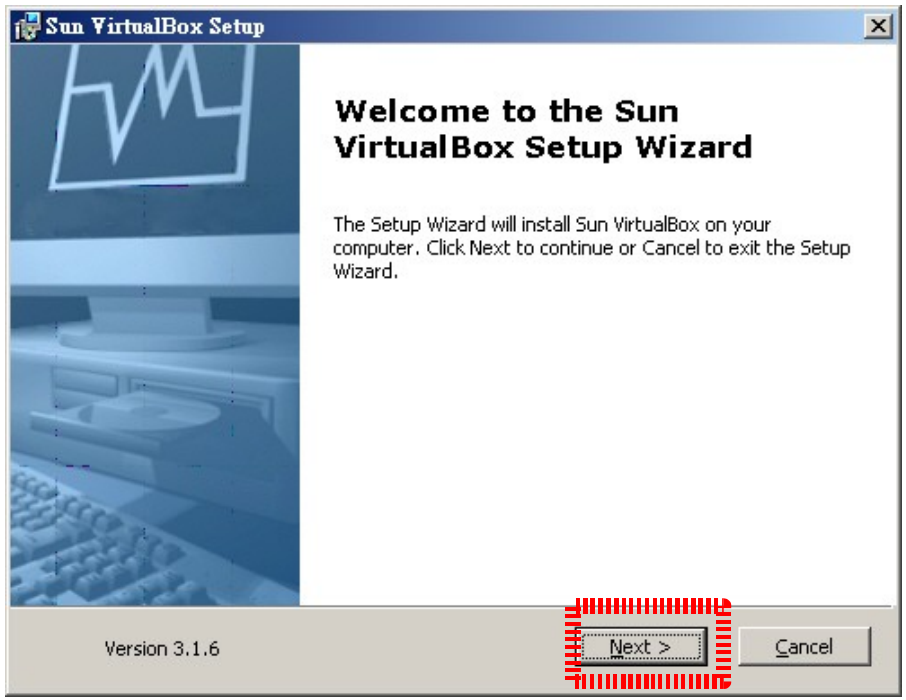
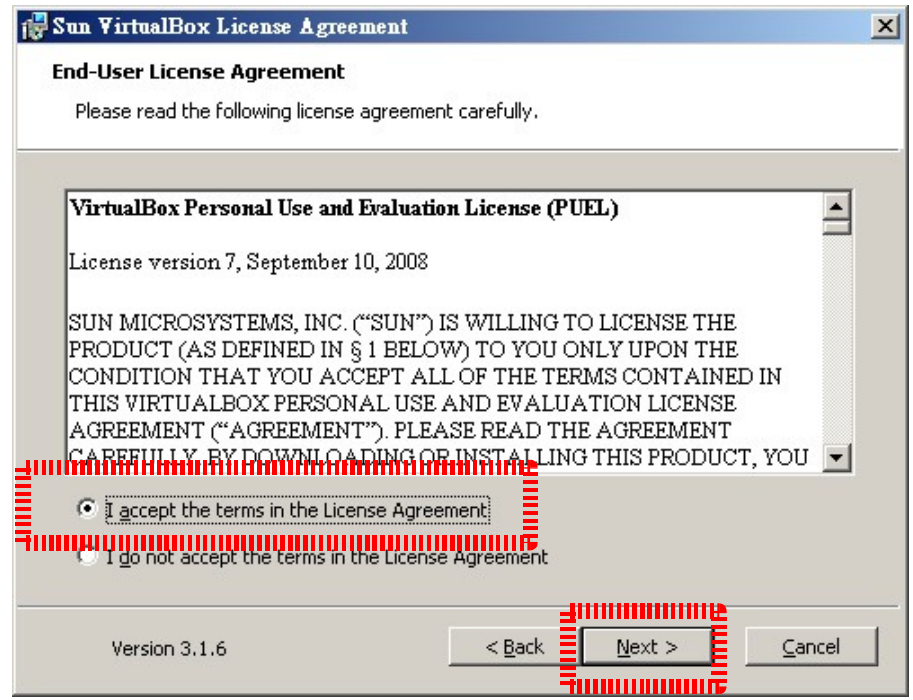
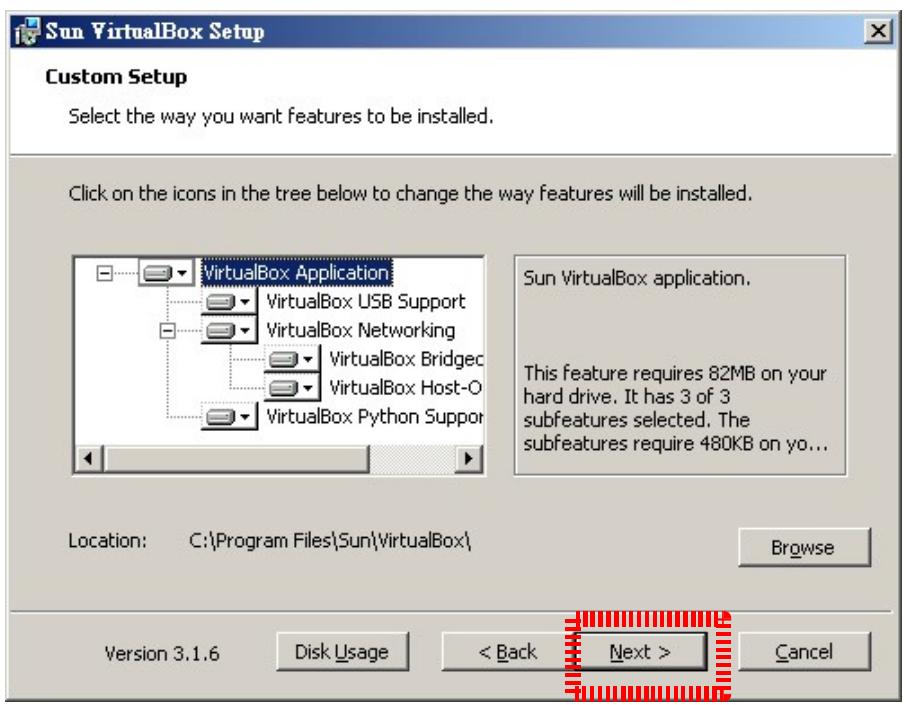
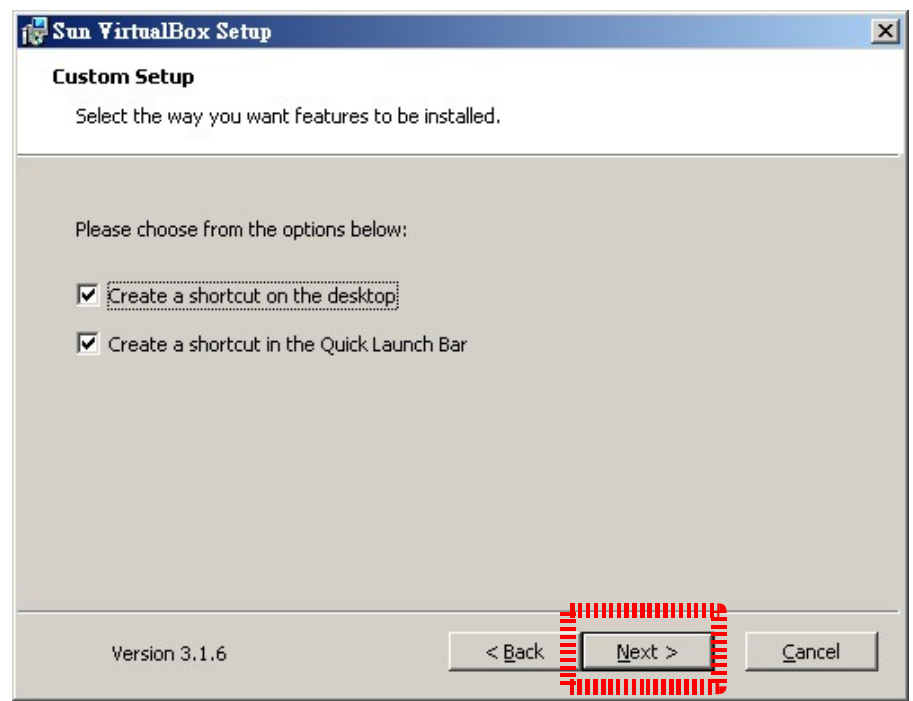
連線至 http://www.virtualbox.org/wiki/Downloads
下載 VirtualBox 虛擬化工具

安装 VirtalBox（1）

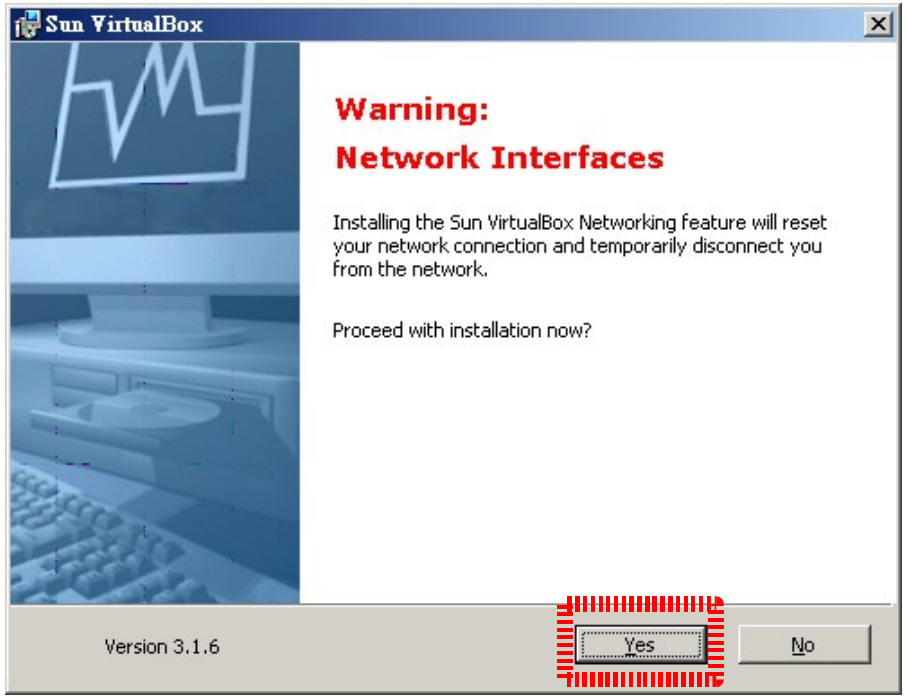


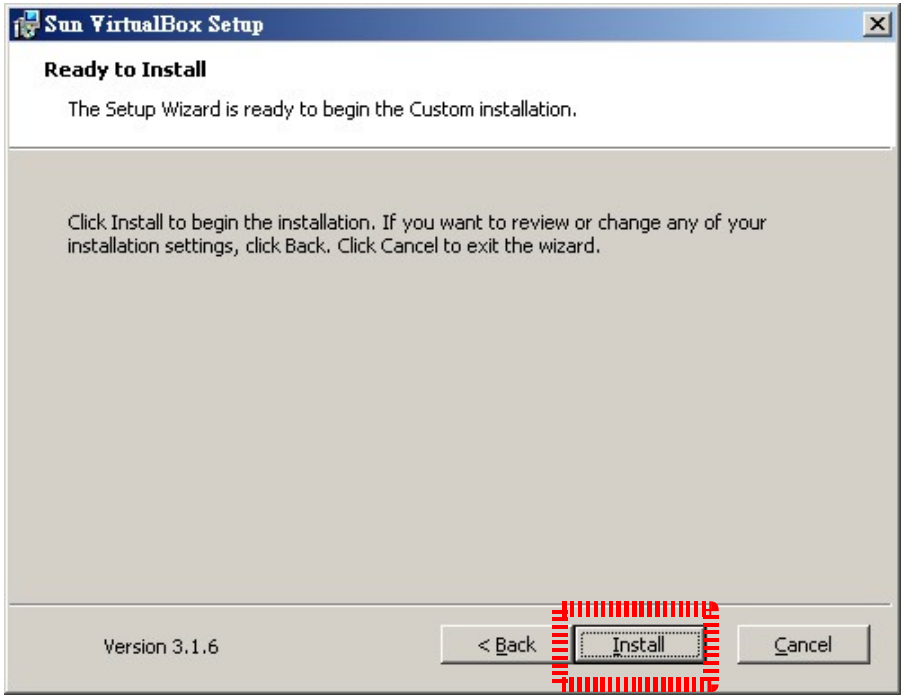


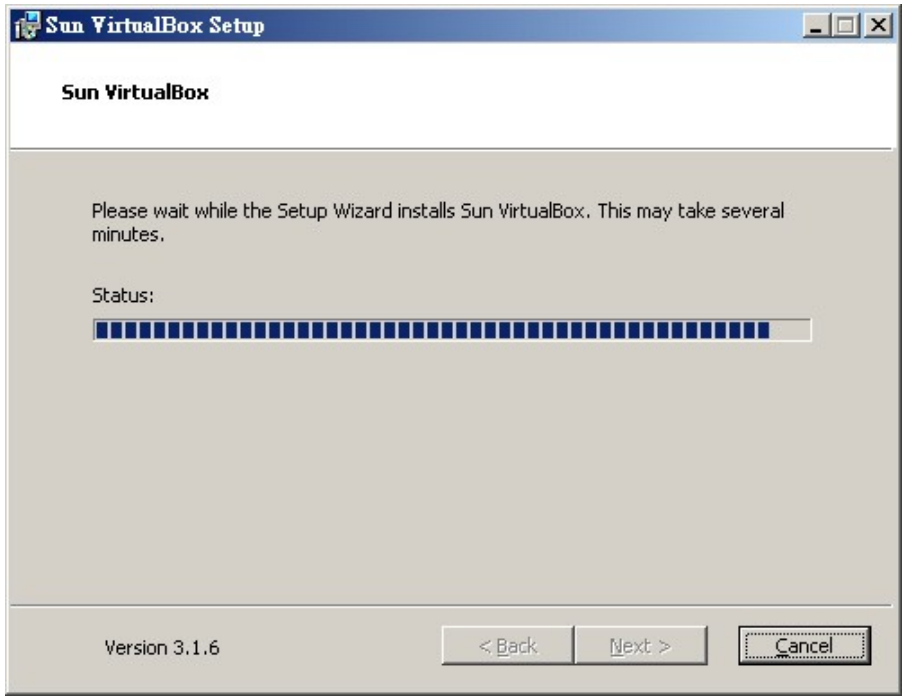


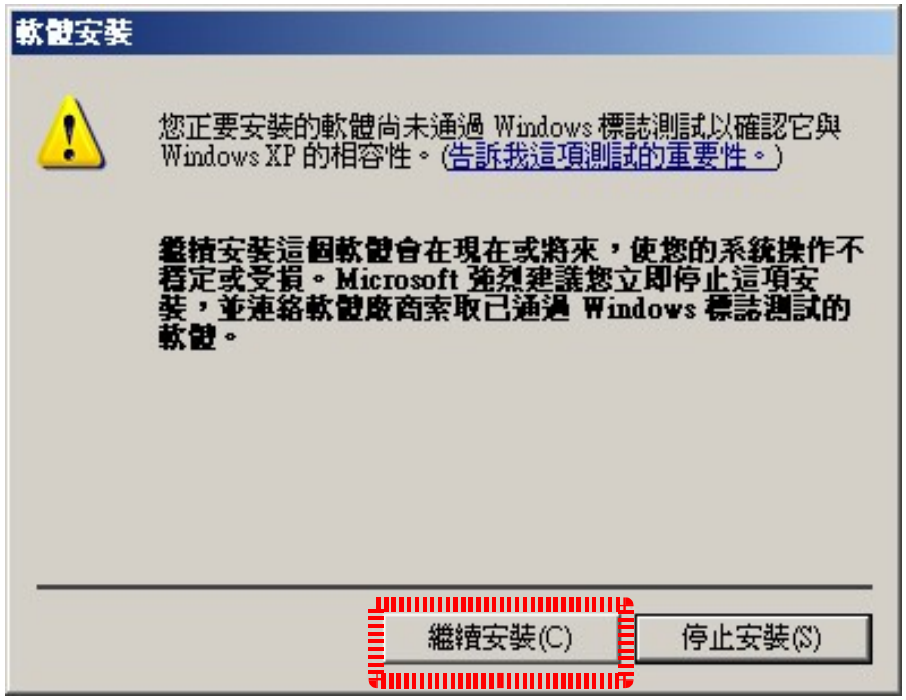


安裝 VirtalBox (2)

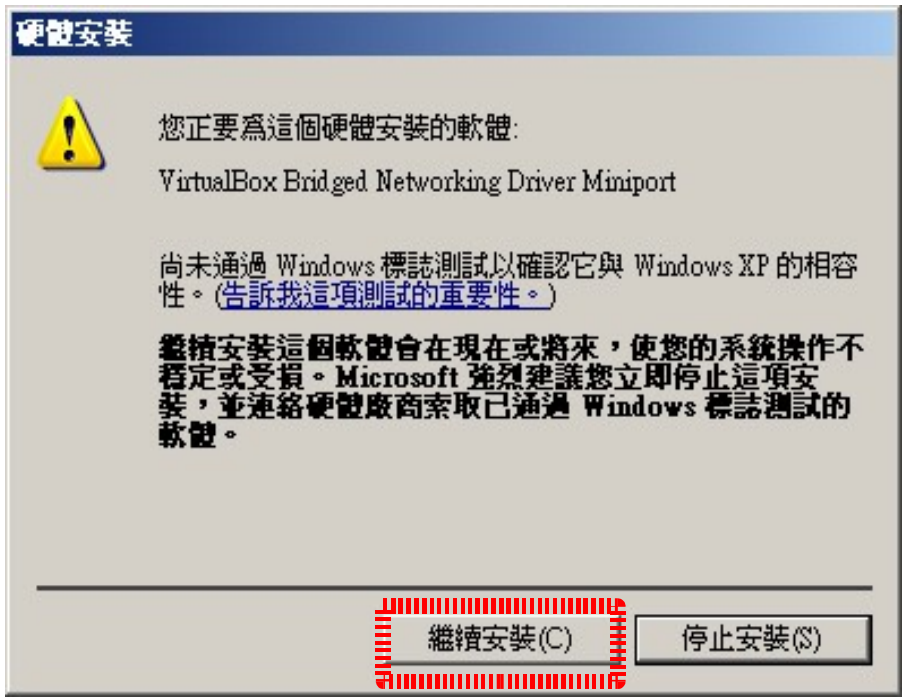


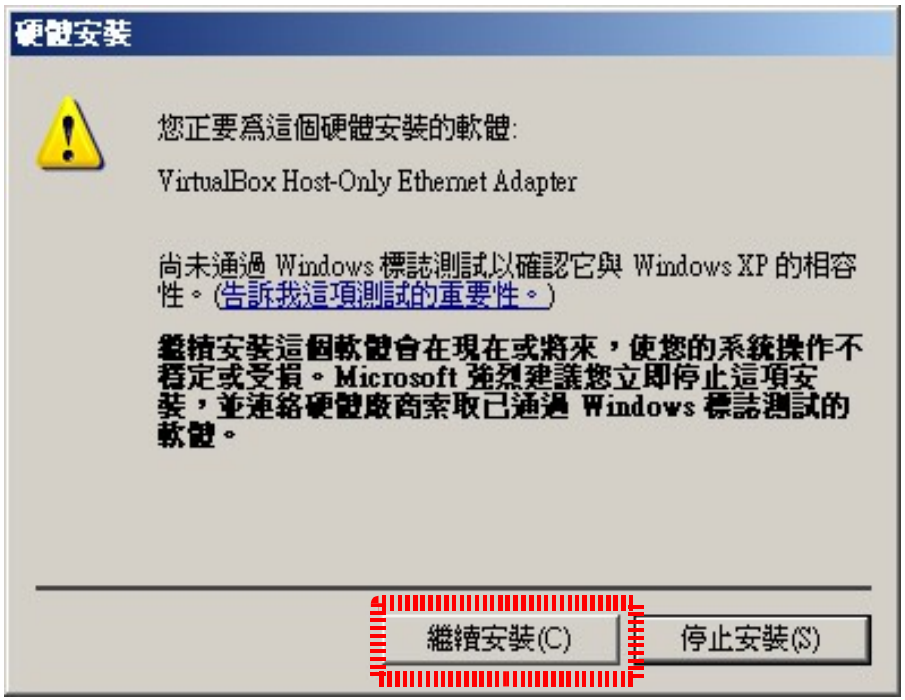


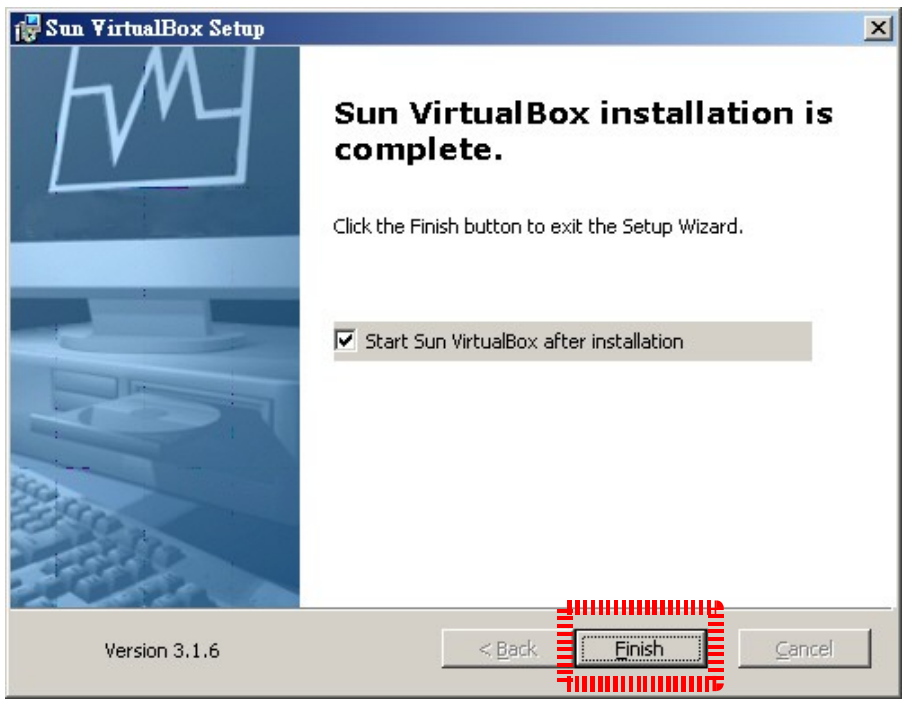


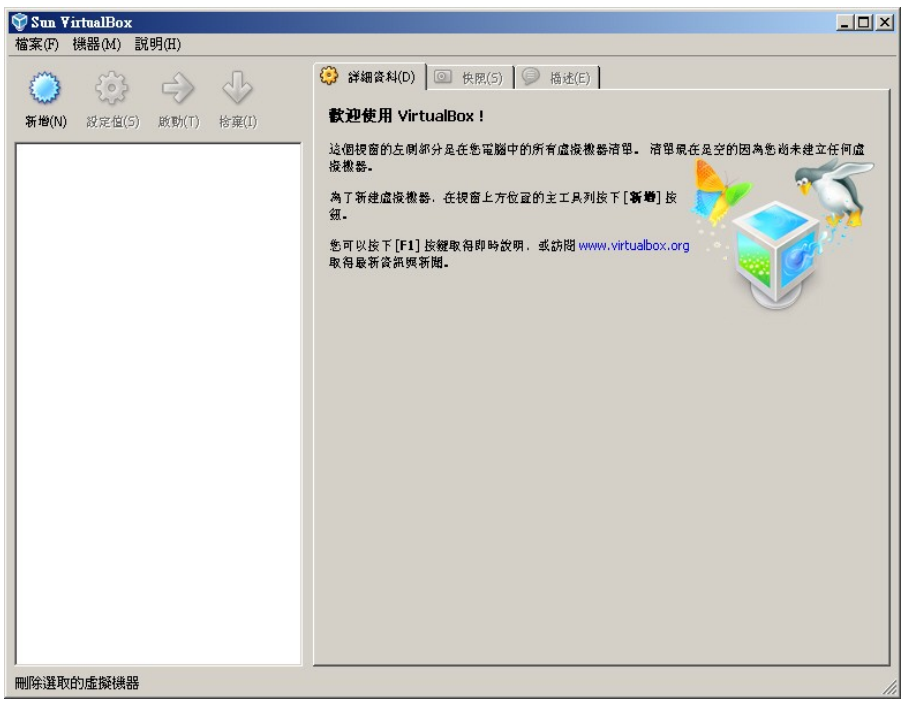


安裝 VirtalBox（3）



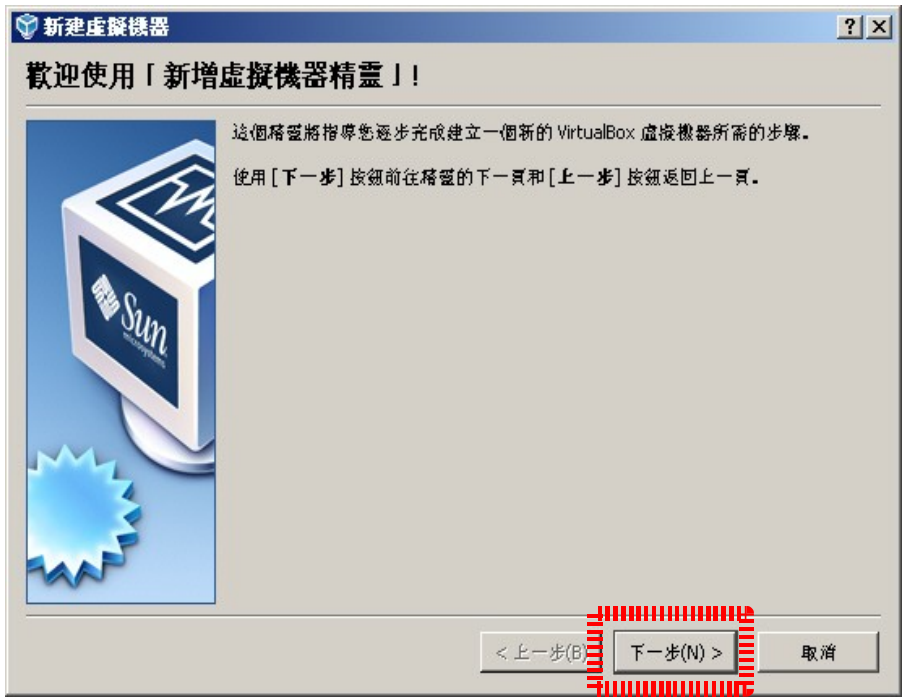


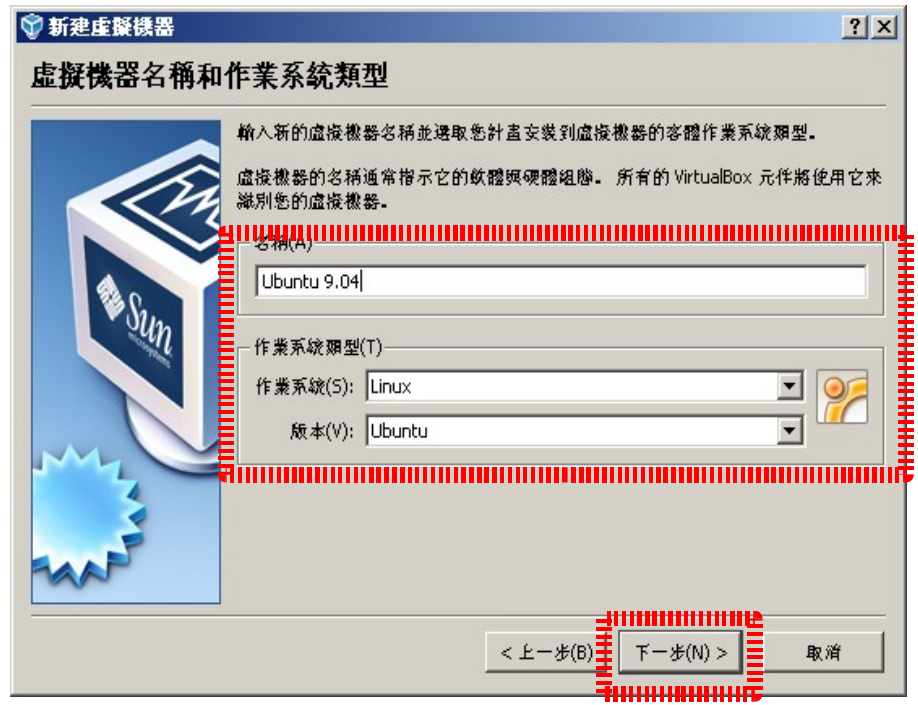


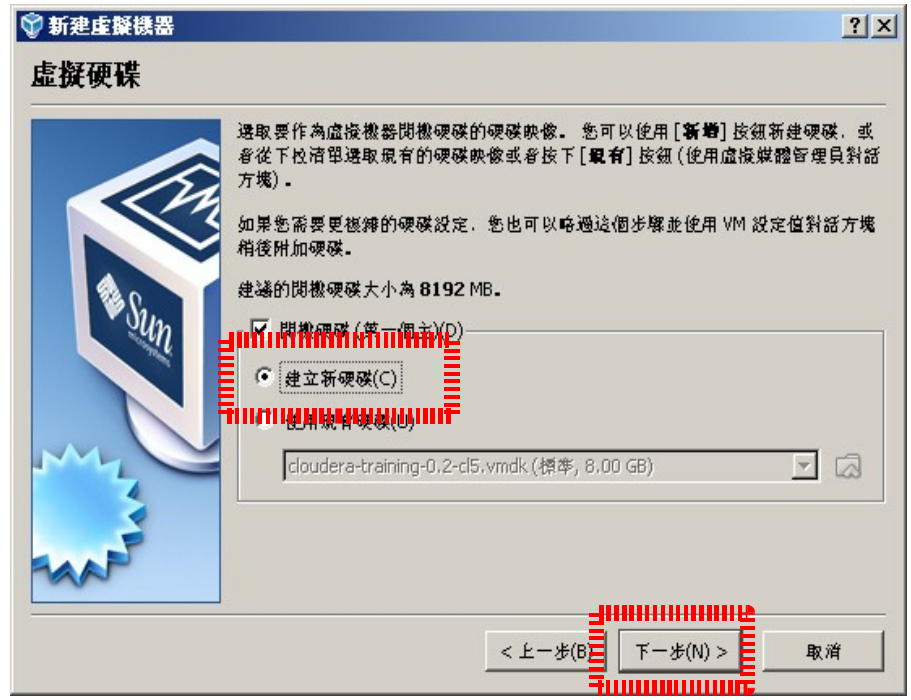


Lab 1 : 安裝 Ubuntu 9.10 GNU/Linux

安裝 Ubuntu 9.10 (1)

