

# Course Information 課程資訊



- 講師介紹：
  - 國網中心 王耀聰 副研究員 / 交大電控碩士
  - [jazz@nchc.org.tw](mailto: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，請減少不必要之講義列印。
- 礙於缺乏實機操作環境，故以影片展示與單機操作為主
  - 若有興趣實機操作，請參考國網中心雲端運算課程錄影
  - <http://trac.nchc.org.tw/cloud>
  - <http://www.classcloud.org/media>
  - <http://www.screentoaster.com/user?username=jazzwang>
- 若需要實驗環境，可至國網中心雲端運算實驗叢集申請帳號
  - <http://hadoop.nchc.org.tw>
- Hadoop 相關問題討論：
  - <http://forum.hadoop.tw>



# 雲端運算的關鍵技術概述

Part 1 : Overview Core Technologies of Cloud Computing

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



Powered by DRBL

# The wisdom of Clouds (Crowds)

雲端序曲：雲端的智慧始終來自於群眾的智慧

2006年8月9日

Google 執行長施密特 ( Eric Schmidt ) 於SES'06會議中首次使用「雲端運算 ( Cloud Computing ) 」來形容無所不在的網路服務

2006年8月24日

Amazon 以 Elastic Compute Cloud 命名其虛擬運算資源服務

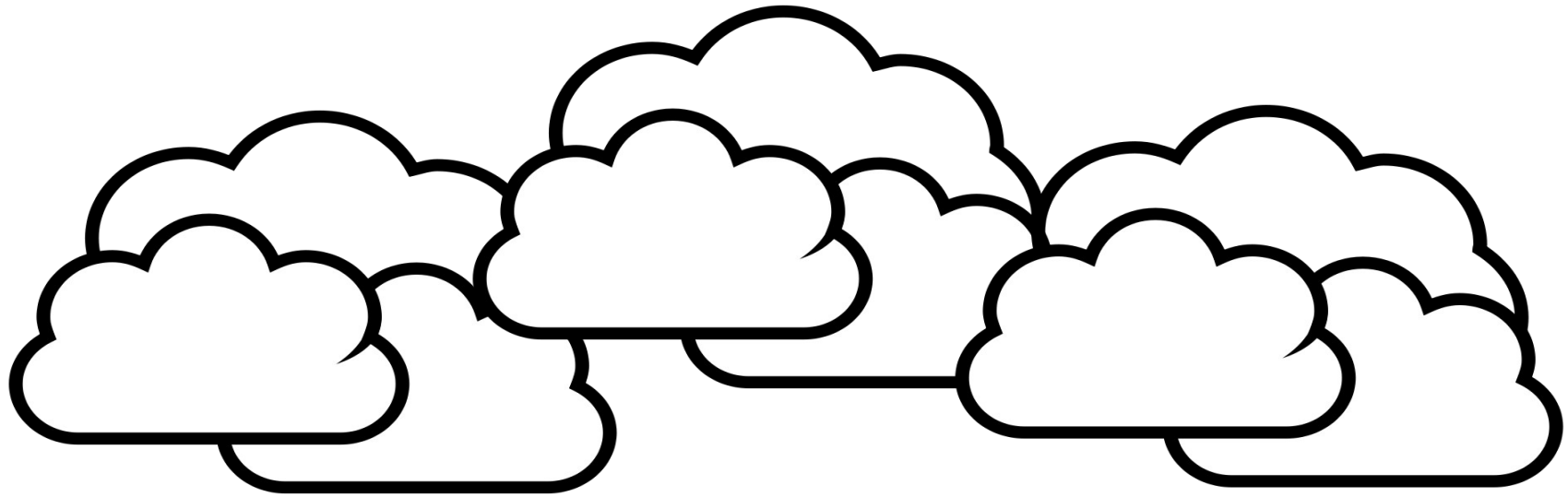
THE WISDOM OF  
CLOUDS

What you need to know  
about cloud computing



# Data is the source of Wisdom !!

用雲掌握資料，加以分析，形成智能給端用



嵌入式的新思維：未來，**端**的智能來自於**雲**

**Devices share the wisdom of Cloud**



# Key Driving Forces of Cloud Computing

## 雲端運算的關鍵驅動力

端

隨需  
行動服務

**Mobile Service**

用任何連網裝置  
都可以存取資料

租賃取代買斷  
動態隨需付費

降低  
經營成本

**Cost Down**

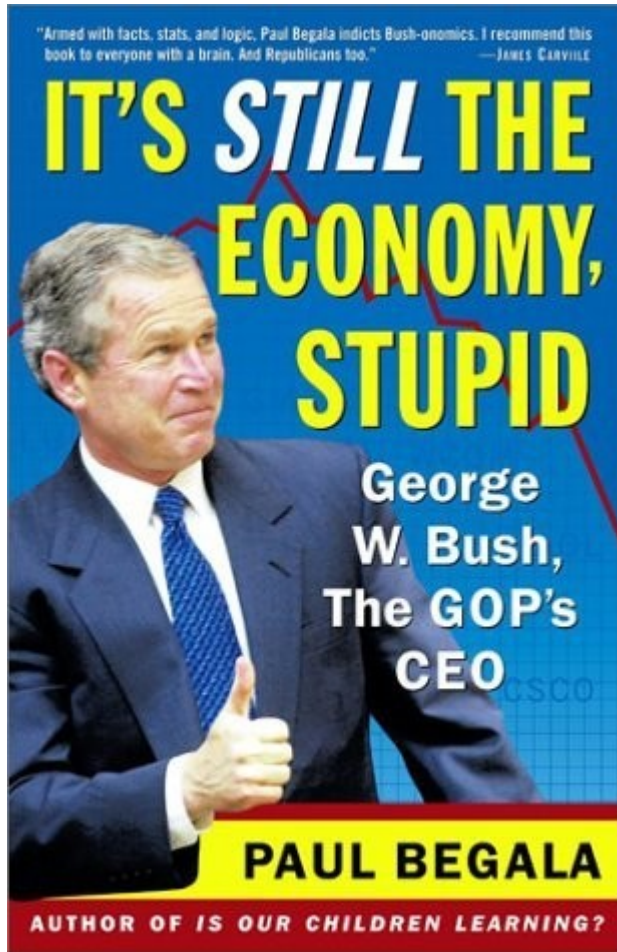
資料往雲擺  
減少資料傳輸

因應  
資料爆炸

**Data Explore**

雲

# IT'S THE DATA, STUPID!



「笨蛋！重點在經濟」

( **"It's the economy, stupid"** )

卡維爾 ( **James Carville** ) 自創這句標語，  
促使柯林頓當上美國第 **42** 屆總統。

- **1992** 年

「笨蛋！重點還是在經濟」

( **"It's STILL the economy, stupid"** )

卻讓小布希被嘲笑是幼稚的總統。

- **2002** 年

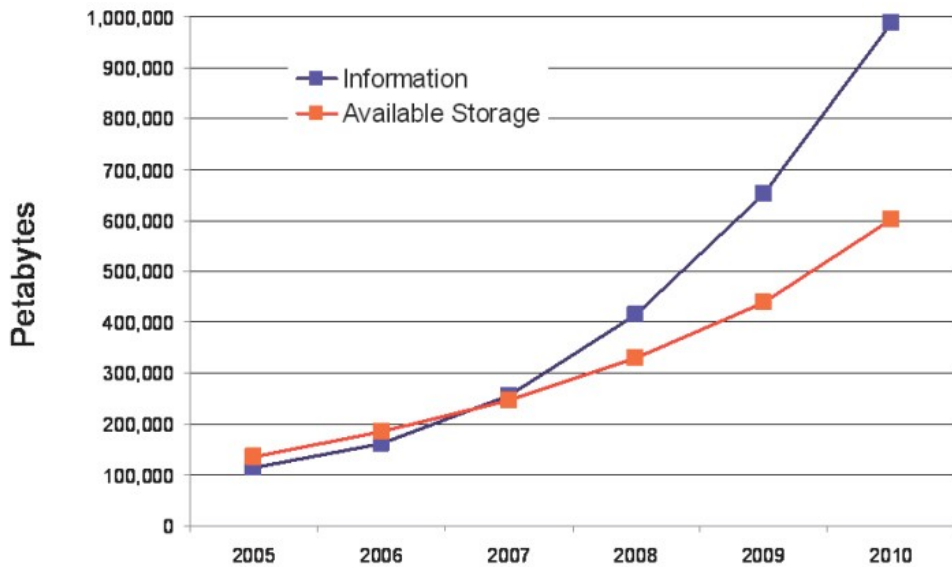
雲端時代，谷歌會說：「笨蛋！重點在資料」

( **"It's the data, stupid"** )

誰掌握了你的資料，就有機會掌握你的荷包  
想想看，電腦、手機掉了，您心疼的是甚麼呢？

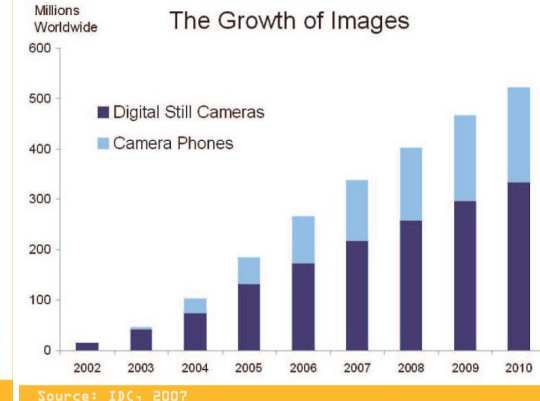
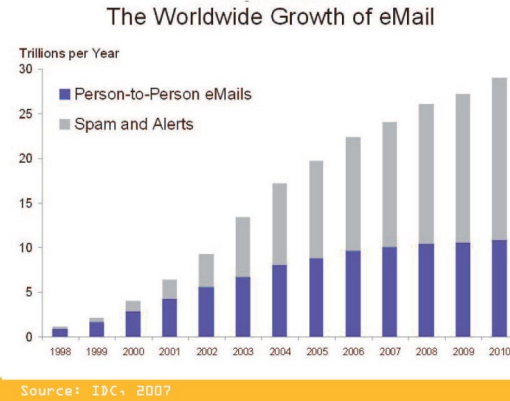
- **2007** 年

# Information Versus Available Storage



# 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**

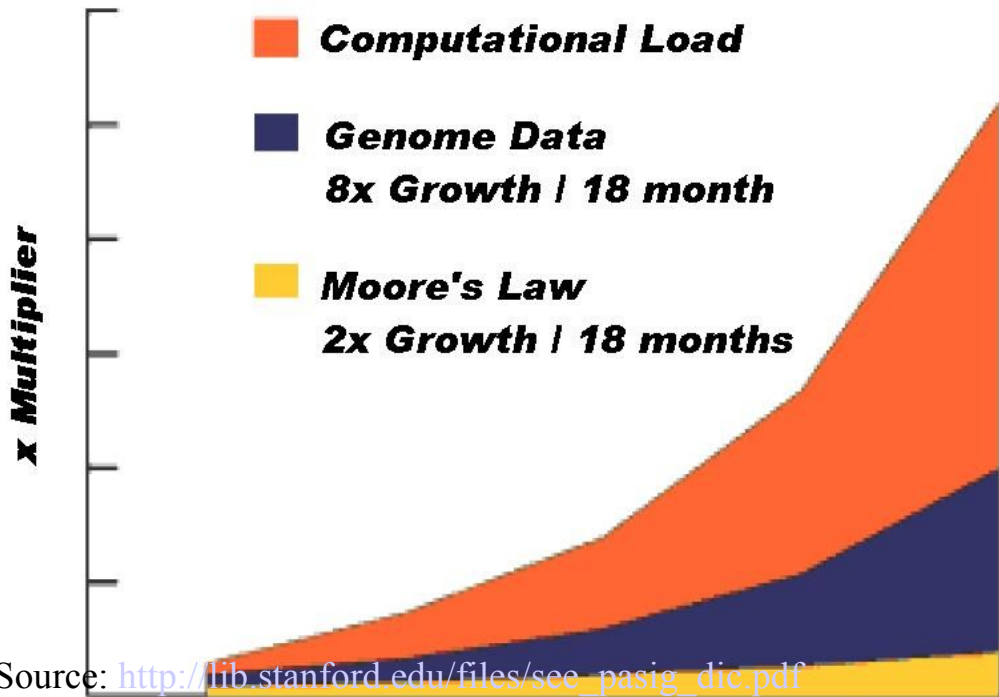


Source: <http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf>

Source: IDC, 2007

Source: IDC, 2007

Source: IDC, 2007



Source: [http://lib.stanford.edu/files/sec\\_pasig\\_dic.pdf](http://lib.stanford.edu/files/sec_pasig_dic.pdf)

Particle Physics Large Hadron Collider (15PB)	Human Genomics (7000PB) 1GB / person 200PB+ captured 200% CAGR	World Wide Web (~1PB)	Wikipedia (10GB) 100% CAGR
Annual Email Traffic, no spam (300PB+)	Internet Archive (1PB+)	Estimated On-line RAM in Google (8PB)	Personal Digital Photos (1000PB+) 100% CAGR
200 of London's Traffic Cams (8TB/day)	2004 Walmart Transaction DB (500TB)	Typical Oil Company (350TB+)	Merck Bio Research DB (1.5TB/qtr)
UPMC Hospitals Imaging Data (500TB/yr)	MIT Babytalk Speech Experiment (1.4PB)	Terashake Earthquake Model of LA Basin (1PB)	One Day of Instant Messaging in 2002 (750GB)
Total digital data to be created this year <b>270,000PB</b> (IDC)			

Phillip B. Gibbons, Data-Intensive Computing Symposium

# 端

平板行動應用

社交溝通協作

多媒體內容

次世代分析

社交分析

情境感知運算

儲存等級記憶體

無所不在的運算

模組化基礎建設

雲端運算

**SaaS :**  
**Web 2.0**

**PaaS :**  
**Big Data**

**IaaS :**  
**Virtualization**

社交網路

評價排行榜

即時搜尋

智慧裝置

大量資訊分析

雲端運算

# 雲



# Two Type of Cloud Architecture ?

雲端架構的兩大陣營？

**SaaS :**  
**Web 2.0**

**IaaS :**  
**Virtualization**



想盡辦法誘你用計算跟網路  
**Computing Intensive**

**SaaS :**  
**Web 2.0**

**PaaS :**  
**Big Data**



想盡辦法誘你提供資料作分析  
**Data Intensive**

# How can we build Cloud Services ??

觀察雲端關鍵提供者，找尋打造雲端服務的模式

Public Cloud  
公用雲端

Target Market  
is **S.M.B.**  
主要客戶為  
中小企業



可敬的對手  
Reference Model

Hybrid  
Cloud

以**大型企業**  
為主要客戶  
**Enterprise** is  
key market

Community Cloud  
社群雲端

Academia **學術**為主



**私有雲端**  
Private Cloud



- Amazon Web Service ( AWS )
- 虛擬伺服器：**Amazon EC2**
  - Small (Default) \$0.085 per hour(L) - \$0.12 per hour(W)
  - All Data Transfer \$0.15 per GB
- 儲存服務：**Amazon S3**
  - \$0.15 per GB – first 50 TB / month of storage used
  - \$0.15 per GB – all data transfer in
  - \$0.01 per 1,000 PUT, COPY, POST, or LIST requests
- 觀念：**Paying for What You Use**

參考來源：  
<http://eblog.cisnet.org.tw/post/Cloud-Computing.aspx>  
<http://aws.amazon.com/ec2/pricing/>  
<http://aws.typepad.com/aws/2010/02/aws-data-transfer-prices-reduced.html>  
<http://aws.amazon.com/s3/#pricing>

# Amazon Web Services 亞馬遜的網頁服務

## E-Commerce

Fulfillment Web Service (FWS)

## Database

SimpleDB, Rational DB Service

## Compute

EC2, Elastic MapReduce (EMR)

## Network

VPC, Elastic Load Balance

## Storage

S3, Elastic Block Storage (EBS)

較外顯的基本服務

## Support

AWS Premium Support

## Payment / Billing

Amazon DevPay

## Messaging

Simple Queue / Notification Service

## Web Traffic

Alexa Web Info / Top Sites

## Content Delivery

Amazon CloudFront

較內隱的流程細節

# Reference Cloud Architecture

## 雲端運算的參考架構

### 應用軟體 Application

Social Computing, Enterprise, ISV, ...

