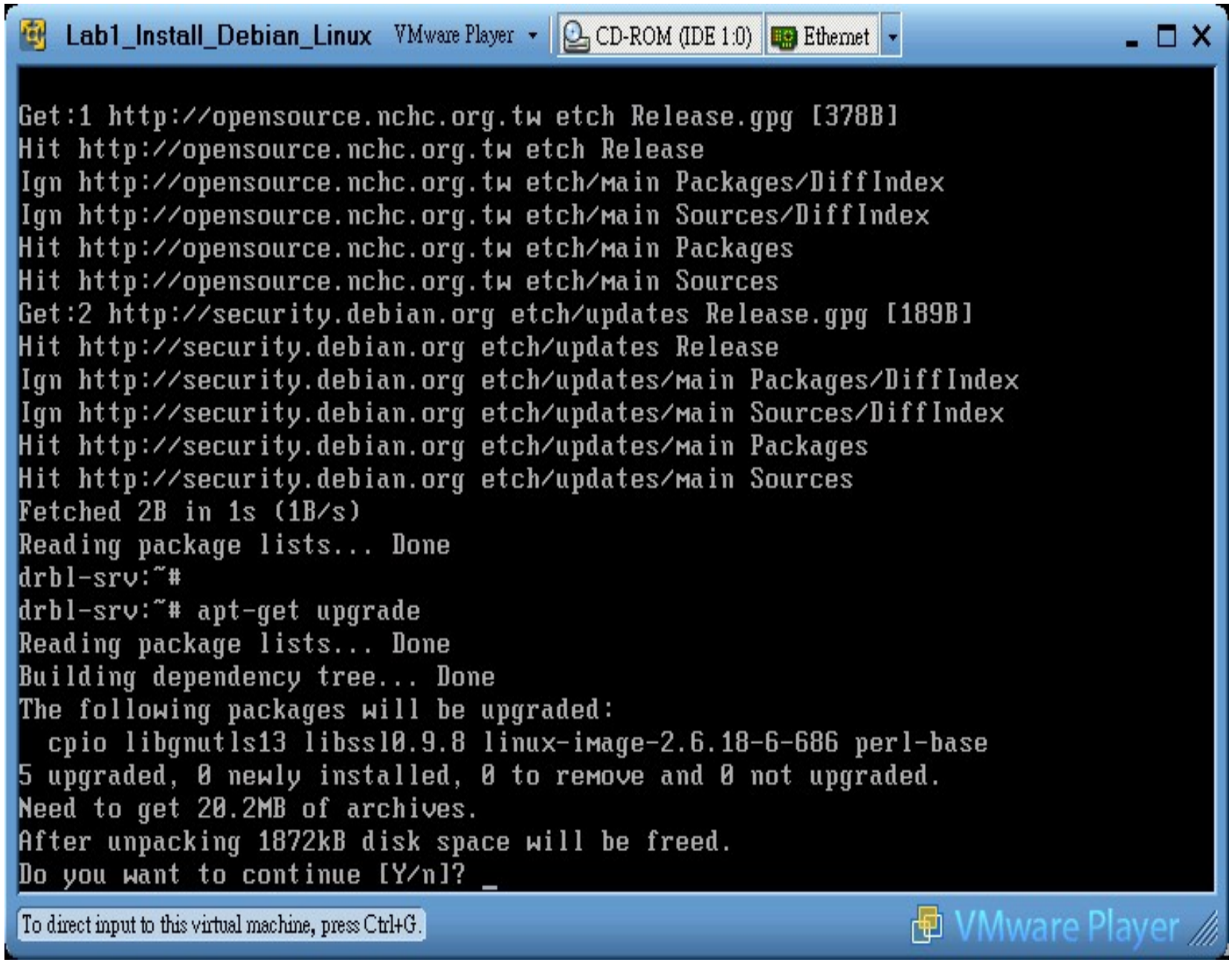
基本管理篇 [1] 用 ifconfig 確認網路連線



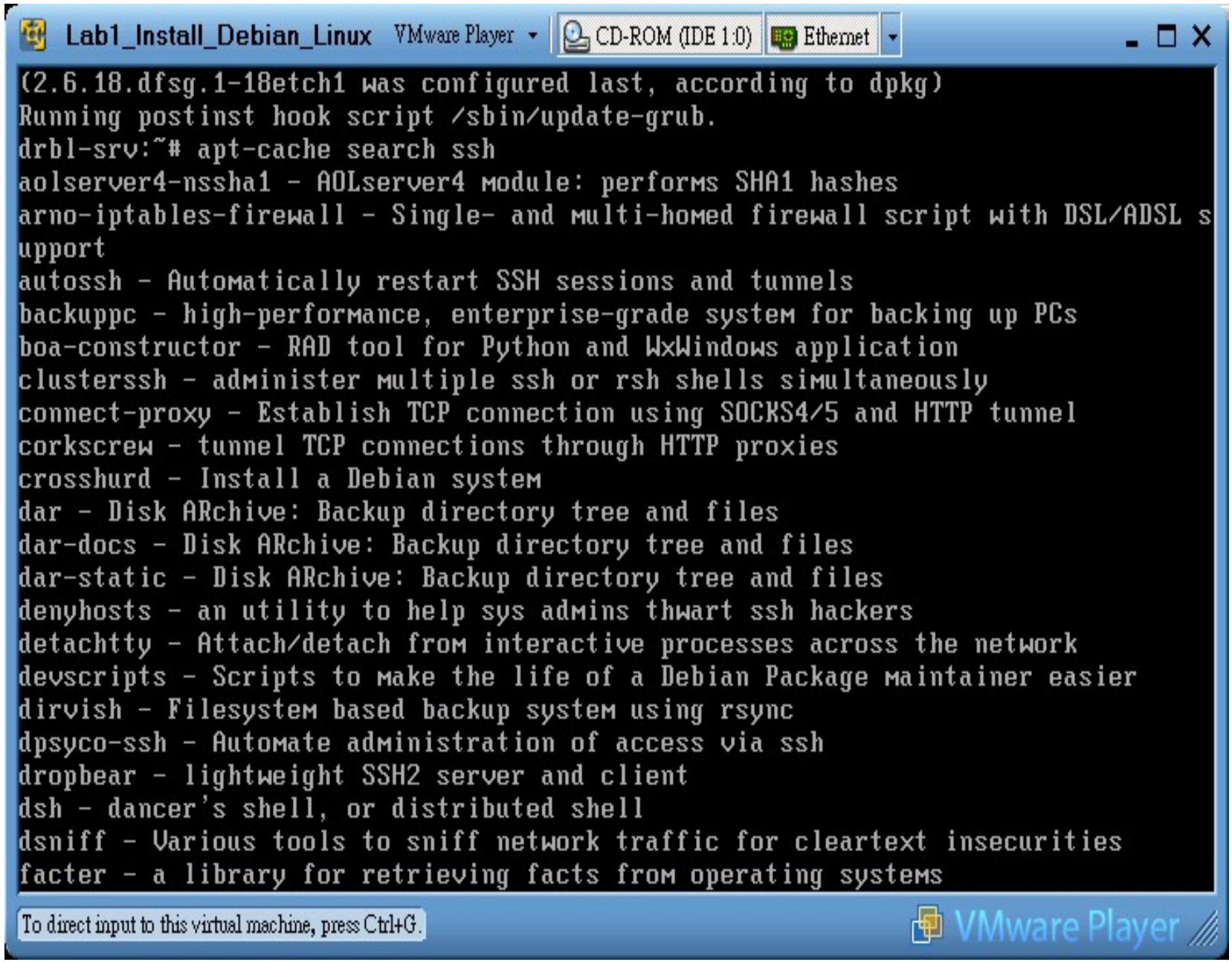
基本管理篇 [2] 用 apt-get update 更新套件



基本管理篇 [3] 用 apt-get upgrade 升級系統

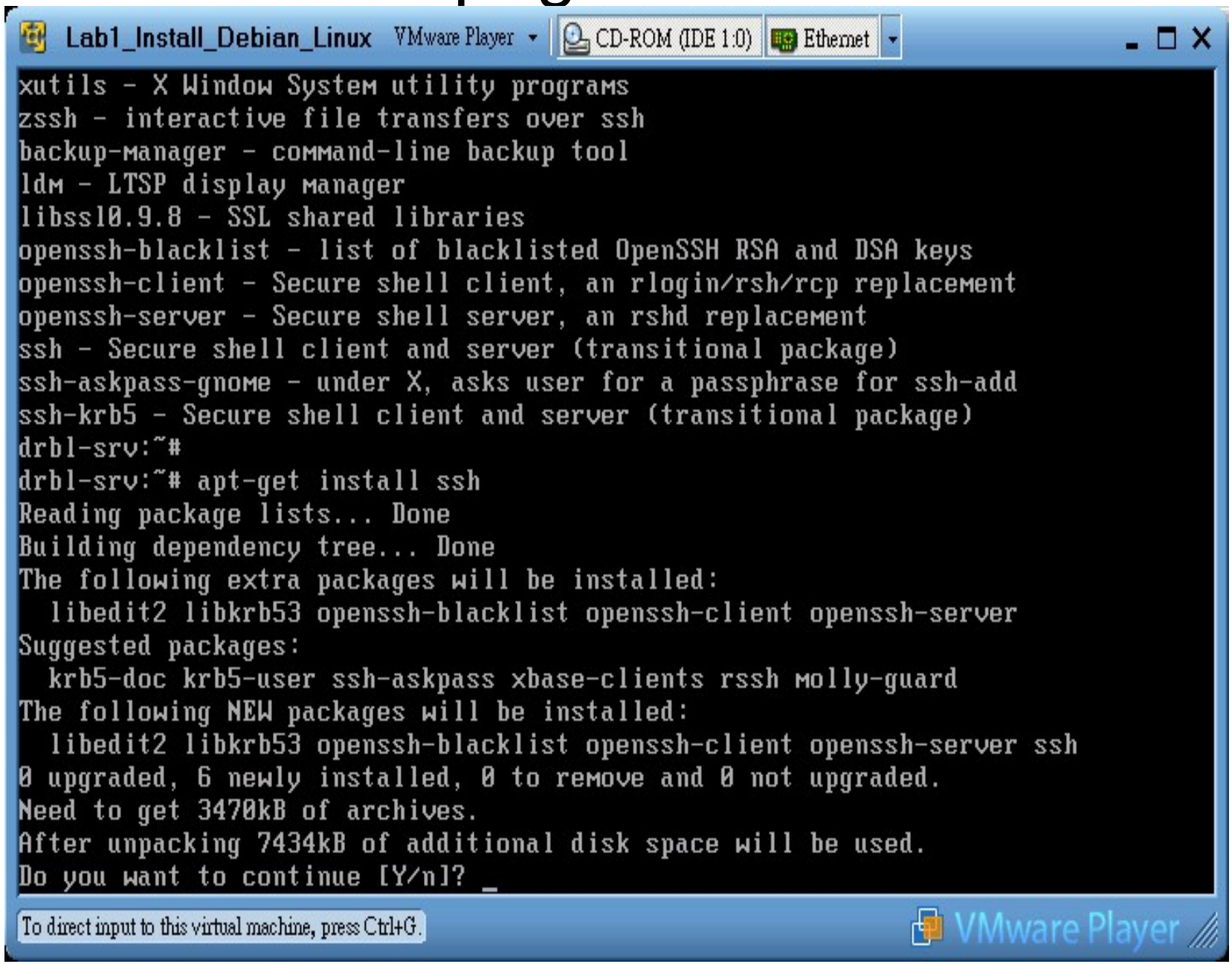


基本管理篇 [4] 用 apt-cache search 搜尋



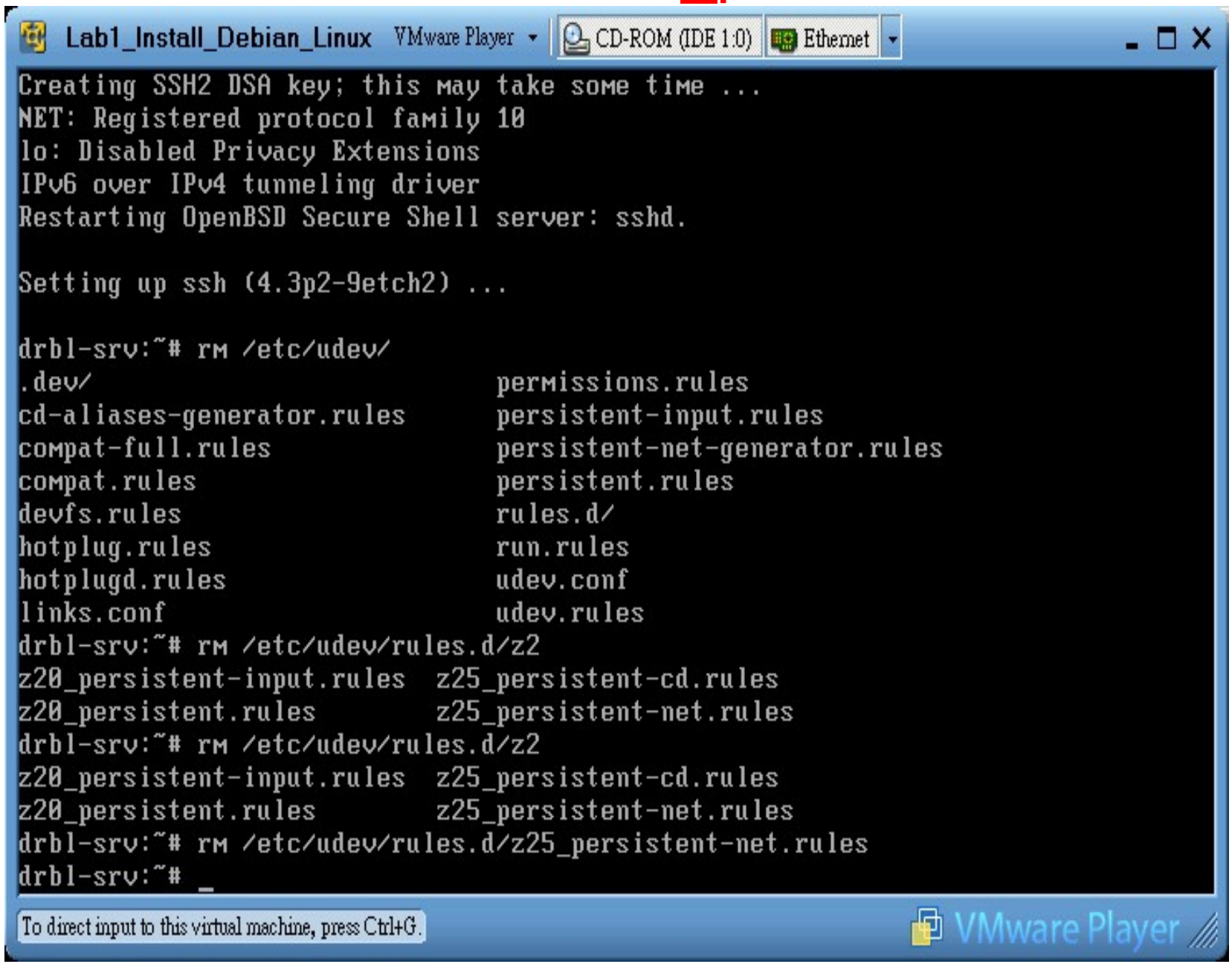
基本管理篇 [5] 用 apt-get install 安裝套件

Ex. apt-get install ssh

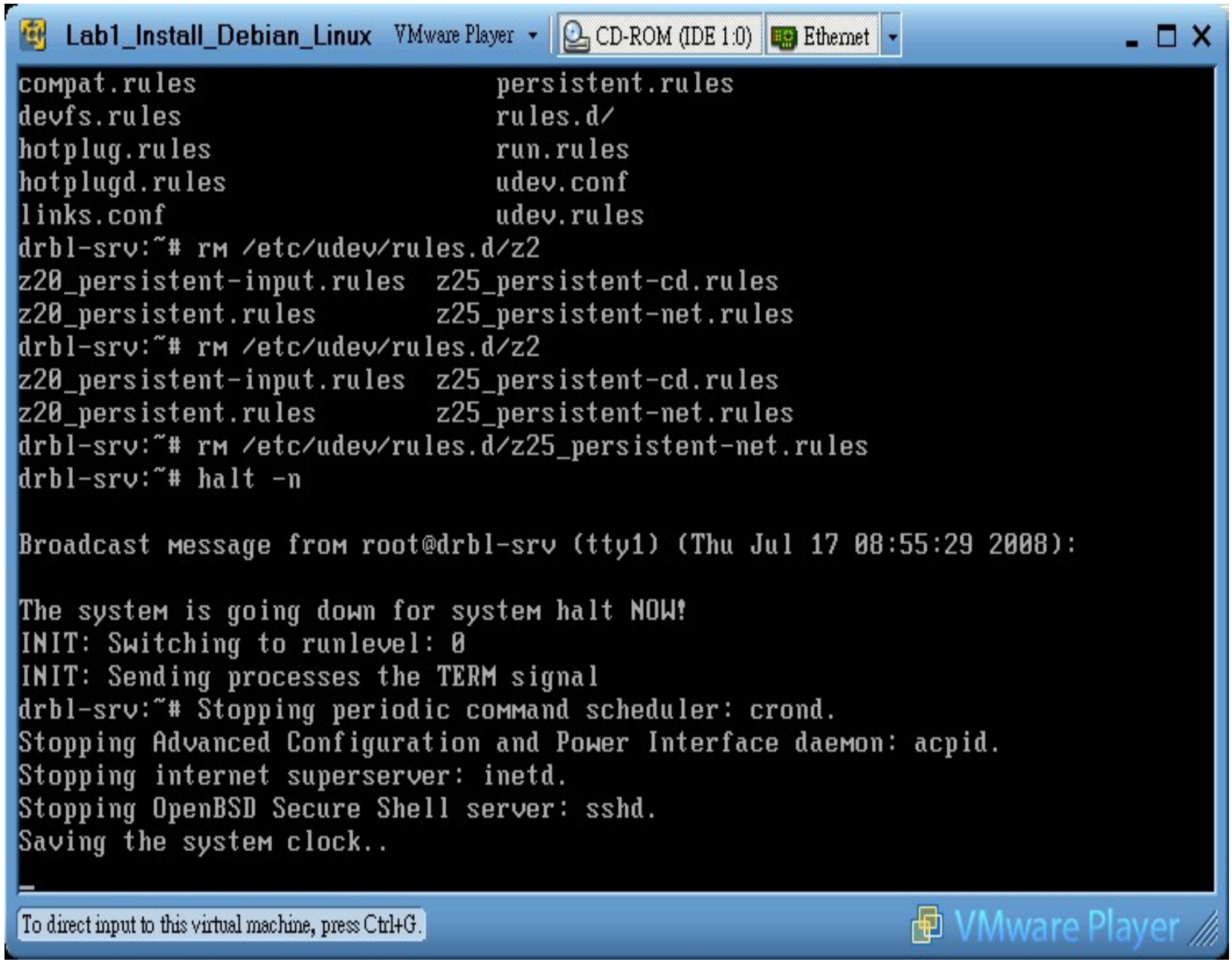


管理番外篇 [*] 關於 udev 鎖住網卡 MAC

rm /etc/udev/rules.d/z25_persistent-net.rules



基本管理篇 [6] 用 halt -n 關機



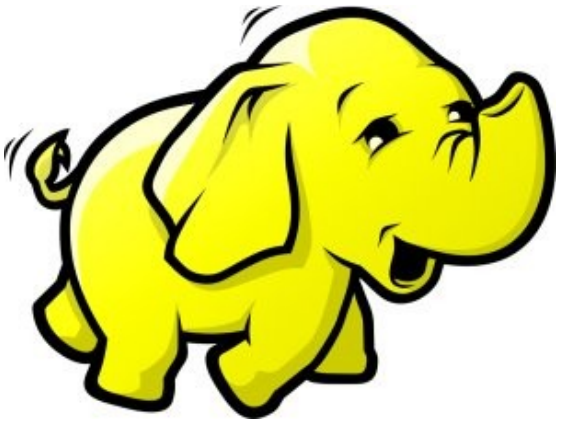
Hadoop 叢集設定解說

Setup Hadoop Fully Distributed Mode

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw



Yahoo's Hadoop Cluster
雅虎的大象軍團

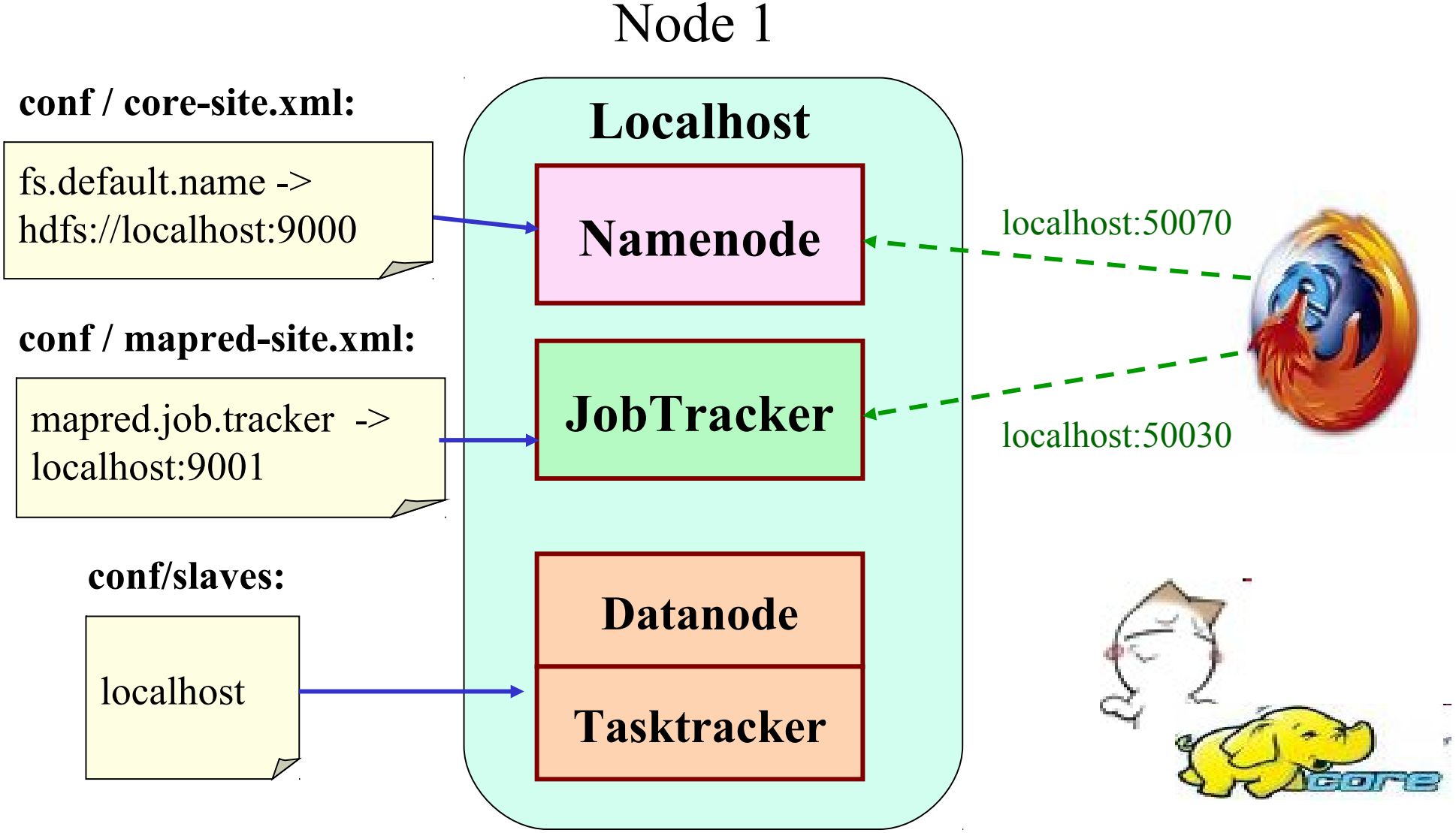
- ~10,000 machines running Hadoop in US
- The largest cluster is currently 2000 nodes
- Nearly 1 petabyte of user data (compressed, unreplicated)
- Running roughly 10,000 research jobs / week

Hadoop Pseudo-Distributed Mode