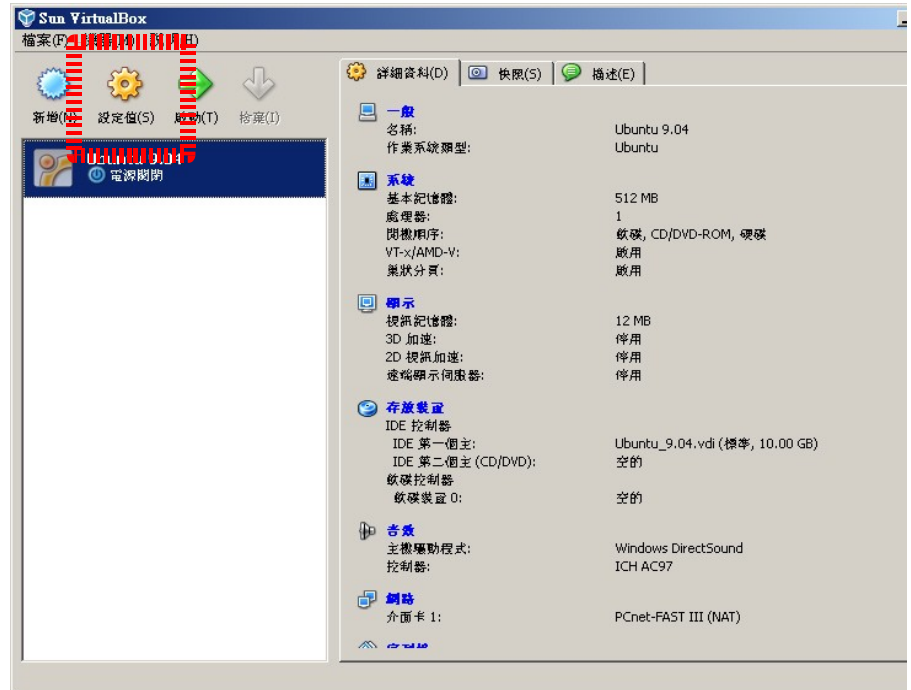


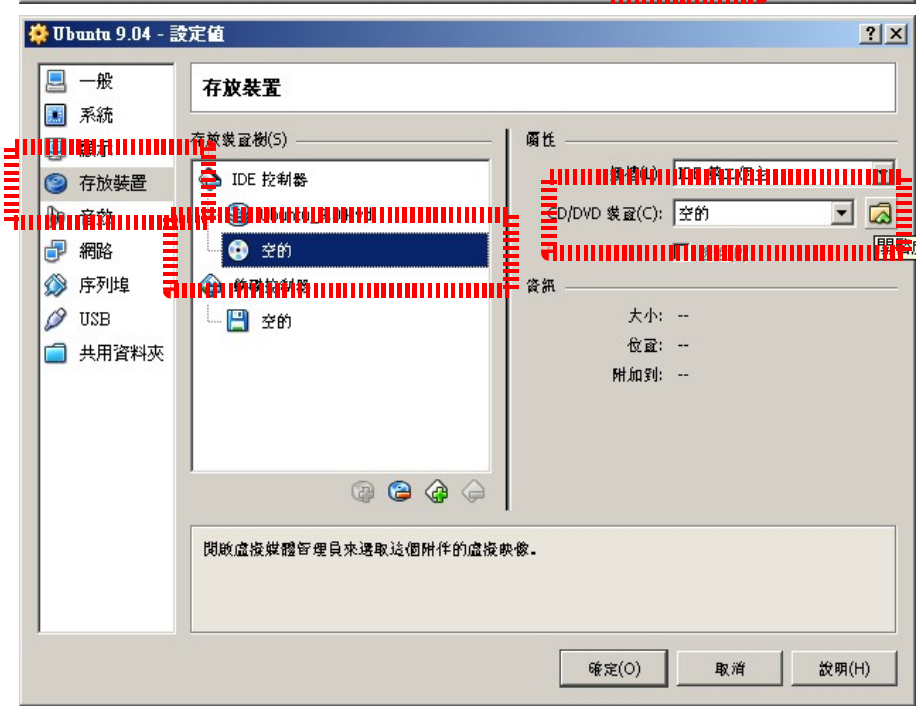


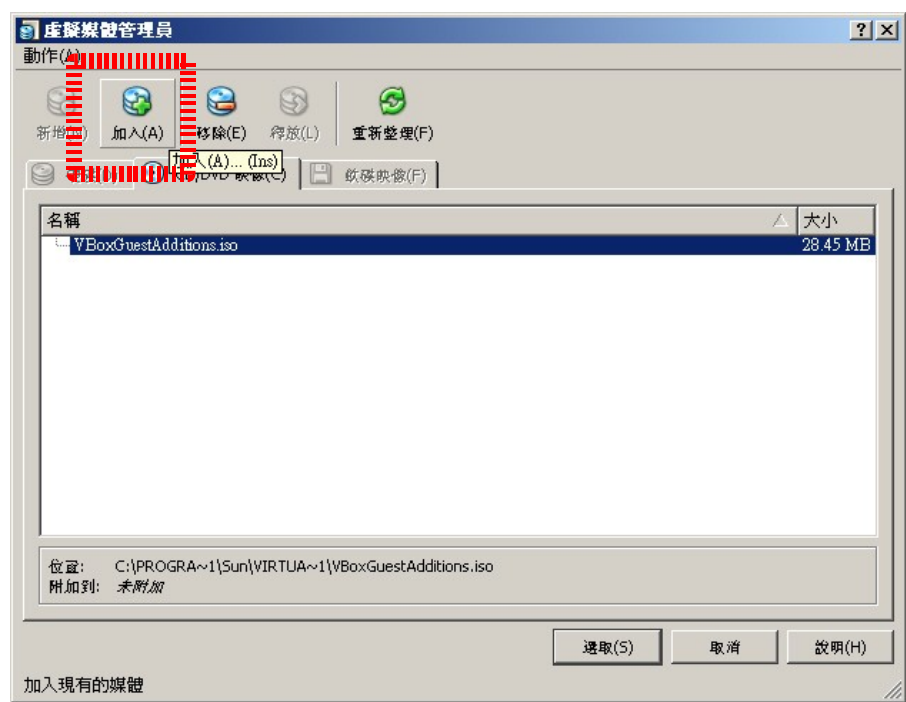


安裝 Ubuntu 9.10 (2)

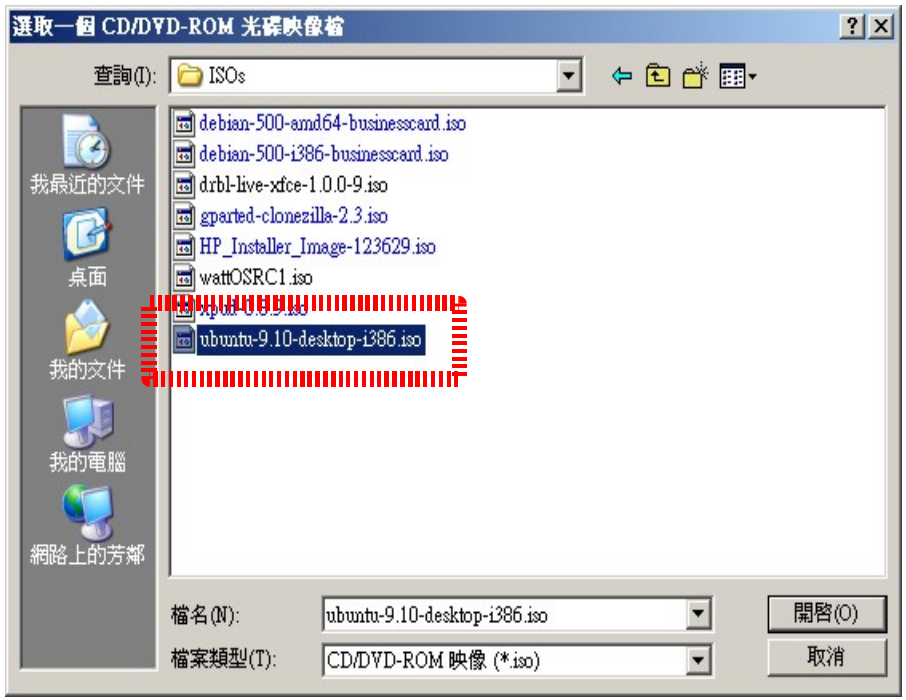
安裝 Ubuntu 9.10 (3)

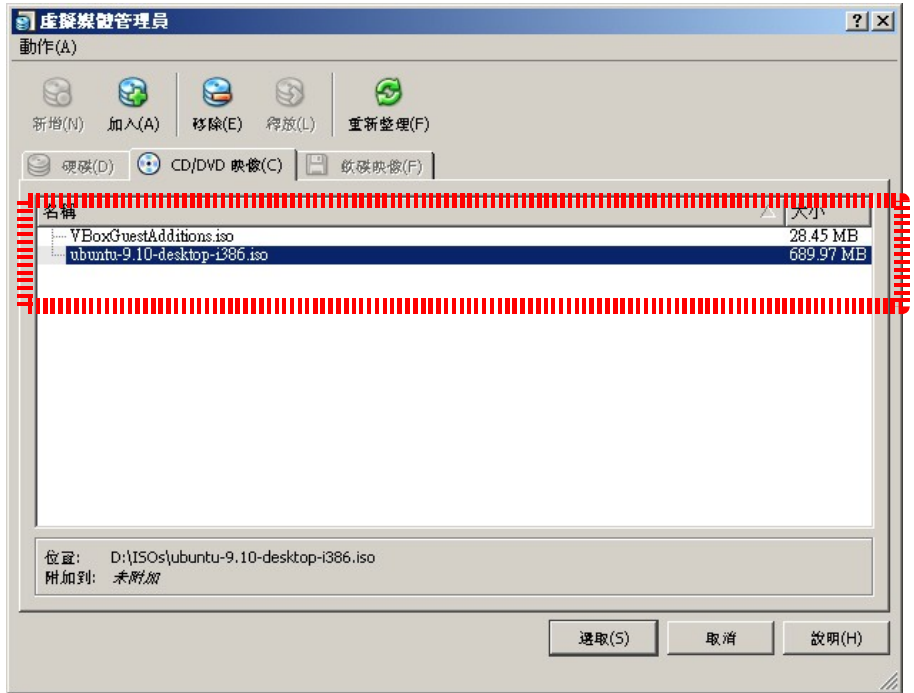


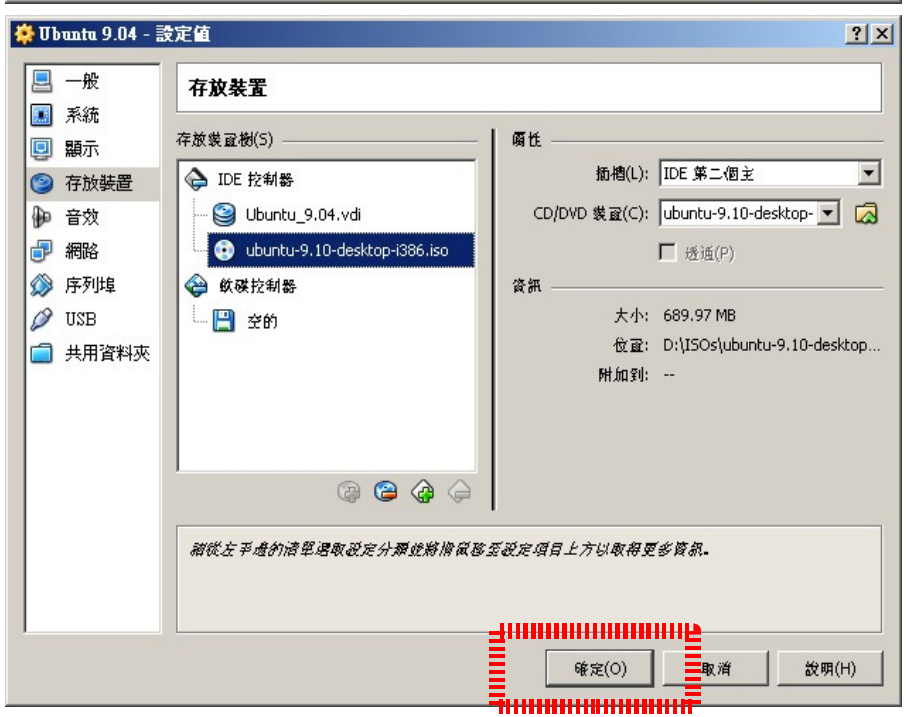


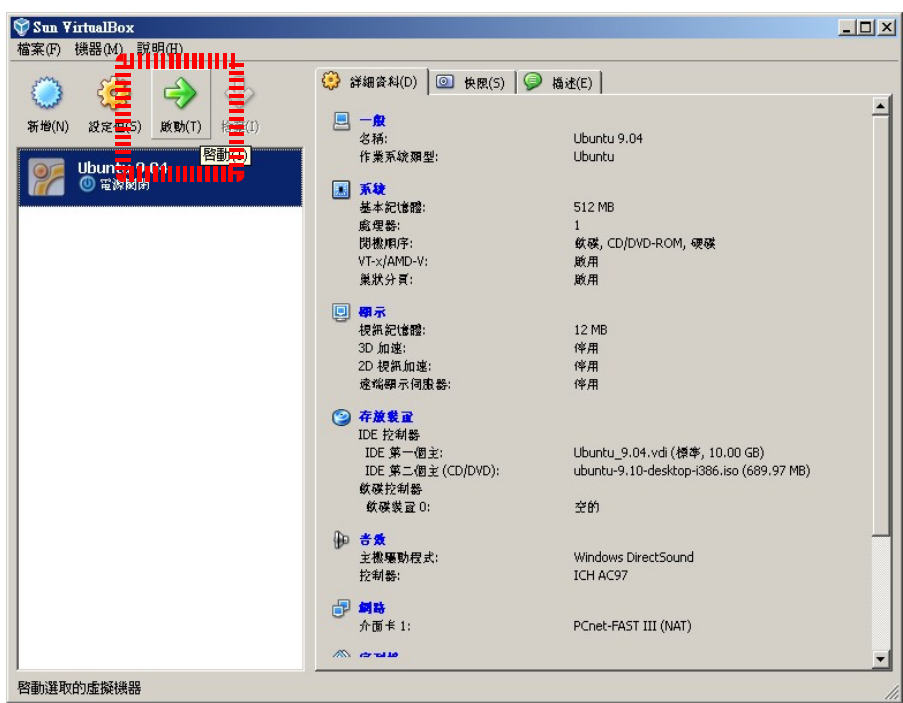


安裝 Ubuntu 9.10 (4)

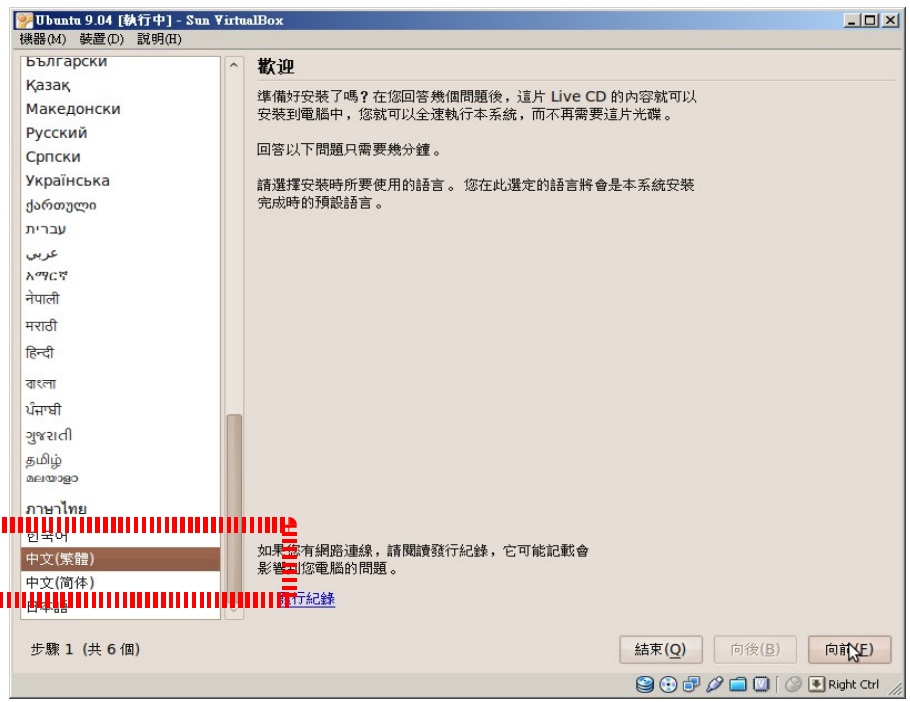


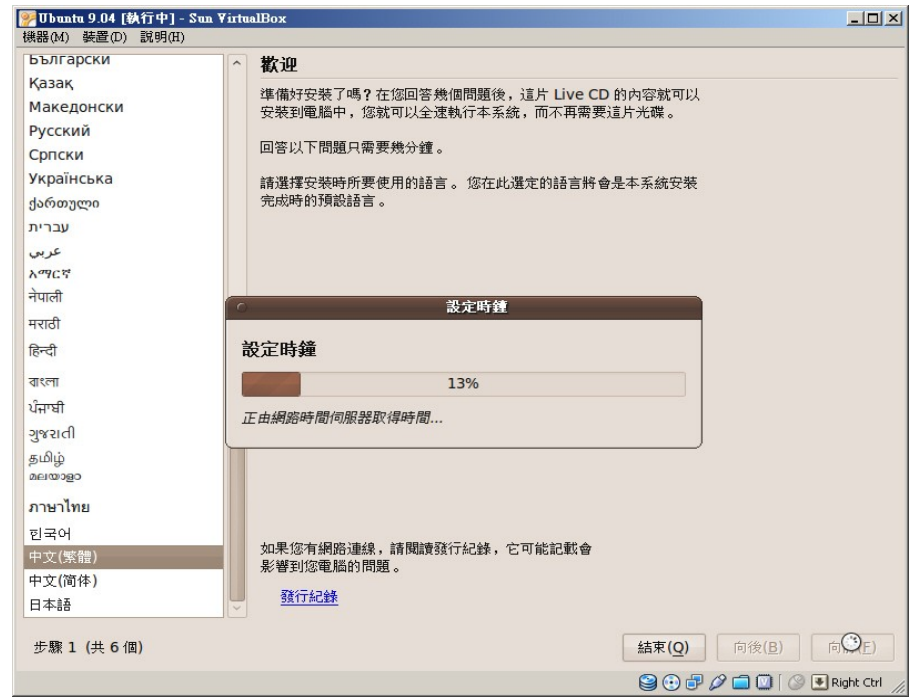


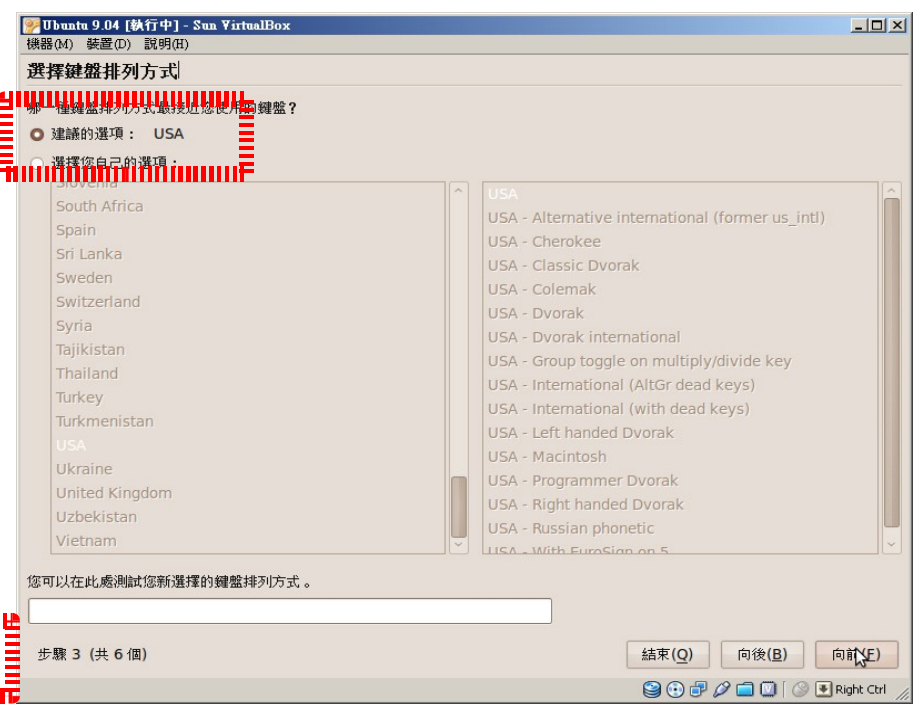




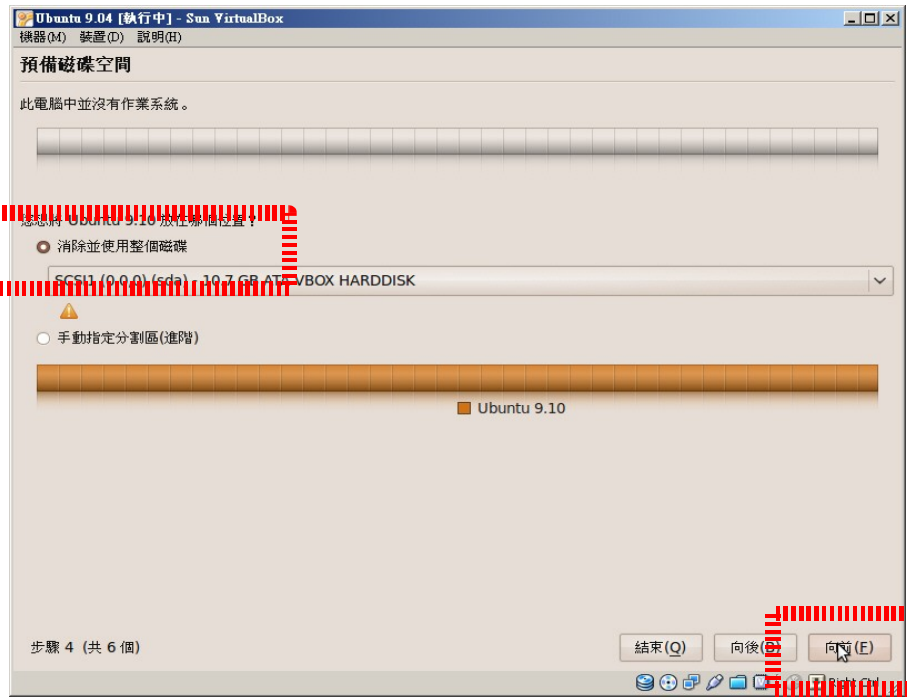
安裝 Ubuntu 9.10 (5)

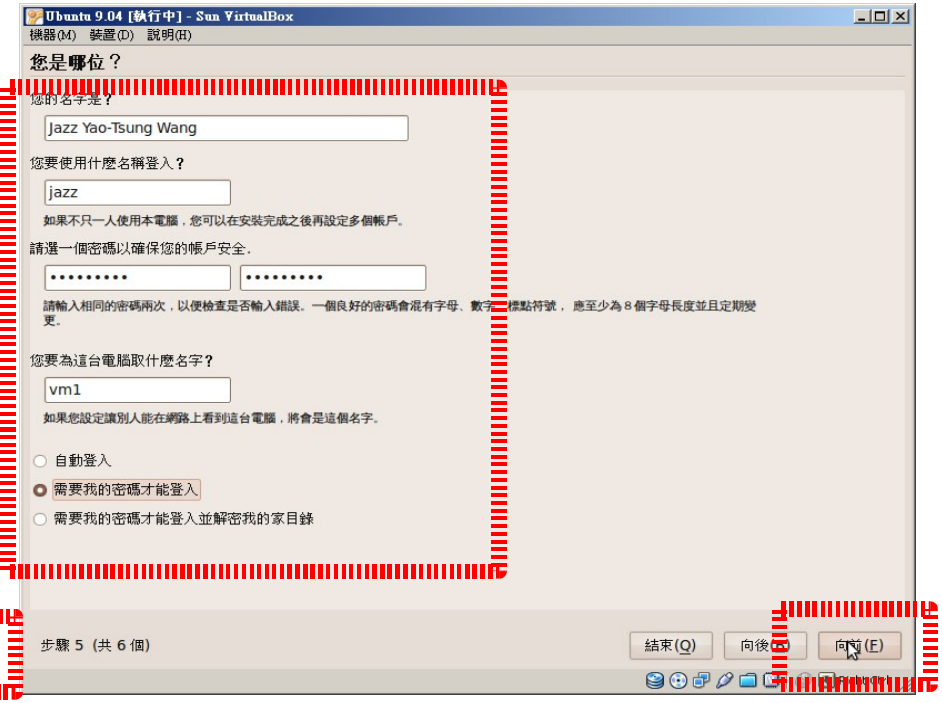


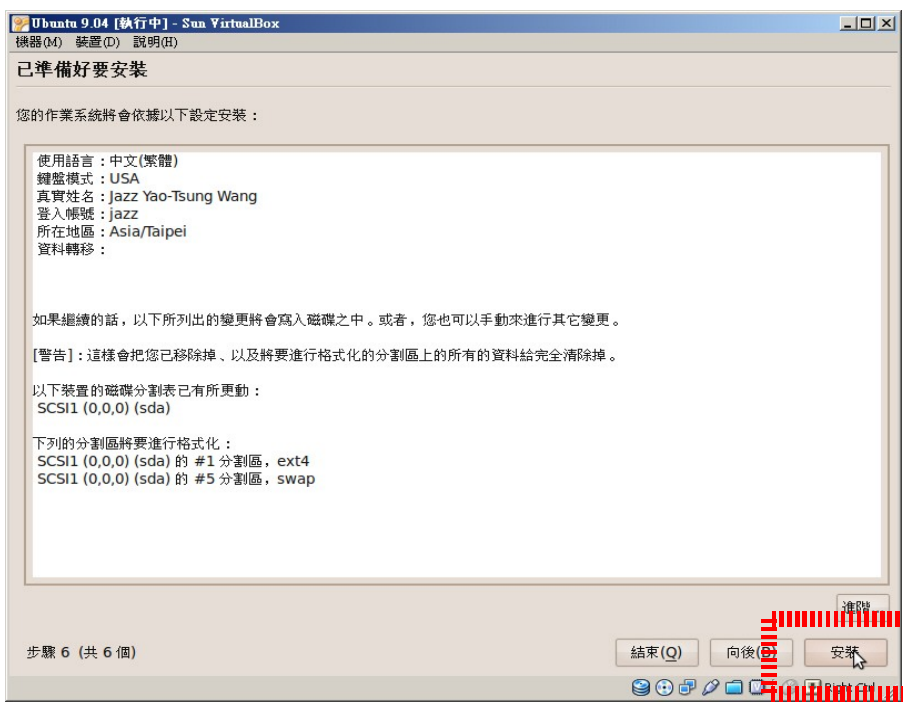




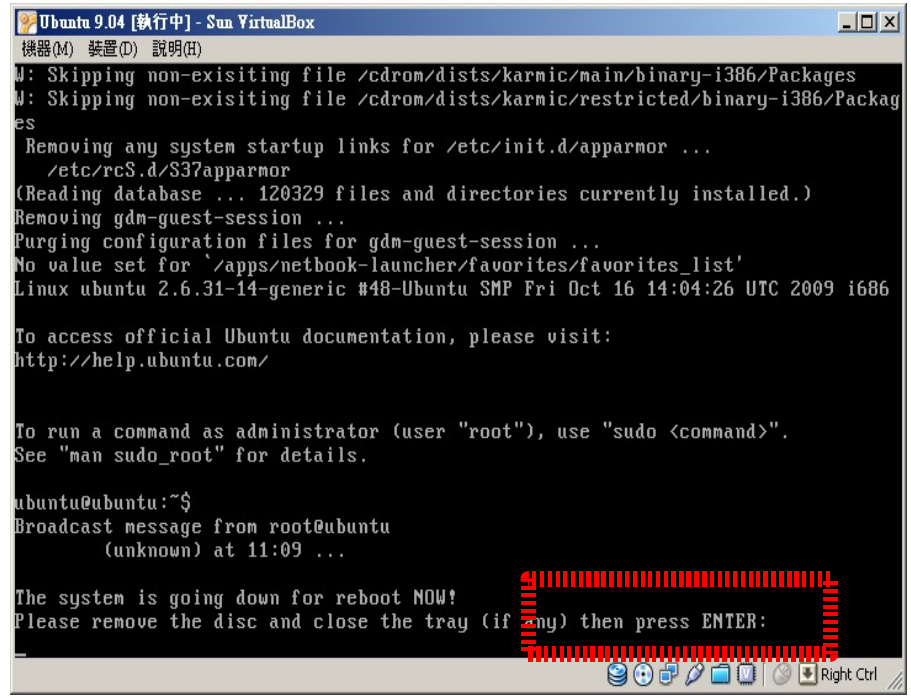
安裝 Ubuntu 9.10 (6)