### 程式語言 Programming

Web 2.0 介面, Mashups, Workflows, ...

### 控制管理 Control

Qos Negotiation, Admission Control, Pricing, SLA Management, Metering...

### 虛擬化 Virtualization

VM, VM management and Deployment

### 硬體設施 Hardware

Infrastructure: Computer, Storage, Network

User-Level

User-Level  
Middleware

Core  
Middleware

System Level

IaaS  
PaaS  
SaaS

# Open Source to build Cloud Service

## 建構雲端服務的 自由軟體

### 應用軟體 Application

Social Computing, Enterprise, ISV, ...

eyeOS, Nutch, ICAS,  
X-RIME, ...

### 程式語言 Programming

Web 2.0 介面, Mashups, Workflows, ...

Hadoop (MapReduce),  
Sector/Sphere, AppScale

### 控制管理 Control

Qos Negotiation, Admission Control,  
Pricing, SLA Management, Metering...

OpenNebula, Enomaly,  
Eucalyptus, OpenQRM, ...

### 虛擬化 Virtualization

VM, VM management and Deployment

Xen, KVM, VirtualBox,  
QEMU, OpenVZ, ...

### 硬體設施 Hardware

Infrastructure: Computer, Storage,  
Network

- Google App Engine (GAE)
- 讓開發者可自行建立網路應用程式於 Google 平台之上。
- 提供：
  - 500MB of storage
  - up to 5 million page views a month
  - 10 applications per developer account
- 限制：
  - 程式設計語言只能用 Python 或 Java
- 計費標準：
  - 連出頻寬 \$0.12 美元/GB, 連入頻寬 \$0.10 美元/GB
  - CPU 時間 \$0.10 美元/時
  - 儲存的資料 \$0.15 美元/GB-每月
  - 電子郵件收件者 \$0.0001 美元/每個收件者



# Three Core Technologies of Google ....

## Google 的三大關鍵技術 ....

- 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 : Simplified 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>





# Open Source Mapping of Google Core Technologies

## Google 三大關鍵技術對應的自由軟體

### BigTable

A huge key-value datastore

HBase, Hypertable  
Cassandra, ....

### MapReduce

To parallel process data

Hadoop MapReduce API  
Sphere MapReduce API, ...

### Google File System

To store petabytes of data

Hadoop Distributed File System (HDFS)  
Sector Distributed File System

更多不同語言的 MapReduce API 實作：

<http://trac.nchc.org.tw/grid/intertrac/wiki%3Ajazz/09-04-14%23MapReduce>

其他值得觀察的分散式檔案系統：

- IBM GPFS - <http://www-03.ibm.com/systems/software/gpfs/>
- Lustre - <http://www.lustre.org/>
- Ceph - <http://ceph.newdream.net/>



# 深入解析雲端虛擬化技術

Part 2 : Deep Dive into Virtualization Technologies

**Jazz Wang**

**Yao-Tsung Wang**

**jazz@nchc.org.tw**



Powered by DRBL

# What is Virtualization ??

## 虛擬化技術有哪些呢??

Application Virtualization 應用程式虛擬化

Desktop Virtualization  
Client Virtualization 桌面虛擬化

Presentation Virtualization 顯示虛擬化

OS-level Virtualization 作業系統虛擬化

Network Virtualization 網路虛擬化

Storage Virtualization 儲存虛擬化

資料庫虛擬化

Database Virtualization

資料虛擬化

Data Virtualization

# Open Source for Virtualization

## 虛擬化技術對應的自由軟體

Application Virtualization  
應用程式虛擬化

Ex. VMWare ThinApp

Desktop Virtualization  
桌面虛擬化

Redhat SPICE

Presentation Virtualization  
顯示虛擬化

VNC, FreeNX

OS-level Virtualization  
作業系統虛擬化

Xen, KVM, OpenVZ

Network Virtualization  
網路虛擬化

OpenFlow vSwitch

Storage Virtualization  
儲存虛擬化

Lessfs, SDFS

# NIST Mapping of Cloud Technologies

## 美國國家標準局的定義主要鎖定虛擬化技術

**5. Measured Service**  
可被監控與量測的服務

Monitoring / AAA  
狀態監控與認證收費機制

**1. On-demand self-service.**  
隨需自助服務

VM Management Tool  
虛擬機器管理平台

**2. Broad network access**  
隨時隨地用任何網路裝置存取

OS-level Virtualization  
作業系統虛擬化

**3. Resource pooling**  
多人共享資源池

Network Virtualization  
網路虛擬化

**4. Rapid elasticity**  
快速重新佈署靈活度

Storage Virtualization  
儲存虛擬化

# NIST Mapping of Open Source Cloud

## 美國國家標準局的定義對應自由軟體技術

Monitoring / AAA  
狀態監控與認證收費機制

Ganglia, Nagios  
/ OpenID, SAML

VM Management Tool  
虛擬機器管理平台

OpenNebula, Eucalyptus

OS-level Virtualization  
作業系統虛擬化

Xen, KVM, OpenVZ

Network Virtualization  
網路虛擬化

OpenFlow vSwitch

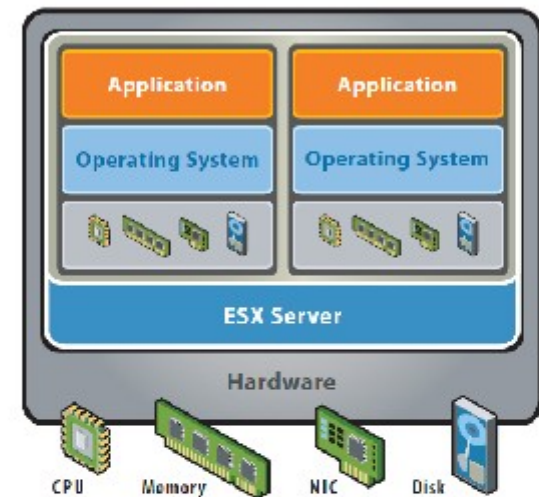
Storage Virtualization  
儲存虛擬化

Lessfs, SDFS, ZFS

# Adjust #1: Virtualization

## 企業導入雲端參考策略 (1) Virtualization

- 導入虛擬化的效益高於雲端運算
- 導入評量指標：
  - 總硬體使用率 (100% CPU 使用 ?)
  - 總電源開銷、空調、管理人力
- 商業解決方案：
  - **VMWare** ESXi / vSphere
  - **Microsoft** Hyper-V
  - **Citrix** XenServer
- 預期效益：
  - 減少伺服器採購成本與營運成本
  - (Server Consolidation)
  - 增加管理彈性與災害復原機制
  - (Ex. 異常斷電造成的服務修復)



VMware ESX Server virtualizes server storage and networking, allowing multiple applications to run in virtual machines on the same physical server.



Windows Server™ 2008  
Hyper-V™



# How to Evaluate the need of Virtualization ?

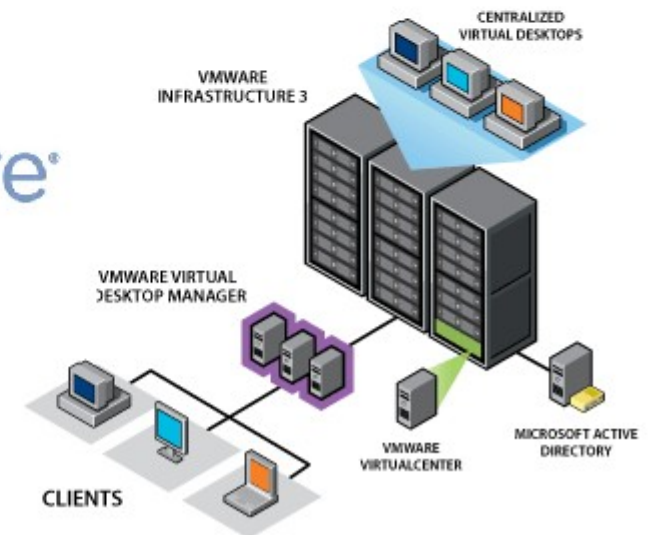
## 如何評估對虛擬化的需求??

- 關鍵驅動因素
  - 預算刪減？節約能源？提高現有硬體使用率？
- 導入評估工具
  - Microsoft Assessment and Planning (MAP) Toolkit for Hyper-V
  - Hyper-V 評估工具：收集 Server 使用情況，產生評估報告等
  - Open Source: MRTG, Ganglia, Nagios, ....
  - 蒐集 CPU 用量、記憶體用量、尖峰負載發生時間、批次排程時間
- 隱含額外成本
  - 若想獲得虛擬化帶來的好處（災害復原與動態負載平衡），需要額外建置共享的儲存設施。



# Thin Client and VDI for next IT procurement ? 未來改採購精簡型電腦與桌面虛擬化??

- 桌面虛擬化的需求越來越高??
- 導入評量指標：
  - 總電源開銷、空調、管理人力
  - 既存共用儲存設施
  - 資料機敏性高，不容外洩
- 商業解決方案：
  - **VMWare** Viewer 4
  - **Microsoft** Hyper-V
  - **Citrix** XenDesktop
- 預期效益：
  - 降低電力損耗 (需評估)
  - 減少資料外洩 (搭配資安政策)



# Another alternative of Virtualization

虛擬化番外篇：改用 **Terminal** + 無碟架構

用綠色自由軟體

**GREEN OSS**

打造綠色工廠

**GREEN FACTORY**

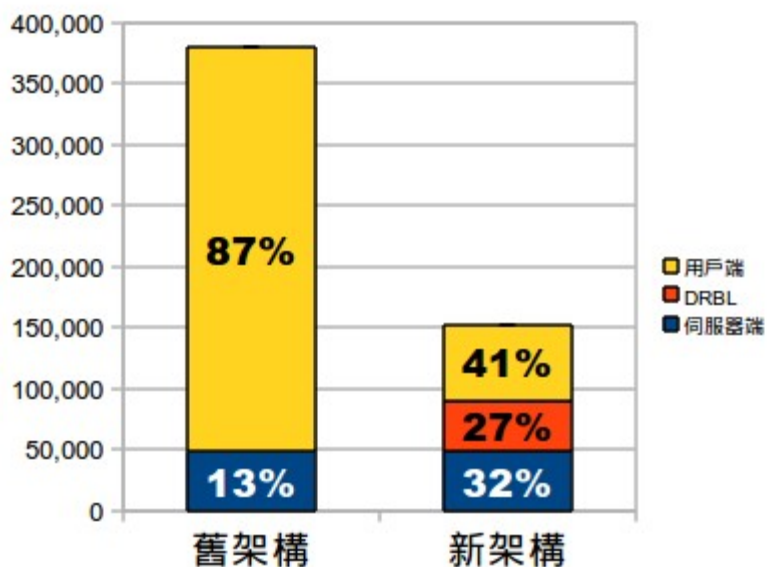
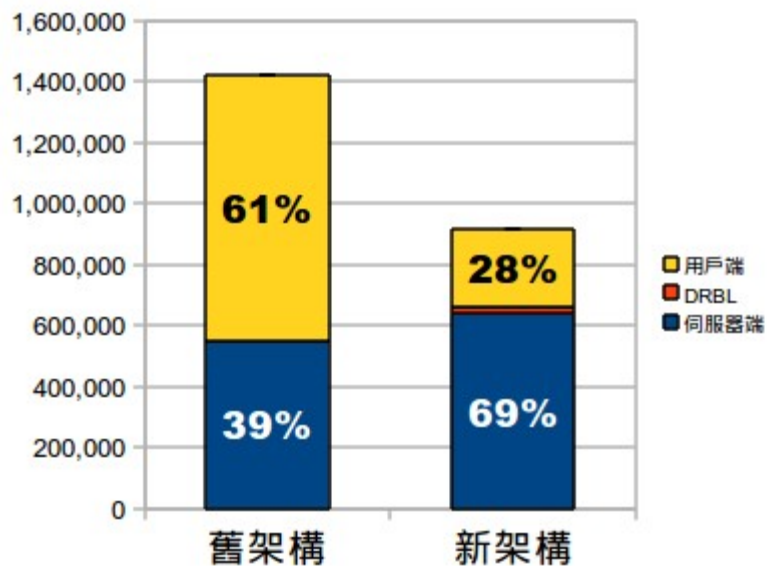
關於企鵝龍 (DRBL) 在金屬加工業的節能案例分享，請參考

[http://trac.nhc.org.tw/cloud/attachment/wiki/jazz/09-10-14/TSMC/08-08-17\\_DRBL\\_Green\\_Computing\\_v2.pdf](http://trac.nhc.org.tw/cloud/attachment/wiki/jazz/09-10-14/TSMC/08-08-17_DRBL_Green_Computing_v2.pdf)

# 修改理念 = 減少不必要的開支與能源消耗

- 降低硬體建置開支
  - 不用 **200W** 的桌上型電腦 (有風扇→易故障)
  - 改用 **20W** 的精簡型電腦 (無風扇→耐油氣)
  - **不裝硬碟** (少一個零件也可減少用電)
- 降低軟體授權開支
  - 購買 **Windows** 作業系統與 **Office** 授權只爲了相當簡單的報工作業並不划算
  - 改用 **Terminal 遠端桌面連線** 模式
- 降低 **MIS** 系統管理人員負擔
  - 故障時直接拿備品精簡電腦替換→**免安裝設定**

# 新舊架構成本比較 (總建置成本 / 年度營運成本)



舊架構	
伺服器端	550,000
一般用戶端	870,000
新架構	
伺服器端	638,000
DRBL	25,000
一般用戶端	255,000
合計金額	
舊架構	1,420,000
新架構	918,000
節省金額	
	502,000

舊架構	
伺服器端	48,198
一般用戶端	331,920
新架構	
伺服器端	48,198
DRBL	41,628
一般用戶端	62,442
合計金額	
舊架構	380,118
新架構	152,268
節省金額	
	227,850

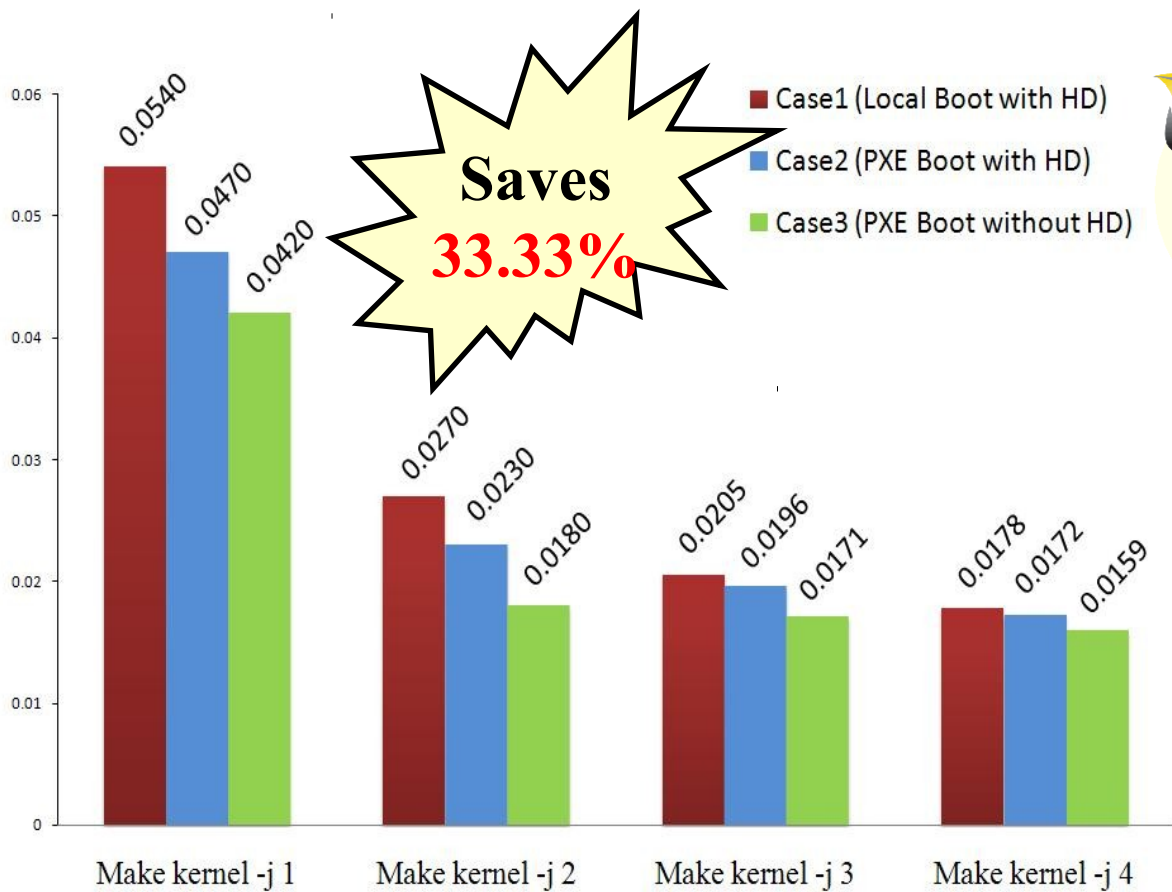
節省 **50 萬**  
總建置成本

省 **17 萬** 硬體  
省 **33 萬** 軟體

節省 **22 萬**  
年度營運成本

省 **3 萬** 電費 (60%)  
省 **19 萬** 工資 (57%)