我們已經實作過單機模式

- Step 1: Setup SSH key exchange
- Step 2. Install Java
- Step 3: Download Hadoop Source Package
- Step 4: Configure hadoop-env.sh
  - export JAVA_HOME=/usr/lib/jvm/java-6-sun
- Step 5: Configure *-site.xml
  - Set Namenode to hdfs://localhost:9000
  - Set Jobtracker to localhost:9001
  - bin/hadoop namenode -format
- Step 6: Format HDFS
- Step 7: Start Hadoop
  - bin/start-all.sh
- Step 8: Complete!! Let's check the status of Hadoop
  - Job admin http://localhost:50030/ HDFS http://localhost:50070/

Diagram of Pseudo-Distributed Mode

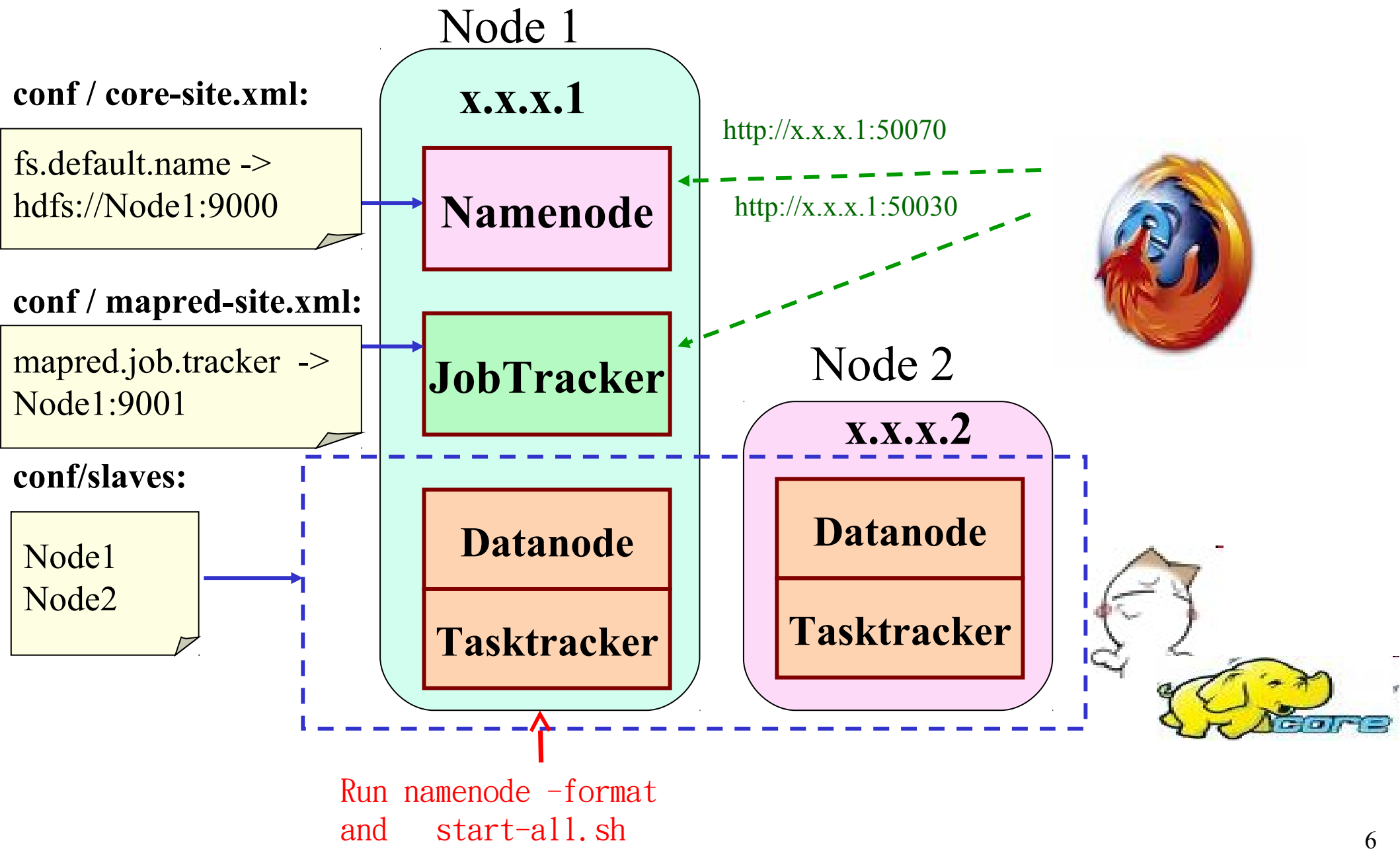
Hadoop 單機環境示意圖



Hadoop Fully-Distributed Mode
我們接著要用兩台電腦實作叢集模式

- Step 1: Setup SSH key exchange
- Step 2. Install Java
- Step 3: Download Hadoop Source Package
- Step 4: Configure hadoop-env.sh
  - export JAVA_HOME=/usr/lib/jvm/java-6-sun
- Step 5: Configure *-site.xml
  - Set Namenode to hdfs://x.x.x.1:9000
  - Set Jobtracker to x.x.x.2:9001