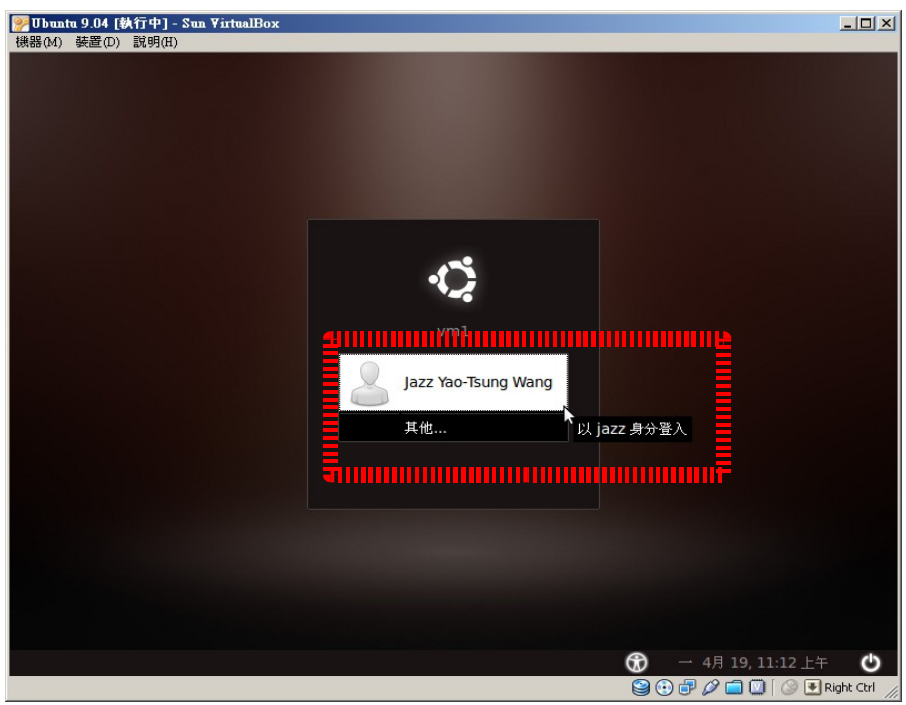


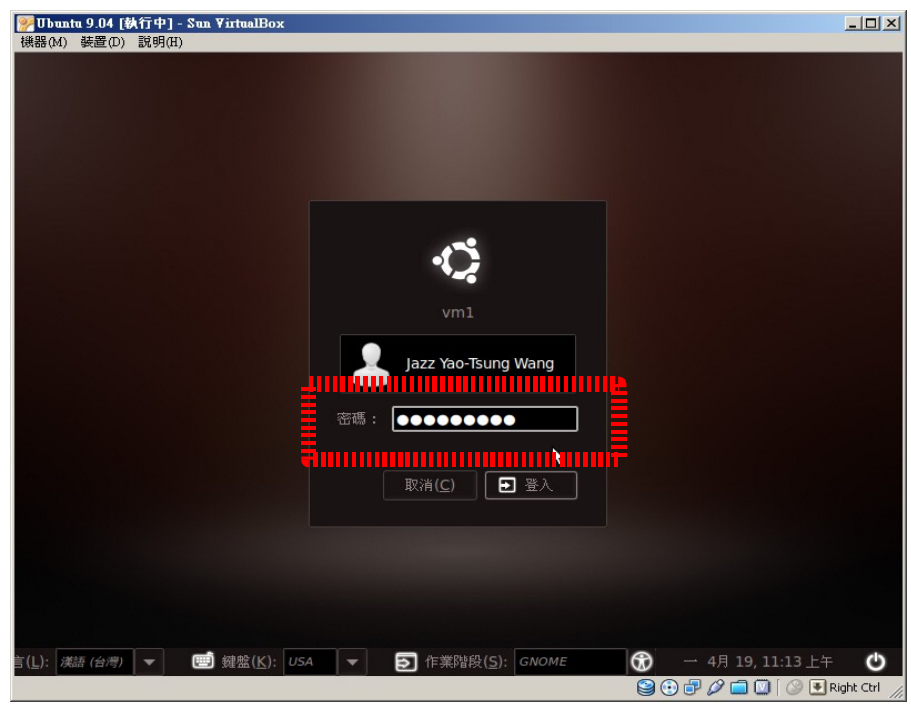




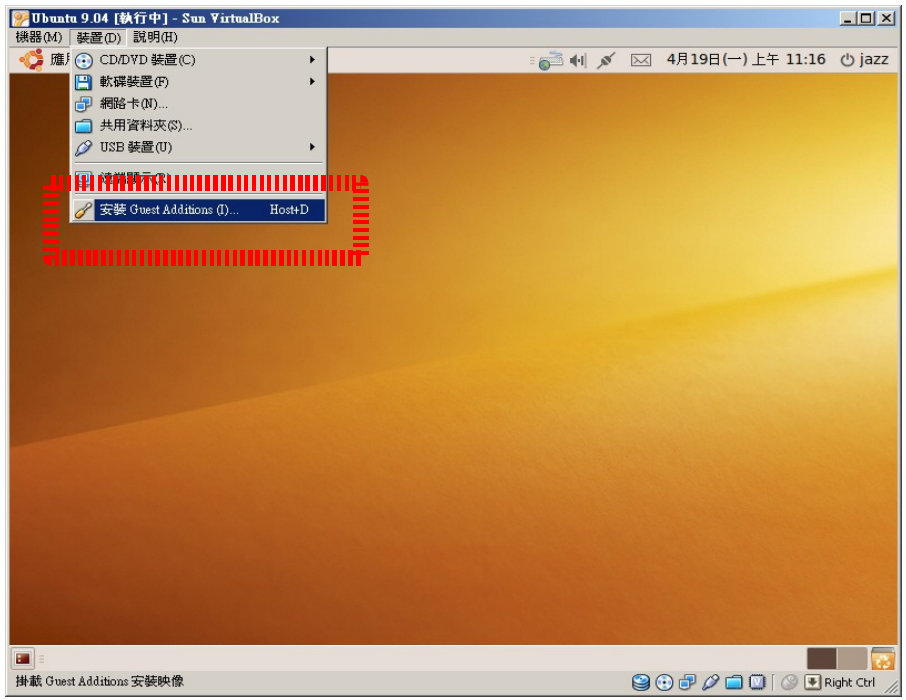
安裝 Ubuntu 9.10 (7)

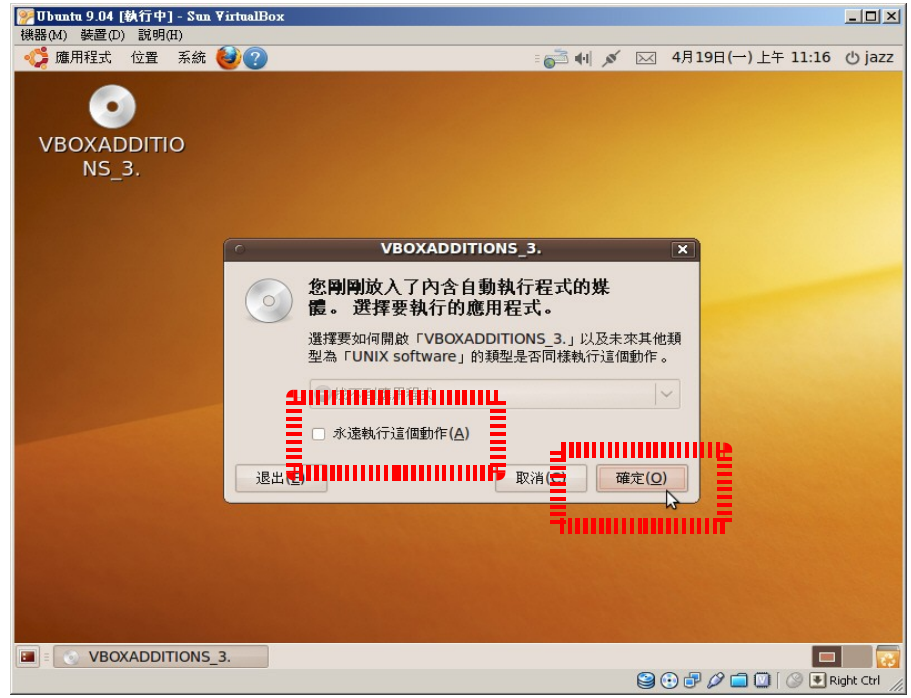


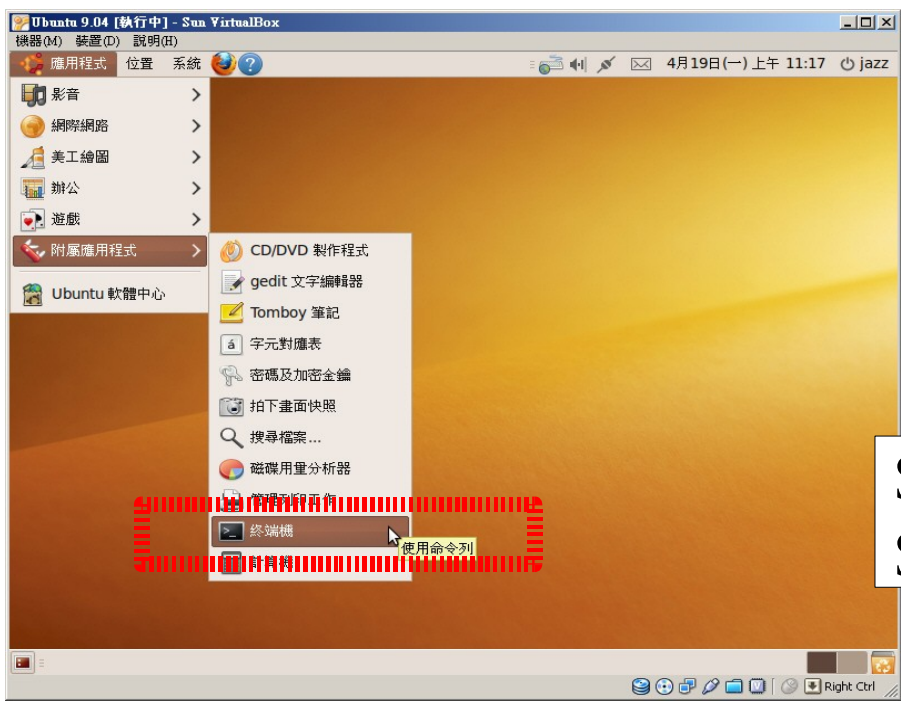


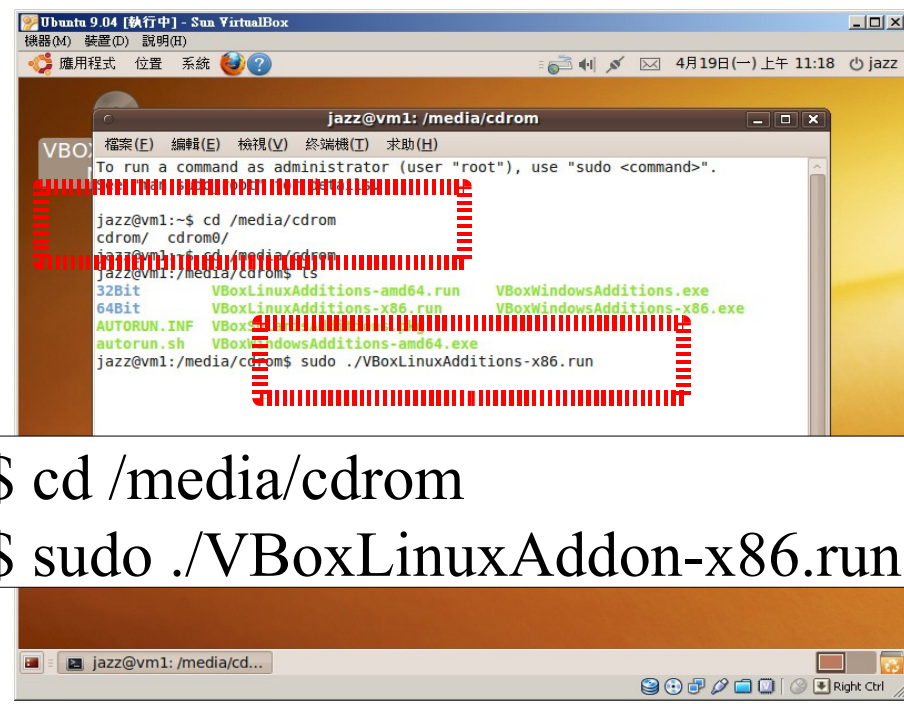


安裝 Ubuntu 9.10 (8)



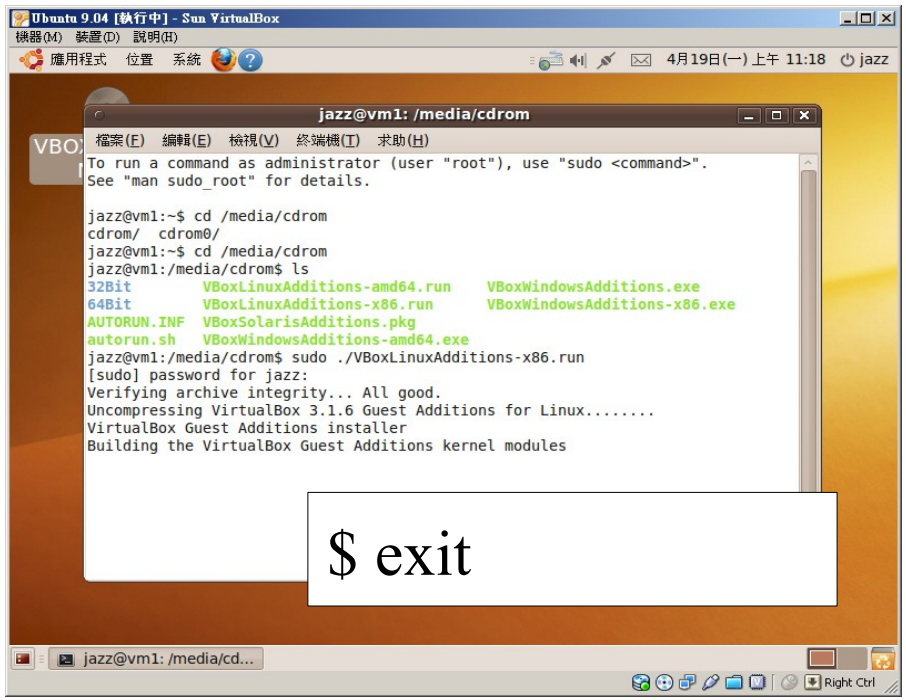


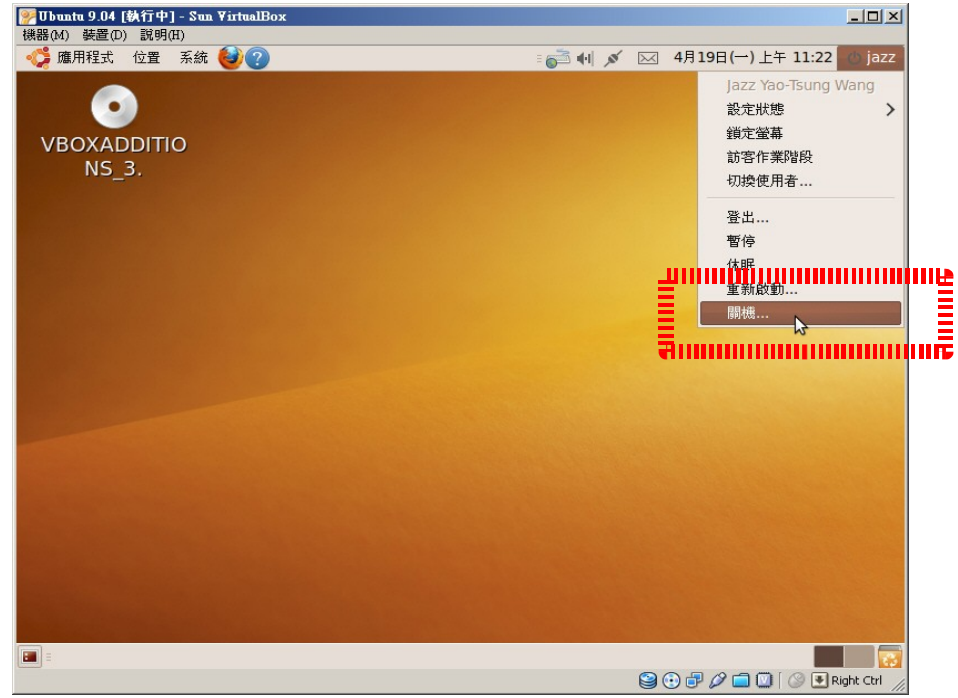





```
$ cd /media/cdrom
$ sudo ./VBoxLinuxAddon-x86.run
```

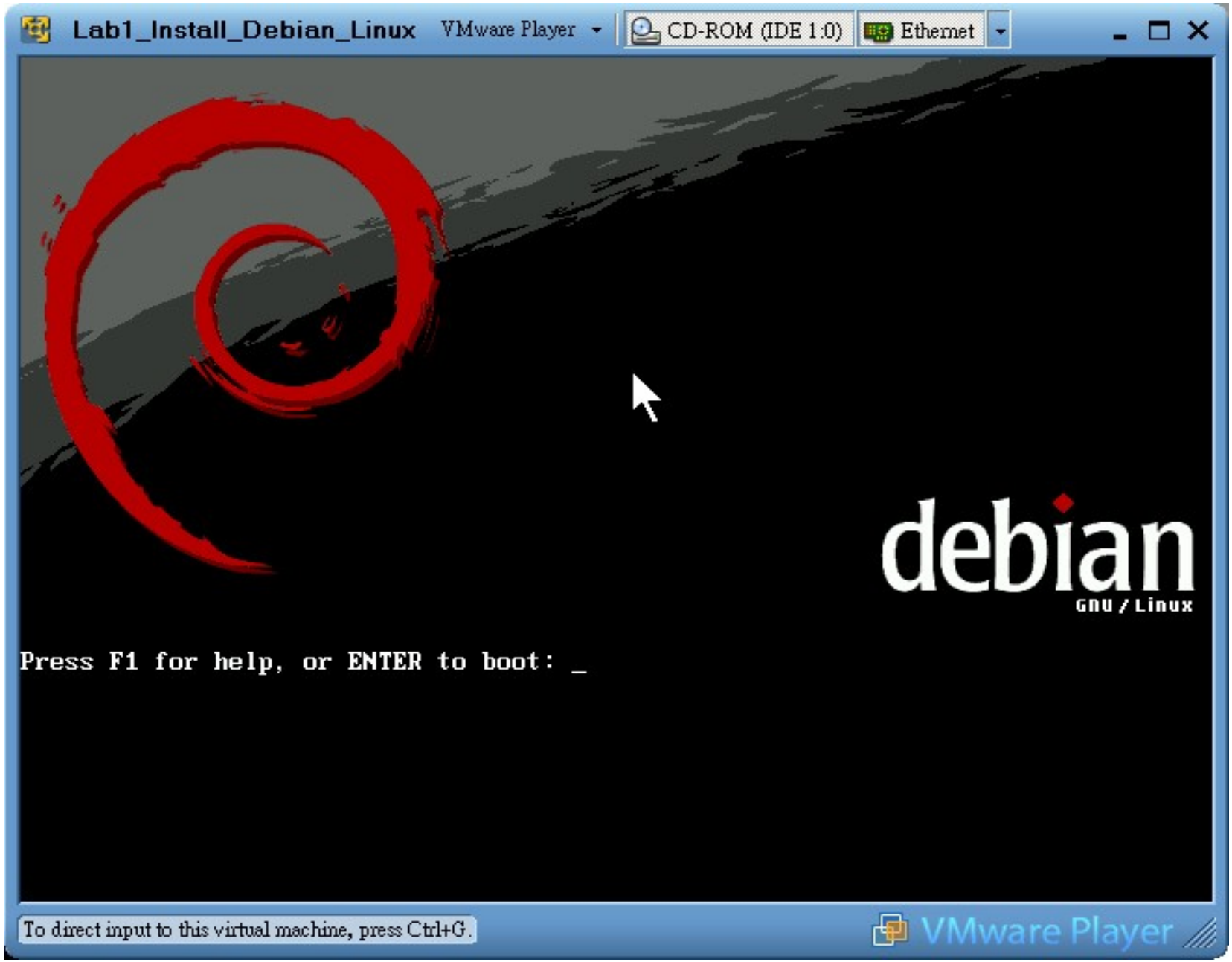
安裝 Ubuntu 9.10 (9)



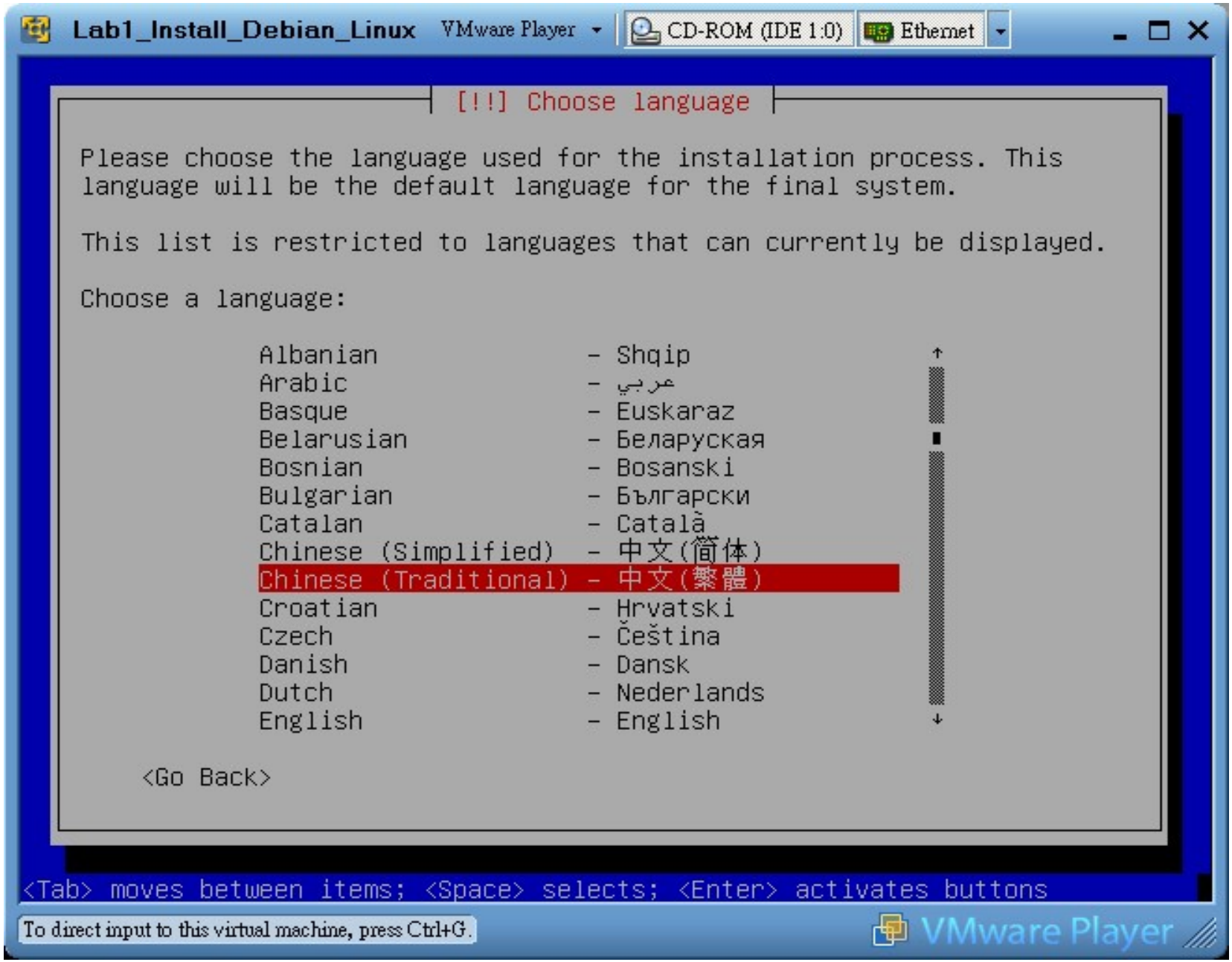


Appendex A :
安裝 Debian 4.0r3 GNU/Linux

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』

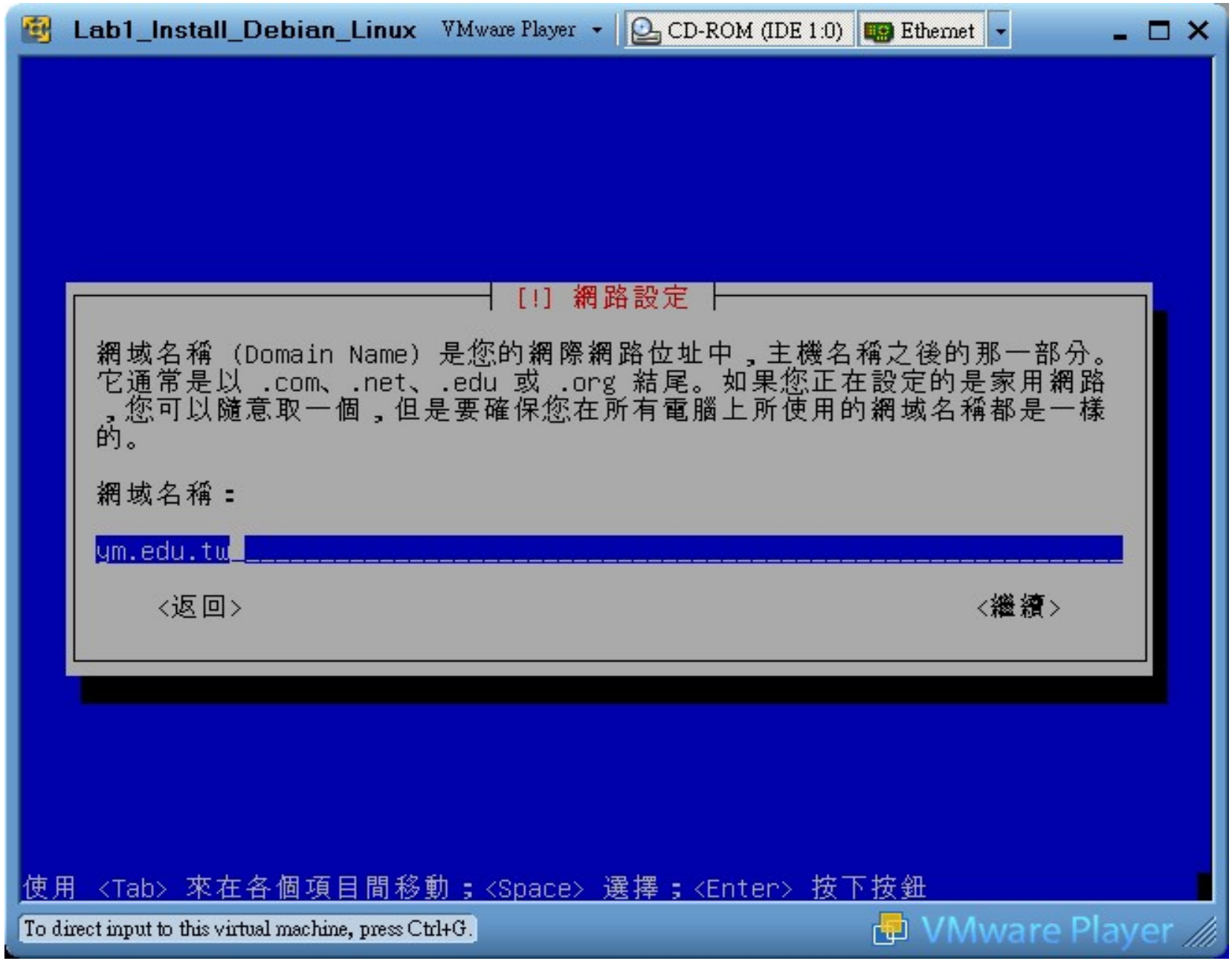

Lab1_Install_Debian_Linux
VMware Player
CD-ROM (IDE 1:0)
Ethernet
[!!] Choose language
根據您所使用的語言，您可能位於以下的國家或地區之一。
請選擇一個國家、領域或地區：
中國
台灣
新加坡
香港
其它
<返回>
使用 <Tab> 來在各個項目間移動；<Space> 選擇；<Enter> 按下按鈕
To direct input to this virtual machine, press Ctrl+G.
VMware Player