# 若無很高的檔案讀寫需求，多善用 RAM Disk，無碟架構不僅省電最高達 33%，效能還比較好



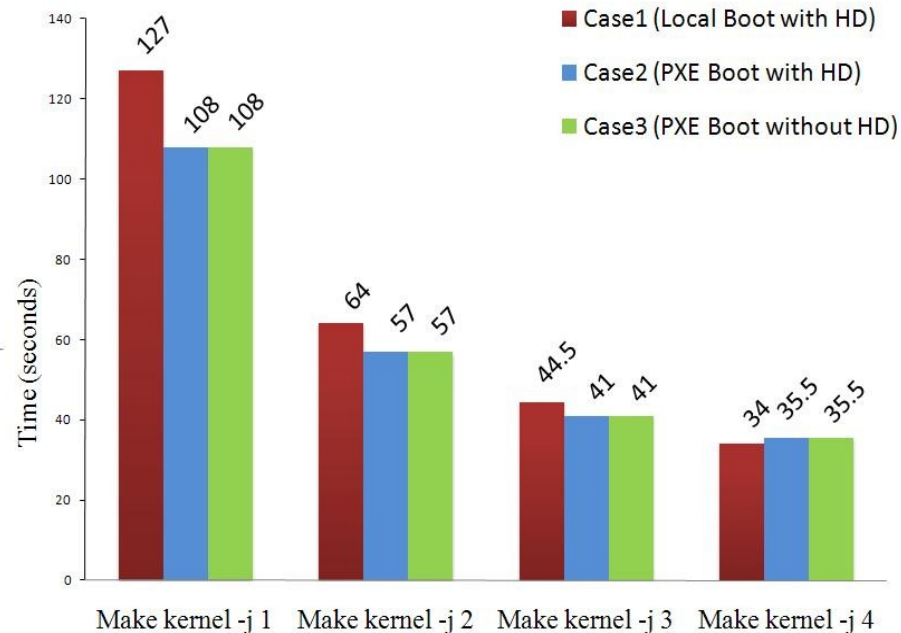
企鵝龍 **DRBL**  
(Diskless Remote Boot in Linux)

適合將整個電腦教室轉換成純自由軟體環境



再生龍 **Clonezilla**

適用完整系統備份、裸機還原或災難復原



關於企鵝龍與再生龍的更多資訊，請參考

<http://drbl.nchc.org.tw> , <http://clonezilla.nchc.org.tw>

關於 E2CC 的更多資訊，請參考

<http://trac.nchc.org.tw/grid/raw-attachment/wiki/deliverable09/1003anav.pdf>

# Adjust #2: Storage Virtualization

## 企業導入雲端參考策略(2) 儲存虛擬化

- **資料整合為跨單位整合的第一步 !!**
- 導入評量指標：
  - 資料所有權 / 散佈權 / 重覆比例
  - 資料成長量 = ?? PB/Year
  - 現有資料副本個數 / 災害復原能力
- 商業硬體方案：EMC、NetApp
- 商業軟體方案：IBM GPFS
- 自由軟體方案：
  - Lustre、ZFS、GlusterFS...
- 預期效益：
  - 減少資料重覆 (Data Deduplication)
  - 提供異地副本備份、災害復原機制



# Building IaaS with Open Source

## 用自由軟體打造 IaaS 服務

應用軟體 Application  
Social Computing, Enterprise, ISV, ...

eyeOS, Nutch, ICAS,  
X-RIME, ...

程式語言 Programming  
Web 2.0 介面, Mashups, Workflows, ...

Hadoop (MapReduce),  
Sector/Sphere, AppScale

控制管理 Control  
Qos Negotiation, Admission Control,  
Pricing, SLA Management, Metering...

OpenNebula, Enomaly,  
Eucalyptus, OpenQRM, ...

虛擬化 Virtualization  
VM, VM management and Deployment

Xen, KVM, VirtualBox,  
QEMU, OpenVZ, ...

硬體設施 Hardware  
Infrastructure: Computer, Storage,  
Network

## VMM Tool #1:

# Eucalyptus

- 原是加州大學聖塔芭芭拉分校 (UCSB) 的研究專案
- **It was a research project of UCSB, USA**
- 目前已轉由 Eucalyptus System 這間公司負責維護
- **Now Eucalyptus System provide technical supports.**
- 創立目的是讓使用者可以**打造自己的 EC2**
- **It designed to help user to build their own Amazon EC2**
- 特色是相容於 Amazon EC2 既有的用戶端介面
- **Its feature is compatible with existing EC2 client.**
- 優勢是 Ubuntu 9.04 已經收錄 Eucalyptus 的套件
- **Ubuntu Enterprise Cloud powered by Eucalyptus in 9.04**
- 目前有提供 Eucalyptus 的官方測試平台供註冊帳號
- **You can register trail account at <http://open.eucalyptus.com/>**
- 缺點：目前仍有部分操作需透過指令模式
- **Cons : you might need to type commands in some case**



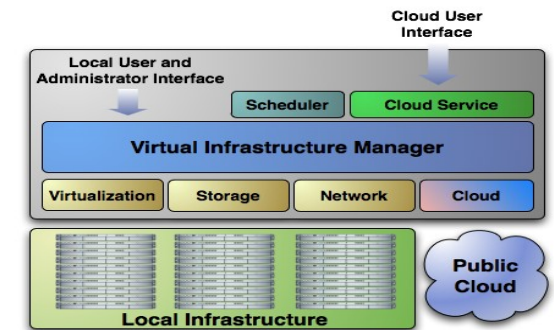
Eucalyptus

關於 Eucalyptus 的更多資訊，請參考  
<http://trac.nchc.org.tw/grid/wiki/Eucalyptus>

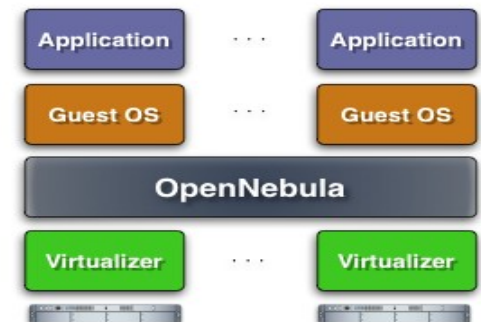


- <http://www.opennebula.org>
- 由歐洲研究學會 (European Union FP7 ) 贊助
- **Sponsor by European Union FP7**
- 將實體叢集轉換成具管理彈性的虛擬基礎設備
- Turn Physical Cluster into Virtual Cluster
- 可管理**虛擬叢集**的狀態、排程、遷徙 (migration)
- **manage status, scheduling and migration of virtual cluster**
- [Ubuntu 9.04 provide package of opennebula](#)
- 缺點：需下指令來進行虛擬機器的遷徙 (migration)。
- **Cons** : You need to type commands to check or migration

OpenNebula.org



關於 OpenNebula 的更多資訊，  
請參考 <http://trac.nchc.org.tw/grid/wiki/OpenNEbula>



# Virtualization for Embedded Device ?

虛擬化只用在「雲」嗎？那「端」呢？



# Motivation for Embedded Virtualization

## 嵌入式軟體採用虛擬化技術的三大動機

多核心  
Multiple  
Cores

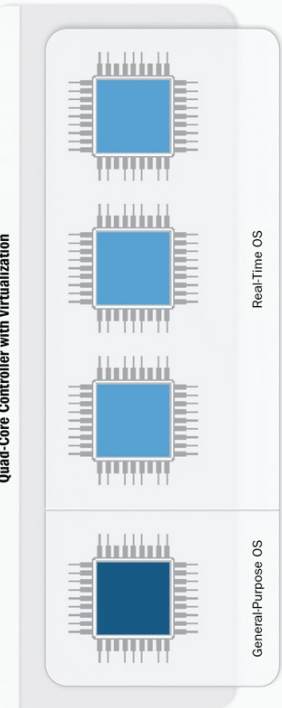
co-existence of different OS environment  
on the same platform

讓同一個硬體平台上共存不同的作業系統

isolating critical components  
from an untrusted OS environment  
在不受信任的作業系統上建立獨立的管制區

an indirection level for remote control  
of OS environments

為作業系統建立一個間接的遠端遙控機制



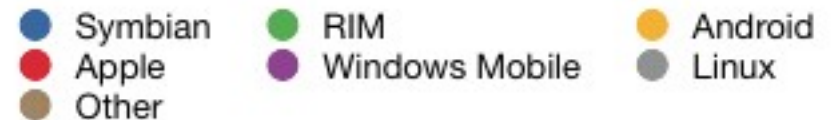
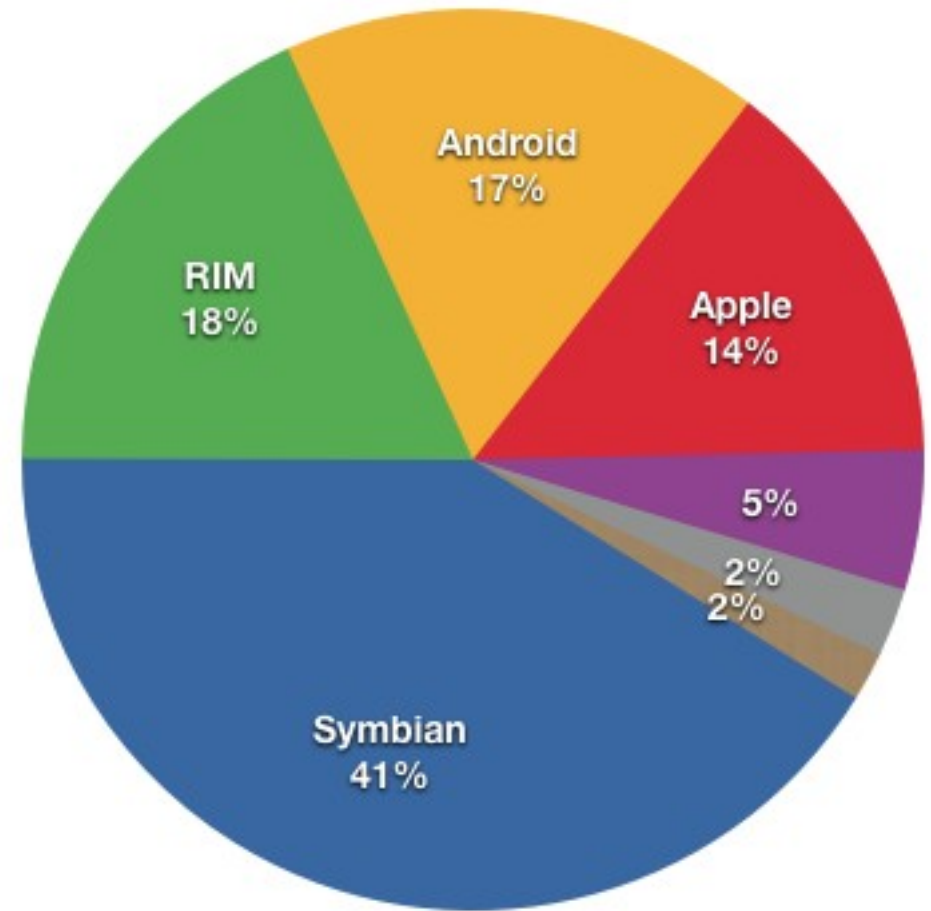
參考來源：**Embedded systems virtualization: Consider a Hypervisor**

<http://www.eetimes.com/design/automotive-design/4016811/Embedded-systems-virtualization-Consider-a-Hypervisor>

參考來源：**10 questions to ask when choosing a virtualization solution**

<http://www.eetimes.com/design/eda-design/4006406/10-questions-to-ask-when-choosing-a-virtualization-solution>

# Application : Key drivers of Multiple OS in one device 應用程式是在同一個裝置跑多作業系統的關鍵驅動力



圖片來源：[http://www.iphonic.tv/iphone\\_screen\\_amaze.png](http://www.iphonic.tv/iphone_screen_amaze.png)

圖片來源：[http://en.wikipedia.org/wiki/File:Smartphone\\_share\\_current.png](http://en.wikipedia.org/wiki/File:Smartphone_share_current.png)

參考來源：**iPhone - Playing mame4iphone with a WiiMote**

<http://www.youtube.com/watch?v=2AqgIWPnr1c>

參考來源：**Android running on iPhone!**

<http://linuxoniphone.blogspot.com/2010/04/ive-been-working-on-this-quietly-in.html>

<http://www.youtube.com/watch?v=5yO2KQHkt4A>

參考來源：**Android on Motorola Rokr E6**

<http://www.youtube.com/watch?v=-QAWd4IjV3g>

參考來源：**Dual Boot Windows Mobile and Android**

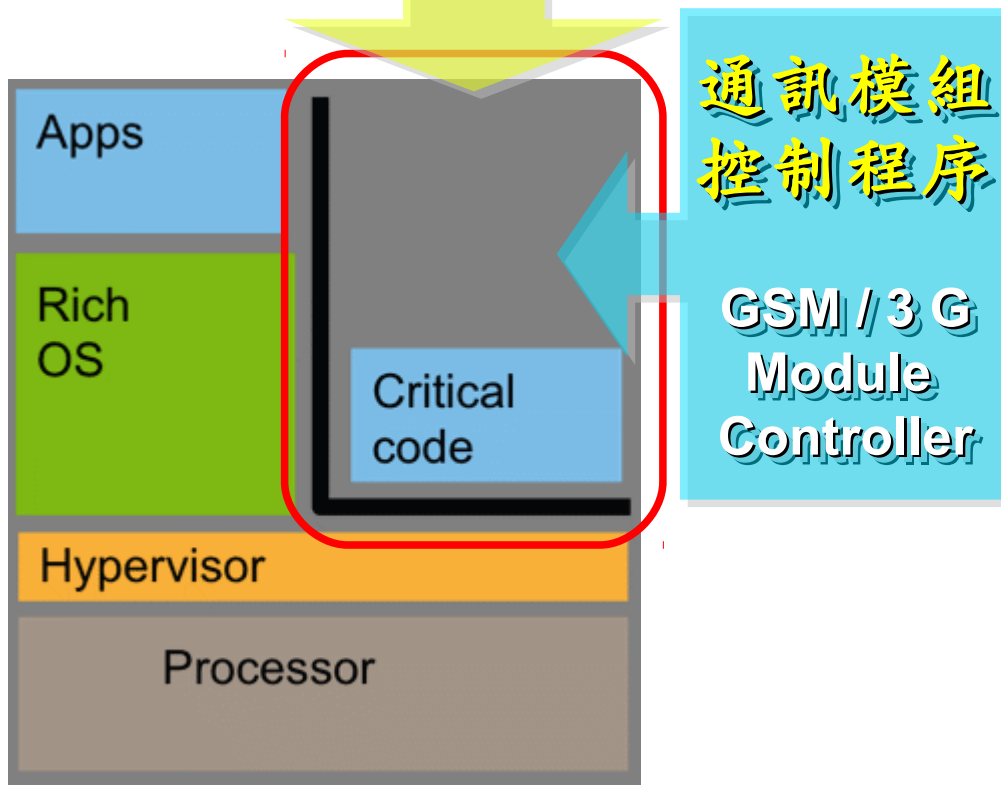
<http://www.redmondpie.com/how-to-dual-boot-windows-mobile-and-android-on-windows-phone-9140407/>