- Step 6: Configure Slaves
- Step 7: Syncronization of all slaves
- Step 8: Format HDFS
  - bin/hadoop namenode -format
- Step 9: Start Hadoop
  - On NameNode： bin/start-dfs.sh
  - On JobTracker： bin/start-mapred.sh
- Step 10: Complete!! Let's check the status of Hadoop
  - Job admin http://x.x.x.2:50030/ HDFS http://x.x.x.1:50070/

Use case #1
設定情境一



Use case #2
設定情境二

Node 1

Node 2

conf / core-site.xml:

fs.default.name -> hdfs://Node1:9000

conf / mapred-site.xml:

mapred.job.tracker -> Node2:9001

conf/slaves:

Node1
Node2

x.x.x.1

Namenode

Datanode

Tasktracker

x.x.x.2

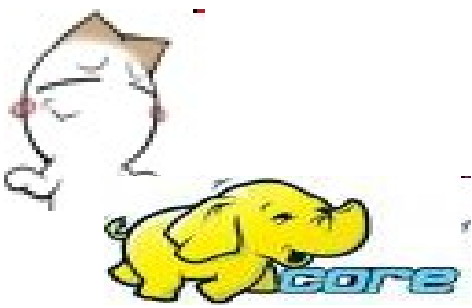
JobTracker

Datanode

Tasktracker

http://x.x.x.1:50070

http://x.x.x.2:50030



Run namenode -format and start-dfs.sh

run start-mapred.sh

Use case #3
設定情境三

conf / core-site.xml:

fs.default.name ->
hdfs://Node1:9000

conf / mapred-site.xml:

mapred.job.tracker ->
Node1:9001

conf/slaves:

Node2
…..
NodeN

Node 1

x.x.x.1

Namenode

JobTracker

http://x.x.x.1:50070

http://x.x.x.1:50030

Node 2

x.x.x.2

Datanode

Tasktracker

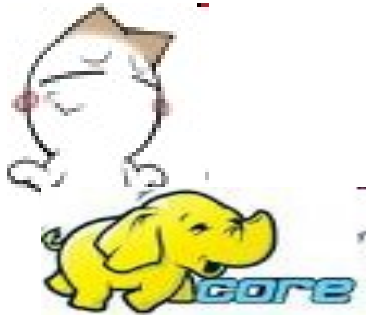
…

Node N

x.x.x.n

Datanode

Tasktracker



Use case #4
設定情境四

conf / core-site.xml:

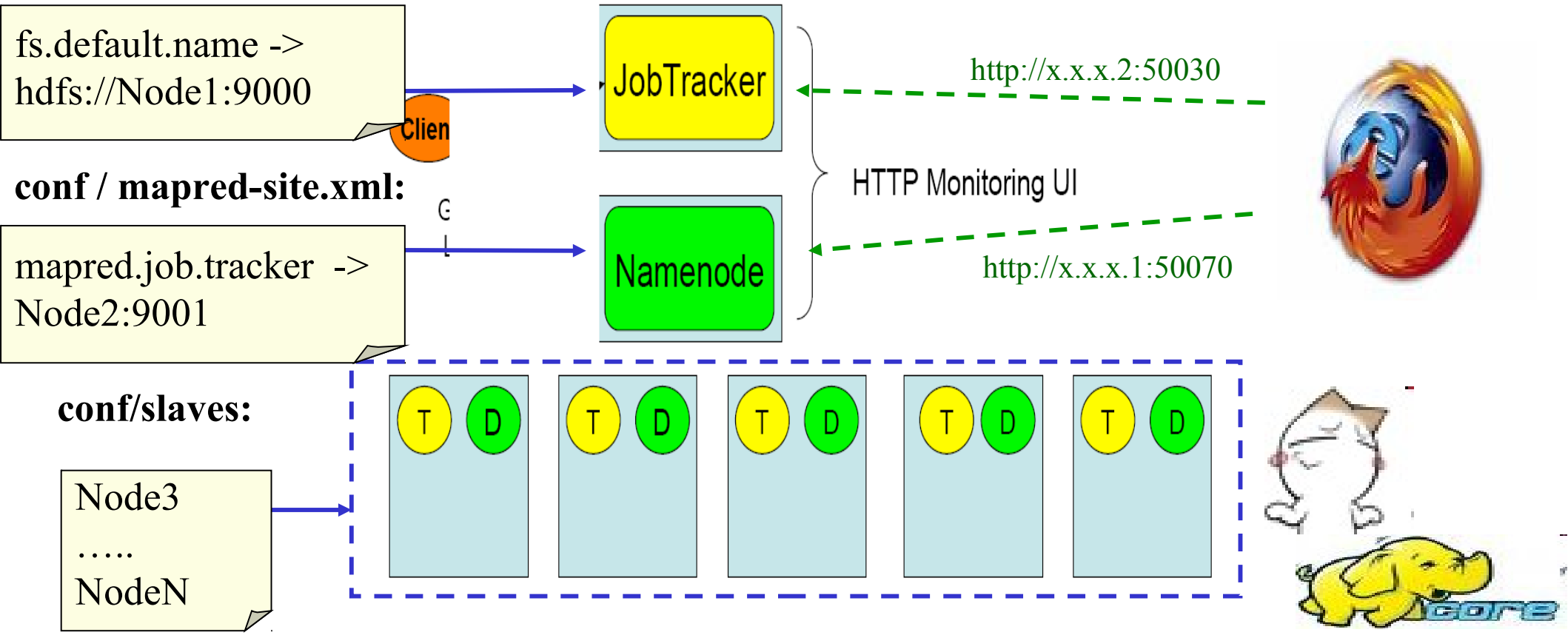
fs.default.name -> hdfs://Node1:9000

conf / mapred-site.xml:

mapred.job.tracker -> Node2:9001

conf/slaves:

Node3
…..
NodeN



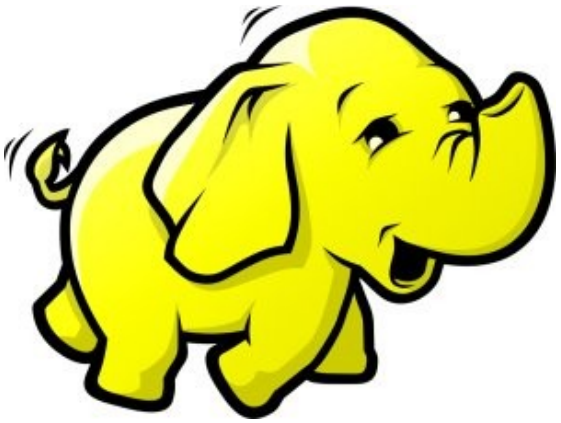
Hadoop 叢集佈署工具

Hadoop Deployment Tool : SmartFog and DRBL

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

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

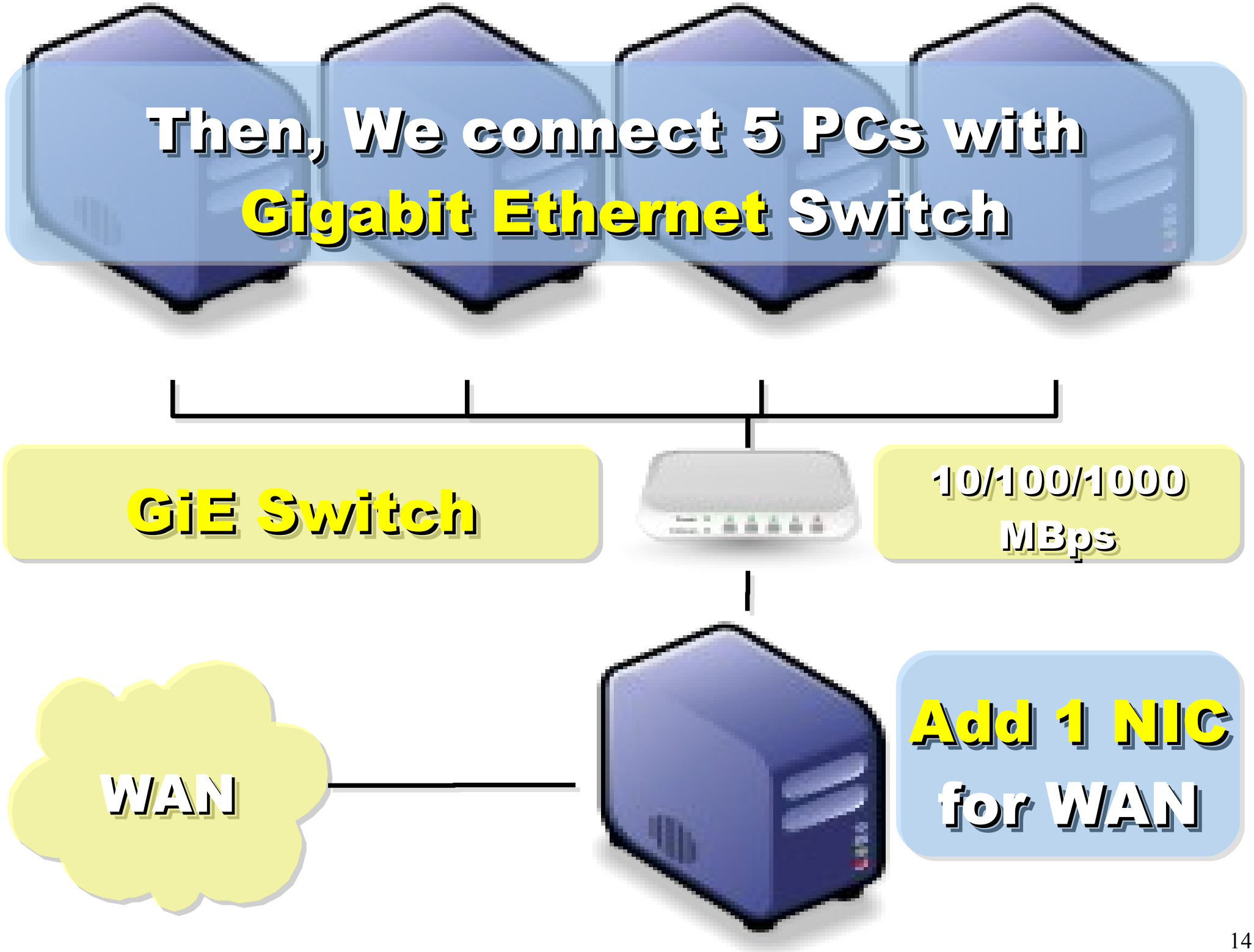
It’d better be

2^n

Manage

Scheduler

Then, We connect 5 PCs with Gigabit Ethernet Switch

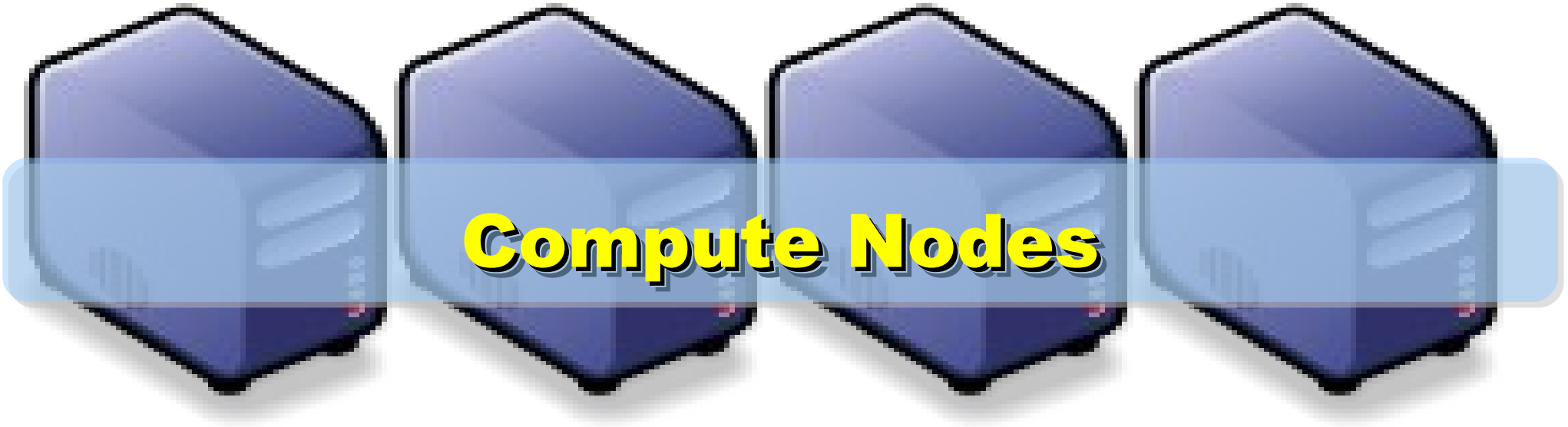


Compute Nodes

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

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

MPICH

Account Mgnt.