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

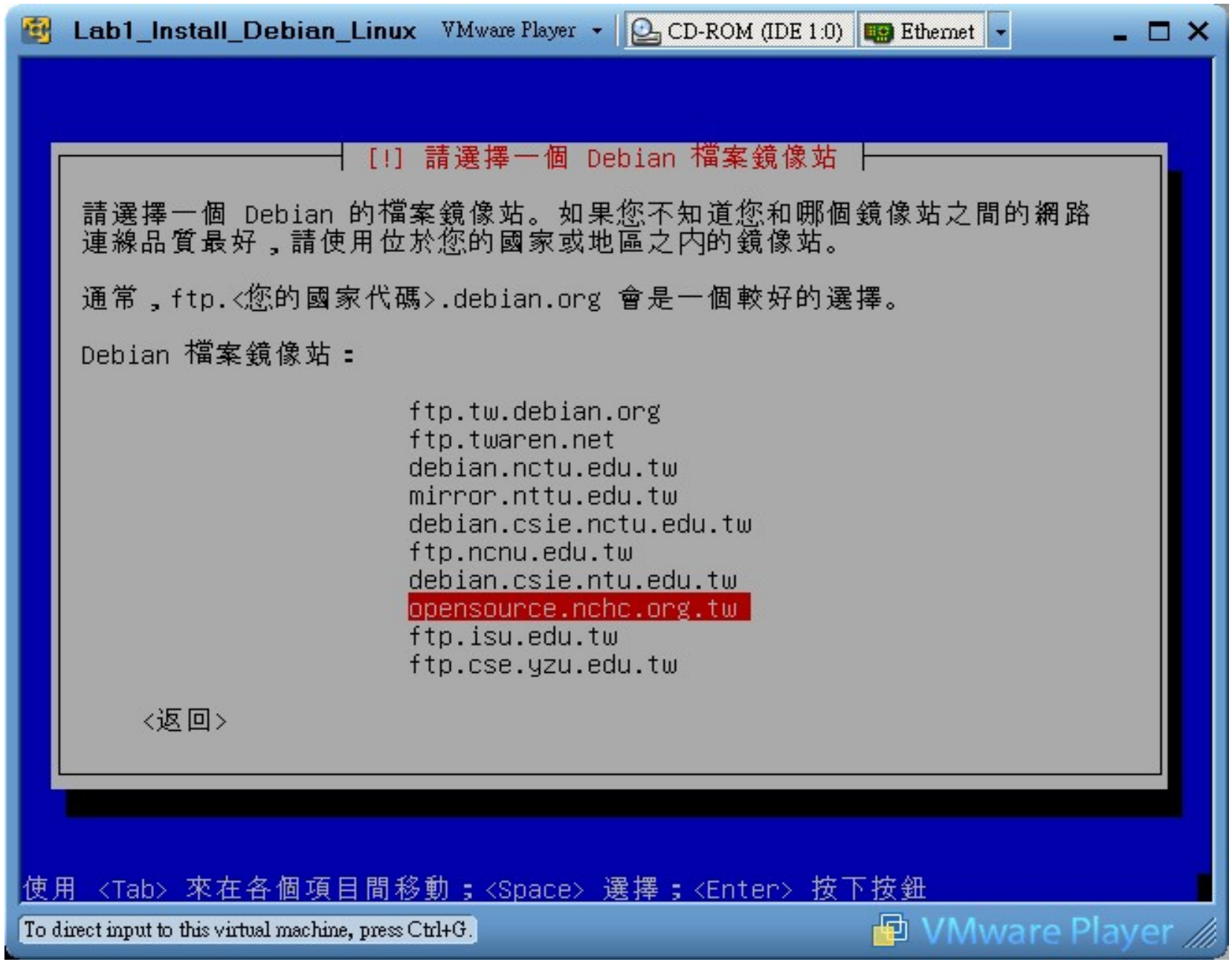
輸入『網域名稱』 = ym.edu.tw



選擇『 Debian 檔案映象站』(1) 台灣區

選擇『 Debian 檔案映象站』(2)

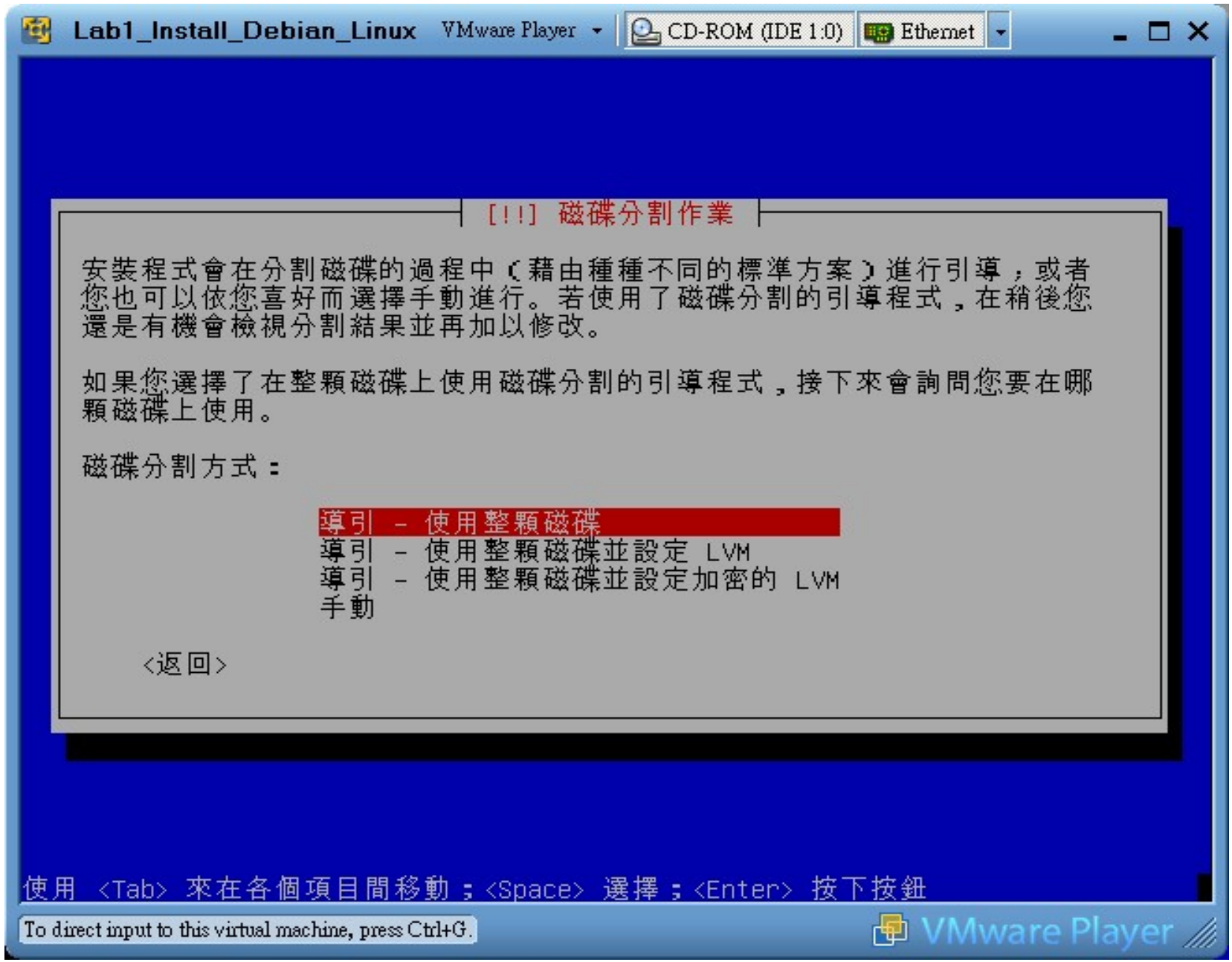
推薦 ftp.twaren.net 或 free.nchc.org.tw



選擇『 Debian 檔案映象站』(3)

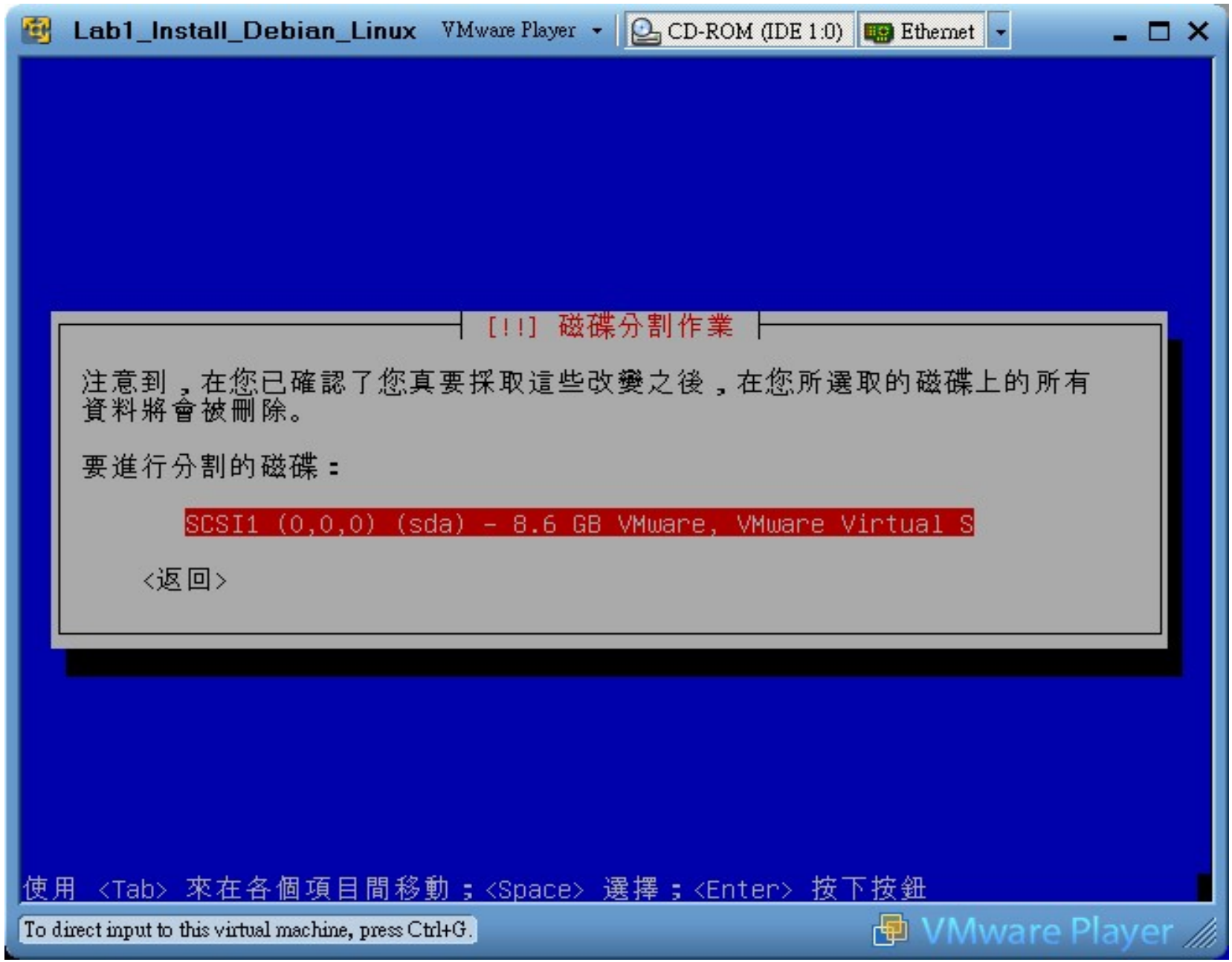
通常不需要輸入 Proxy ，除非網路環境有限制

選擇『磁碟分割方式』
初學者建議採用第一種『引導 - 整顆硬碟』

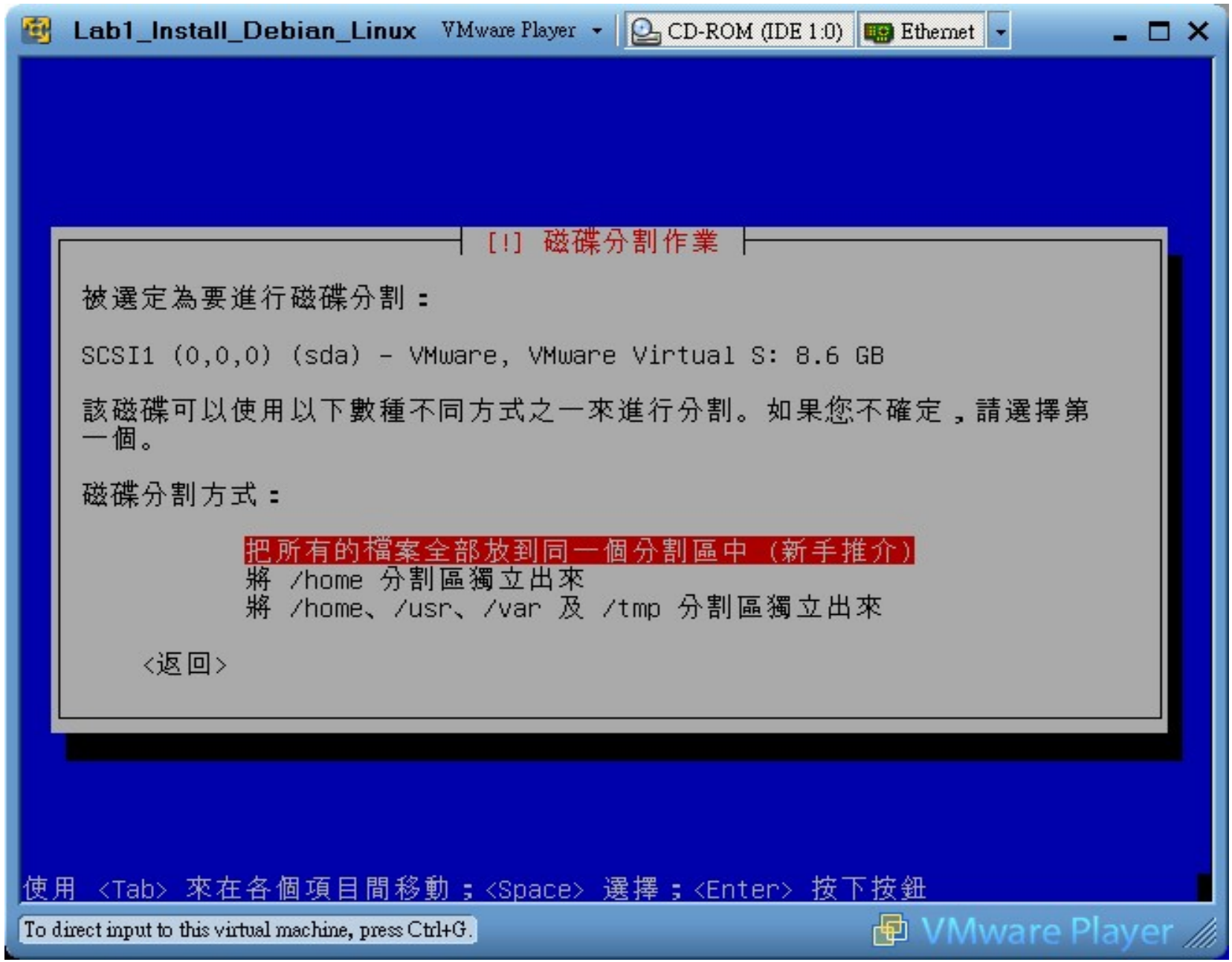


選擇『欲分割磁碟』

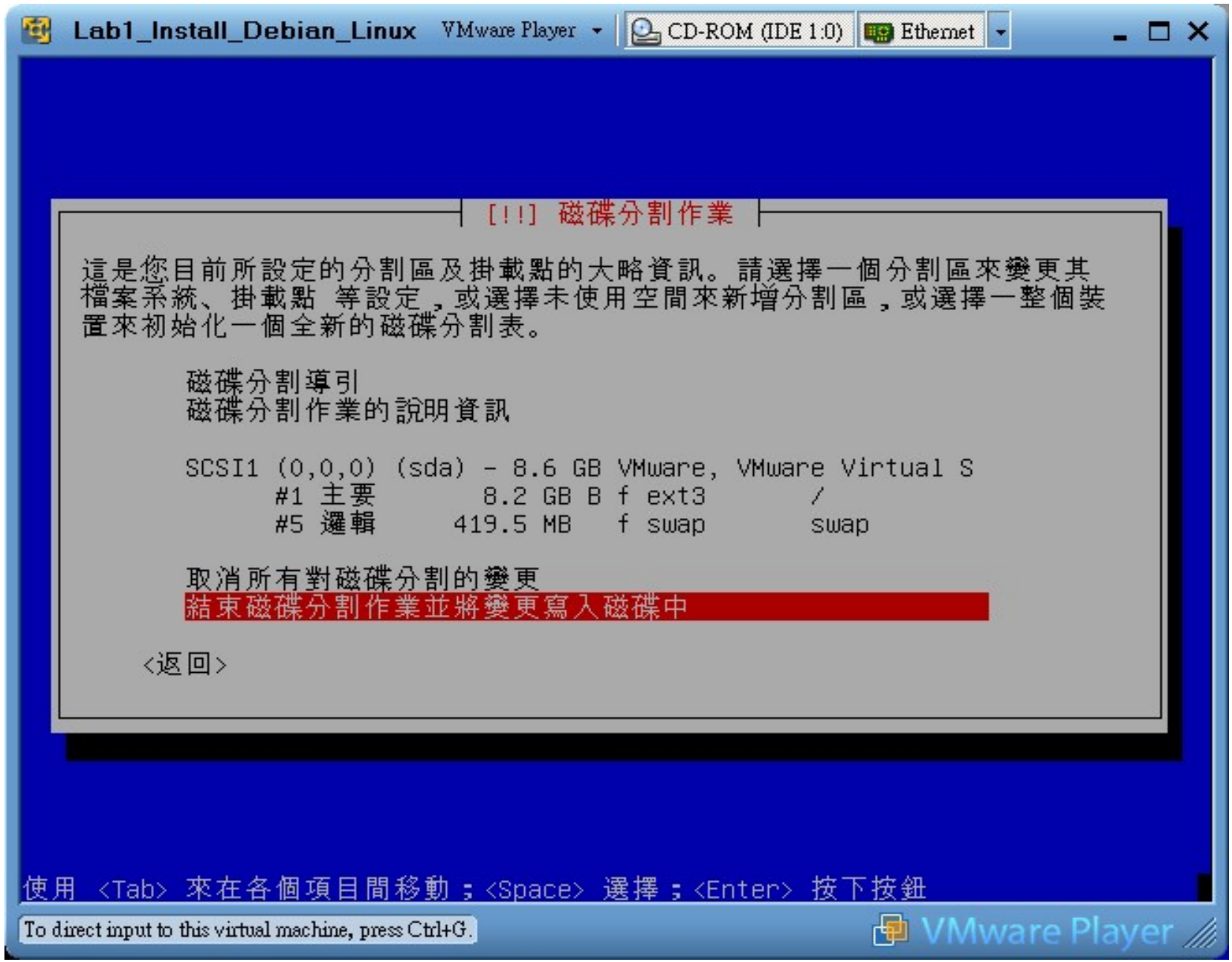
目前只有虛擬一顆硬碟，若實機有兩顆請小心！



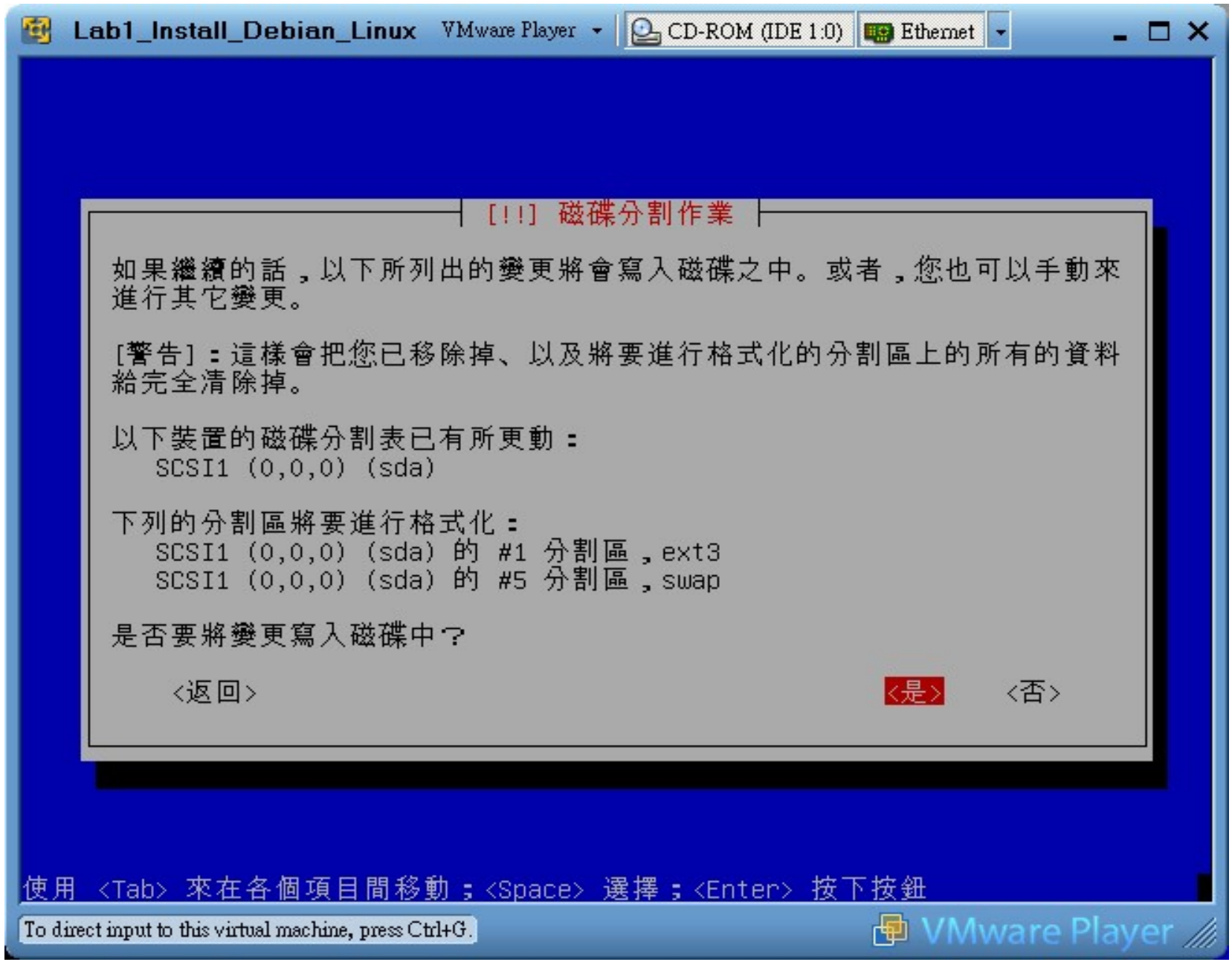
練習時可用同一個分割區，
服務用系統請至少獨立出 /home 分割區



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)