<http://www.youtube.com/watch?v=Nvj4ObHmxCI>

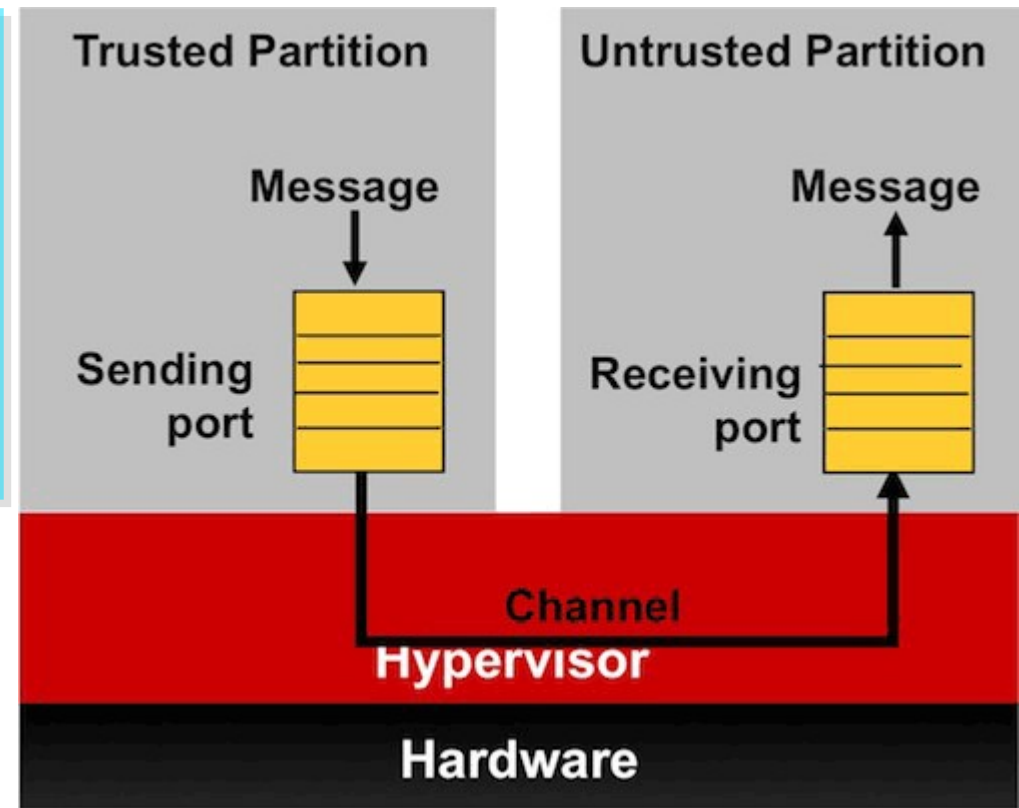
# Virtualization for Embedded Security

## 爲了安全性，必須建立獨立的管制區

DRM 音樂授權  
Digital Rights Management



區域間用訊息傳遞方式溝通  
Communicate with Message Channel



參考來源：**Embedded systems virtualization: Consider a Hypervisor**

<http://www.eetimes.com/design/automotive-design/4016811/Embedded-systems-virtualization-Consider-a-Hypervisor>

參考來源：**Securing Smart Grid Devices - Using Virtualization to Protect the Grid**

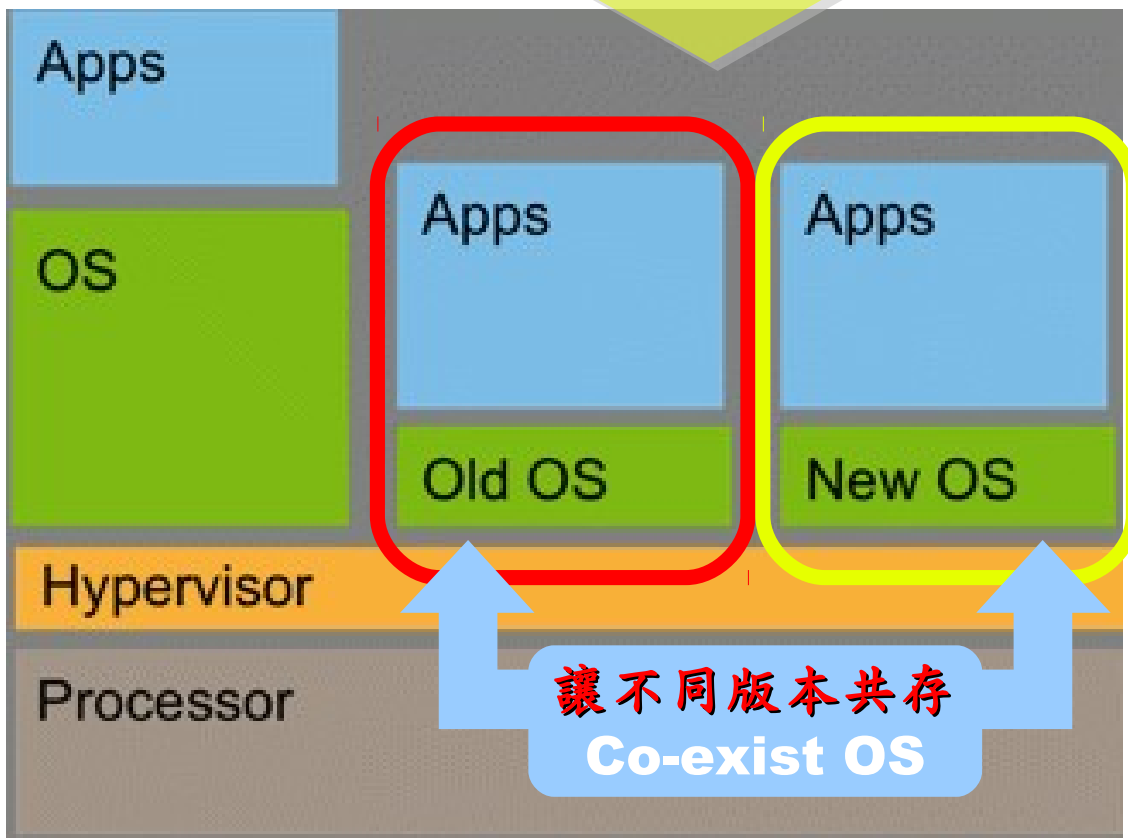
<http://embeddedinnovator.com/securing-smart-grid-devices>

# Upgrade and Data Security : Key drivers of remote control

## 有助於作業系統升級並保護手機被偷時的資料安全

FOTA 韌體升級管制  
firmware-over-the-air

遠距資料清除  
Remote control to wipe data



**\*#06#**

參考來源：**Embedded systems virtualization: Consider a Hypervisor**

<http://www.eetimes.com/design/automotive-design/4016811/Embedded-systems-virtualization-Consider-a-Hypervisor>

# Open Source for Embedded Virtualization

## 嵌入式虛擬化技術的自由軟體

### Embedded **XEN** on ARM platforms

<http://embeddedxen.sf.net>

### **KVM** for ARM

<http://www.ncl.cs.columbia.edu>

sourceforge FIND AND DEVELOP OPEN SOURCE SOFTWARE

[Find Software](#) [Develop](#) [Create Project](#) [Blog](#) [Site Support](#) [About](#)

SourceForge.net > Find Software > Embedded XEN on ARM platforms

 Embedded XEN on ARM platforms Alpha by rossierd

[Summary](#) [Files](#) [Support](#) [Develop](#)

EmbeddedXEN is an academic project and consists of a single multi-kernel ARM-target binary image with bedded

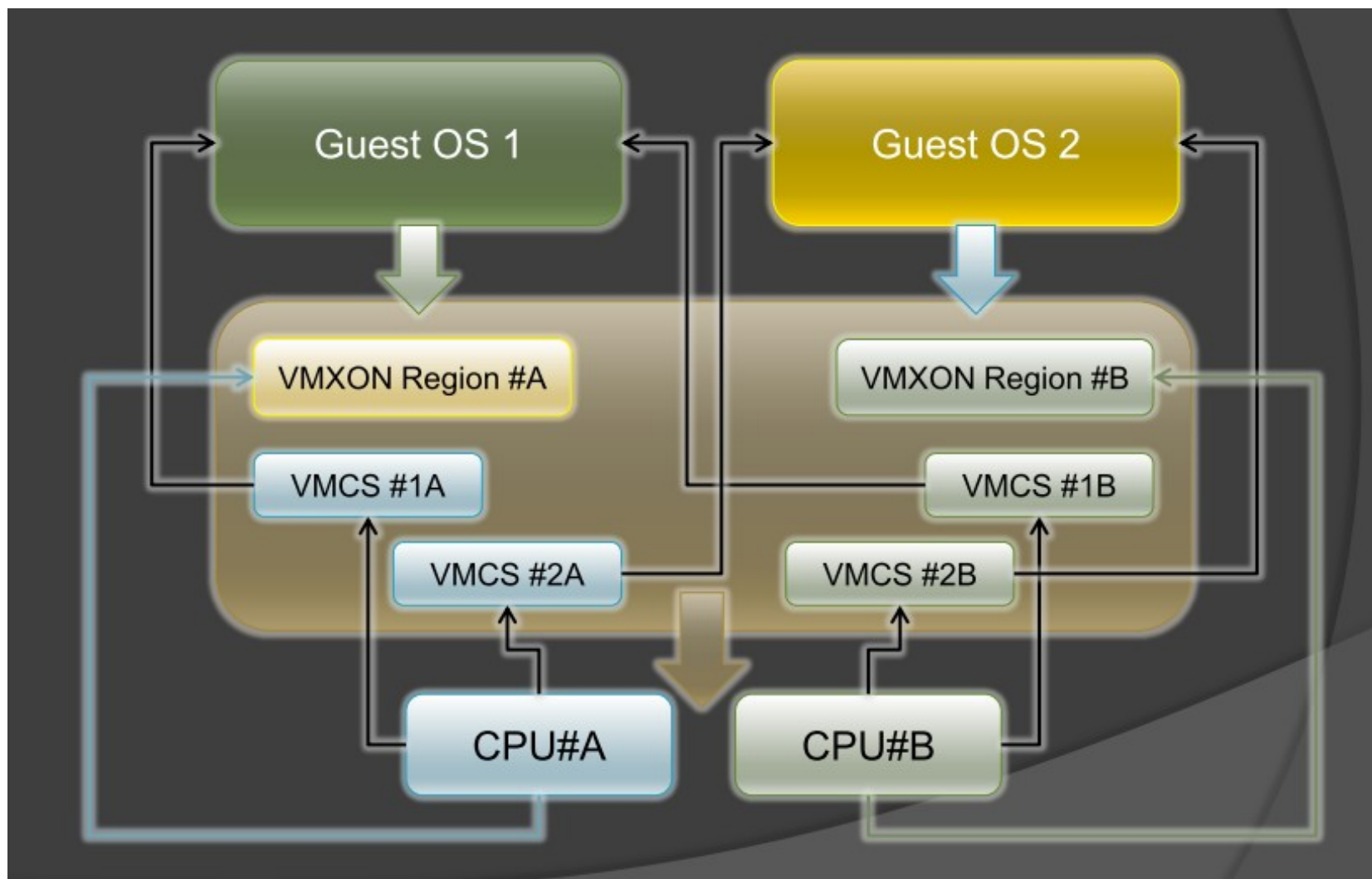
EDIT

[view project details](#)

Requirement	KVM	Xen	Embedded Hypervisor
Embedded Processor Support	no	yes	yes
Real-time	no	no	yes
Fast communication	no	no	yes
Footprint	>10MB	16MB	<64kB
Code size	100s kLoC	100s kLoC	10 kLoC

# 虛擬化衍生的新興資安問題

透過虛擬機器，竊取鍵盤輸入、植入後門 .....



圖片來源： Hacks in Taiwan Conference 2010

[http://www.hitcon.org/hit2010/download/6\\_New%20Battlefield%20For%20Malware%20Game.pdf](http://www.hitcon.org/hit2010/download/6_New%20Battlefield%20For%20Malware%20Game.pdf)

王大寶 & PK / Hypervisor - New Battlefield For Malware Game 虛擬機 - 惡意程式攻防的新戰場





# 深入解析雲端大量資料分析技術

Part 3 : Deep Dive into Data Science Technologies

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



Powered by DRBL

# Open Source Mapping of Google Core Technologies

## Google 三大關鍵技術對應的自由軟體

### BigTable

A huge key-value datastore

HBase, Hypertable  
Cassandra, ....

### MapReduce

To parallel process data

Hadoop MapReduce API  
Sphere MapReduce API, ...

### Google File System

To store petabytes of data

Hadoop Distributed File System (HDFS)  
Sector Distributed File System

更多不同語言的 MapReduce API 實作：

<http://trac.nchc.org.tw/grid/intertrac/wiki%3Ajazz/09-04-14%23MapReduce>

其他值得觀察的分散式檔案系統：

- IBM GPFS - <http://www-03.ibm.com/systems/software/gpfs/>
- Lustre - <http://www.lustre.org/>
- Ceph - <http://ceph.newdream.net/>

# Building PaaS with Open Source

## 用自由軟體打造 PaaS 雲端服務

應用軟體 Application  
Social Computing, Enterprise, ISV, ...

eyeOS, Nutch, ICAS,  
X-RIME, ...

程式語言 Programming  
Web 2.0 介面, Mashups, Workflows, ...

Hadoop (MapReduce),  
Sector/Sphere, AppScale

控制管理 Control  
Qos Negotiation, Admission Control,  
Pricing, SLA Management, Metering...

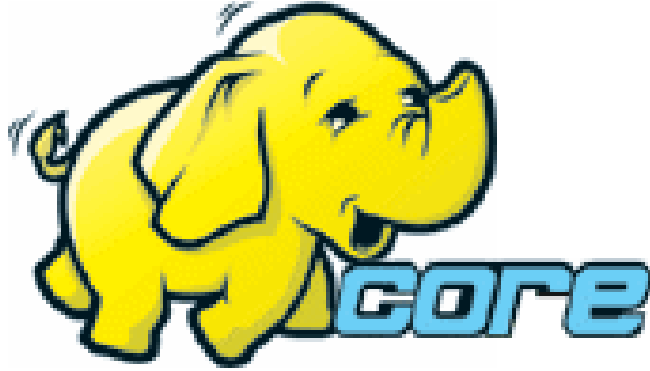
OpenNebula, Enomaly,  
Eucalyptus, OpenQRM, ...

虛擬化 Virtualization  
VM, VM management and Deployment

Xen, KVM, VirtualBox,  
QEMU, OpenVZ, ...