SSHD

NIS

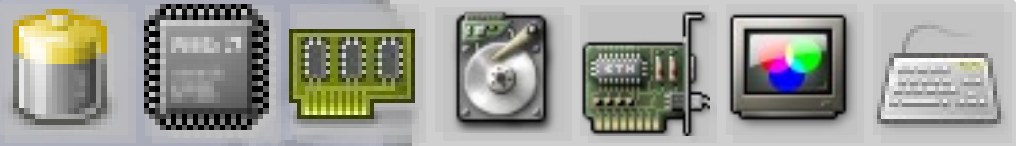
YP

GCC

GNU Libc

Bash

Perl



Kernel Module

Linux Kernel

Boot Loader

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

- 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

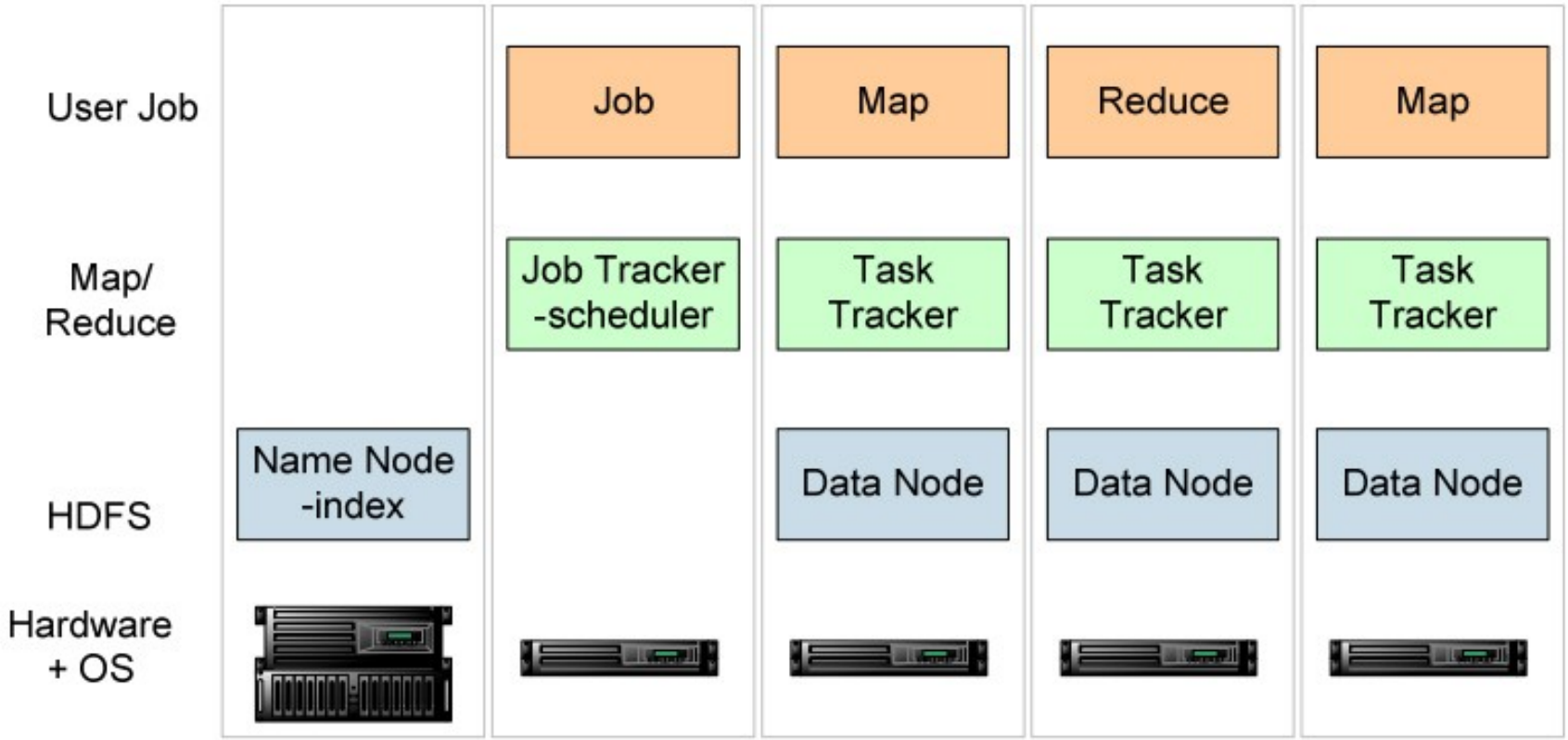
- Language for describing systems to deploy —everything from datacentres to test cases
- Runtime to create *components* from the model
- Components have a lifecycle
- LGPL Licensed, Java 5+