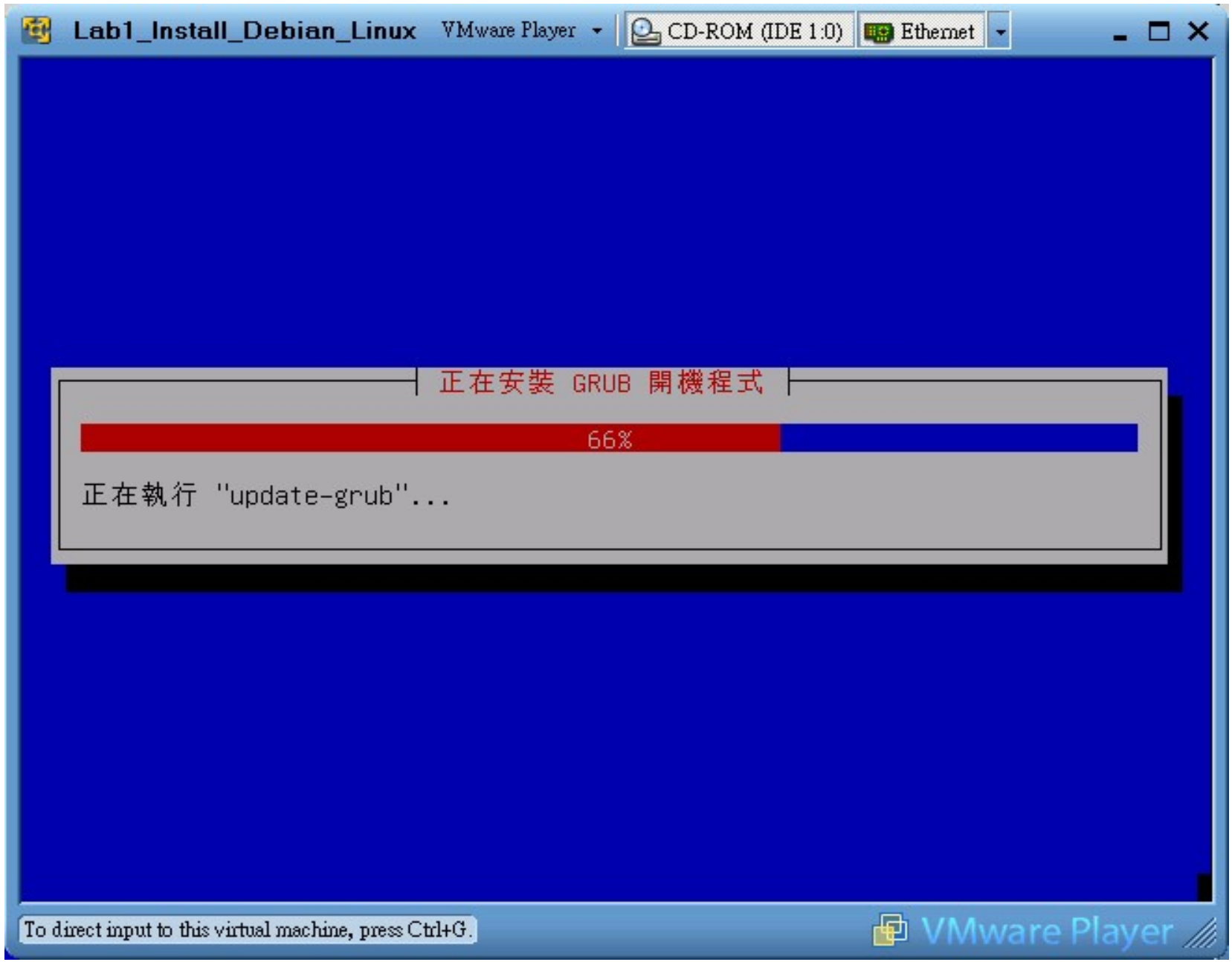
確認將目前的磁區分割表寫入硬碟
請用 TAB 選擇 < 是 >



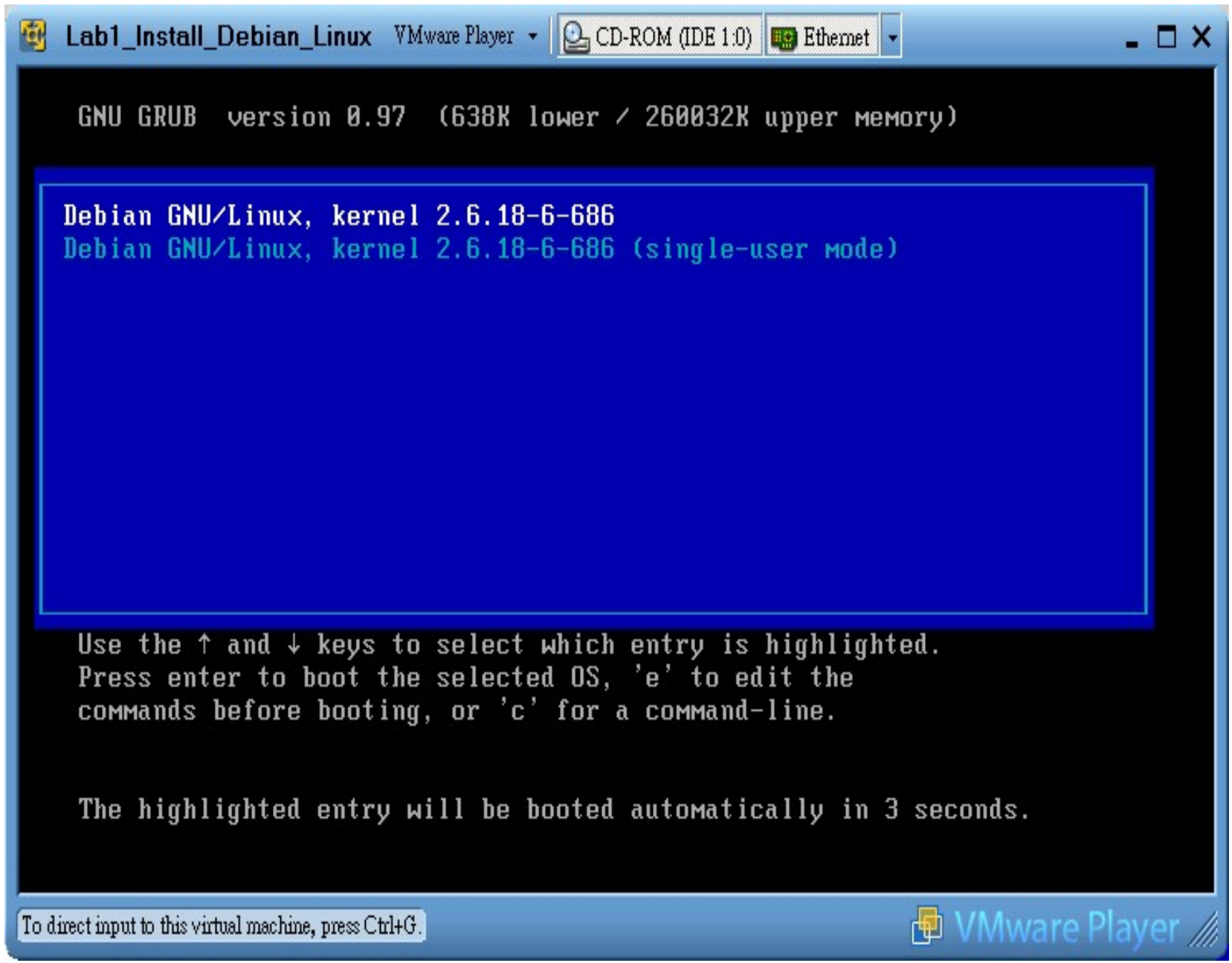
設定最高權限 root 管理者密碼

建立自己的帳號，習慣上不用 root 登入

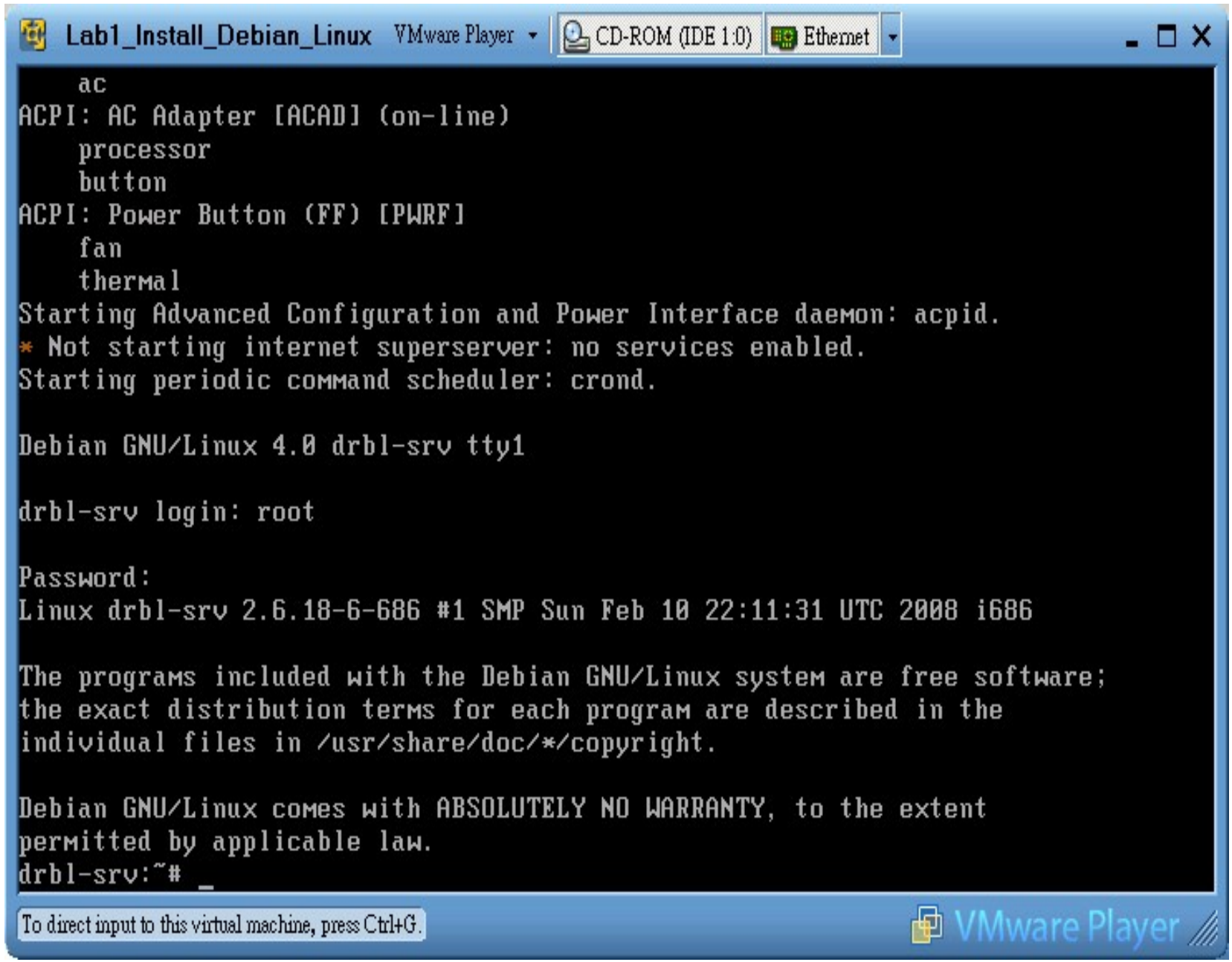
確認安裝 Grub Boot Loader



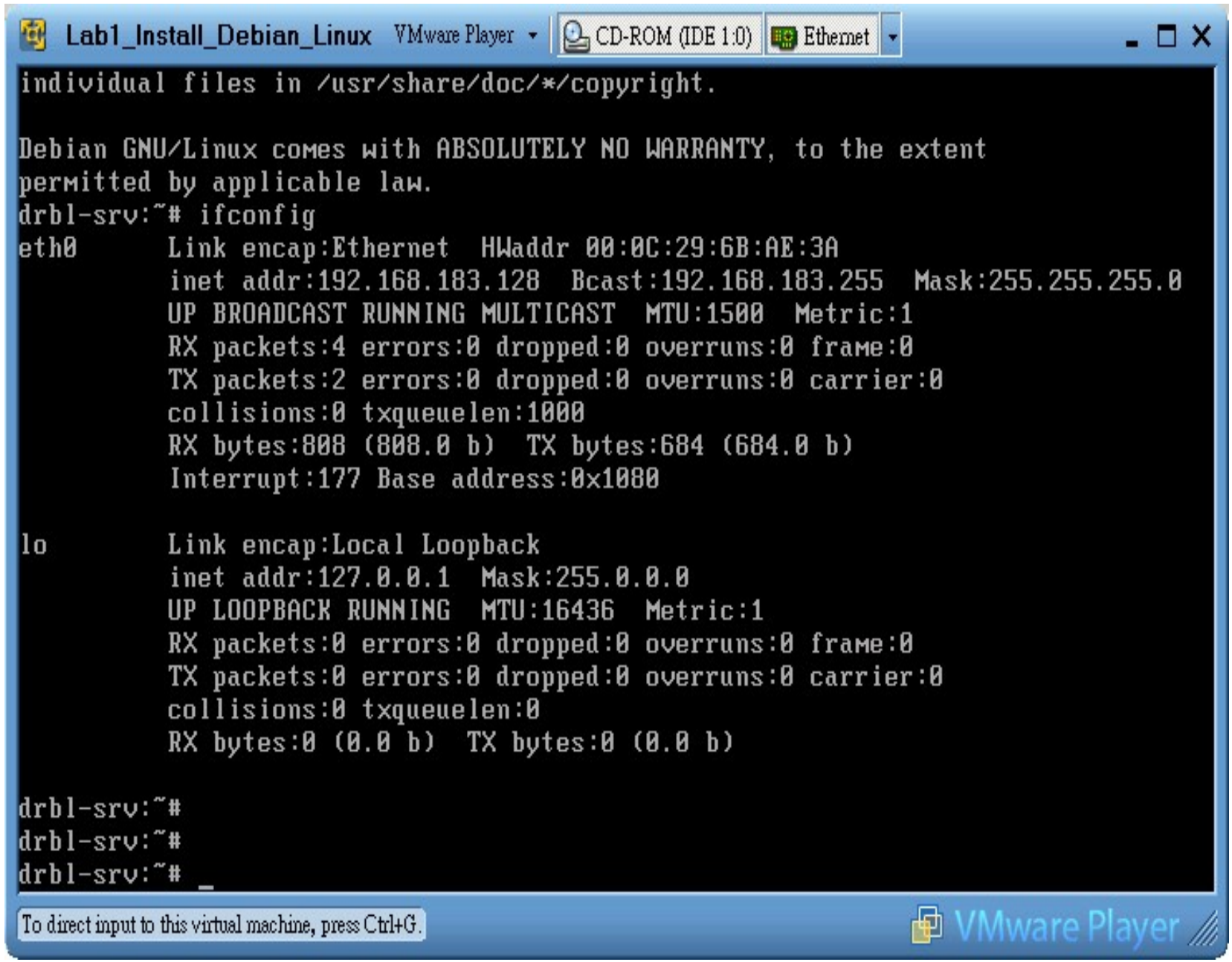
第一次系統開機


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

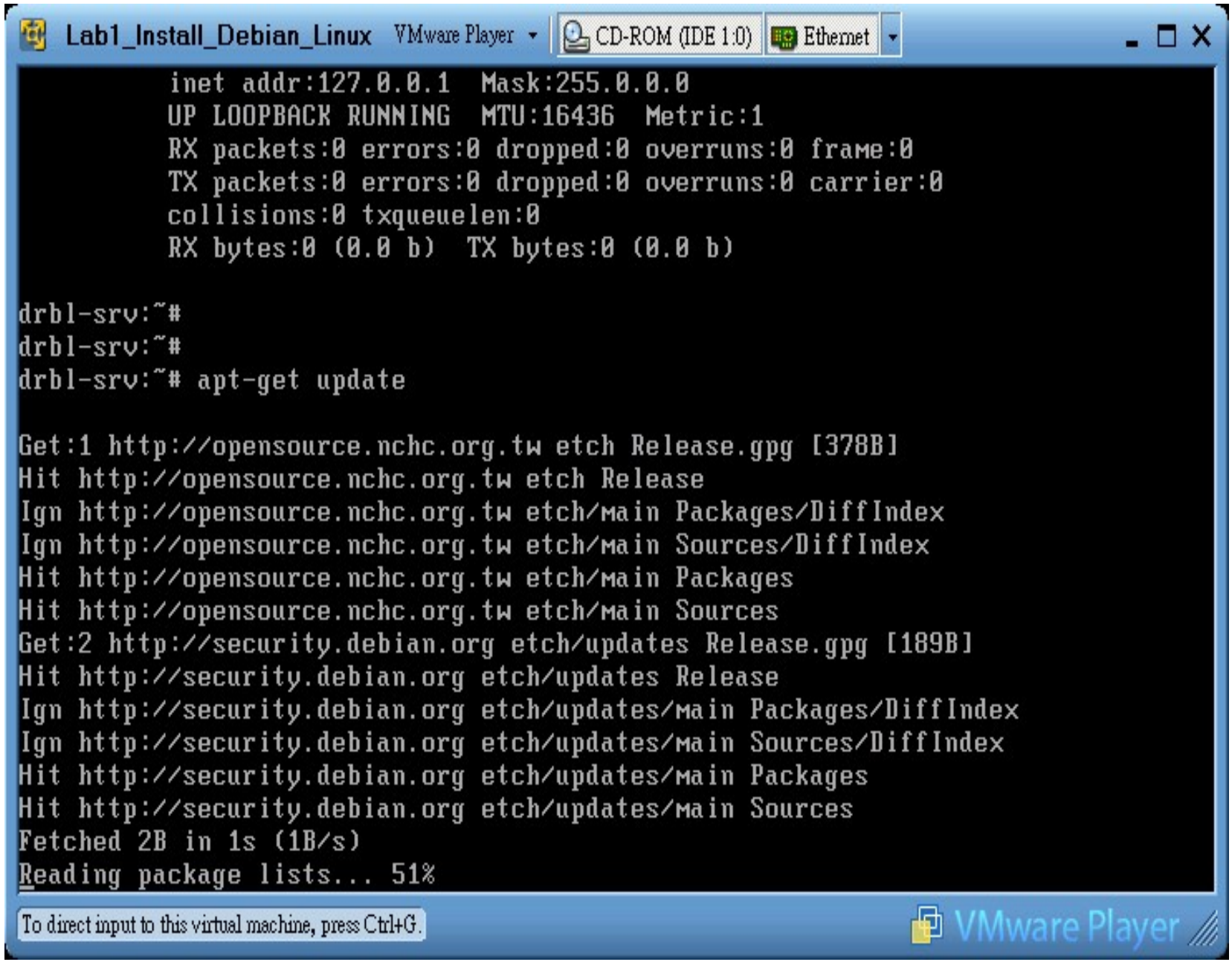
首次可使用 root 登入，
日後建議用一般身分登入



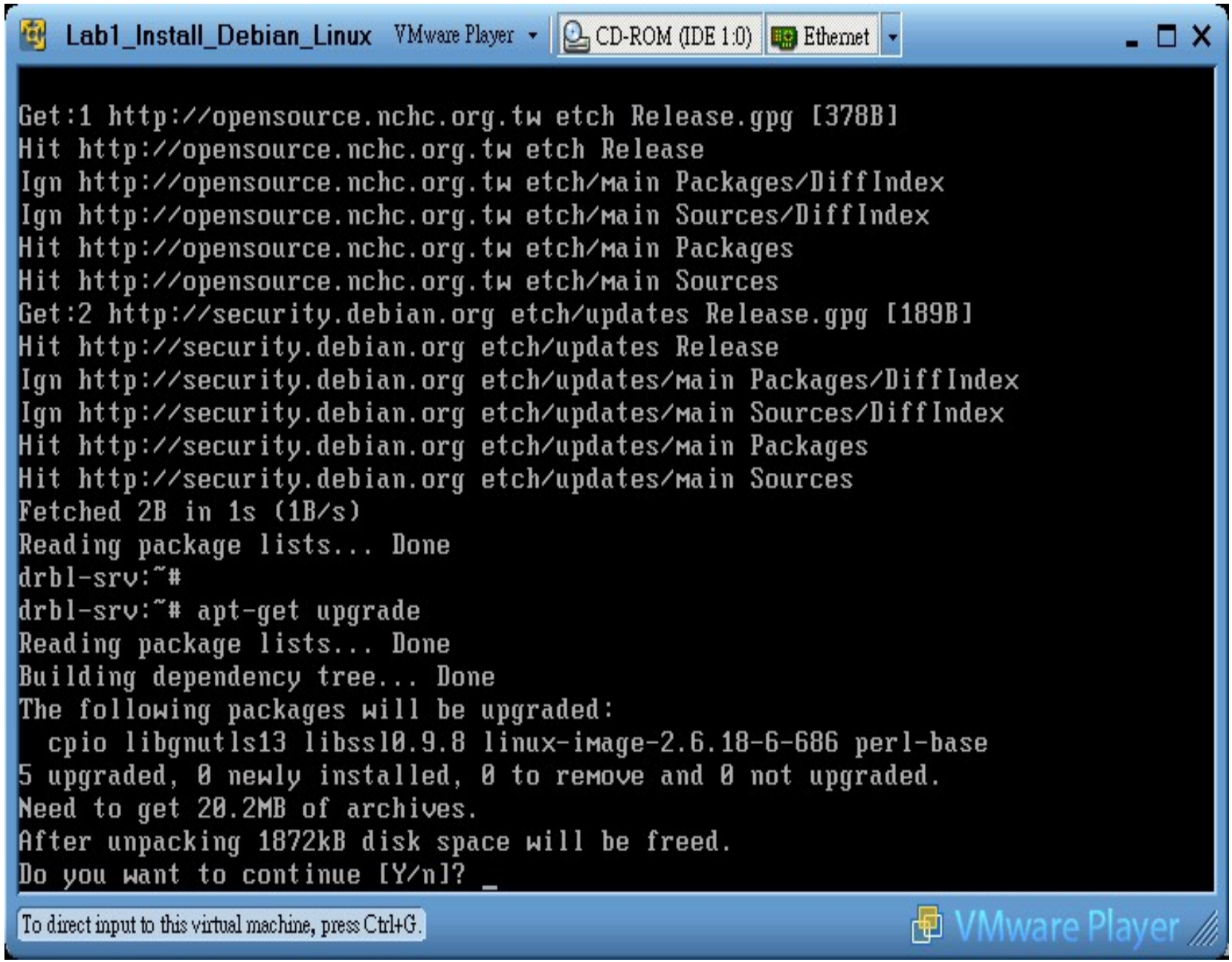
基本管理篇 [1] 用 ifconfig 確認網路連線



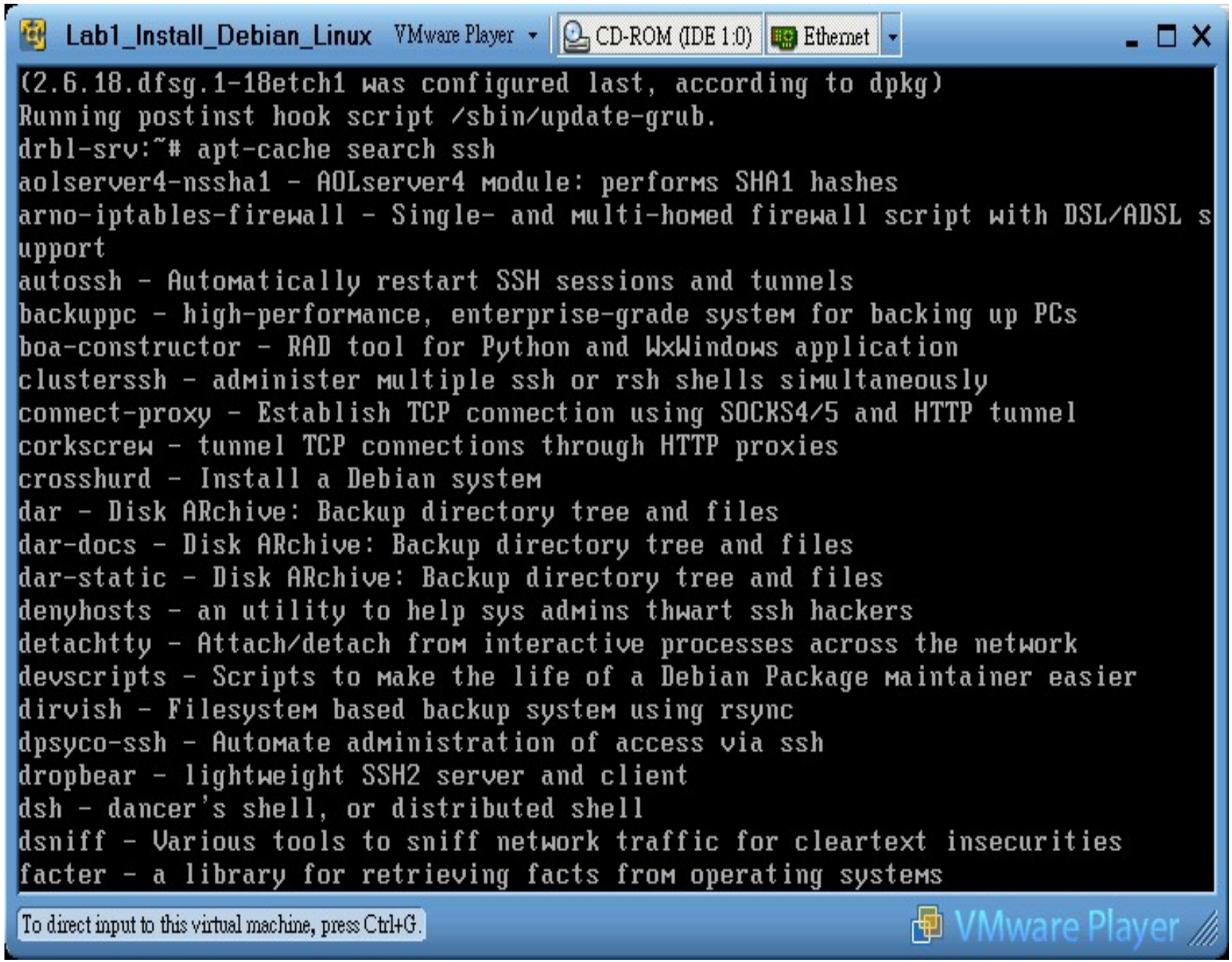
基本管理篇 [2] 用 apt-get update 更新套件



基本管理篇 [3] 用 apt-get upgrade 升級系統

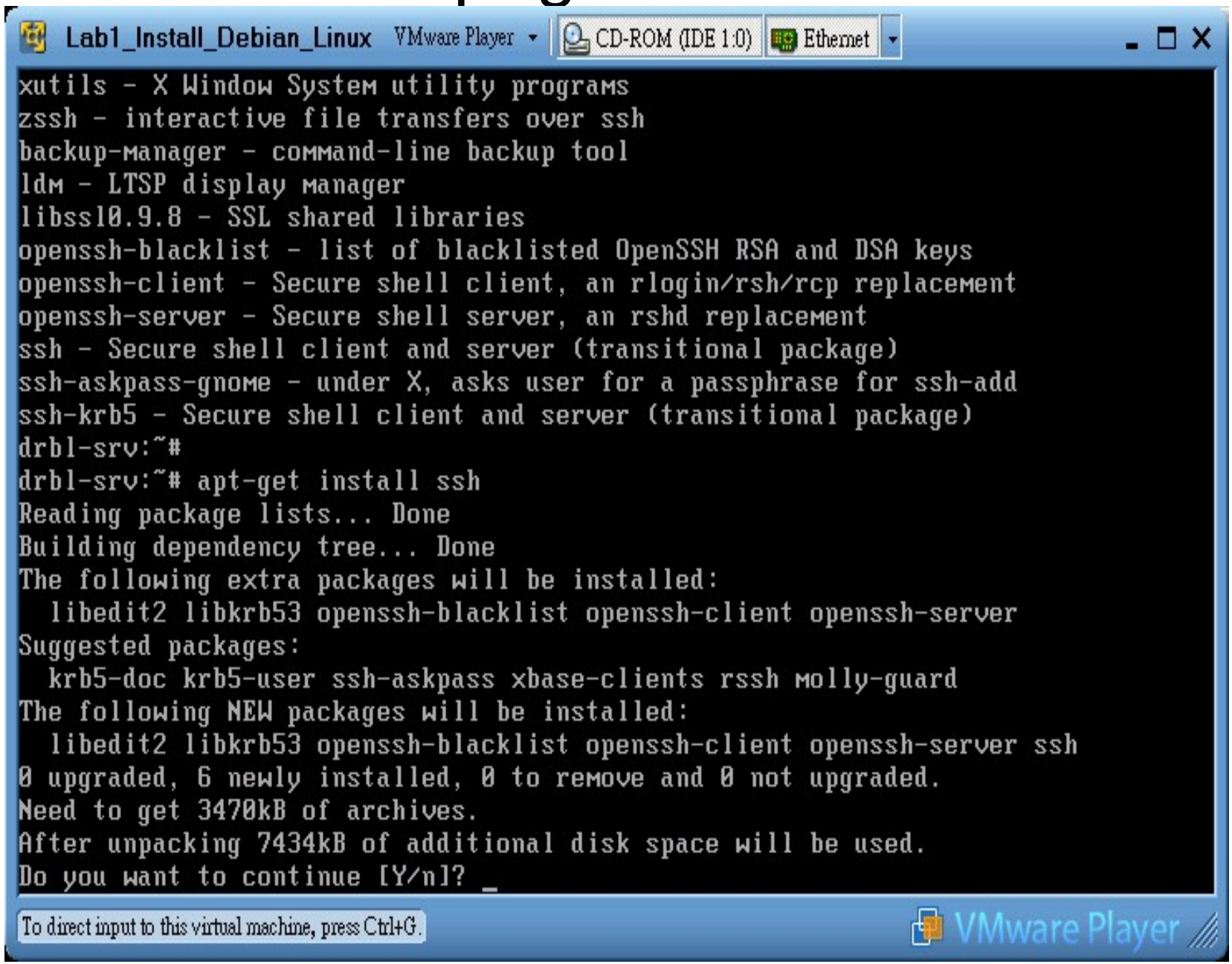


基本管理篇 [4] 用 apt-cache search 搜尋



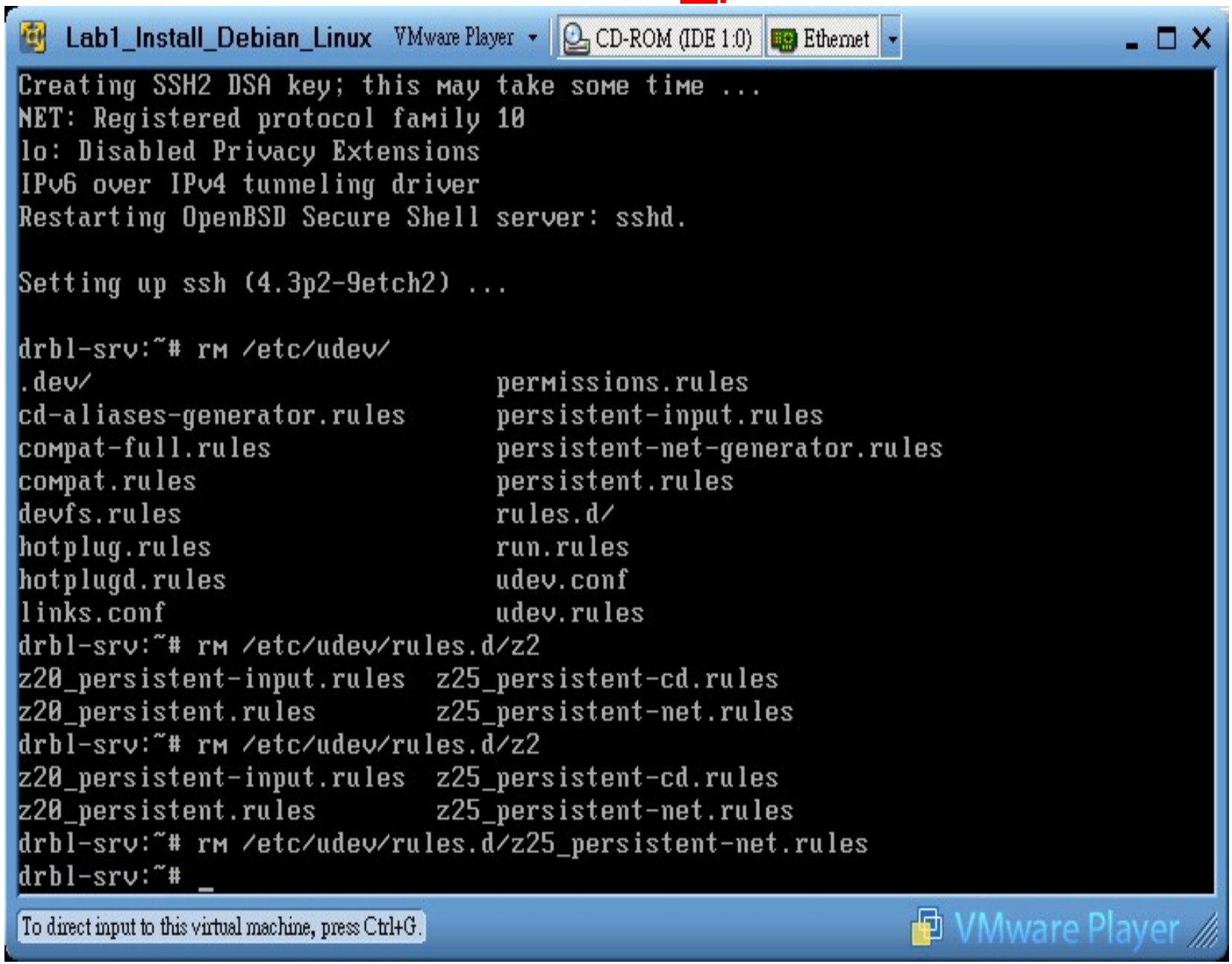
基本管理篇 [5] 用 apt-get install 安裝套件

Ex. apt-get install ssh

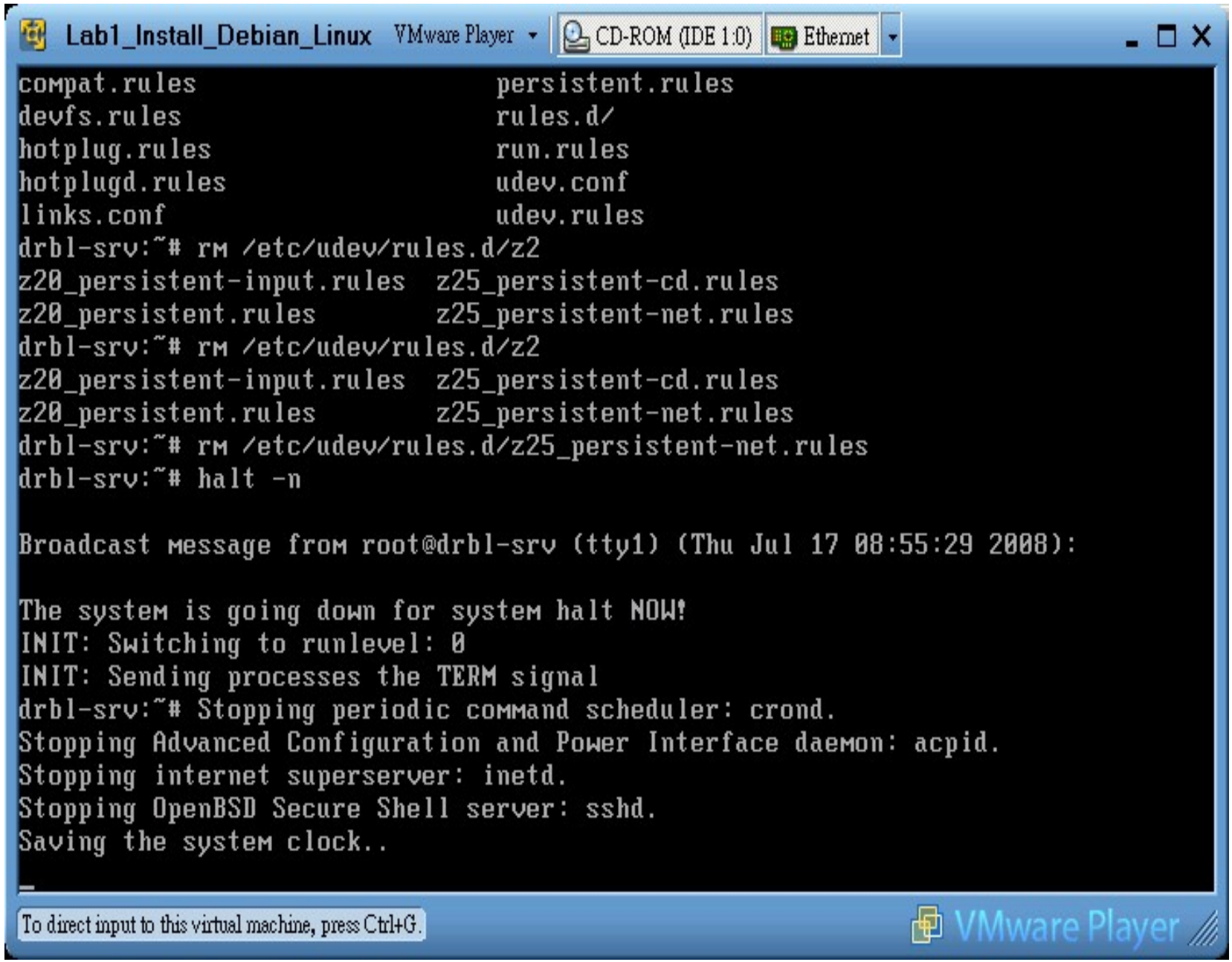


管理番外篇 [*] 關於 udev 鎖住網卡 MAC

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



基本管理篇 [6] 用 halt -n 關機



財團法人國家實驗研究院

國家高速網路與計算中心

NATIONAL CENTER FOR HIGH-PERFORMANCE COMPUTING

設置Hadoop環境

王耀聰 陳威宇 楊順發

jazz@nchc.org.tw

waue@nchc.org.tw

shunfa@nchc.org.tw

國家高速網路與計算中心(NCHC)

自由軟體實驗室

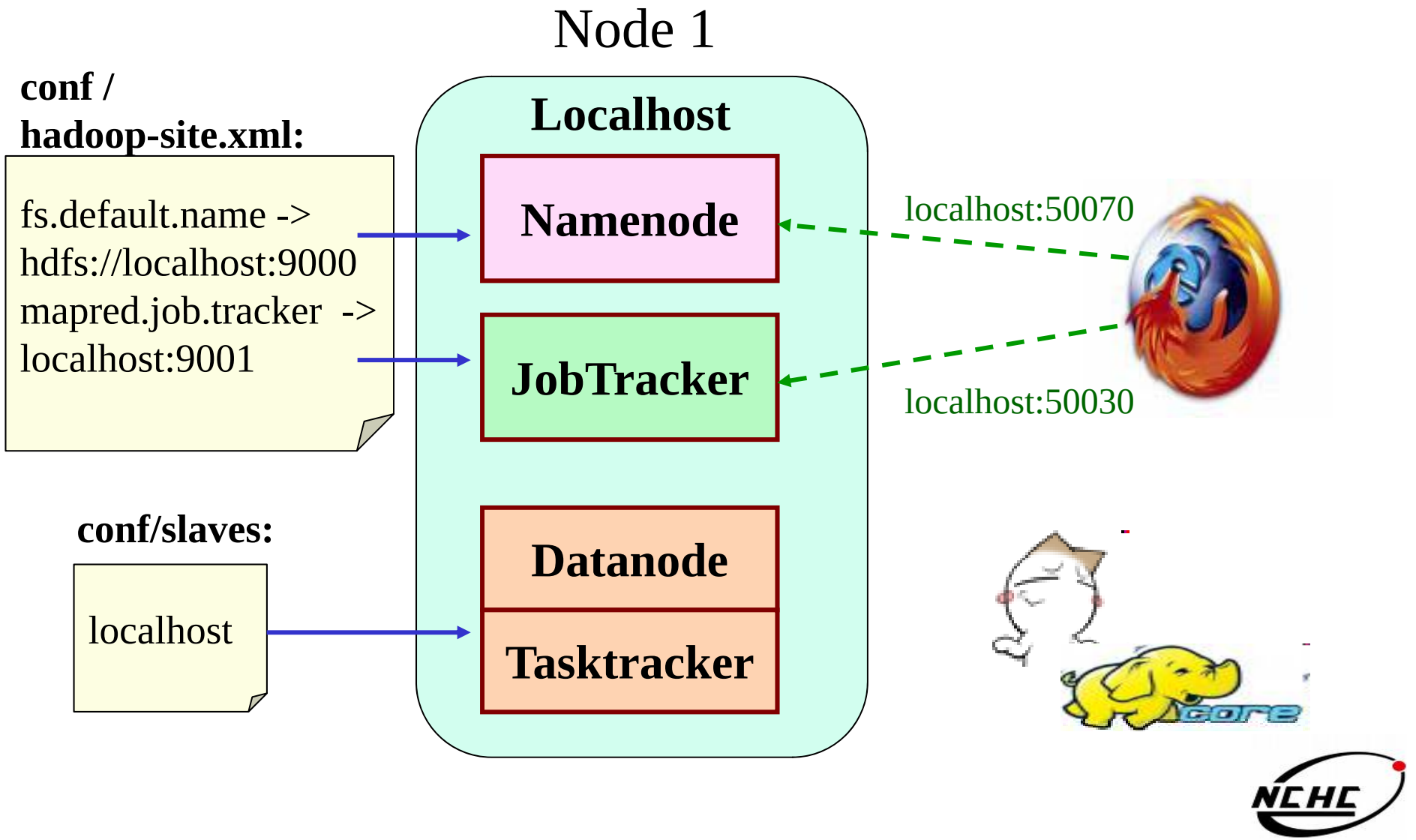
Yahoo's Hadoop Cluster

- ~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 單機設定與啟動

- step 1. 設定登入免密碼
- step 2. 安裝java
- step 3. 下載安裝Hadoop
- step 4.1 設定 hadoop-env.sh
  - export JAVA_HOME=/usr/lib/jvm/java-6-sun
- step 4.2 設定 hadoop-site.xml
  - 設定Namenode-> hdfs://localhost:9000
  - 設定Jobtracker -> localhost:9001
- step 5.1 格式化HDFS
  - bin/hadoop namenode -format
- step 5.2 啟動Hadoop
  - bin/start-all.sh
- step 6. 完成！檢查運作狀態
  - Job admin http://localhost:50030/ HDFS http://localhost:50070/

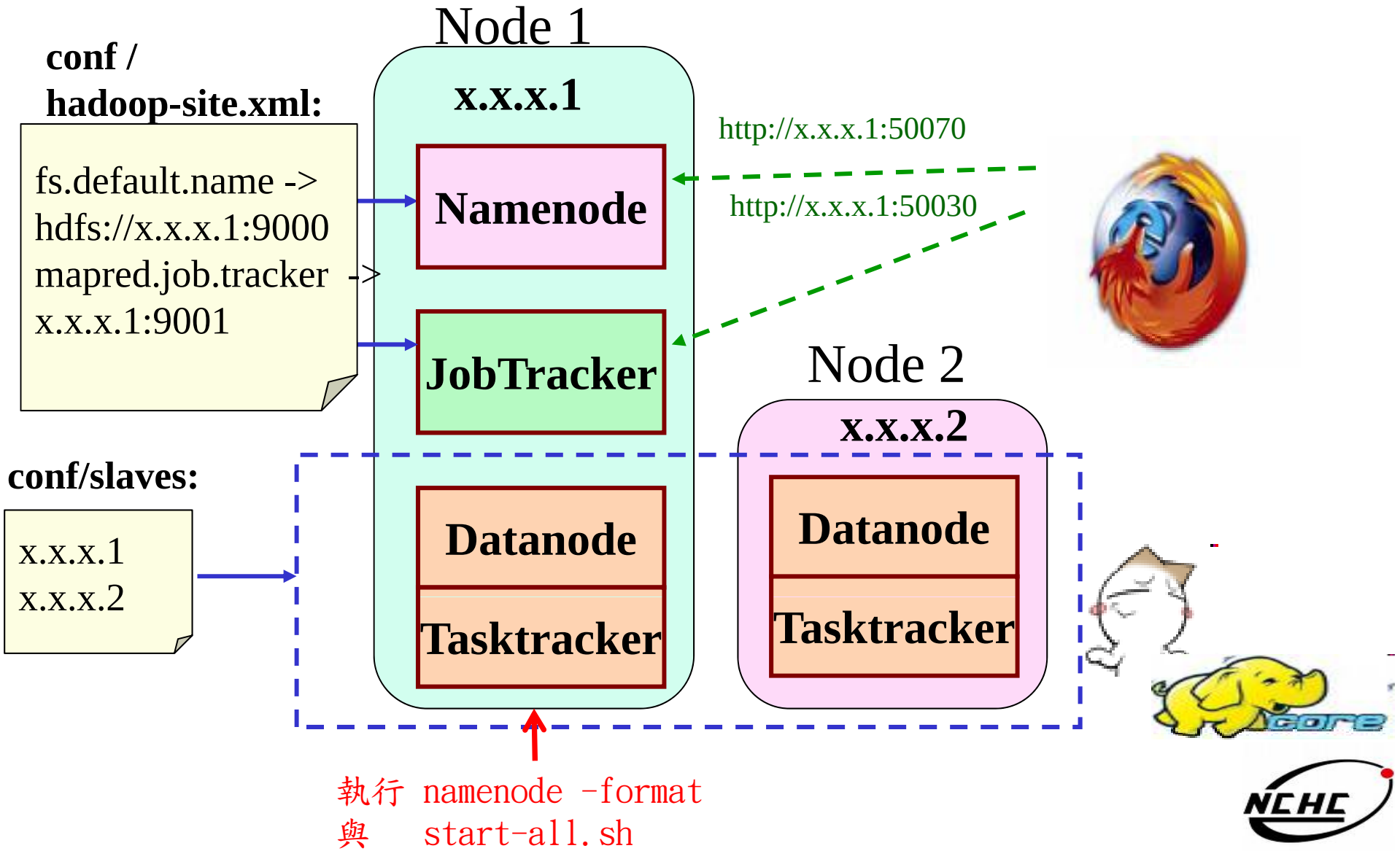
Hadoop 單機環境示意圖



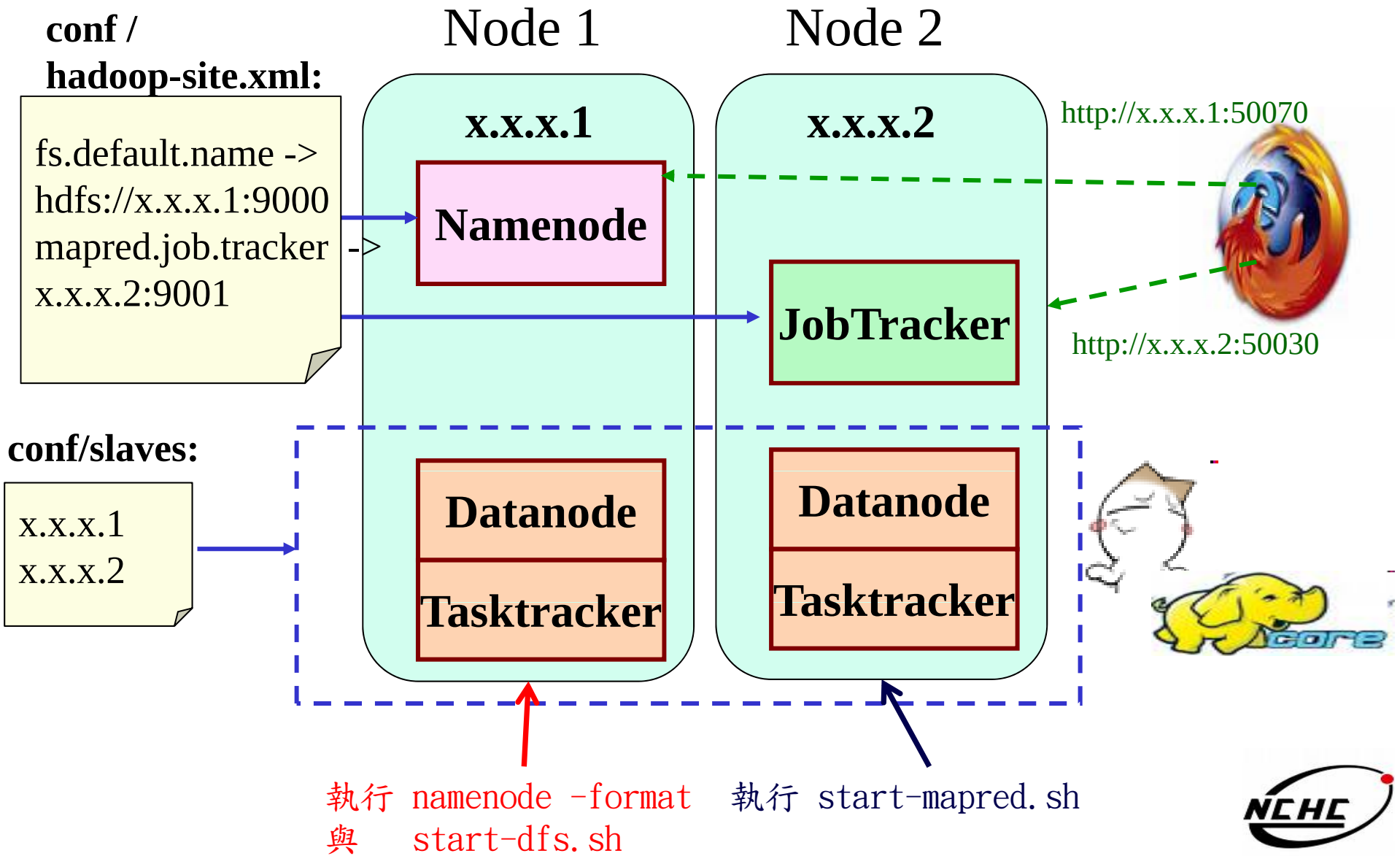
Hadoop 叢集設定與啟動

- step 1. 設定登入免密碼
- step 2. 安裝java
- step 3. 下載安裝Hadoop
- step 4.1 設定 hadoop-env.sh
  - export JAVA_HOME=/usr/lib/jvm/java-6-sun
- step 4.2 設定 hadoop-site.xml
  - 設定Namenode-> hdfs://x.x.x.1:9000
  - 設定Jobtracker -> x.x.x.2:9001
- step 4.3 設定slaves 檔
- step 4.4 將叢集內的電腦Hadoop都做一樣的配置
- step 5.1 格式化HDFS
  - bin/hadoop namenode -format
- step 5.2 啟動Hadoop
  - nodeN執行： bin/start-dfs.sh ； nodeJ執行： bin/start-mapred.sh
- step 6. 完成！檢查運作狀態