硬體設施 Hardware  
Infrastructure: Computer, Storage, Network

# Hadoop

- <http://hadoop.apache.org>
  - Hadoop 是 Apache Top Level 開發專案
  - **Hadoop is Apache Top Level Project**
  - 目前主要由 Yahoo! 資助、開發與運用
  - **Major sponsor is Yahoo!**
  - 創始者是 Doug Cutting，參考 Google Filesystem
  - **Developed by Doug Cutting, Reference from Google Filesystem**
  - 以 Java 開發，提供 HDFS 與 MapReduce API。
  - **Written by Java, it provides HDFS and MapReduce API**
  - 2006 年使用在 Yahoo 內部服務中
  - **Used in Yahoo since year 2006**
  - 已佈署於上千個節點。
  - **It had been deploy to 4000+ nodes in Yahoo**
  - 處理 Petabyte 等級資料量。
  - **Design to process dataset in Petabyte**
- 
- Facebook、Last.fm  
、Joost are also  
powered by Hadoop**

# Sector / Sphere

- <http://sector.sourceforge.net/>
- 由美國資料探勘中心研發的自由軟體專案。
- **Developed by National Center for Data Mining, USA**
- 採用 C/C++ 語言撰寫，因此效能較 Hadoop 更好。
- **Written by C/C++, so performance is better than Hadoop**
- 提供「類似」Google File System 與 MapReduce 的機制
- **Provide file system similar to Google File System and MapReduce API**
- 基於UDT高效率網路協定來加速資料傳輸效率
- **Based on UDT which enhance the network performance**
- Open Cloud Testbed有提供測試環境，並開發Ma1Stone效能評比軟體
- **Open Cloud Consortium provide Open Cloud Testbed and develop Ma1Stone toolkit for benchmark**

**Sector-Sphere**

National Center for Data Mining  
University of Illinois at Chicago



Open Data Group

<http://www.opendatagroup.com/>

# 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。

# Two Key Elements of Operating System

## 作業系統兩大關鍵組成元素

**Scheduler**  
程序排程

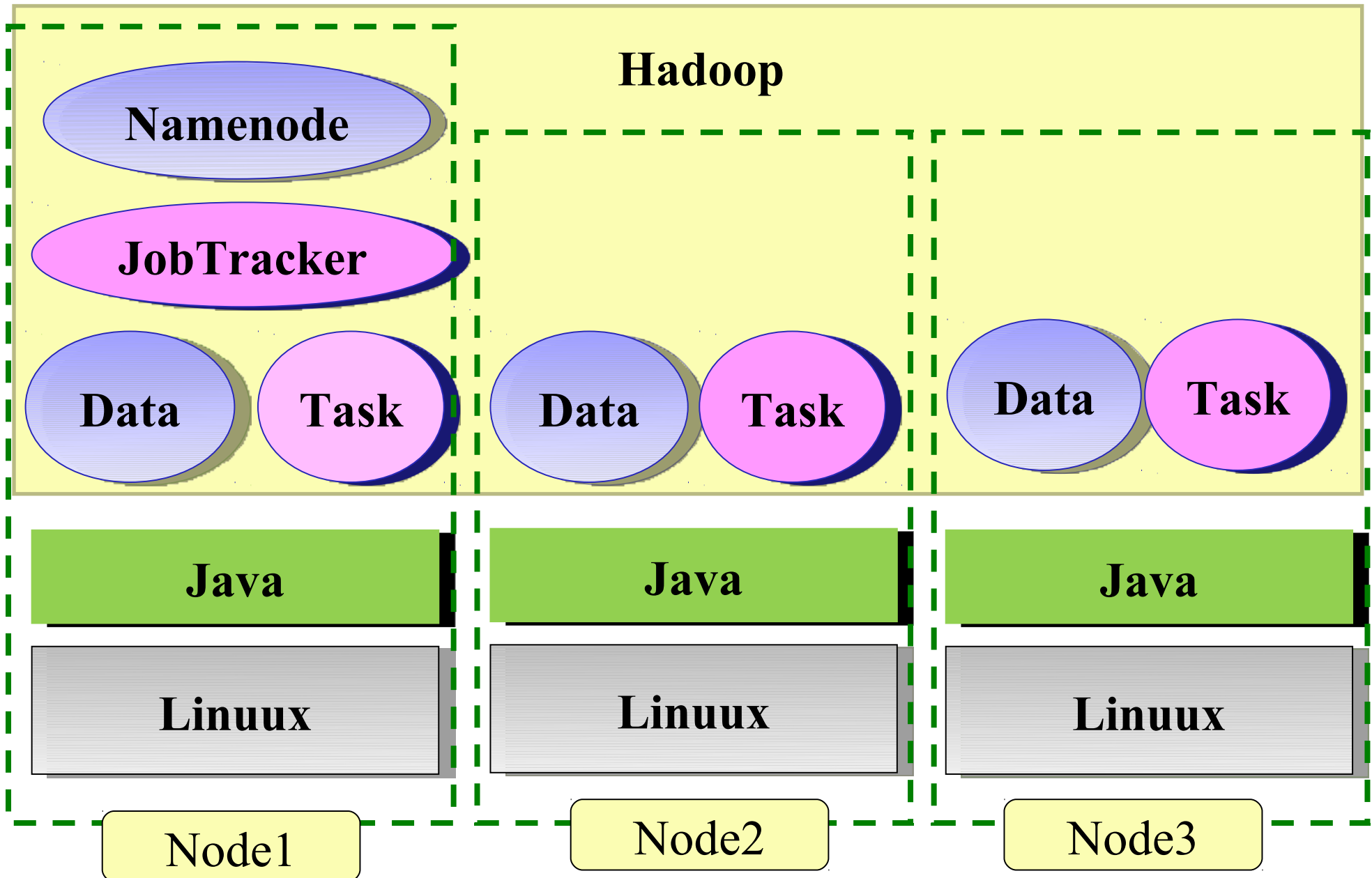


**File System**  
檔案系統



# Distributed Operating System of Hadoop

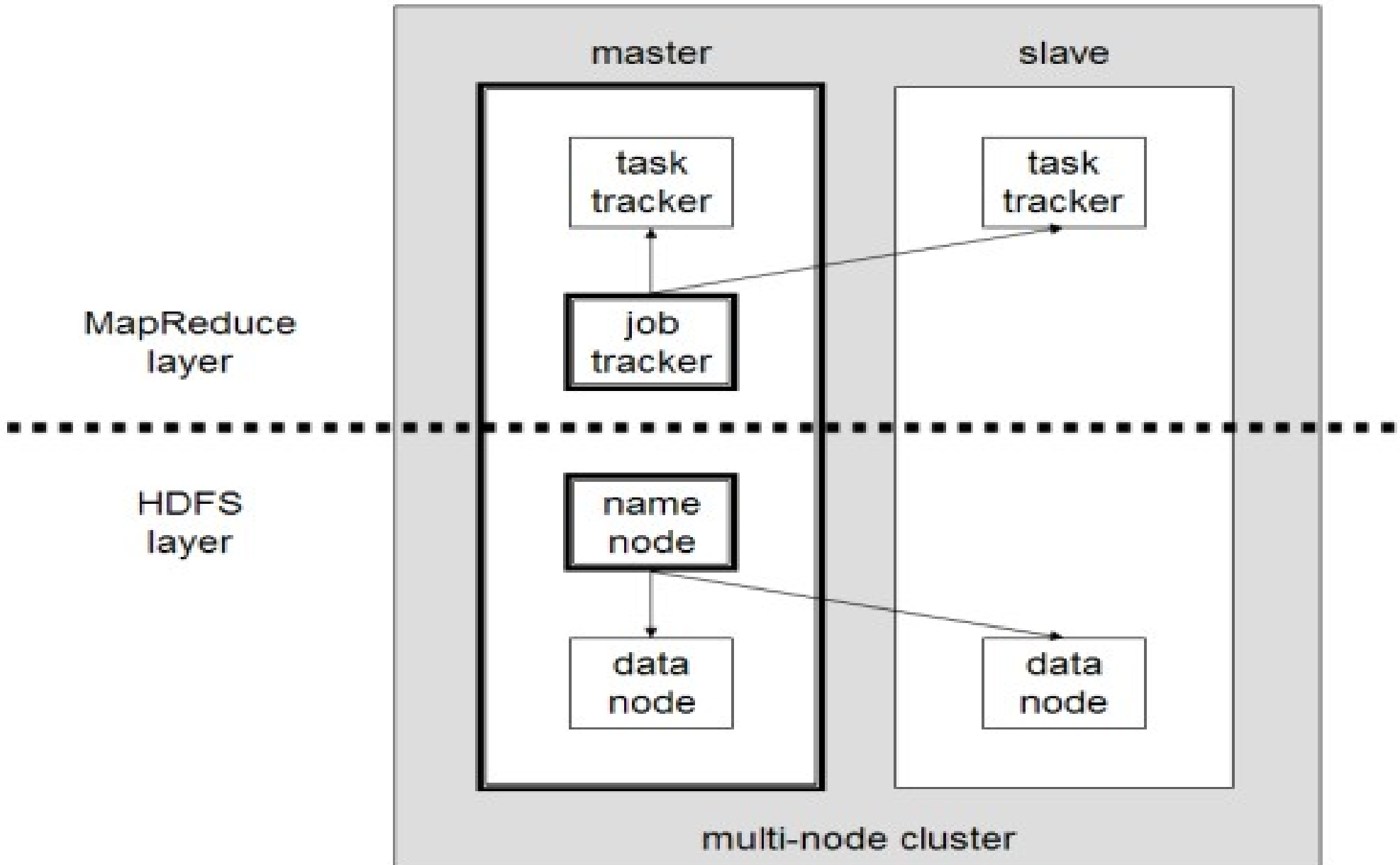
**Hadoop** 建構成一個分散式作業系統





# Different Roles of Hadoop Architecture

## *Hadoop* 軟體架構中的不同角色



# Two Key Roles of HDFS

## HDFS 軟體架構的兩種關鍵角色

### 名稱節點 **NameNode**

- **Master**
- 管理 **HDFS** 的名稱空間
- 控制對檔案的讀 / 寫
- 配置副本策略
- 對名稱空間作檢查及紀錄
- 只能有一個

### 資料節點 **DataNode**

- **Workers**
- 執行讀 / 寫動作
- 執行 **Namenode** 的副本策略
- 可多個

# Two Key Roles of Job Scheduler

## 程序排程的兩種關鍵角色

### JobTracker

- **Master Node**
- 使用者發起工作
- 指派工作給 Tasktrackers
- 排程決策、工作分配、錯誤處理
- 只能有一個

### TaskTracker

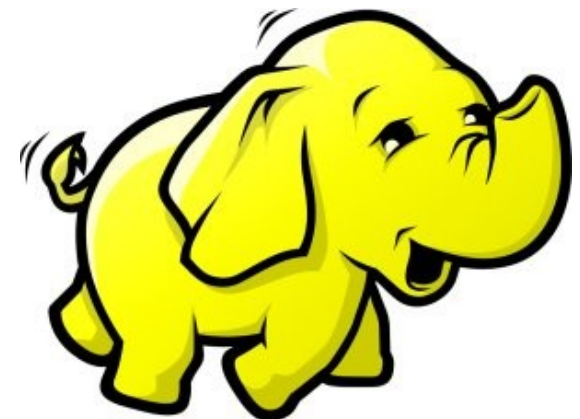
- **Worker Nodes**
- 運作 Map/Reduce 的工作
- 管理儲存、回覆運算結果
- 可多個



# HDFS 簡介

*Introduction to Hadoop Distributed File System*

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



# ***What is HDFS ??***

## ***什麼是 HDFS ??***

- **Hadoop Distributed File System**

- 實現類似 Google File System 分散式檔案系統
- Reference from Google File System.
- 一個易於擴充的分散式檔案系統，目的為對大量資料進行分析
- **A scalable distributed file system for large data analysis .**
- 運作於廉價的普通硬體上，又可以提供容錯功能
- **based on commodity hardware with high fault-tolerant.**
- 給大量的用戶提供總體性能較高的服務
- **It have better overall performance to serve large amount of users.**

# Features of HDFS ...

## HDFS 的特色是 ...

- **硬體錯誤容忍能力 Fault Tolerance**
  - 硬體錯誤是正常而非異常
  - Failure is the norm rather than exception
  - 自動恢復或故障排除
  - automatic recovery or report failure
- **串流式的資料存取 Streaming data access**
  - 批次處理多於用戶交互處理
  - Batch processing rather than interactive user access.
  - 高 Throughput 而非低 Latency
  - High aggregate data bandwidth (throughput)

# Features of HDFS ...

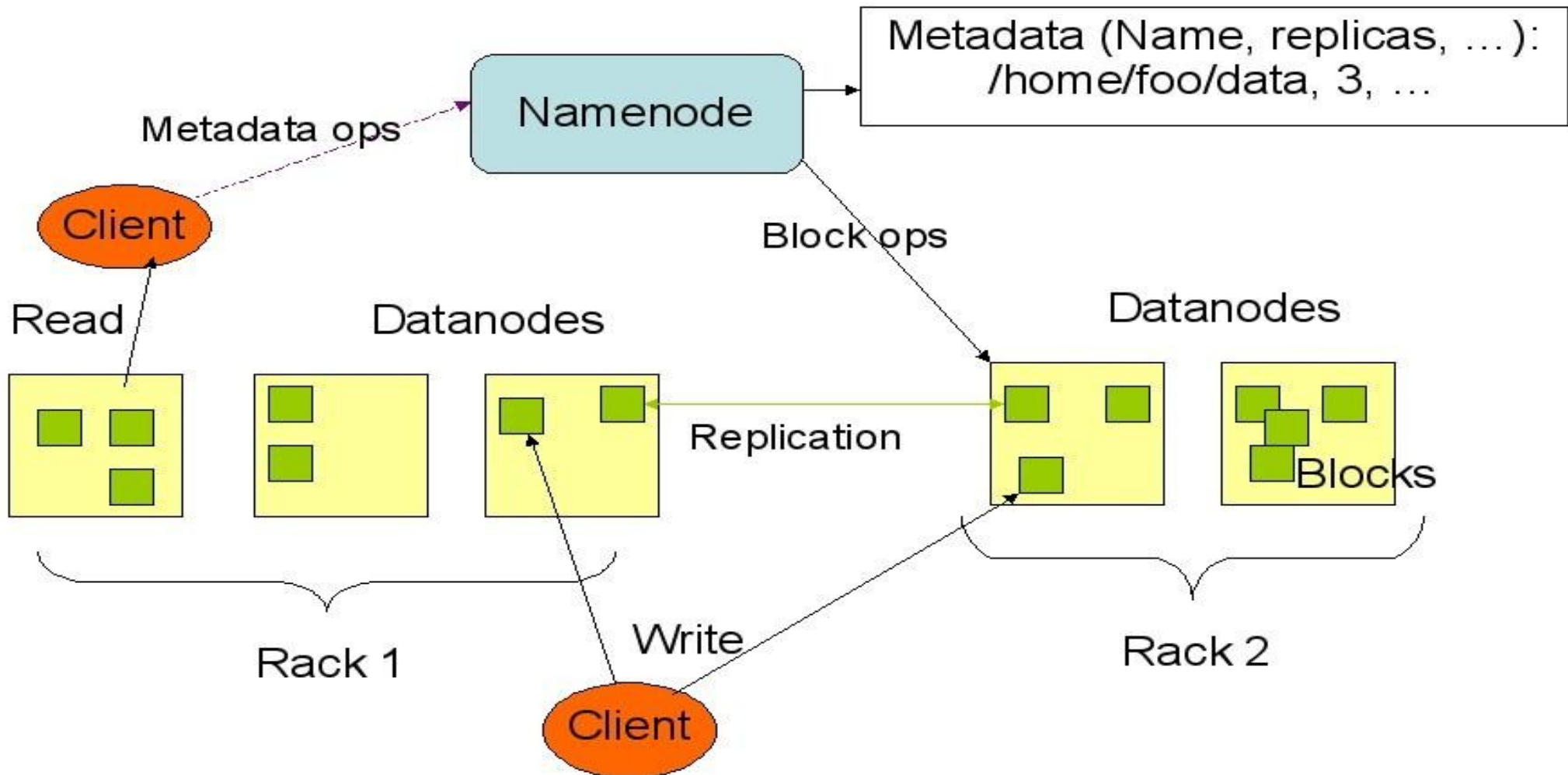
## HDFS 的特色是 ...