http://smartfrog.org/

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

LABS hp

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

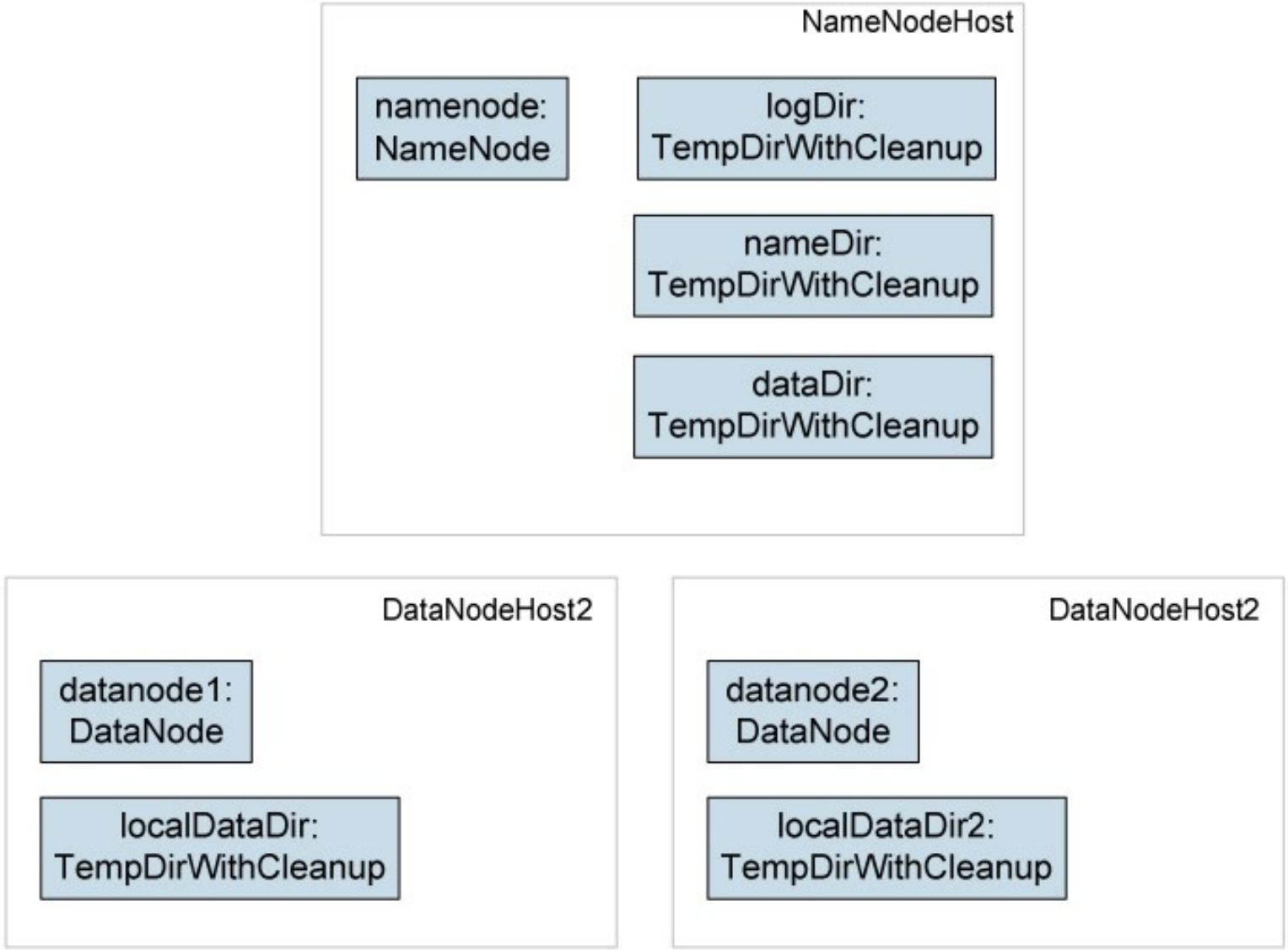
Model the system in the SmartFrog language

```
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 :

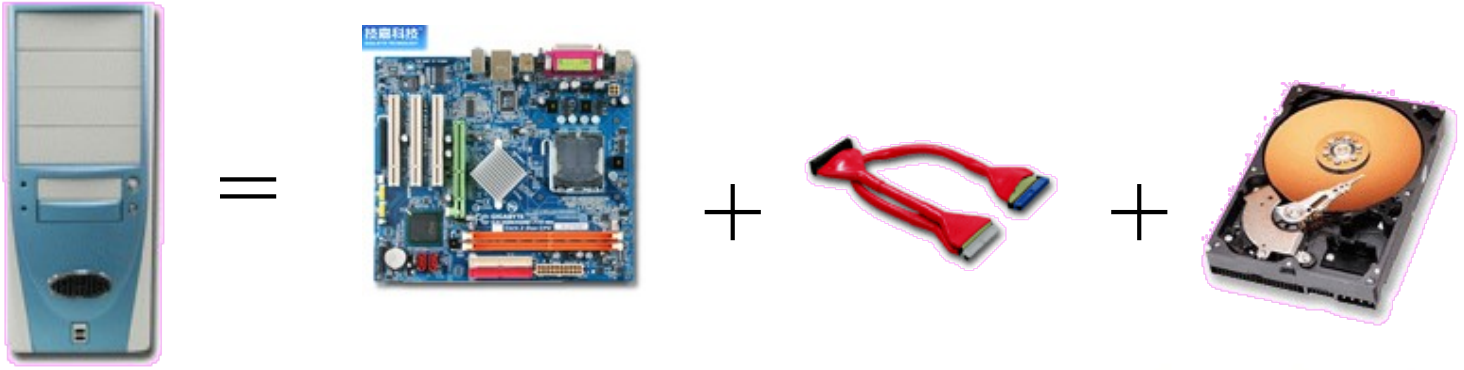
Introduction to DRBL

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

What is DRBL ??

- **Diskless Remote Boot in Linux**
- Network is cheap, and our time is expansive
- In simple words, DRBL is .....
  - Replace IDE/SATA cable with network cable
  - 40+ student PCs connected to one DRBL server

Diskfull PC



Diskless PC

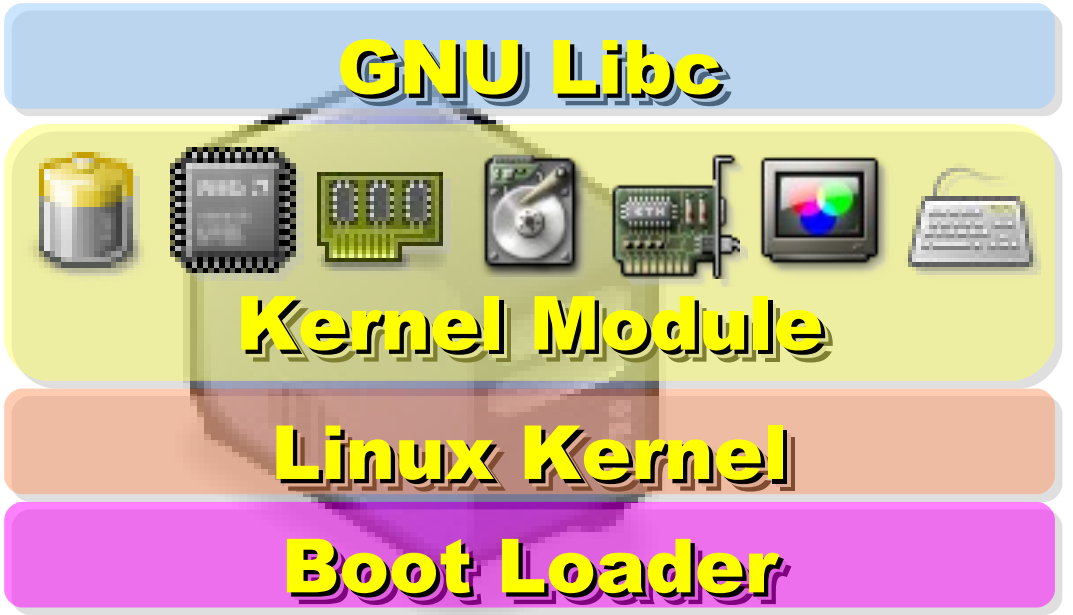
Server