  - Job admin http://x.x.x.2:50030/ HDFS http://x.x.x.1:50070/

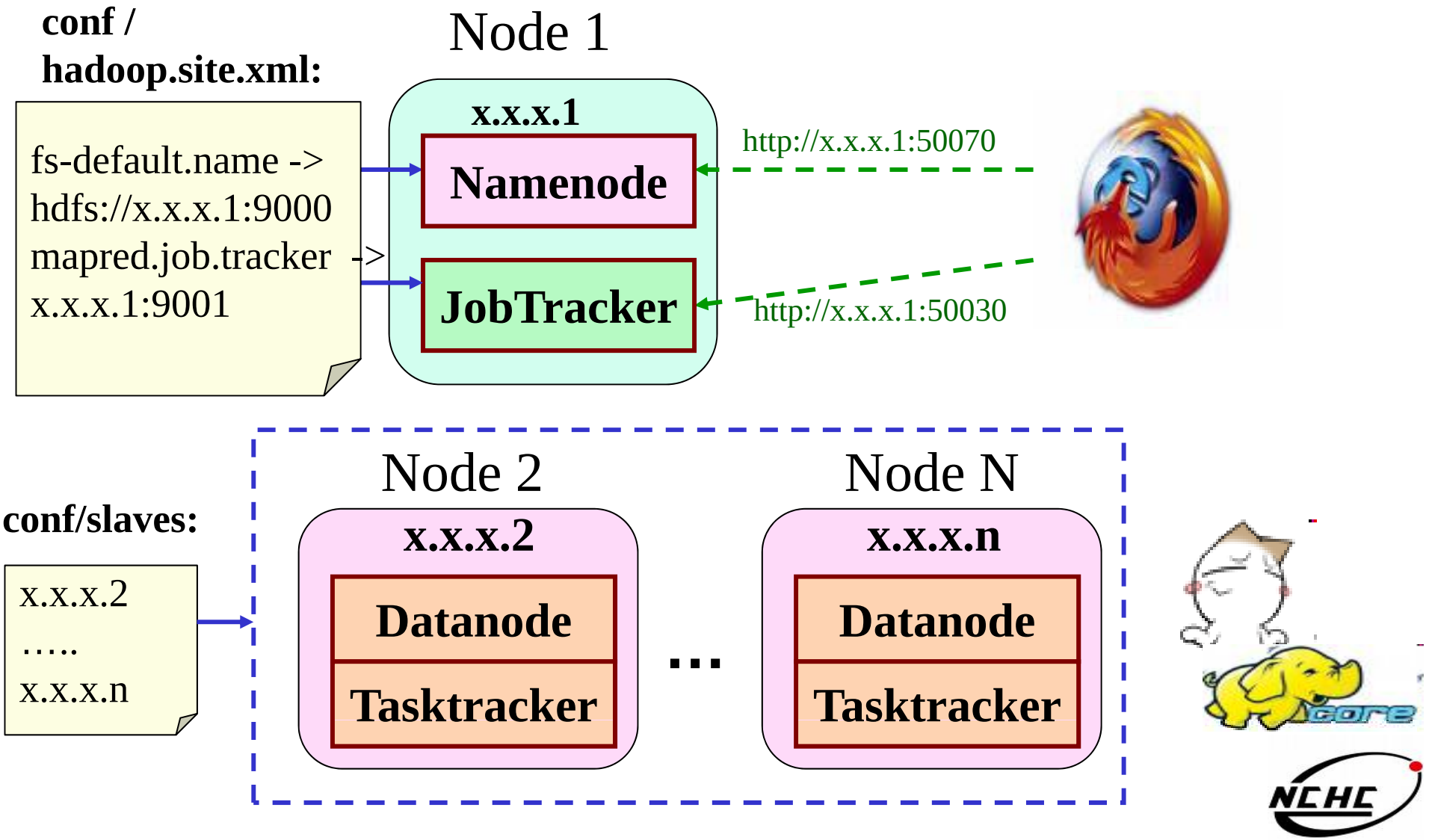
情況一


Node 1
x.x.x.1
conf /
hadoop-site.xml:
fs.default.name ->
hdfs://x.x.x.1:9000
mapred.job.tracker ->
x.x.x.1:9001
Namenode
JobTracker
http://x.x.x.1:50070
http://x.x.x.1:50030
Node 2
x.x.x.2
conf/slaves:
x.x.x.1
x.x.x.2
Datanode
Tasktracker
Datanode
Tasktracker
執行 namenode -format
與 start-all.sh
NCHC

情況二


conf / hadoop-site.xml:
fs.default.name -> hdfs://x.x.x.1:9000
mapred.job.tracker -> x.x.x.2:9001
Node 1
Node 2
x.x.x.1
x.x.x.2
Namenode
JobTracker
http://x.x.x.1:50070
http://x.x.x.2:50030
conf/slaves:
x.x.x.1
x.x.x.2
Datanode
Tasktracker
Datanode
Tasktracker
執行 namenode -format
與 start-dfs.sh
執行 start-mapred.sh

情況三


conf /
hadoop.site.xml:
fs-default.name ->
hdfs://x.x.x.1:9000
mapred.job.tracker ->
x.x.x.1:9001
Node 1
x.x.x.1
Namenode
JobTracker
http://x.x.x.1:50070
http://x.x.x.1:50030
conf/slaves:
x.x.x.2
…..
x.x.x.n
Node 2
x.x.x.2
Datanode
Tasktracker
…
Node N
x.x.x.n
Datanode
Tasktracker
NCHC

情況四

conf /
hadoop-site.xml:

mapred.job.tracker->
x.x.x.2:9001
fs.default.name ->
hdfs://x.x.x.1:9000

JobTracker

Namenode

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

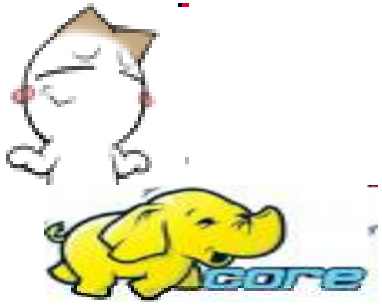
HTTP Monitoring UI

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

conf/slaves:

x.x.x.3
…….
x.x.x.n

T D T D T D T D T D



NCHC

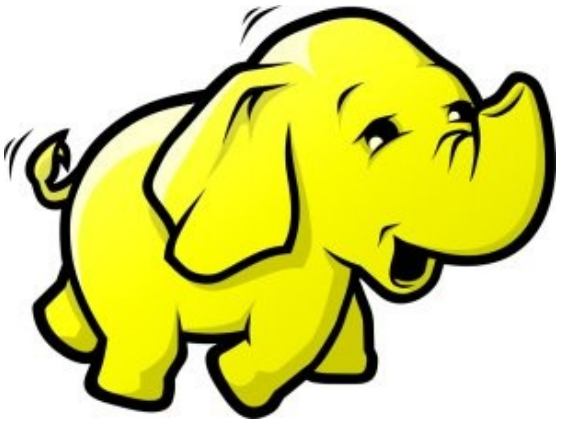
Hadoop 相關計畫

Hadoop Ecosystem

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

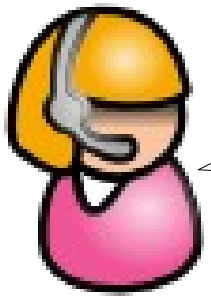


Hadoop 只支援用 **Java** 開發嘛？

Is Hadoop only support Java ?

總不能全部都重新設計吧？如何與舊系統相容？

Can Hadoop work with existing software ?



可以跟資料庫結合嘛？

Can Hadoop work with Databases ?

開發者們有聽到大家的需求.....

Yes, we hear the feedback of developers ...

Is Hadoop only support Java ?

- Although the Hadoop framework is implemented in Java™, Map/Reduce applications need not be written in Java.
- Hadoop Streaming is a utility which allows users to create and run jobs with any executables (e.g. shell utilities) as the mapper and/or the reducer.
- Hadoop Pipes is a SWIG-compatible C++ API to implement Map/Reduce applications (non JNI™ based).