- **大規模資料集 Large data sets and files**
  - 支援 Petabytes 等級的磁碟空間
  - Support Petabytes size
- **一致性模型 Coherency Model**
  - 一次寫入，多次存取 Write-once-read-many
  - 簡化一致性處理問題 This assumption simplifies coherency
- **在地運算 Data Locality**
  - 到資料的節點上計算 > 將資料從遠端複製過來計算
  - “move compute to data” > “move data to compute”
- **異質平台移植性 Heterogeneous**
  - 即使硬體不同也可移植、擴充
  - HDFS could be deployed on different hardware

# How HDFS manage data ...

## HDFS 如何管理資料 ...

### HDFS Architecture





# How does HDFS work ...

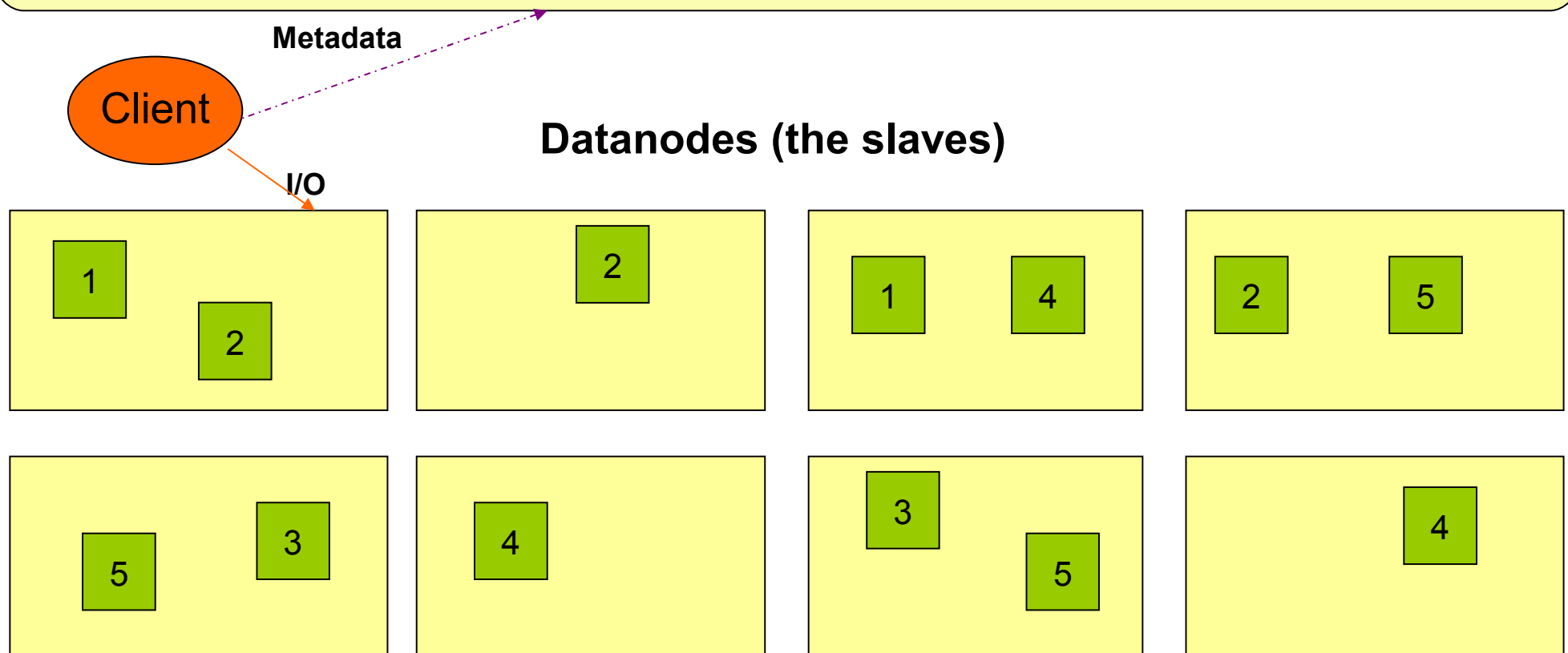
## HDFS 如何運作 ...

Namenode (the master)

Path and Filename – Replication , blocks

name:/users/joeYahoo/myFile - copies:2, blocks:{1,3}

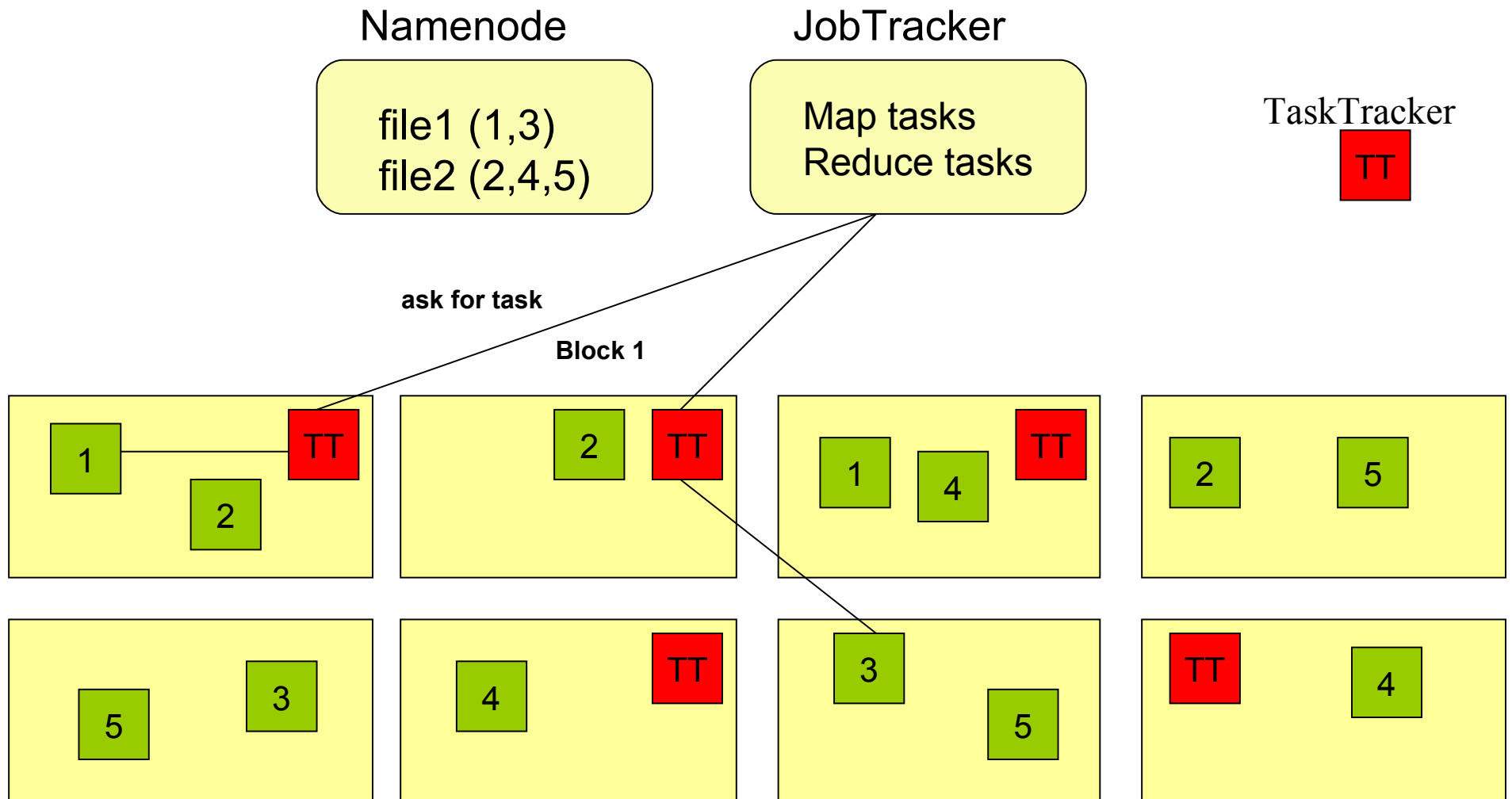
name:/users/bobYahoo/someData.gzip, copies:3, blocks:{2,4,5}



# About Data Locality ...

## HDFS 如何達成在地運算 ...

- Increase reliability and read bandwidth
  - robustness : read replication while found any failure
  - High read bandwidth : distribute read ( but increase write bottleneck )



# About Fault Tolerance ...

## HDFS 如何達成容錯機制 ...

資料崩毀  
Data Corrupt

網路或資料  
節點失效  
Network Fault  
DataNode Fault

名稱節點錯誤  
NameNode Fault

- 資料完整性 Data integrity
  - checked with CRC32
  - 用副本取代出錯資料
  - Replace corrupt block with replication one
- Heartbeat
  - Datanode send **heartbeat** to Namenode
- Metadata
  - FSImage、Editlog 為核心印象檔及日誌檔
  - FSImage – core file system mapping image
  - Editlog – like. SQL transaction log
  - 多份儲存，當名稱節點故障時可以手動復原
  - Multiple backups of FSImage and Editlog
  - Manually recovery while NameNode Fault

# Coherency Model and Performance of HDFS

## HDFS 的一致性機制與效能 ...

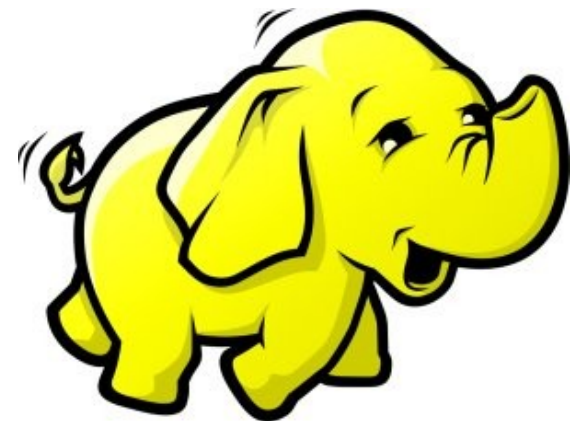
- **檔案一致性機制 Coherency model of files**
  - 刪除檔案 \ 新增寫入檔案 \ 讀取檔案皆由名稱節點負責
  - NameNode handle the operation of write, read and delete.
- **巨量空間及效能機制 Large Data Set and Performance**
  - 預設每個區塊大小以 64MB 為單位
  - By default, the block size is 64MB
  - 大區塊可提高存取效率
  - Bigger block size will enhance read performance
  - 檔案有可能大過一顆磁碟
  - Single file stored on HDFS might be larger than single physical disk of DataNode.
  - 區塊均勻散佈各節點以分散讀取流量
  - Fully distributed blocks increase throughput of reading.



# MapReduce 簡介

*Introduction to MapReduce*

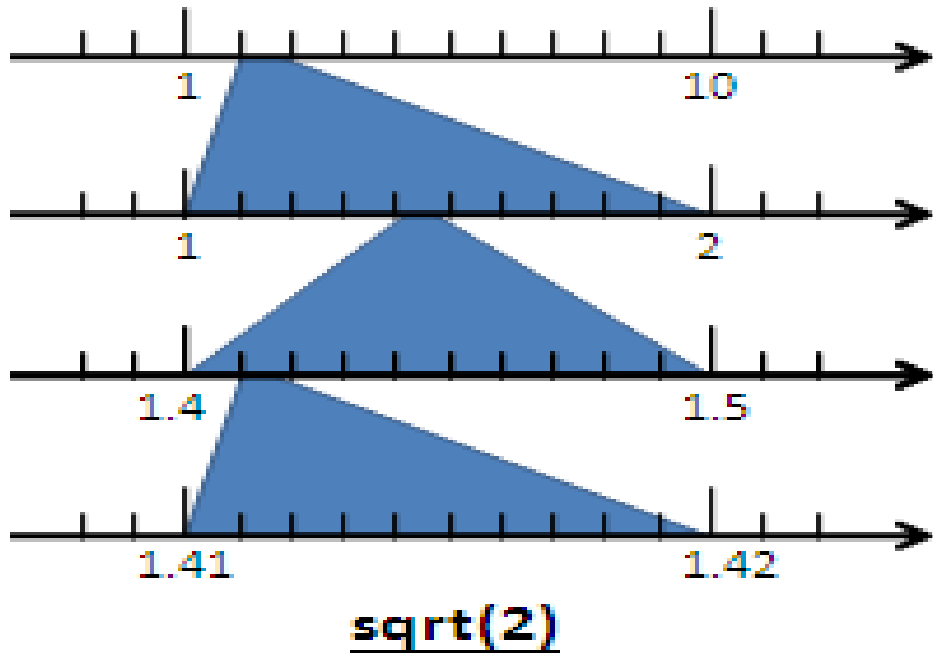
***Jazz Wang***  
***Yao-Tsung Wang***  
***jazz@nchc.org.tw***



# Divide and Conquer Algorithms

## 分而治之演算法

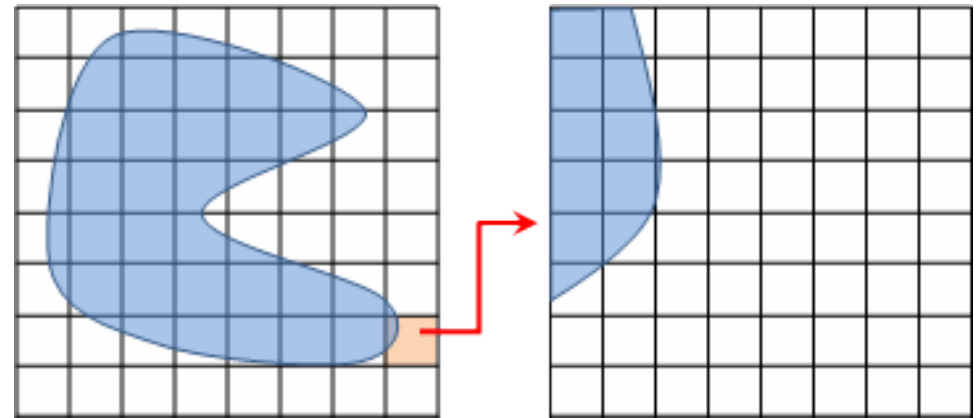
Example 1:



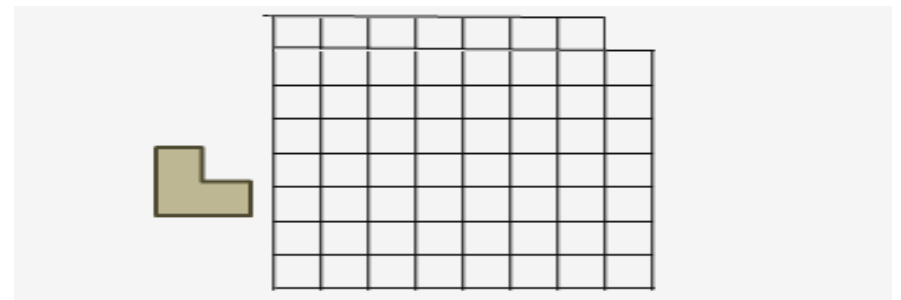
Example 4: The way to climb 5 steps stair within 2 steps each time. 眼前有五階樓梯，每次可踏上一階或踏上兩階，那麼爬完五階共有幾種踏法？

Ex : (1,1,1,1,1) or (1,2,1,1)

Example 2:



Example 3:



# What is MapReduce ??

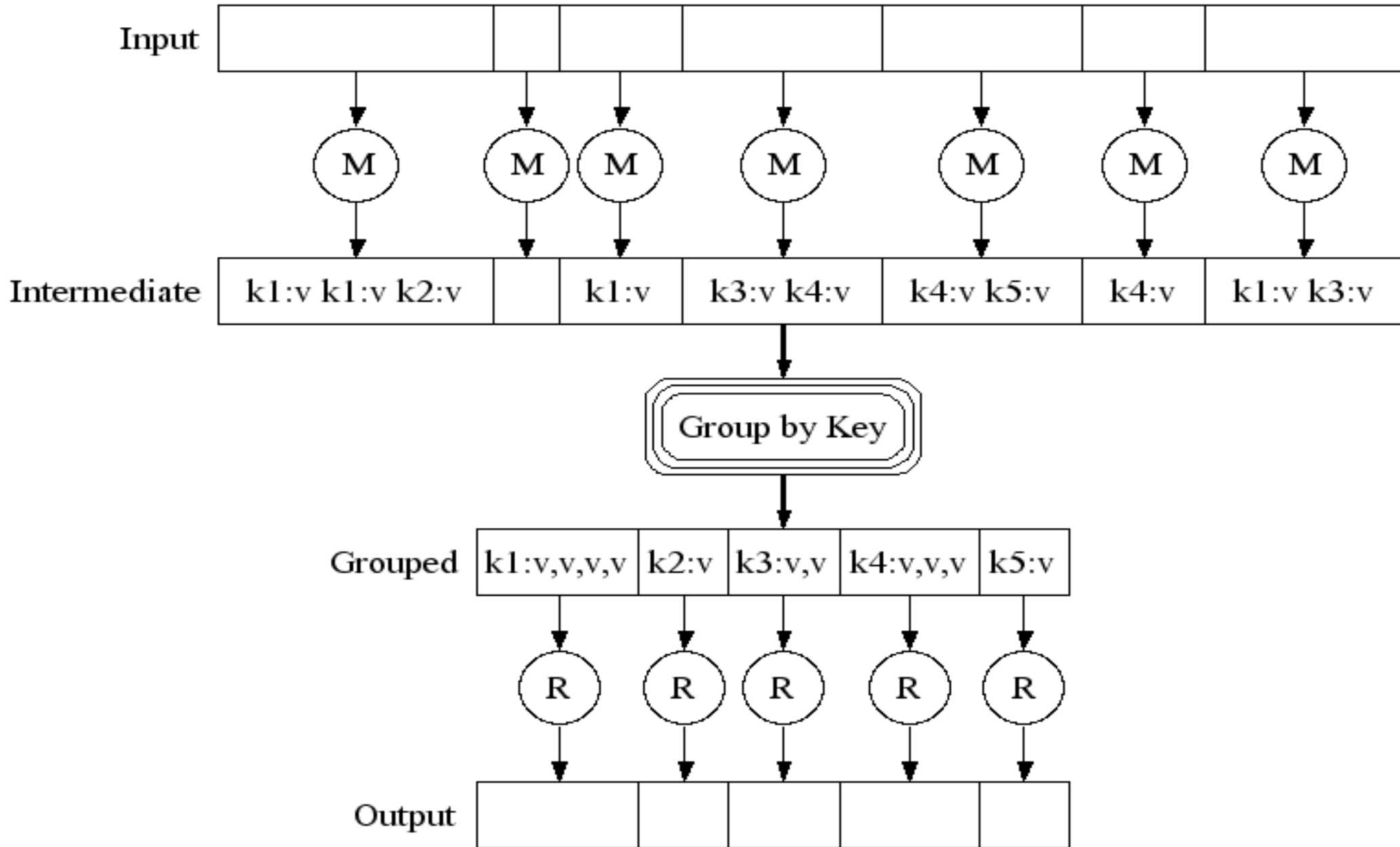
## 什麼是 MapReduce ??

- MapReduce 是 Google 申請的軟體專利，主要用來處理大量資料
- MapReduce is a **patented** software framework introduced by **Google** to support distributed computing on large data sets on clusters of computers.
- 啟發自函數編程中常用的 map 與 reduce 函數。
- The framework is inspired by **map** and **reduce** functions commonly used in **functional programming**, although their purpose in the MapReduce framework is not the same as their original forms
  - Map(...):  $N \rightarrow N$ 
    - Ex. [ 1,2,3,4 ] – (**\*2**) -> [ 2,4,6,8 ]
  - Reduce(...):  $N \rightarrow 1$ 
    - [ 1,2,3,4 ] - (**sum**) -> 10
- **Logical view of MapReduce**
  - **Map(k1,v1) -> list(k2,v2)**
  - **Reduce(k2, list (v2)) -> list(v3)**

Source: <http://en.wikipedia.org/wiki/MapReduce>

# Google's MapReduce Diagram

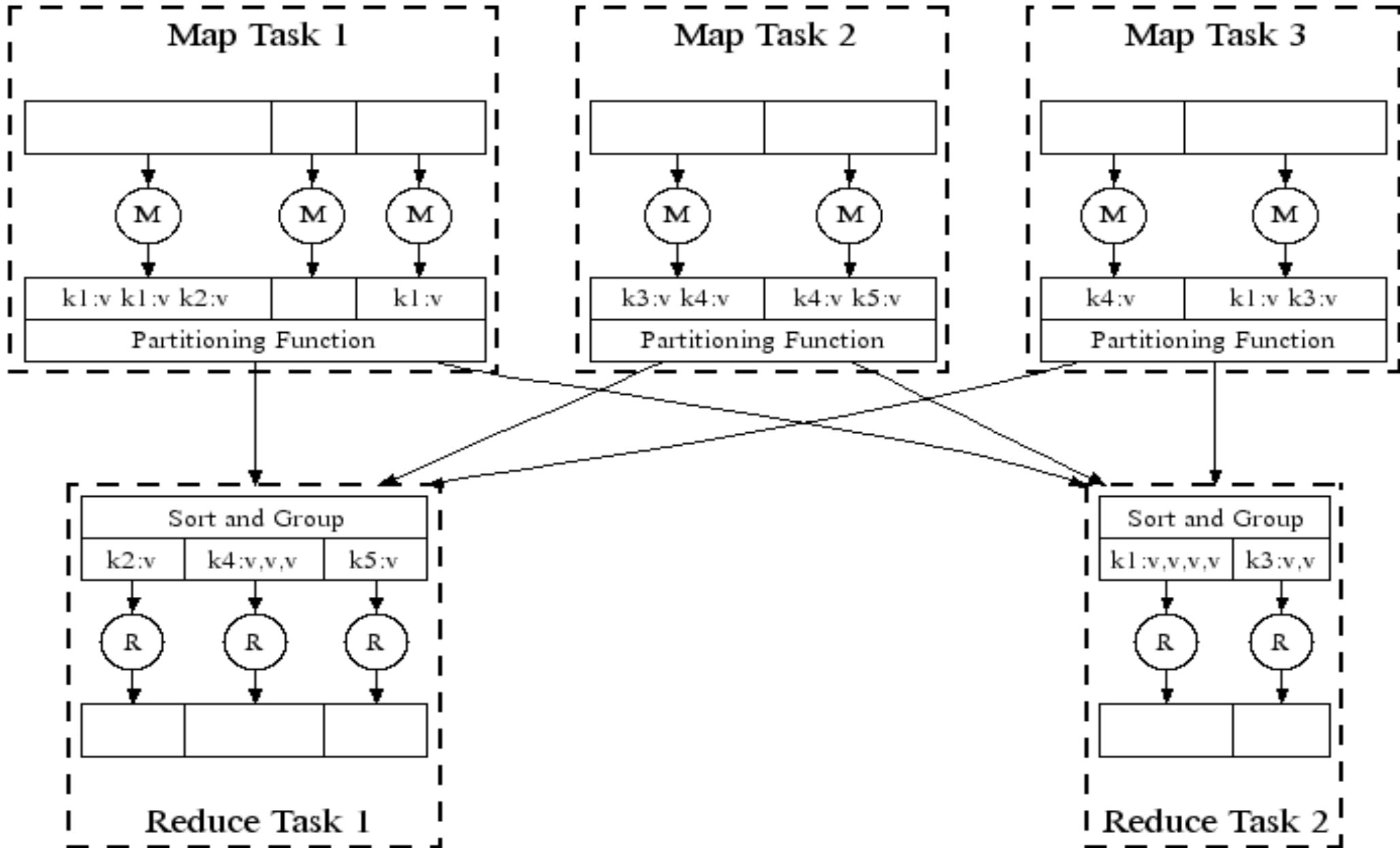
## Google 的 MapReduce 圖解





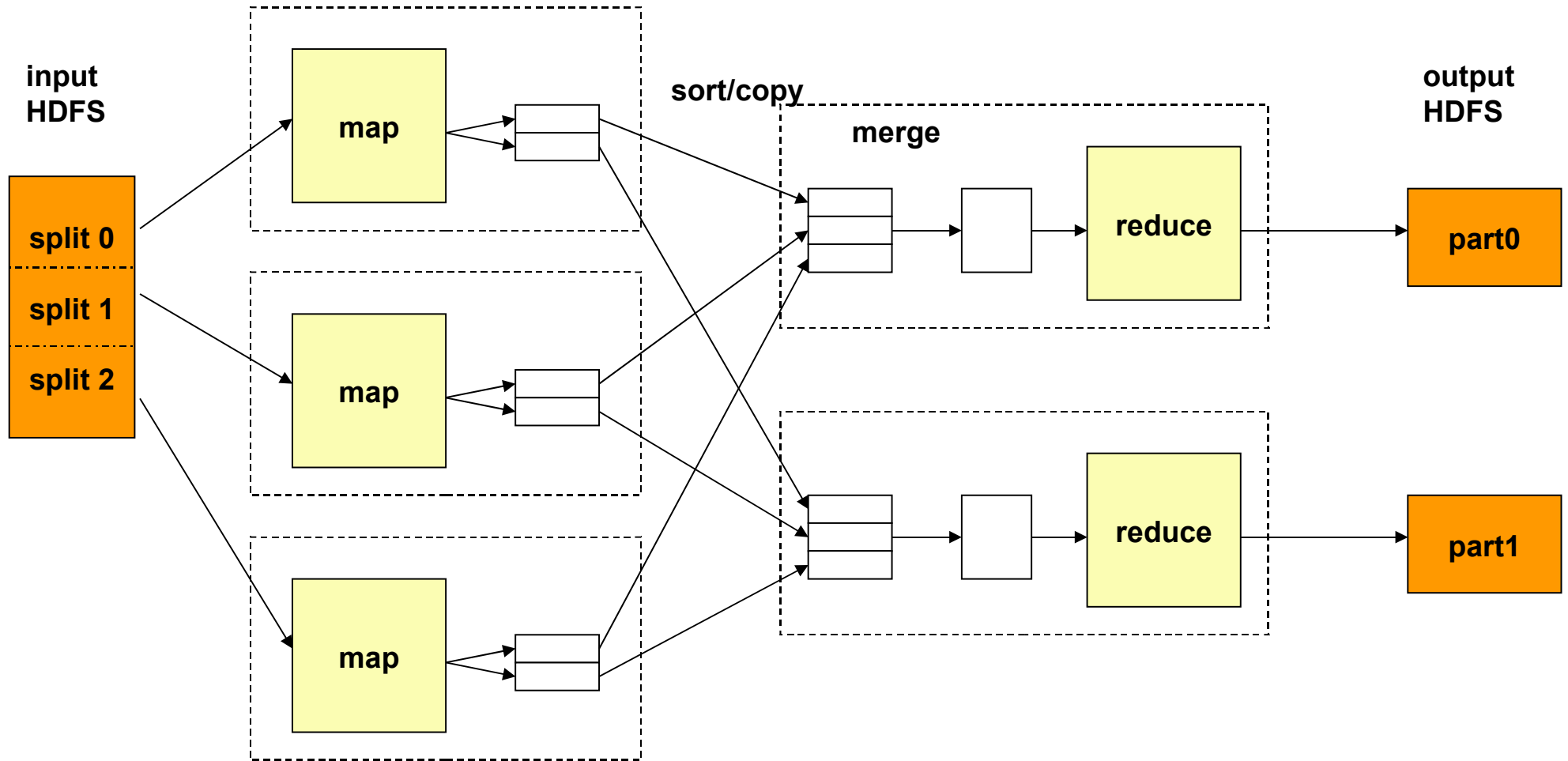
# Google's MapReduce in Parallel

## Google 的 MapReduce 平行版圖解



# How does MapReduce work in Hadoop

## Hadoop MapReduce 運作流程



JobTracker 跟 NameNode 取得需要運算的 blocks

JobTracker 選數個 TaskTracker 來作 Map 運算，產生些中間檔案

JobTracker 將中間檔案整合排序後，複製到需要的 TaskTracker 去

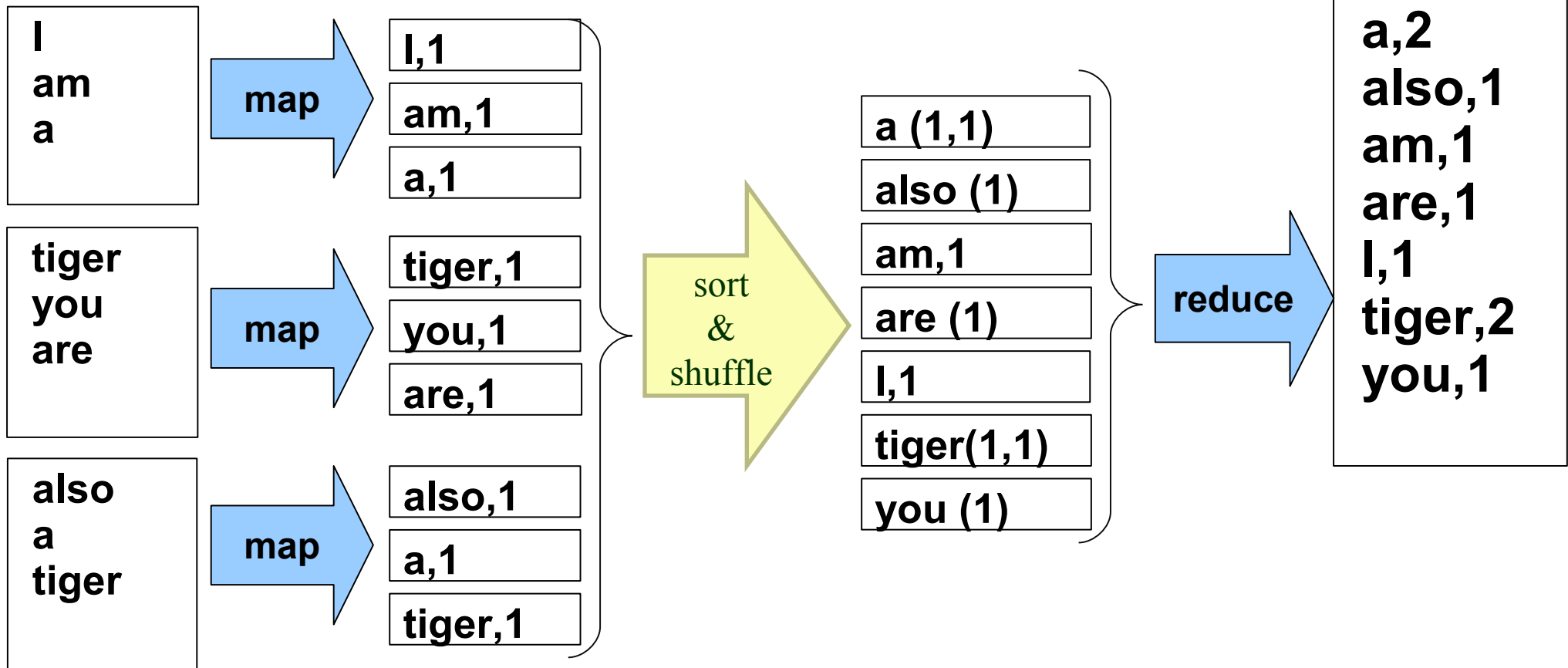
JobTracker 派遣 TaskTracker 作 reduce

reduce 完後通知 JobTracker 與 Namenode 以產生 output

# MapReduce by Example (1)

## MapReduce 運作實例 (1)

I am a tiger, you are also a tiger



JobTracker 先選了三個 Tracker 做 map

Map 結束後，hadoop 進行中間資料的重組與排序

JobTracker 再選一個 TaskTracker 作 reduce

# MapReduce by Example (2)

## MapReduce 運作實例 (2)

$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \rightarrow \begin{bmatrix} \text{sqrt}(a + b) \\ \text{sqrt}(c + d) \end{bmatrix}$

$\begin{bmatrix} 1.0 & 0.0 & 3.0 \\ 3.2 & 0.8 & 32.0 \\ 1.0 & 14.0 & 1.0 \end{bmatrix} \rightarrow ?$

Input File

```
0 0 1.0 // A[0][1] = 1.0
0 1 0.0 // A[0][1] = 0.0
0 2 3.0 // A[0][2] = 3.0
1 0 3.2 // A[1][0] = 3.2
1 1 0.8 // A[1][1] = 0.8
```

map

```
(0, 1.0)
(0, 0.0)
(0, 3.0)
(1, 3.2)
(1, 0.8)
```

```
1 2 32.0 // A[1][2] = 32.0
2 0 1.0 // A[2][0] = 1.0
2 1 14.0 // A[2][1] = 14.0
2 2 1.0 // A[2][2] = 1.0
```

map

```
(1, 32.0)
(2, 1.0)
(2, 14.0)
(2, 1.0)
```

sort /  
merge

```
(0, {1.0, 0.0, 3.0})
(1, {3.2, 0.8, 32.0})
(2, {1.0, 14.0, 1.0})
```

reduce

```
(0, sqrt(1.0 + 0.0 + 3.0))
(1, sqrt(3.2 + 0.8 + 32.0))
(2, sqrt(1.0 + 14.0 + 1.0))
```

# *MapReduce is suitable to ....*

## **MapReduce** 合適用於 ....

- 大規模資料集
- **Large Data Set**
- 可拆解
- **Parallelization**
- Text tokenization
- Indexing and Search
- Data mining
- machine learning
- ...