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

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

You can choose:

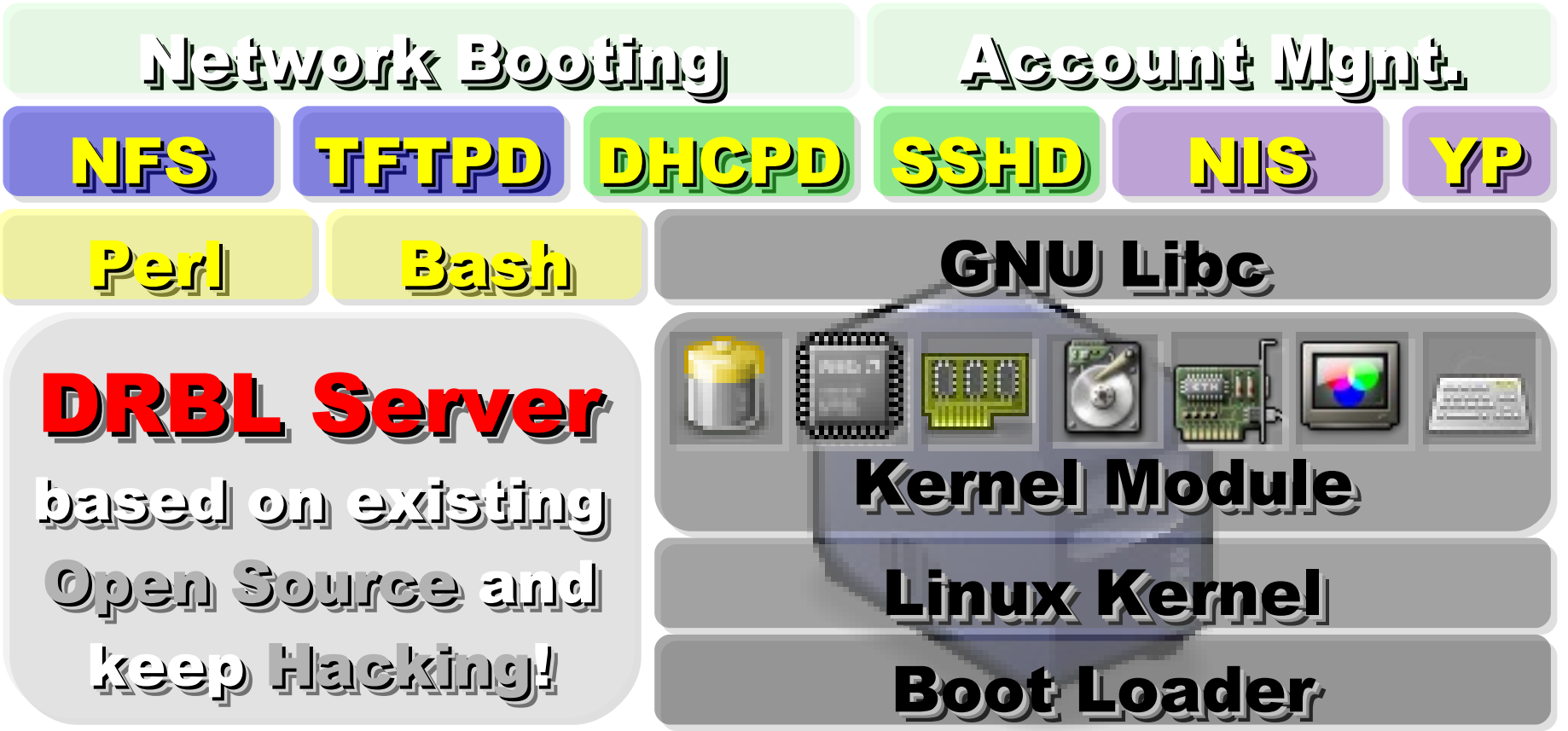
Redhat, Fedora, CentOS, Mandriva, Ubuntu, Debian, ...



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

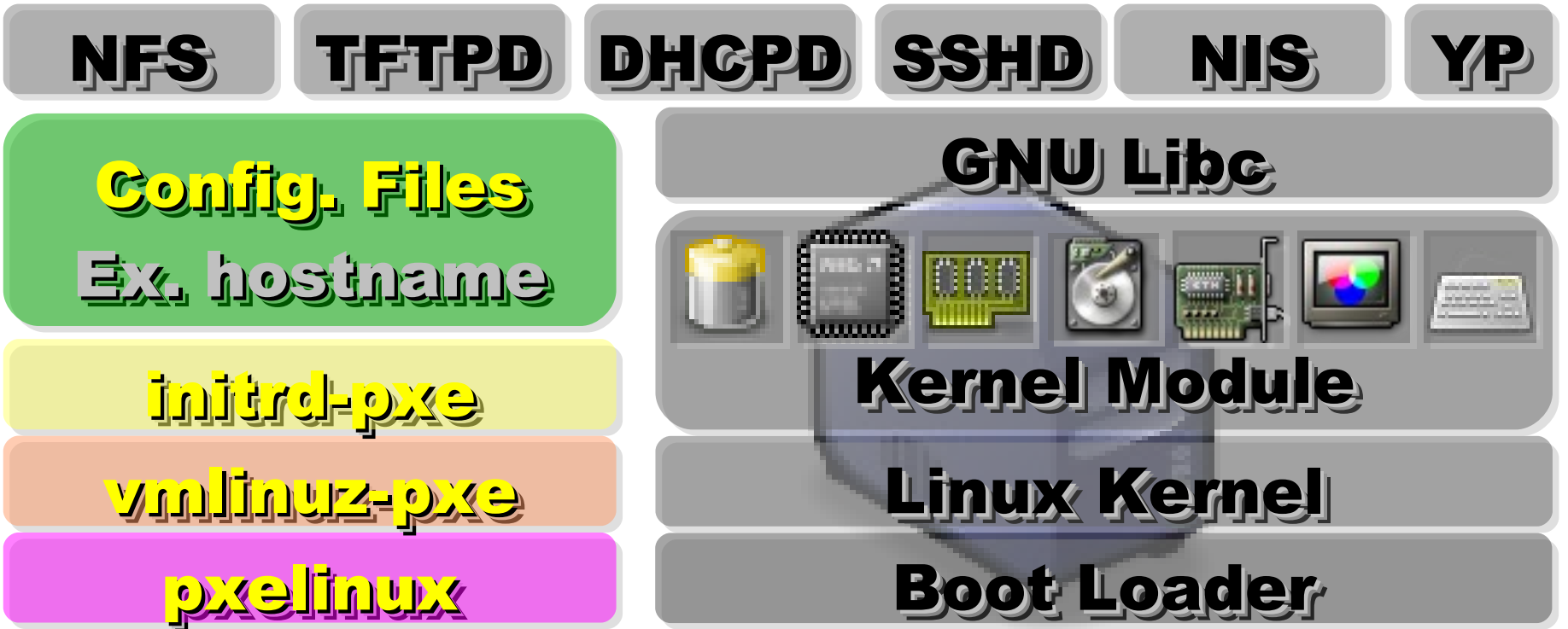
There are lots of service needed:

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



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

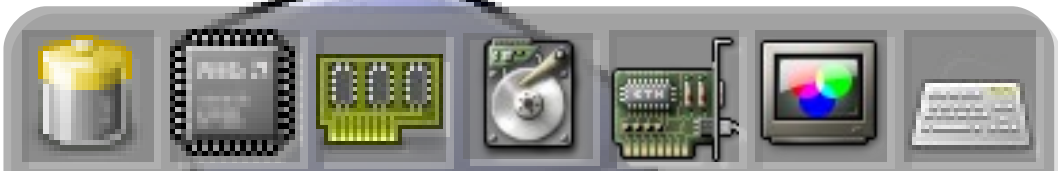
Config. Files
Ex. hostname

initrd-pxe

vmlinuz-pxe

pxelinux

GNU Libc

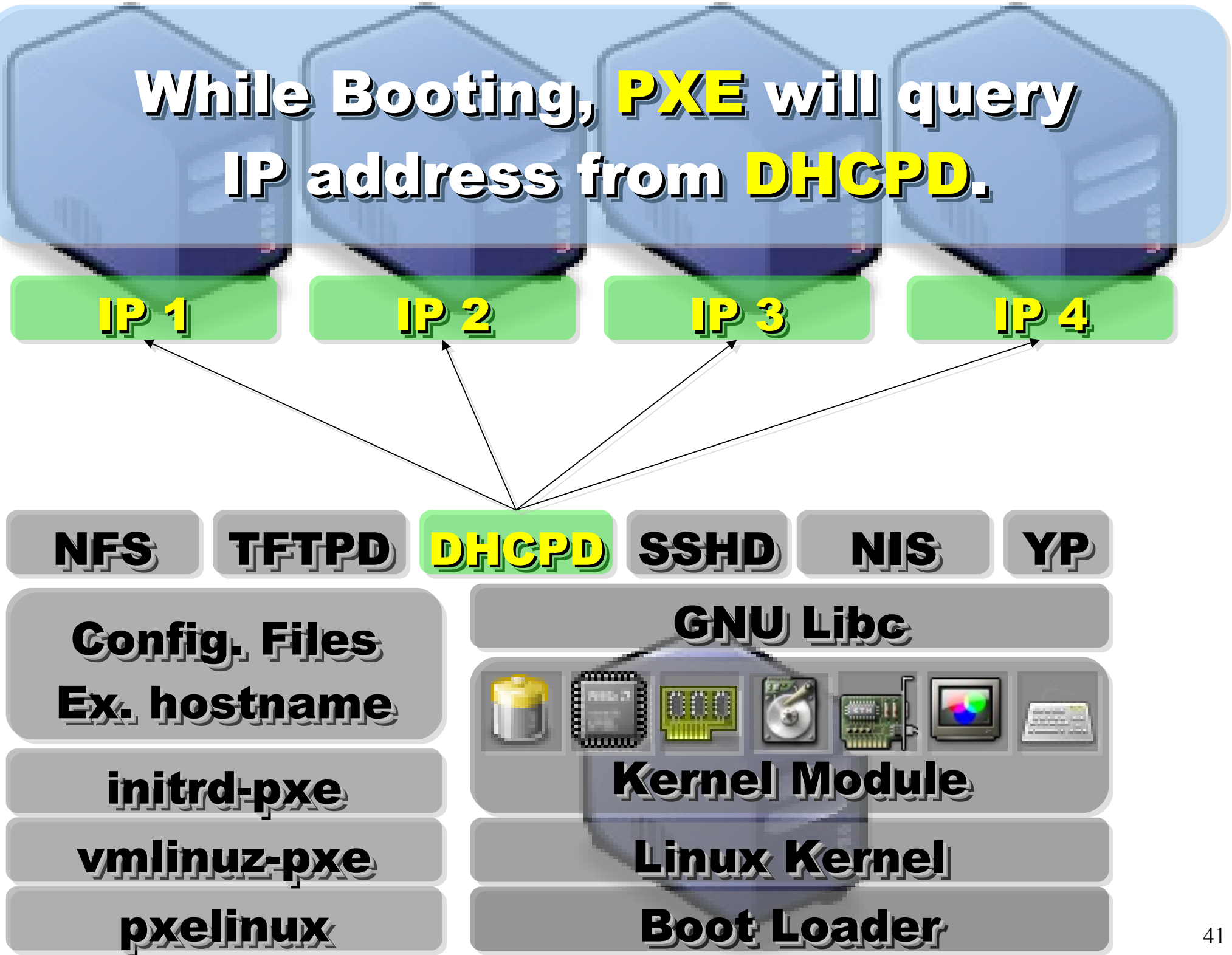


Kernel Module

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

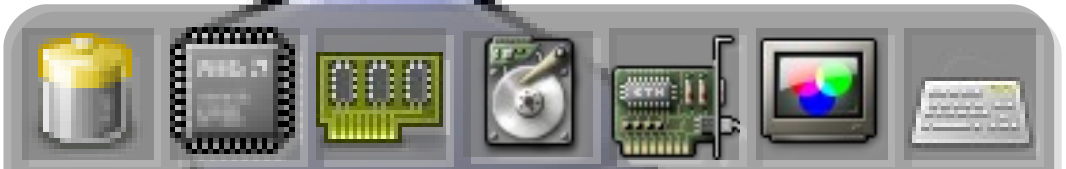
Config. Files