Hadoop Pipes (C++, Python)

- Hadoop Pipes allows C++ code to use Hadoop DFS and map/reduce.
- The C++ interface is "swigable" so that interfaces can be generated for python and other scripting languages.
- For more detail, check the API Document of org.apache.hadoop.mapred.pipes
- You can also find example code at

  hadoop-*/src/examples/pipes
- About the pipes C++ WordCount example code:

  http://wiki.apache.org/hadoop/C++WordCount

Hadoop Streaming

- Hadoop Streaming is a utility which allows users to create and run Map-Reduce jobs with any executables (e.g. Unix shell utilities) as the mapper and/or the reducer.
- It's useful when you need to run existing program written in shell script, perl script or even PHP.
- Note: both the mapper and the reducer are executables that read the input from STDIN (line by line) and emit the output to STDOUT.
- For more detail, check the official document of Hadoop Streaming

Running Hadoop Streaming

```
jazz@hadoop:~$ hadoop jar hadoop-streaming.jar -help
10/08/11 00:20:00 ERROR streaming.StreamJob: Missing required option -input
Usage: $HADOOP_HOME/bin/hadoop [--config dir] jar \
          $HADOOP_HOME/hadoop-streaming.jar [options]
Options:
  -input    <path>     DFS input file(s) for the Map step
  -output   <path>     DFS output directory for the Reduce step
  -mapper   <cmd|JavaClassName>      The streaming command to run
  -combiner <JavaClassName> Combiner has to be a Java class
  -reducer  <cmd|JavaClassName>      The streaming command to run
  -file     <file>     File/dir to be shipped in the Job jar file
  -dfs    <h:p>|local  Optional. Override DFS configuration
  -jt     <h:p>|local  Optional. Override JobTracker configuration
  -additionalconfspec specfile  Optional.
  -inputformat TextInputFormat(default)|SequenceFileAsTextInputFormat|
JavaClassName Optional.
  -outputformat TextOutputFormat(default)|JavaClassName  Optional.
```

… More …

Hadoop Streaming with shell commands (1)

```
hadoop:~$ hadoop fs -rmr input output
hadoop:~$ hadoop fs -put /etc/hadoop/conf input
hadoop:~$ hadoop jar hadoop-streaming.jar -input
input -output output -mapper /bin/cat -reducer /
usr/bin/wc
```

Hadoop Streaming with shell commands (2)

```
hadoop:~$ echo "sed -e \"s/ /\n/g\" | grep ." > streamingMapper.sh
hadoop:~$ echo "uniq -c | awk '{print \$2 \"\t\" \$1}'" > streamingReducer.sh
hadoop:~$ chmod a+x streamingMapper.sh
hadoop:~$ chmod a+x streamingReducer.sh
hadoop:~$ hadoop fs -put /etc/hadoop/conf input
hadoop:~$ hadoop jar hadoop-streaming.jar -input input -output output -mapper streamingMapper.sh -reducer streamingReducer.sh -file streamingMapper.sh -file streamingReducer.sh
```

Homework #3

- Try to run hadoop streaming job on http://hadoop.nchc.org.tw
  - Mapper: /bin/cat
  - Reducer: /usr/bin/wc
  - Input: /etc/hadoop/conf
  - Output: /user/${id}/output
- Run with extra parameter with -numReduceTasks
- Please take screenshot of both results and compare the results.
- Compare the time of 1 reducer with 8 reducers.
- Send a report of this homework to jazz_AT_nchc_DOT_org_DOT_tw

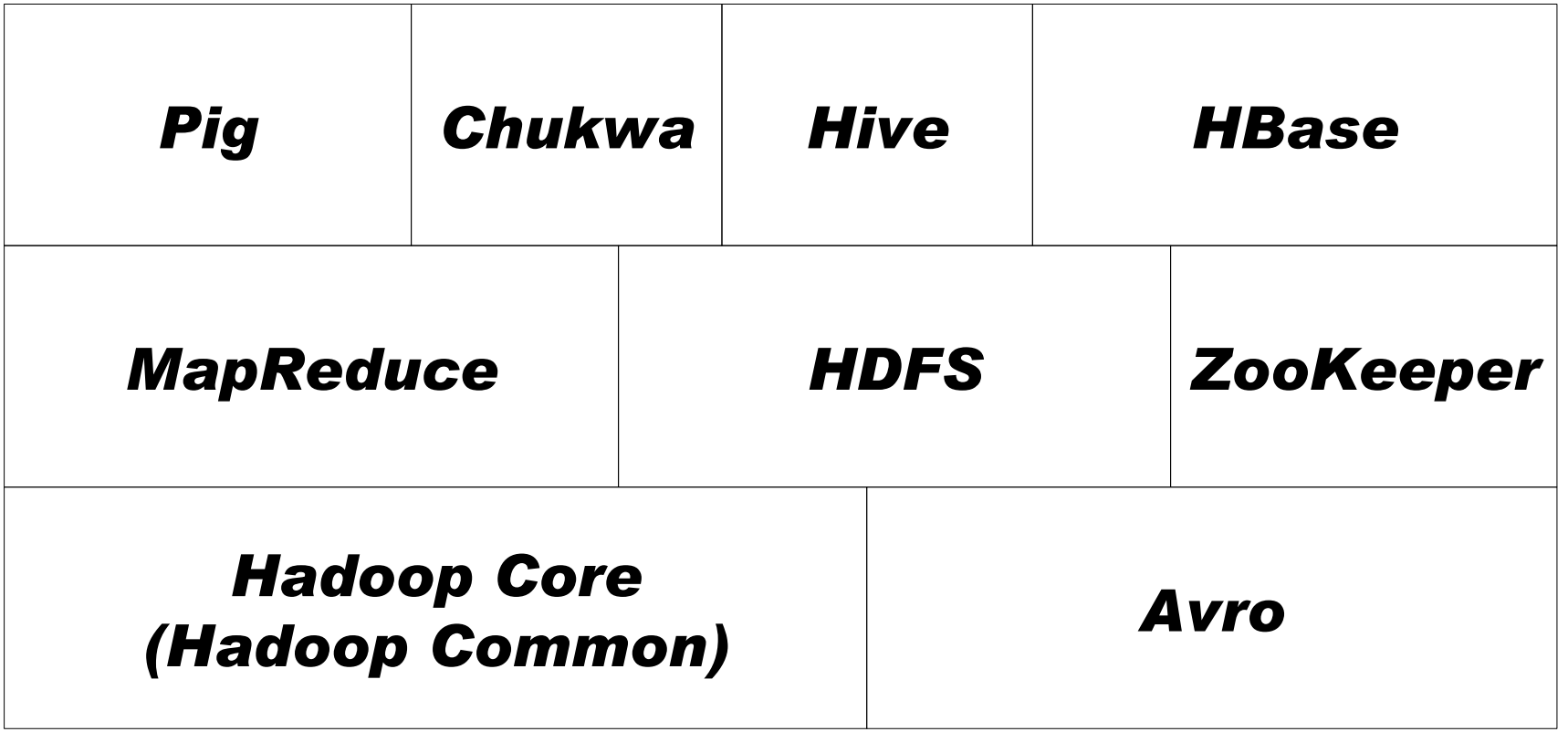
There are serveral Hadoop subprojects

- Hadoop Common: The common utilities that support the other Hadoop subprojects.
- HDFS: A distributed file system that provides high throughput access to application data.
- MapReduce: A software framework for distributed processing of large data sets on compute clusters.

Other Hadoop related projects

- Chukwa: A data collection system for managing large distributed systems.
- HBase: A scalable, distributed database that supports structured data storage for large tables.
- Hive: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- Pig: A high-level data-flow language and execution framework for parallel computation.
- ZooKeeper: A high-performance coordination service for distributed applications.

Hadoop Ecosystem



Source: *Hadoop: The Definitive Guide*

Avro

- Avro is a data serialization system.
- It provides:
  - *Rich data structures.*
  - *A compact, fast, binary data format.*
  - *A container file, to store persistent data.*
  - *Remote procedure call (RPC).*
  - *Simple integration with dynamic languages.*
- Code generation is not required to read or write data files nor to use or implement RPC protocols. Code generation as an optional optimization, only worth implementing for statically typed languages.
- For more detail, please check the official document: http://avro.apache.org/docs/current/

Zoo Keeper

- http://hadoop.apache.org/zookeeper/
- ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or another by distributed applications.
- *Each time they are implemented there is a lot of work that goes into fixing the bugs and race conditions that are inevitable. Because of the difficulty of implementing these kinds of services, applications initially usually skimp on them ,which make them brittle in the presence of change and difficult to manage. Even when done correctly, different implementations of these services lead to management complexity when the applications are deployed.*

Pig

- http://hadoop.apache.org/pig/
- Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs.
- Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs
- Pig's language layer currently consists of a textual language called Pig Latin, which has the following key properties:
  - Ease of programming
  - Optimization opportunities
  - Extensibility

Hive

- http://hadoop.apache.org/hive/
- Hive is a data warehouse infrastructure built on top of Hadoop that provides tools to enable easy data summarization, adhoc querying and analysis of large datasets data stored in Hadoop files.
- Hive QL is based on SQL and enables users familiar with SQL to query this data.

Chukwa

- http://hadoop.apache.org/chukwa/
- Chukwa is an open source data collection system for monitoring large distributed systems.
- built on top of HDFS and Map/Reduce framework
- includes a flexible and powerful toolkit for displaying, monitoring and analyzing results to make the best use of the collected data.

Mahout

- http://mahout.apache.org/
- Mahout is a scalable machine learning libraries.
- implemented on top of Apache Hadoop using the map/reduce paradigm.
- Mahout currently has
  - Collaborative Filtering
  - User and Item based recommenders
  - K-Means, Fuzzy K-Means clustering
  - Mean Shift clustering
  - More ...

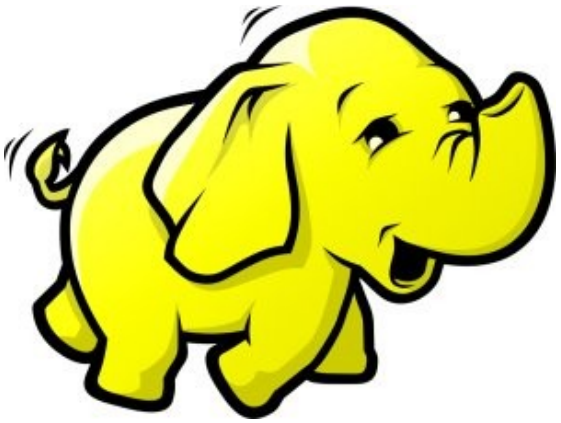
HBase 雲端資料庫

Introduction to HBase

Jazz Wang

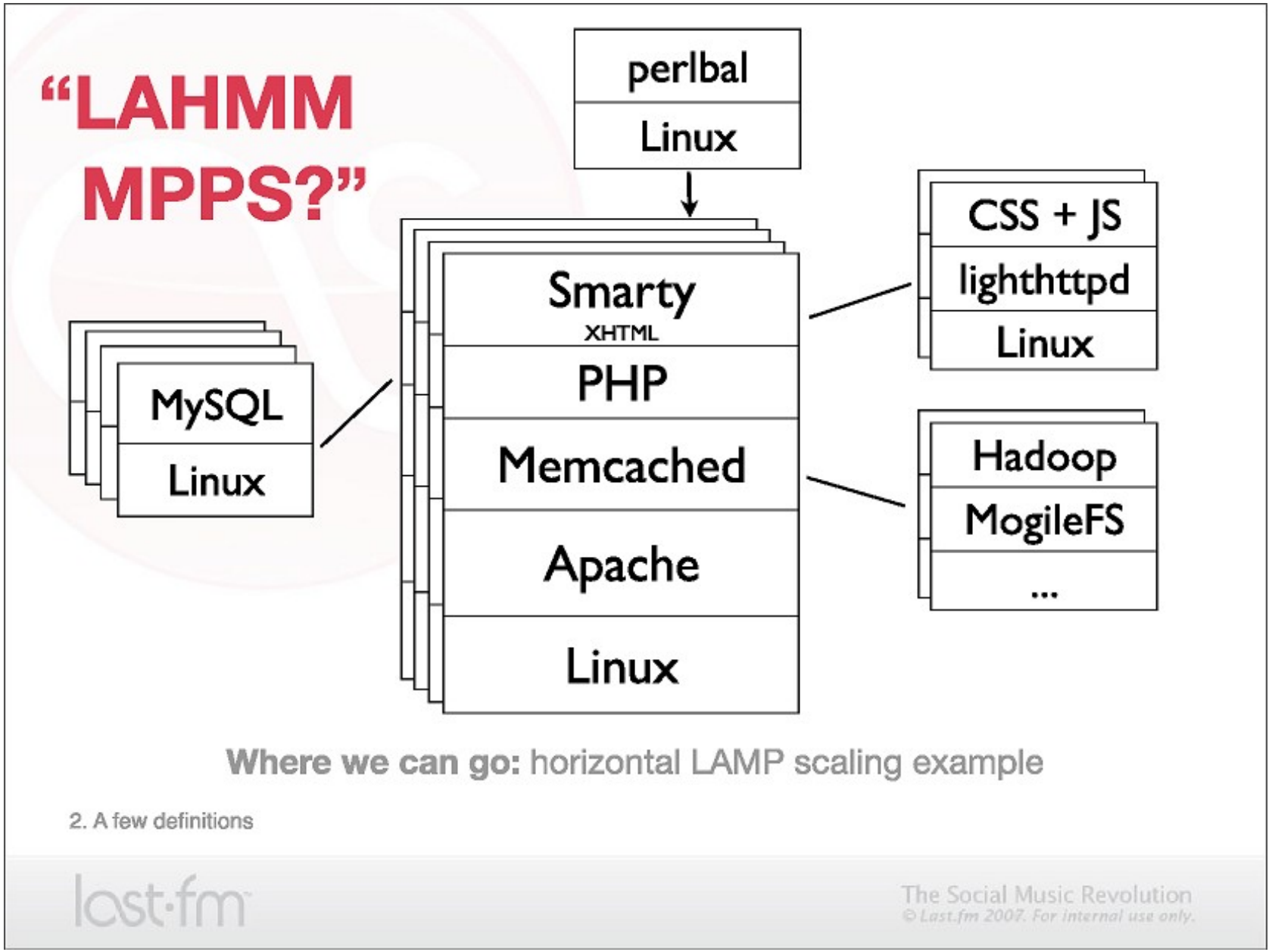
Yao-Tsung Wang

jazz@nchc.org.tw



It's all about SCALE!!

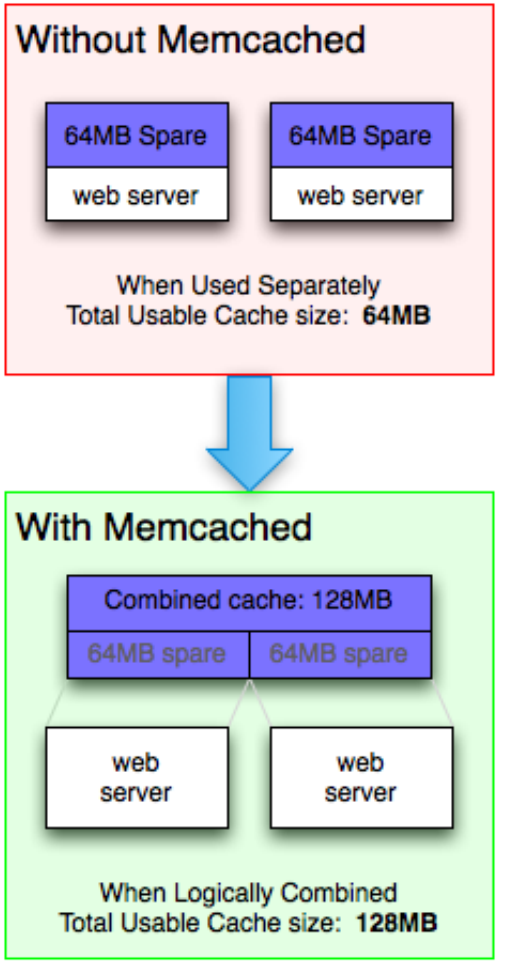
How to scale up web service in the past ?



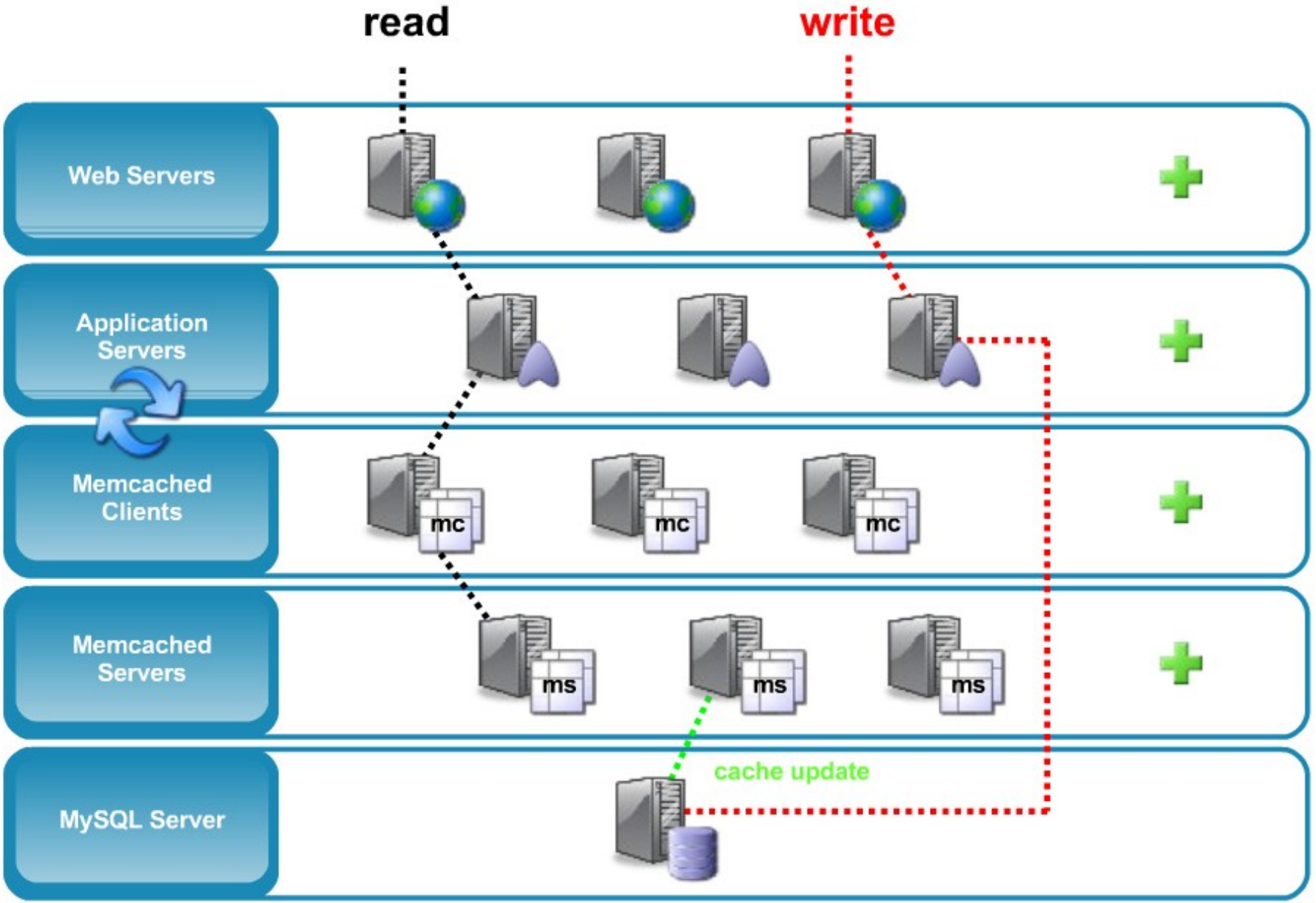
Source: http://www.slideshare.net/mallipeddi/scalable-lamp-development-for-growing-web-apps

Tools used by large scale websites

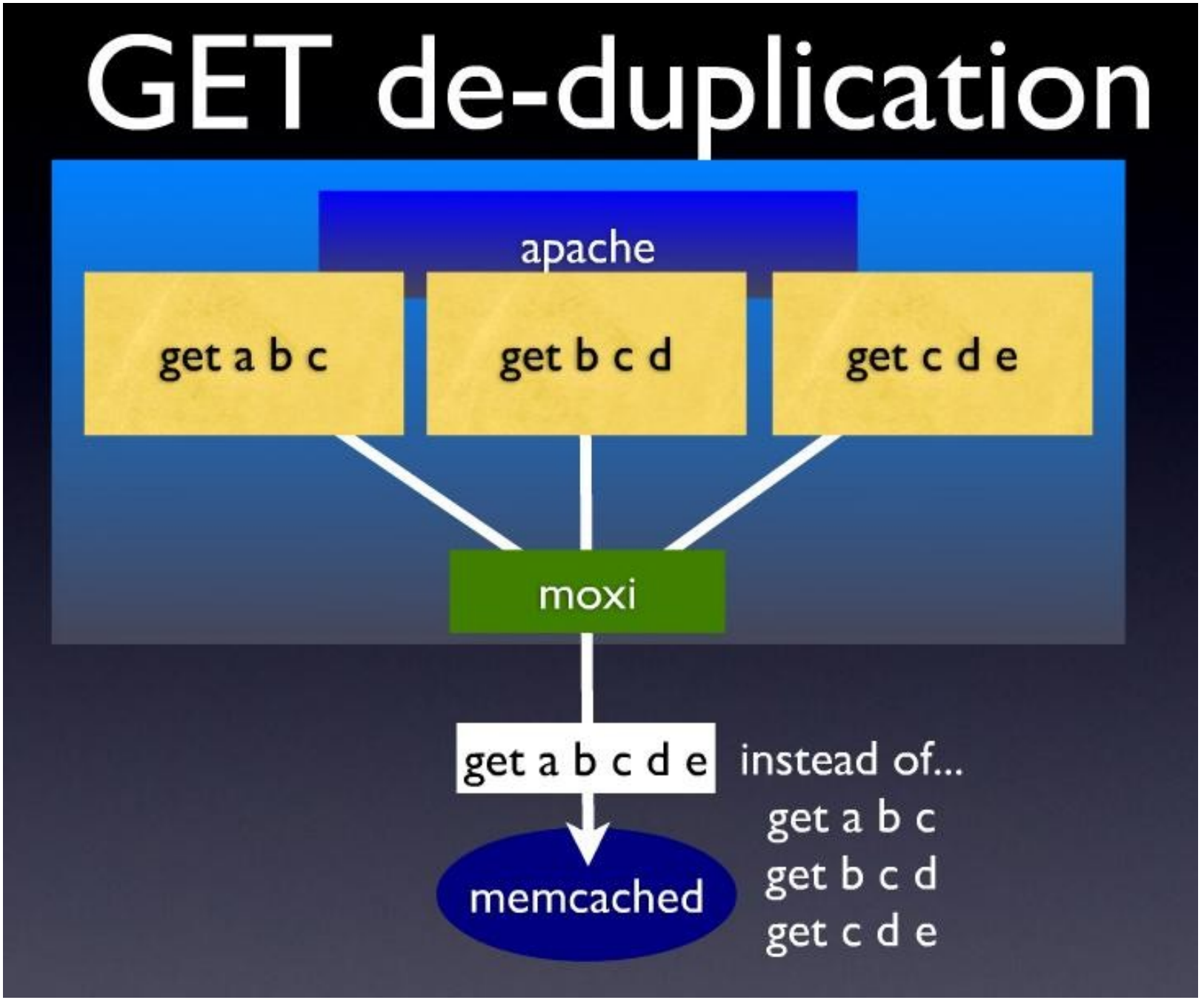
- Perlbal - http://www.danga.com/perlbal/
  - 多個網頁伺服器的負載平衡
  - Load balancer
- MogileFS - http://www.danga.com/mogilefs/
  - 分散式檔案系統
  - Distributed File System fo small files
  - 有公司認為 MogileFS 比起 Hadoop 適合拿來處理小檔案
- memcached - http://memcached.org/
  - 共享記憶體 ??
  - Share Memory
  - 把資料庫或經常讀取的部分，用記憶體快取 (Cache) 方式存放
- Moxi - http://code.google.com/p/moxi/
  - Memcache 的 PROXY
- More Resource:
  - http://code.google.com/p/memcached/wiki/HowToLearnMoreScalability
  - http://www.slideshare.net/techdude/scalable-web-architectures-common-patterns-and-approaches



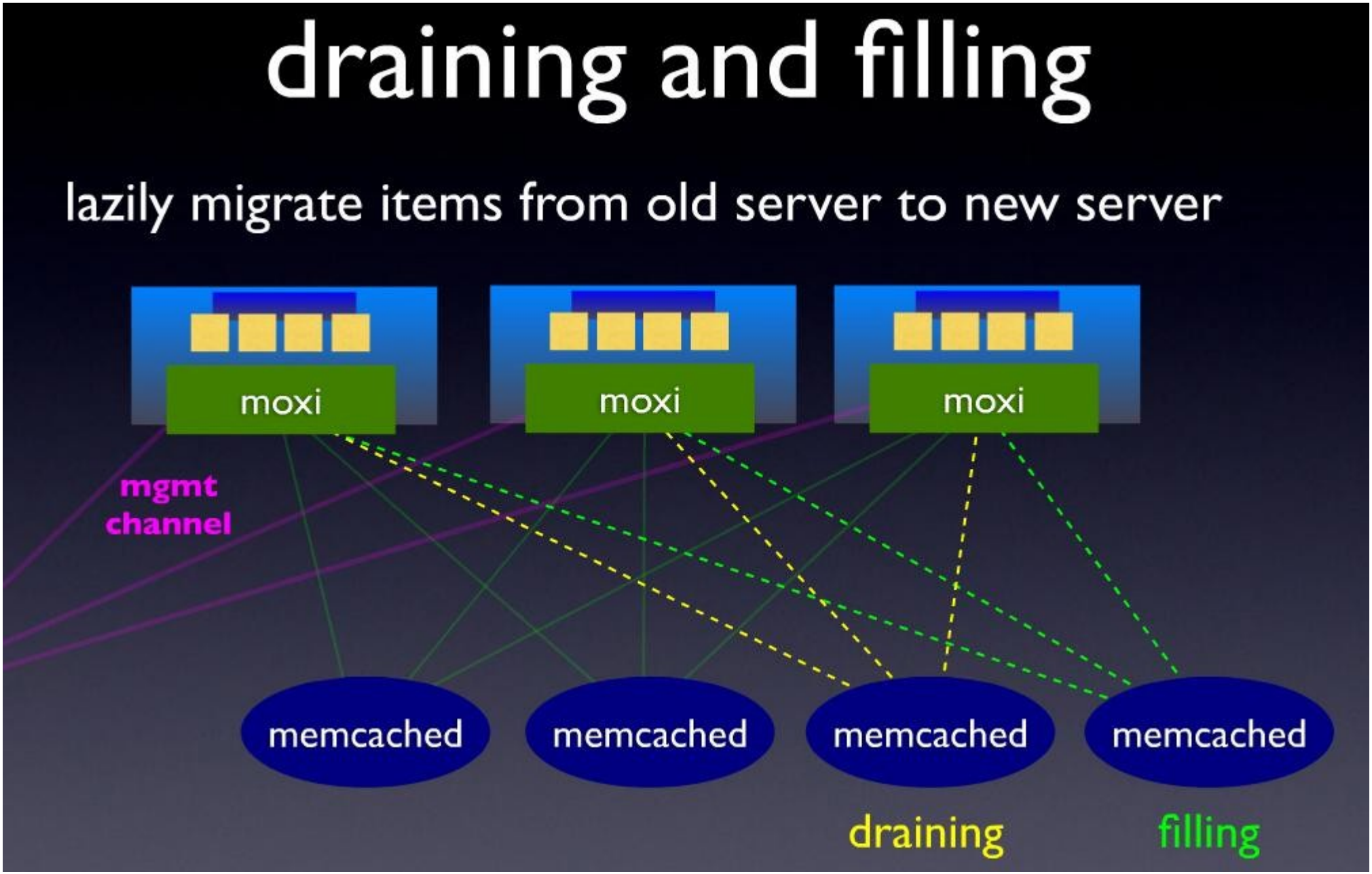
Memcached & MySQL



Source: http://mashraqi.com/2008/07/memcached-for-mysql-advanced-use-cases_09.html



Source: http://www.slideshare.net/northscale/moxi-memcached-proxy



Source: http://www.slideshare.net/northscale/moxi-memcached-proxy

HBase is ..

- HBase is a distributed column-oriented database built on top of HDFS.
- A distributed data store that can scale horizontally to 1,000s of commodity servers and petabytes of indexed storage.
- Designed to operate on top of the Hadoop distributed file system (HDFS) or Kosmos File System (KFS, aka Cloudstore) for scalability, fault tolerance, and high availability.
- Integrated into the Hadoop map-reduce platform and paradigm.

Benefits

- Distributed storage
- Table-like in data structure
  - multi-dimensional map
- High scalability
- High availability
- High performance

Who use HBase

- Adobe
  - 內部使用 (Structure data)
- Kalooga
  - 圖片搜尋引擎 http://www.kalooga.com/
- Meetup
  - 社群聚會網站 http://www.meetup.com/
- Streamy
  - Migrate from MySQL to Hbase http://www.streamy.com/
- Trend Micro
  - 雲端掃毒架構 http://trendmicro.com/
- Yahoo!
  - 儲存文件 fingerprint 避免重複 http://www.yahoo.com/
- More - http://wiki.apache.org/hadoop/Hbase/PoweredBy

Backdrop

- Started toward by Chad Walters and Jim
- 2006.11
  - Google releases paper on BigTable
- 2007.2
  - Initial HBase prototype created as Hadoop contrib.
- 2007.10
  - First useable HBase
- 2008.1
  - Hadoop become Apache top-level project and HBase becomes subproject
- 2008.10~
  - HBase 0.18, 0.19 released

HBase Is Not …

- Tables have one primary index, the *row key*.
- No join operators.
- Scans and queries can select a subset of available columns, perhaps by using a wildcard.
- There are three types of lookups:
  - Fast lookup using row key and optional timestamp.
  - Full table scan
  - Range scan from region start to end.

HBase Is Not …(2)

- Limited atomicity and transaction support.
  - HBase supports multiple batched mutations of single rows only.
  - Data is unstructured and untyped.
- No accessed or manipulated via SQL.
  - Programmatic access via Java, REST, or Thrift APIs.
  - Scripting via JRuby.

Why Bigtable?

- Performance of RDBMS system is good for transaction processing but for very large scale analytic processing, the solutions are commercial, expensive, and specialized.
- Very large scale analytic processing
  - Big queries – typically range or table scans.
  - Big databases (100s of TB)

Why Bigtable? (2)

- Map reduce on Bigtable with optionally Cascading on top to support some relational algebras may be a cost effective solution.
- Sharding is not a solution to scale open source RDBMS platforms
  - Application specific
  - Labor intensive (re)partitionaing

Why HBase ?

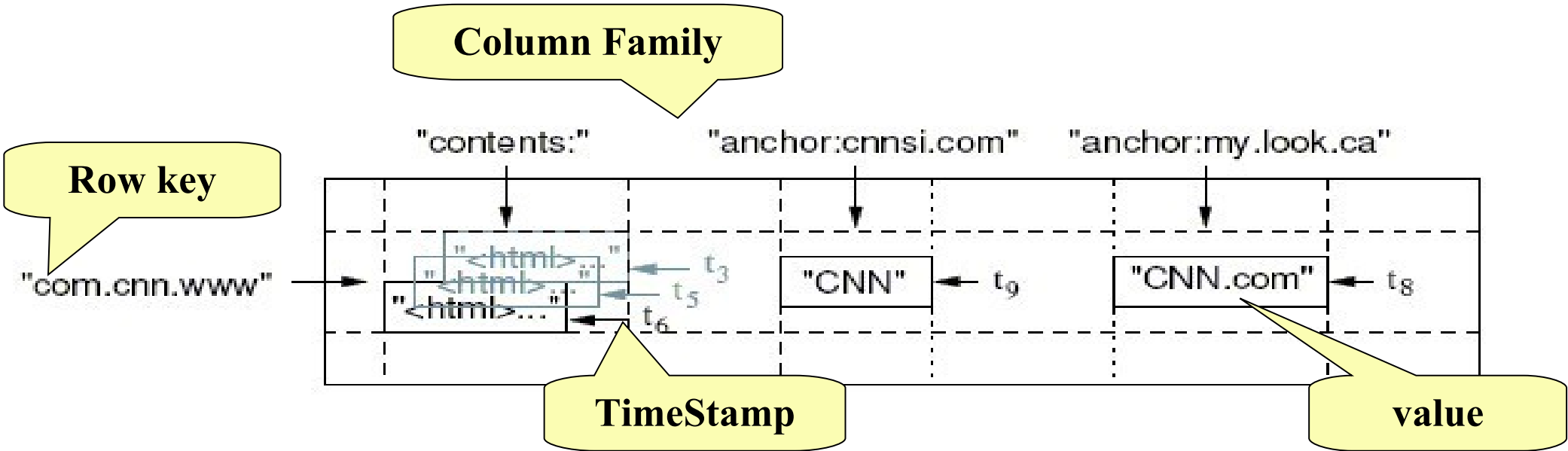
- HBase is a Bigtable clone.
- It is open source
- It has a good community and promise for the future
- It is developed on top of and has good integration for the Hadoop platform, if you are using Hadoop already.
- It has a Cascading connector.

HBase benefits than RDBMS

- *No real indexes*
- *Automatic partitioning*
- *Scale linearly and automatically with new nodes*
- *Commodity hardware*
- *Fault tolerance*
- *Batch processing*

Data Model

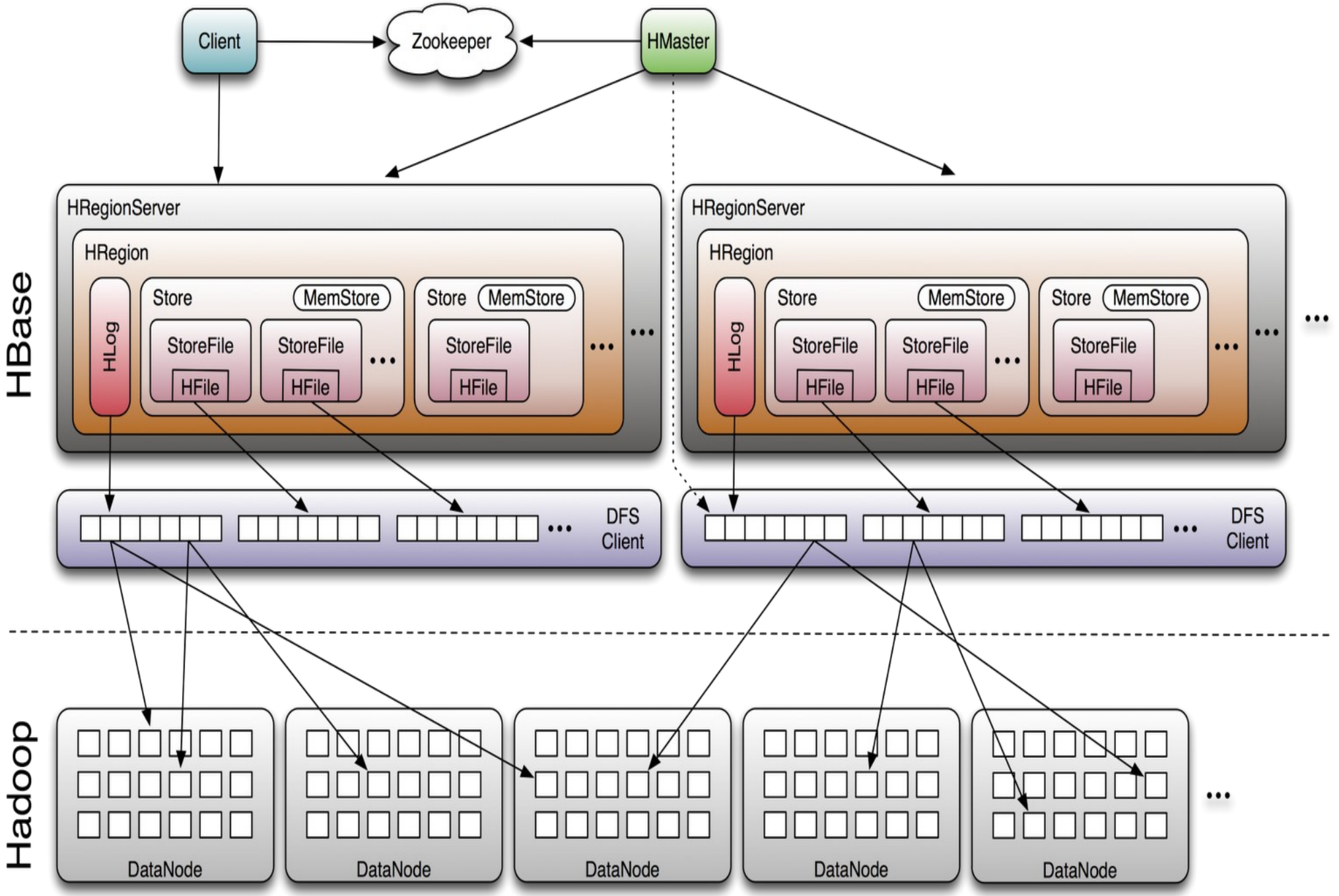
- Tables are sorted by Row
- Table schema only define it's *column families* .
  - Each family consists of any number of columns
  - Each column consists of any number of versions
  - Columns only exist when inserted, NULLs are free.
  - Columns within a family are sorted and stored together
- Everything except table names are byte[]
- (Row, Family: Column, Timestamp) → Value



Members

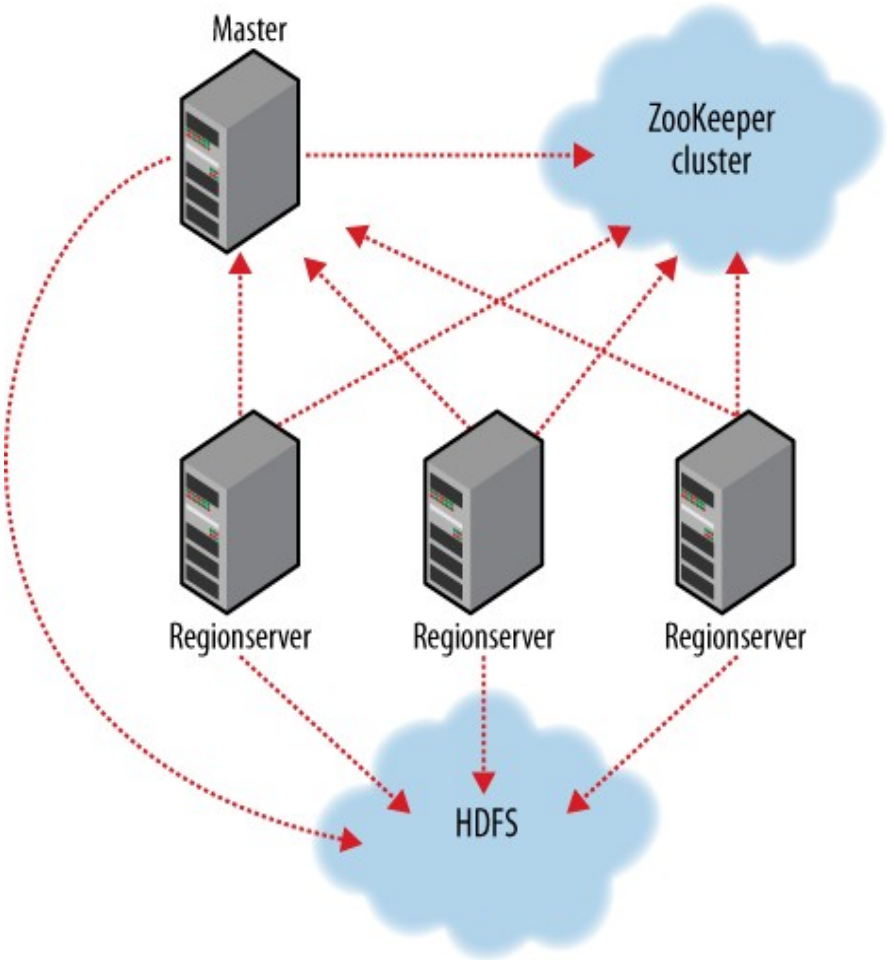
- *Master*
  - Responsible for monitoring region servers
  - Load balancing for regions
  - Redirect client to correct region servers
  - The current SPOF
- *regionserver* slaves
  - Serving requests(Write/Read/Scan) of Client
  - Send HeartBeat to Master
  - Throughput and Region numbers are scalable by region servers

Architecture



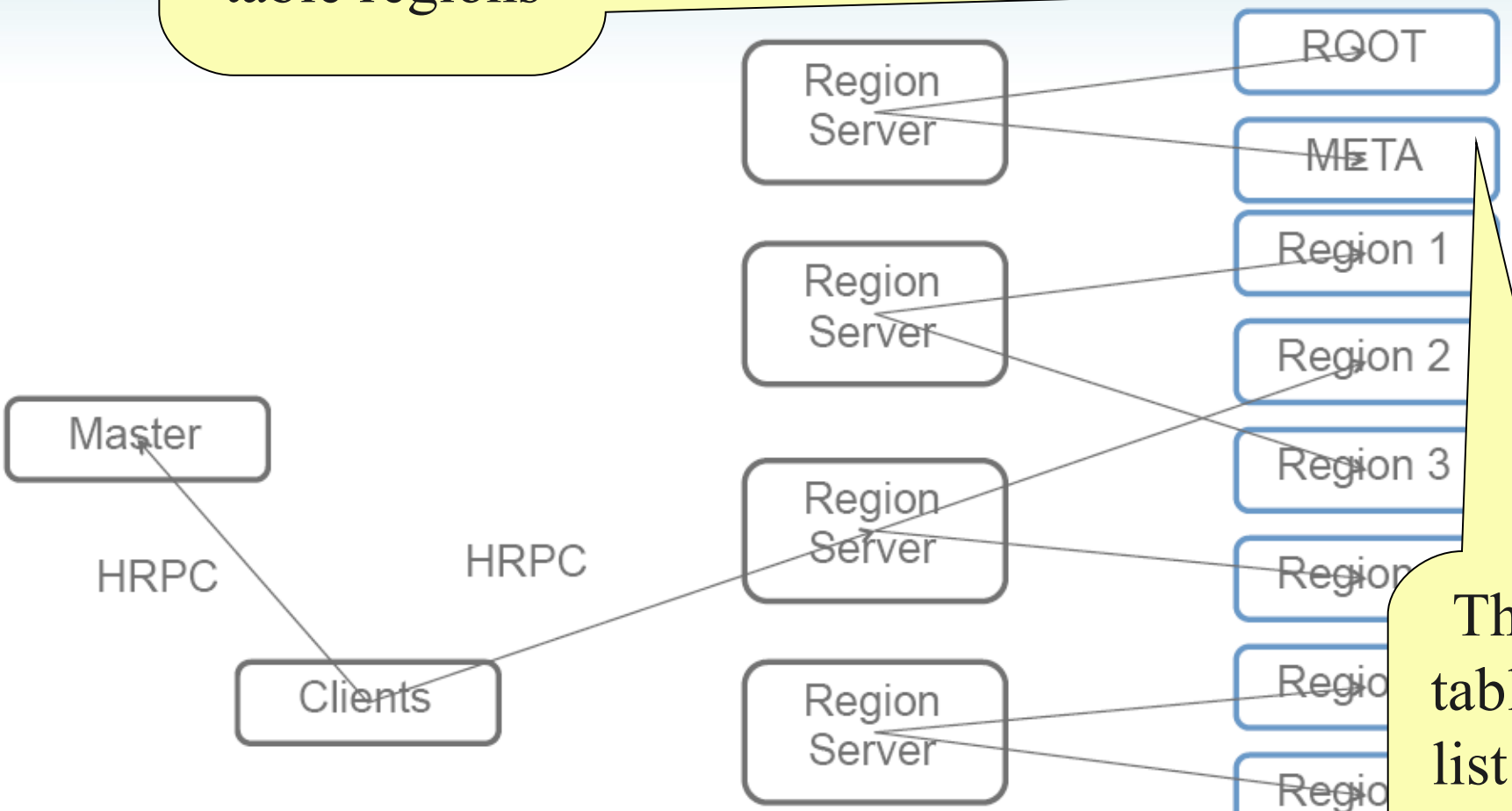
ZooKeeper

- HBase depends on ZooKeeper (Chapter 13) and by default it manages a ZooKeeper instance as the authority on cluster state



Operation

The -ROOT- table holds the list of .META. table regions



The .META. table holds the list of all user-space regions.

Questions?

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

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

Powered by DRBL

輕鬆入手的叢集式搜尋引擎 -

Crawlzilla Develop Team
Free Software Lab @ NCHC

TAIWAN www.nchc.org.tw National Applied Research Laboratories NCHC

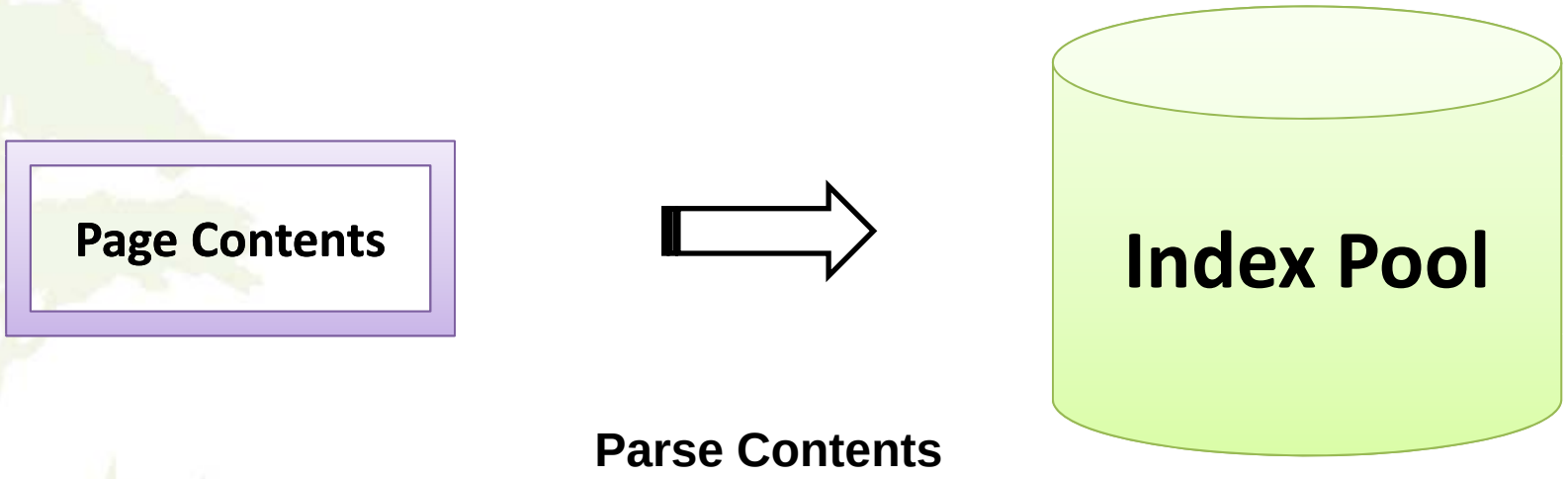
搜尋引擎運作原理 – Phase1

- Crawling the Web

Crawler visits the web pages of the links

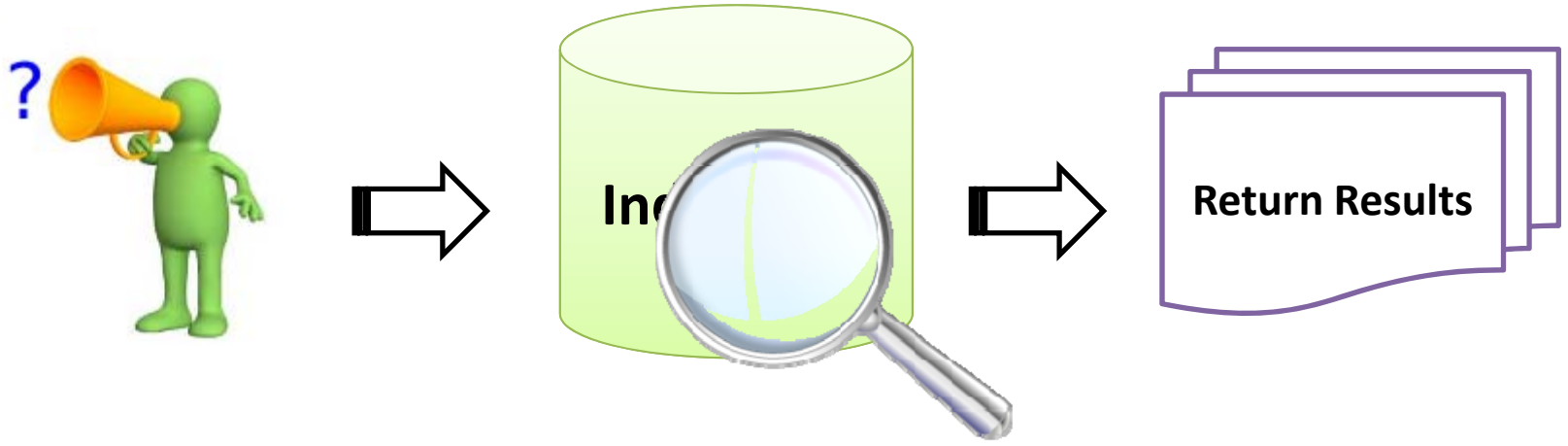
搜尋引擎運作原理 – Phase2

- Building the Index Pool



搜尋引擎運作原理 – Phase3

- Serving Queries



User Sent a Query Search from Index Pool

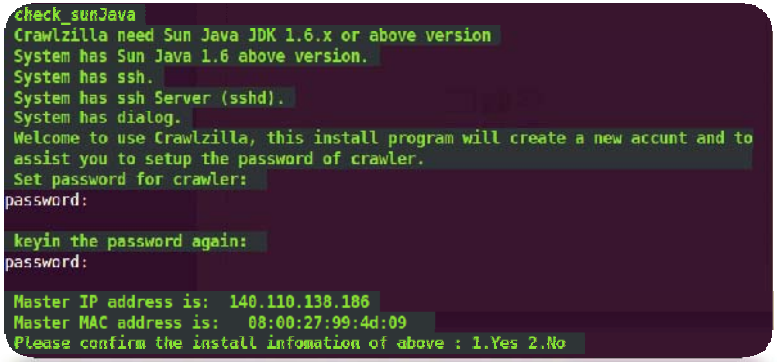
What is Crawlzilla?

- Crawlzilla 簡介
  - 於2009推出實驗版
  - Crawlzilla 於2010更名並延續實驗版開發更多新功能
  - 提供簡單安裝及操作管理介面，輕鬆建立搜尋引擎的套件工具
  - 提供索引資料庫瀏覽功能，搜尋引擎資料庫資訊一目了然

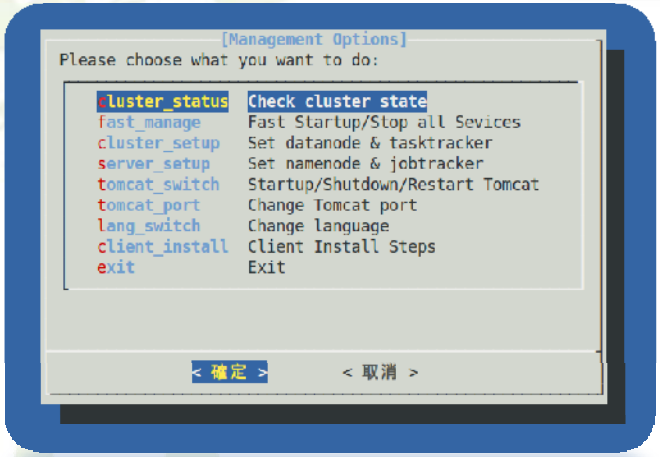
Why Crawlzilla?

- 開放式搜尋引擎不適用於企業內部網站
- 使用Opensource建立搜尋引擎的技術門檻太高
- 叢集環境架設不易
- 使用Crawlzilla優點
  - Opensource專案，使用者可依自己的需求修改源始碼
  - 使用簡單，可輕鬆建立叢集環境
  - 友善的操作環境，節省適應系統時間
  - 支援中文分詞，提高搜尋精準度

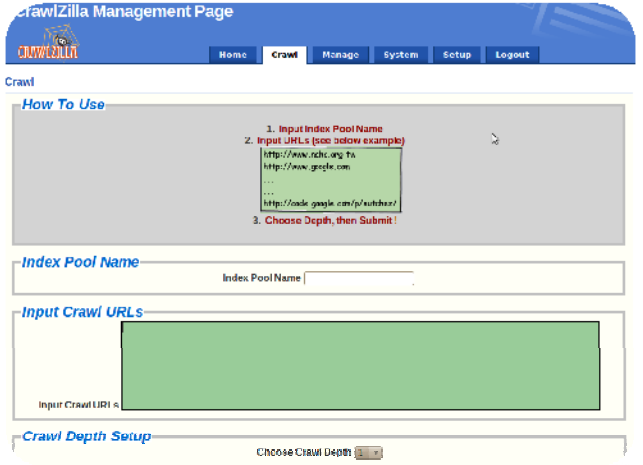
Crawlzilla 操作介面特色



(1) Easy to Deploy Crawling Cluster Environment



(2) Easy to Manage

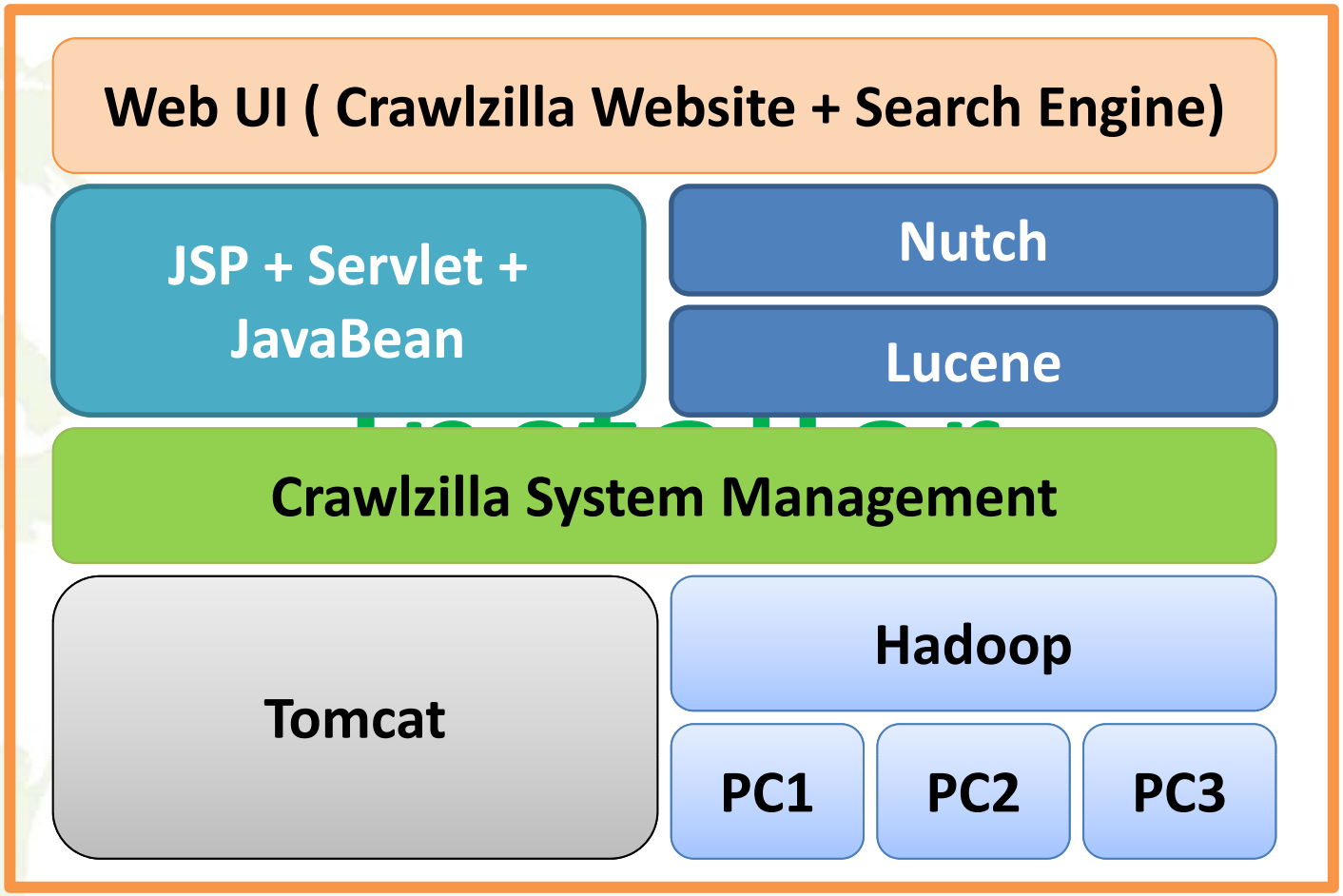


(3) Easy to Use

Crawlzilla 系統功能

- 支援叢集運算及顧全安全性
- 支援中文分詞功能
- 支援多工網頁爬取
- 支援多重搜尋引擎
- 即時瀏覽資料庫資訊
- 解決中文亂碼及中文支援
- 支援多國語言
- 網頁管理

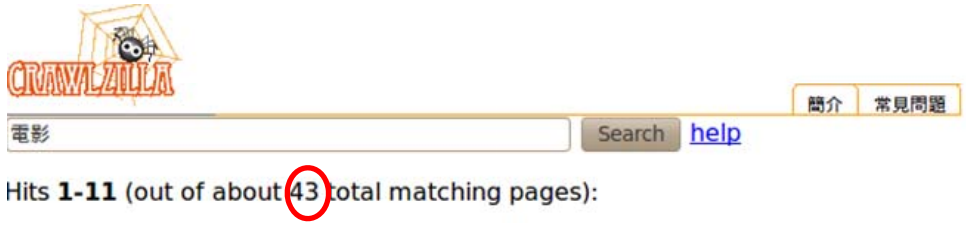
系統架構


Web UI (Crawlzilla Website + Search Engine)
JSP + Servlet + JavaBean
Nutch
Lucene
Crawlzilla System Management
Tomcat
Hadoop
PC1
PC2
PC3

搜尋引擎加入中文分詞功能

- 索引資料庫會以中文字詞為基本單位建立索引
- 加入中文分詞針對同一網站爬取進行搜尋
  - 搜尋引擎**無**中文分詞功能時，搜尋關鍵字 - 電影
    - *760* 筆搜尋結果

  - 搜尋引擎**加入**中文分詞功能時，搜尋關鍵字 - 電影
    - *43* 筆搜尋結果
    - 可提高搜尋的精準度



Crawlzilla - 叢集環境需求

- 如果你覺得...
  - 一台電腦無法滿足你的運算需求
  - 閒置電腦太多
  - 解：讓多台電腦分工運算
- 但是...
  - 架設叢集環境很麻煩!?
  - 解：Crawlzilla 提供叢集安裝模式，只要三分鐘即可建立叢集式搜尋引擎!!!

Resources

- Crawlzilla @ Google Code Project Hosting (中文說明頁)
  - http://code.google.com/p/crawlzilla/
- Crawlzilla @ SourceForge(英文說明頁)
  - http://sourceforge.net/p/crawlzilla/home/
- Crawlzilla User Group @ Google
  - http://groups.google.com/group/crawlzilla-user
- NCHC Cloud Computing Research Group
  - http://trac.nchc.org.tw/cloud