Ex. hostname

initrd-pxe

vmlinuz-pxe

pxelinux

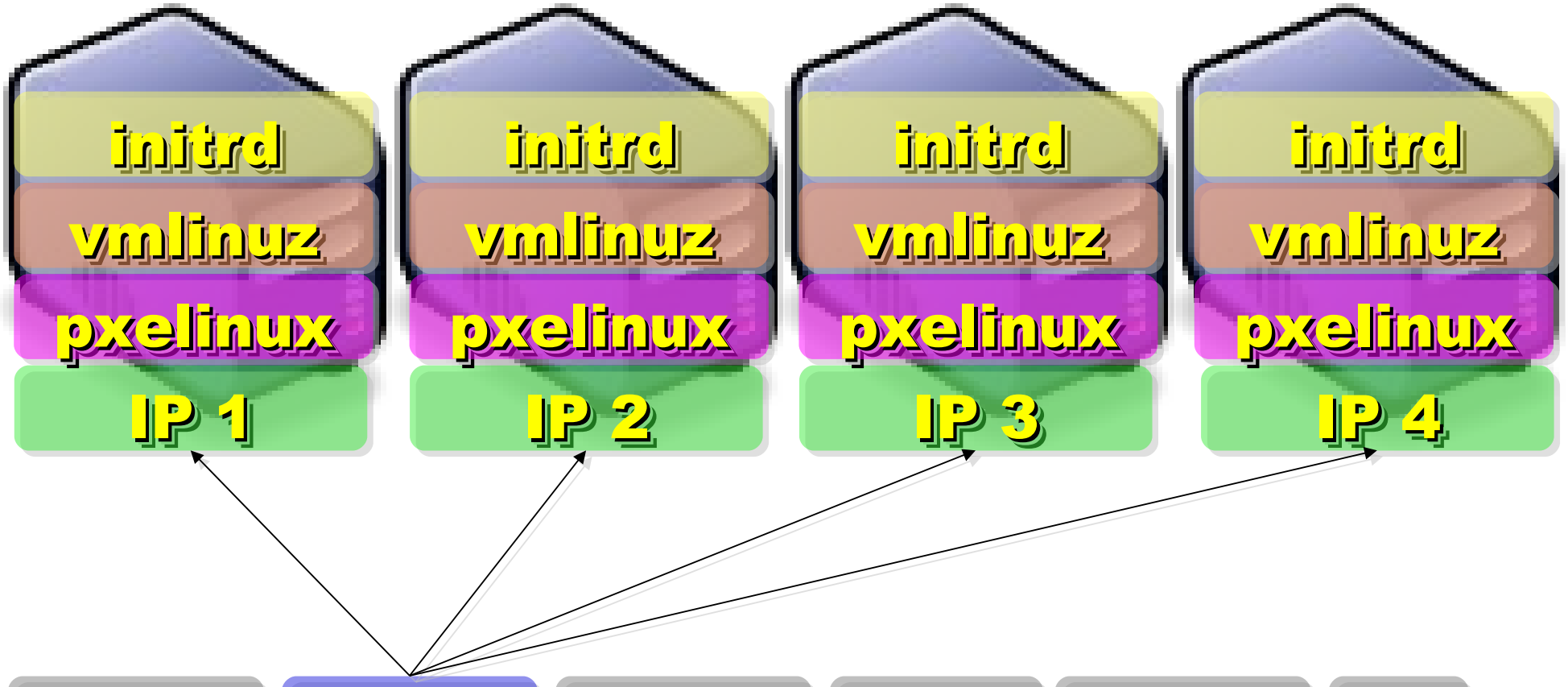
GNU Libc



Kernel Module

Linux Kernel

Boot Loader



NFS | TFTPD | DHCPD | SSHD | NIS | YP

Config. Files
Ex. hostname

initrd-pxe

vmlinuz-pxe

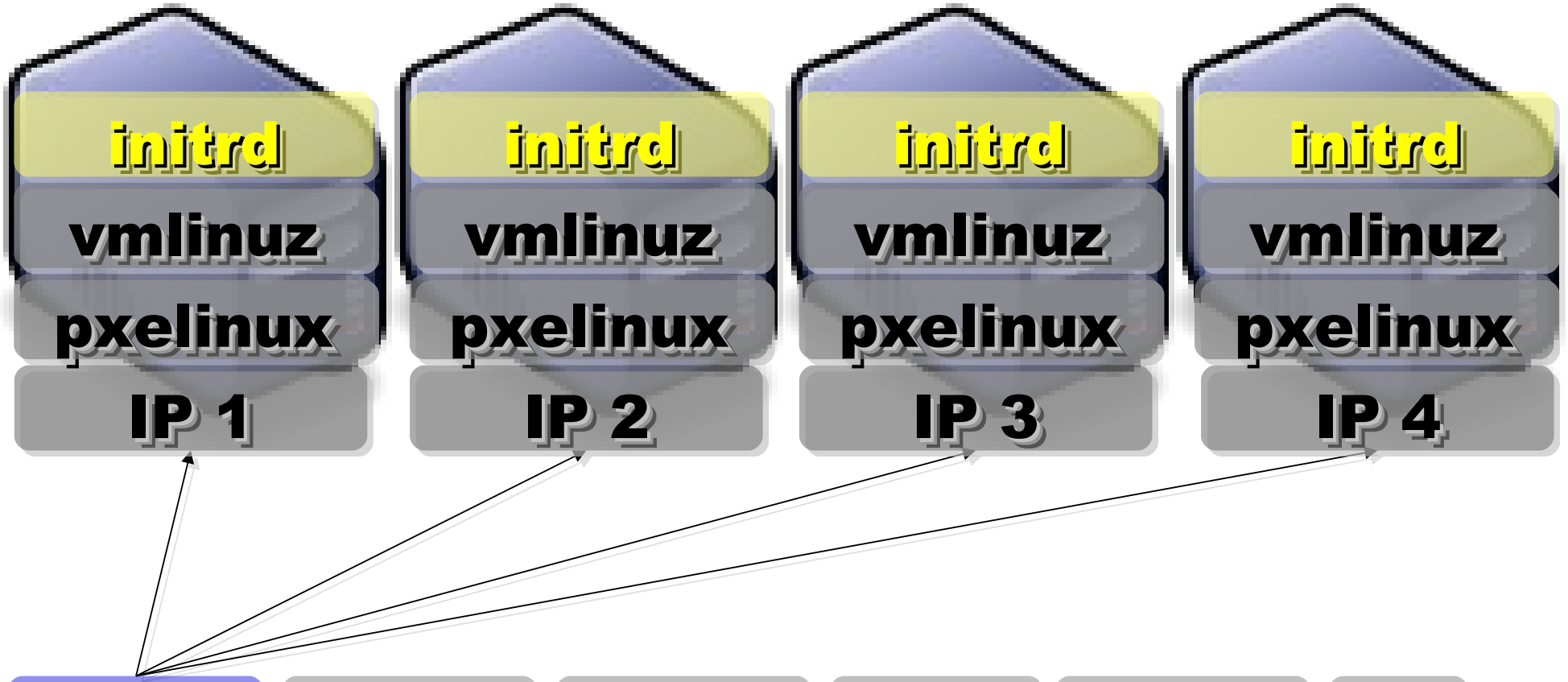
pxelinux

GNU Libc

Kernel Module

Linux Kernel

Boot Loader



NFS | TFTPD | DHCPD | SSHD | NIS | YP

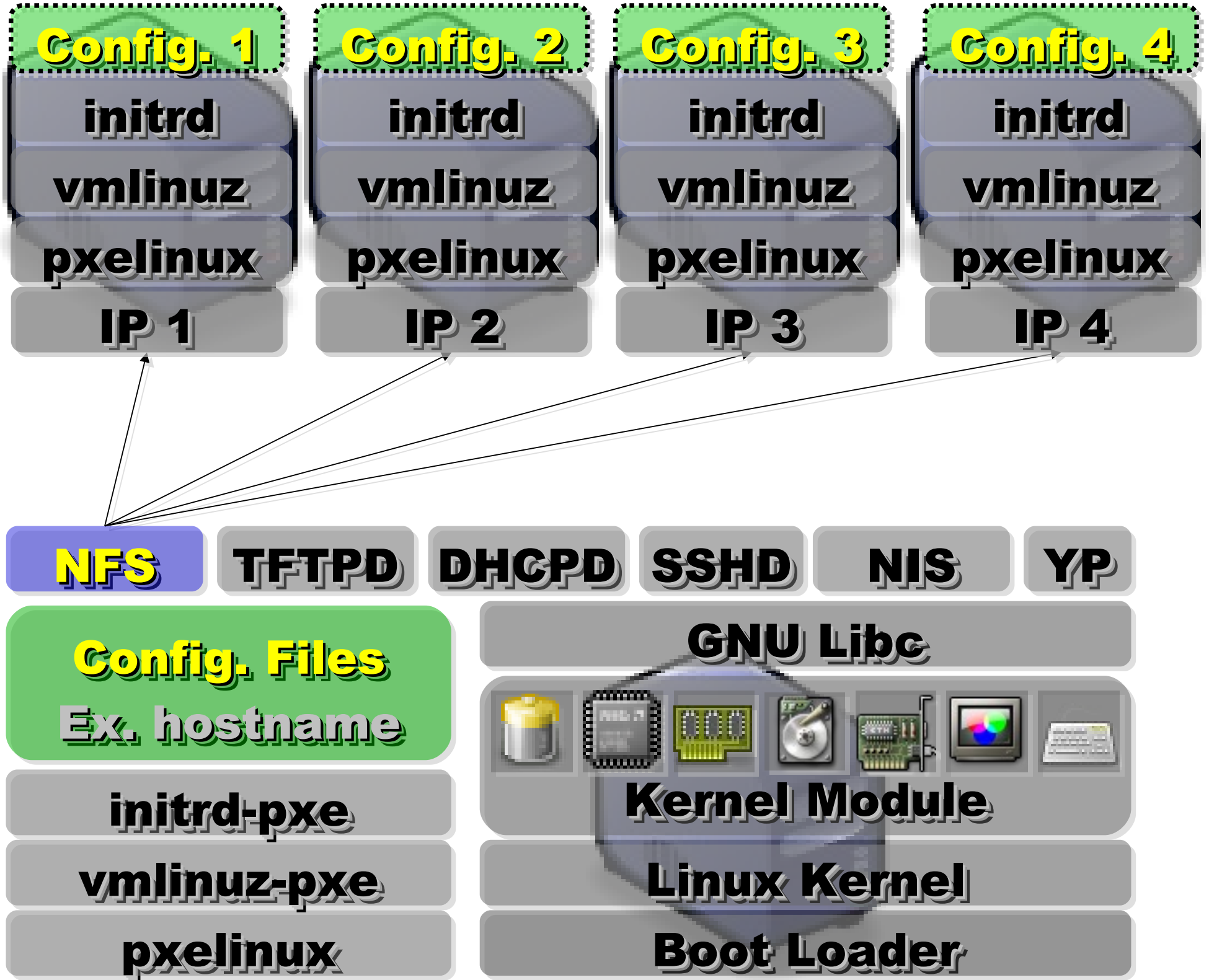
Config. Files

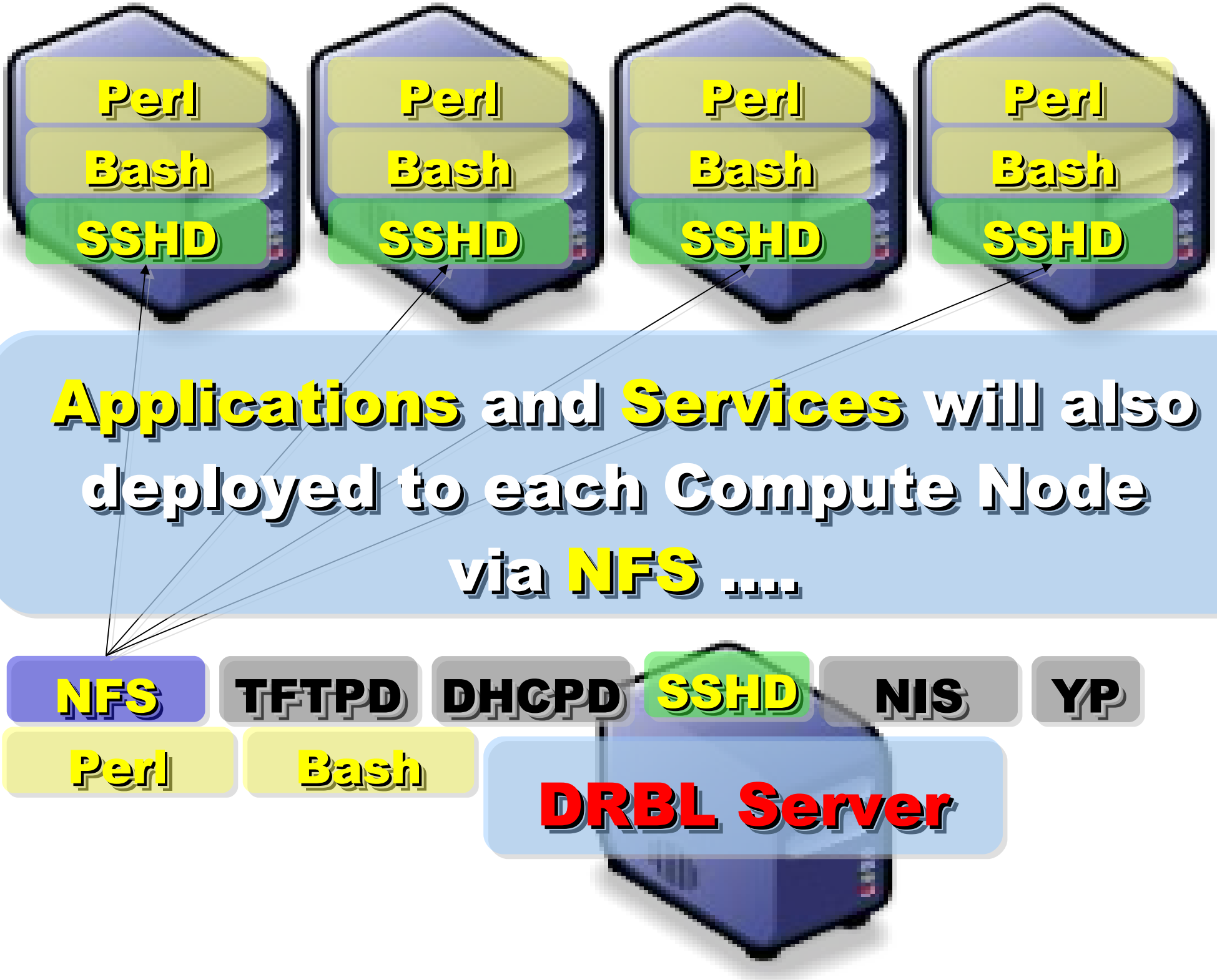
GNU Libc

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

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

Questions?

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

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

Powered by DRBL