- <http://www.dbms2.com/2008/08/26/known-applications-of-mapreduce/>
- <http://wiki.apache.org/hadoop/PoweredBy>



# Hadoop 進階課程

# HBase 資料庫應用

< V 0.20 >

王耀聰 陳威宇

Jazz@nchc.org.tw

waue@nchc.org.tw



財團法人國家實驗研究院

國家高速網路與計算中心

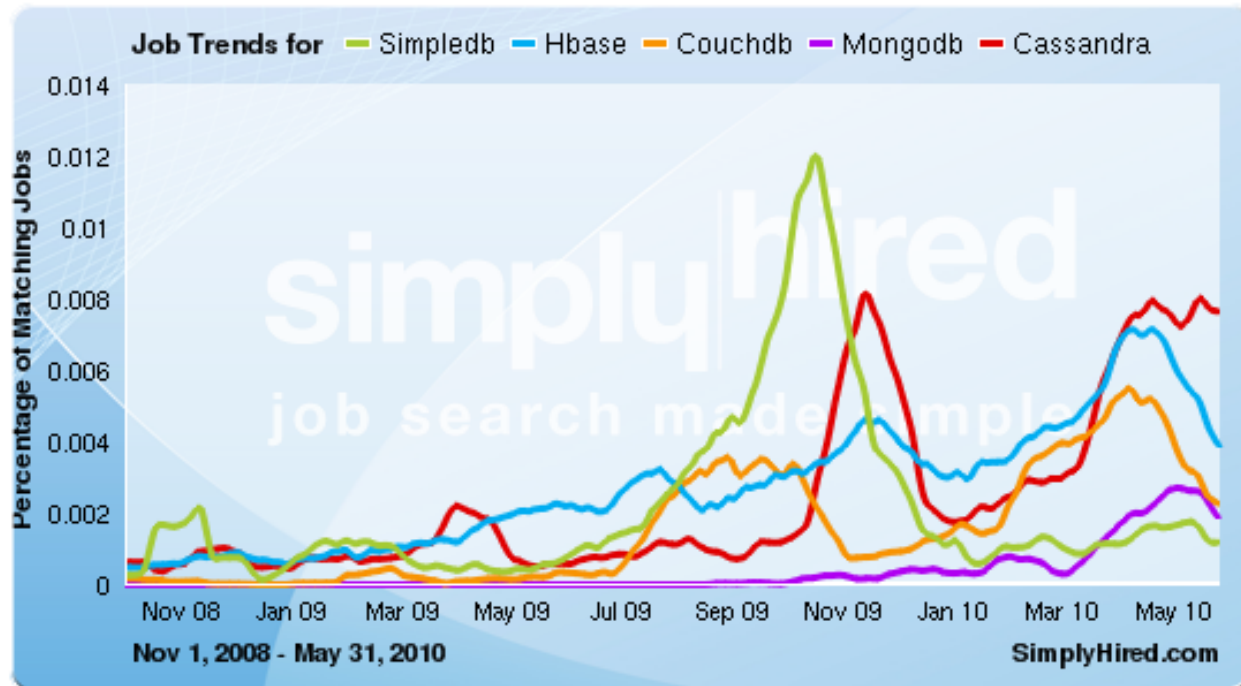
NATIONAL CENTER FOR HIGH-PERFORMANCE COMPUTING



# 2010 年 NoSQL 職缺排行榜

Simpledb, Hbase, Couchdb, Mongoddb, Cassandra Trends

1. Cassandra
2. HBase
3. CouchDB
4. MongoDB
5. SimpleDB



## Simpledb, Hbase, Couchdb, Mongoddb, Cassandra 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:

- [Simpledb jobs](#) increased 357%
- [Hbase jobs](#) increased 745%
- [Couchdb jobs](#) did not change or there is no data available
- [Mongoddb jobs](#) increased 18,480%
- [Cassandra jobs](#) did not change or there is no data available

(2010-07-25)

# 一、HBase 介紹

介紹HBase如何而來，它的 Why, What, How ....，以及它的架構

HBase, *Hadoop database*, is an open-source, distributed, versioned, column-oriented store modeled after Google' Bigtable. Use it when you need random, realtime read/write access to your Big Data.



# BigTable ?

- Bigtable: 一個結構化數據的分佈式存儲系統
- Google Style的數據庫，使用結構化的文件來存儲數據
- 不支持關聯或是類似於SQL的高級查詢。
- 大規模處理、高容錯性
- PB級的存儲能力
- 每秒數百萬的讀寫操作

# HBase

- 設計概念與結構類似Bigtable
- HBase 以 Hadoop 分散式檔案系統 (HDFS) 為基礎，提供類Bigtable 功能
- HBase 是具有以下特點的儲存系統：
  - ◆ 類似表格的資料結構 (Multi-Dimensional Map)
  - ◆ 分散式
  - ◆ 高可用性、高效能
  - ◆ 很容易擴充容量及效能
- HBase 適用於利用數以千計的一般伺服器上，來儲存Petabytes級的資料。
- HBase同時提供Hadoop MapReduce程式設計。

# 開發歷程

- 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
- 2010.3
  - ◆ HBase graduates from Hadoop sub-project to Apache Top Level Project
- 2010.7
  - ◆ HBase 0.20.6 released

# 誰使用 HBase

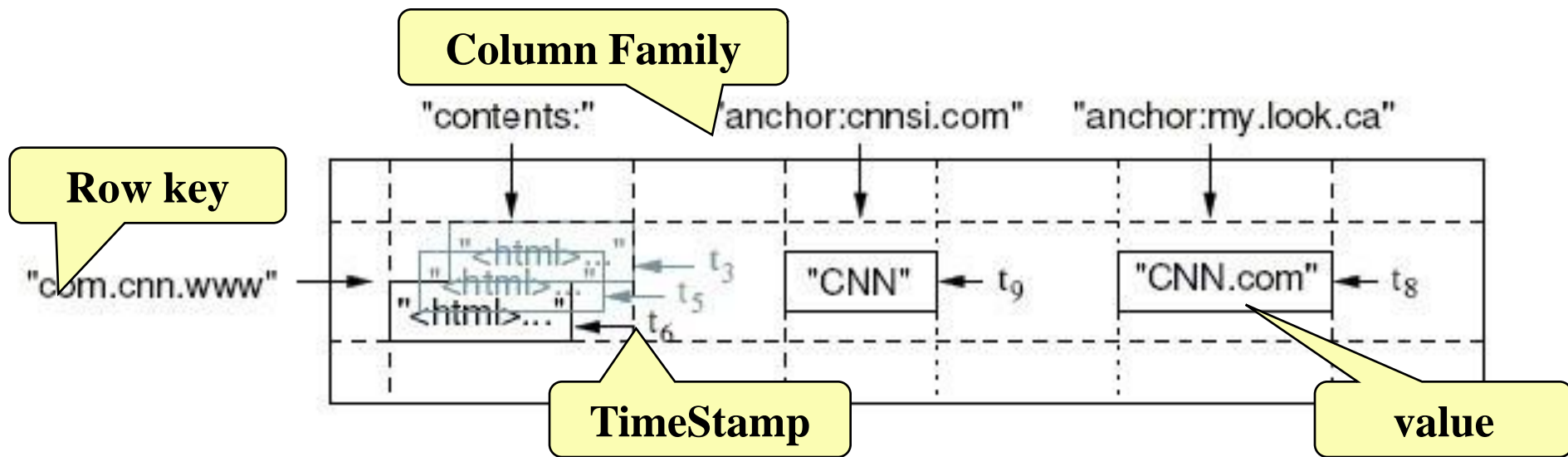
- Adobe
  - ◆ 內部使用 (Structure data)
- Kalooga
  - ◆ 圖片搜尋引擎 <http://www.kalooga.com/>
- Meetup
  - ◆ 社群聚會網站 <http://www.meetup.com/>
- Streamy
  - ◆ 成功從 MySQL 移轉到 Hbase <http://www.streamy.com/>
- Trend Micro
  - ◆ 雲端掃毒架構 <http://trendmicro.com/>
- Yahoo!
  - ◆ 儲存文件 fingerprint 避免重複 <http://www.yahoo.com/>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

# 為什麼使用HBase?

- 不是關聯式(Relational)資料庫系統
  - ◆ 表格(Table)只有一個主要索引 (primary index) 即 row key.
  - ◆ 不提供 join
  - ◆ 不提供 SQL 語法。
- 提供Java函式庫, 與 REST與Thrift等介面。
- 提供 getRow(), Scan() 存取資料。
  - ◆ getRow()可以取得一筆row range的資料, 同時也可以指定版本 (timestamp)。
- Scan()可以取得整個表格的資料或是一組row range (設定start key, end key)
- 有限的單元性(Atomicity)與交易 (transaction)功能.
- 只有一種資料型態 (bytes)
- 可以配合MapReduce框架, 進行複雜的分析與查詢

# Data Model

- Table依 *row key* 來自動排序
- Table schema 只要定義 *column families*.
  - ◆ 每個column family 可有無限數量的 columns
  - ◆ 每個column的值可有無限數量的時間版本(timestamp)
  - ◆ Column可以動態新增，每個row可有不同數量的columns。
  - ◆ 同一個column family的columns會群聚在一個實體儲存單元上，且依column 的名稱排序。
  - ◆ byte[] 是唯一的資料型態(Row, Family: Column, Timestamp) Value



# Data Model

- HBase實際上儲存Table時，是以column family為單位來存放

Row Key	Time Stamp	Column (Family) “content:”
com.cnn.www	t9	“<html>...”
	t6	“<html>...”

Row Key	Time Stamp	Column (Family) “anchor:”
com.cnn.www	t9	“anchor:cnnsi. com” “CNN”
	t8	“anchor:cnnsi. com” “CNN”
		“anchor:my.loc com” “MyLook”

# HTable 成員

Table, Family, Column, Qualifier, Row, TimeStamp

		Contents	Department		
			news	bid	sport
t1	com.yahoo.news.t w	“撿到學雜費，硬要分三成”	“tech”		
t2		“科研論文評比 5校進500 大”	“tech”		
t3		“罰蹲立300下！班長「住 院」 師懊悔”	“tech”		
t1	com.yahoo.bid.tw	“… iphone 4G 9/17 日上 市”		“3C”	
t1	com.yahoo.sport.t w	“Nadal 大滿貫”			“MBA”

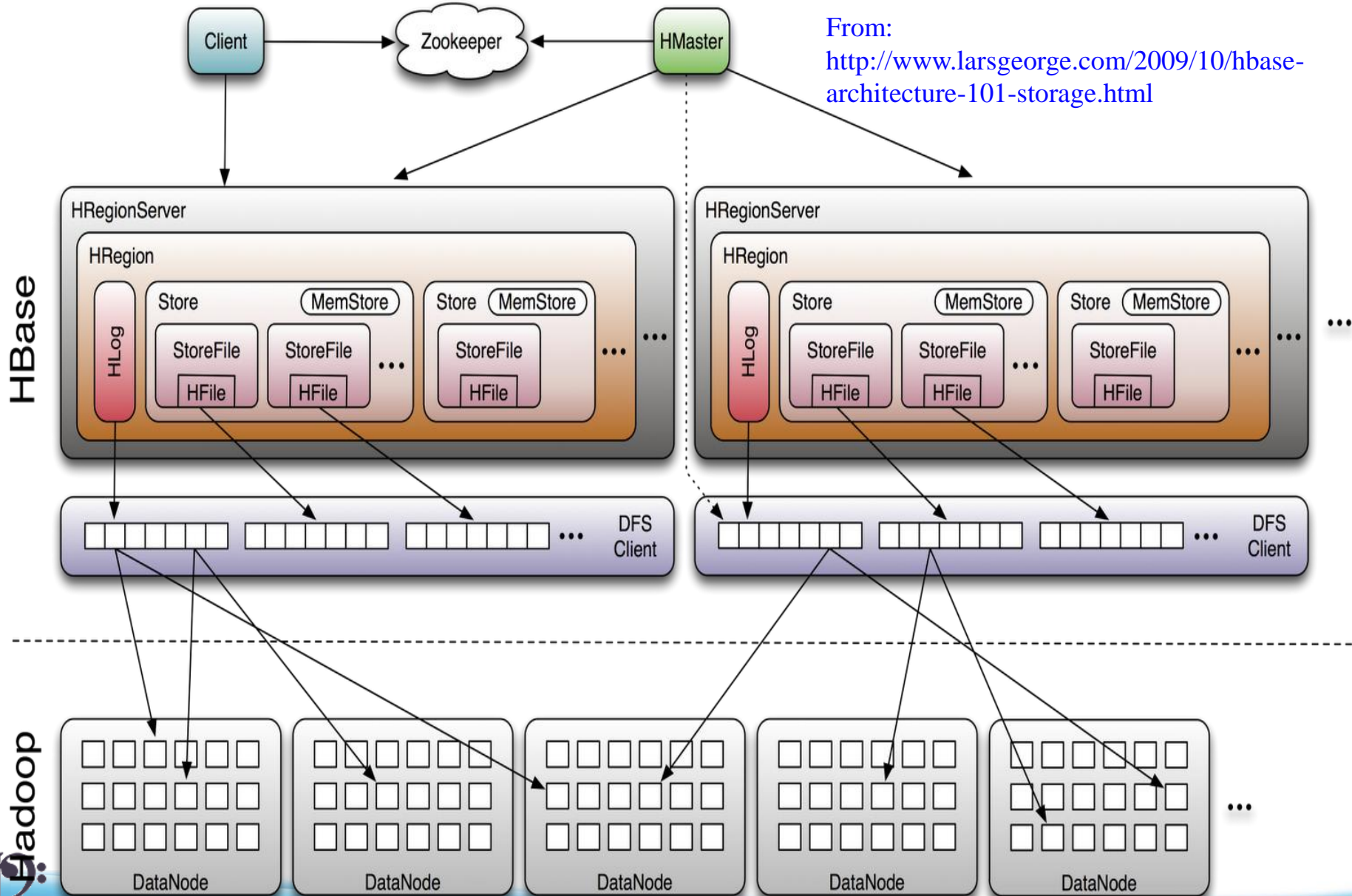


# Regions

- 表格是由一或多個 region 所構成
  - ◆ Region 是由其 startKey 與 endKey 所指定
- 每個 region 可能會存在於多個不同節點上，而且是由數個HDFS 檔案與區塊所構成，這類 region 是由 Hadoop 負責複製

Region	Row Keys	Column Family “Content”
Region 1	00000	...
	00001	...
	...	...
	09999	...
Region 2	10000	...
	...	...
	...	...
	29999	...

# HBase 與 Hadoop 搭配的架構

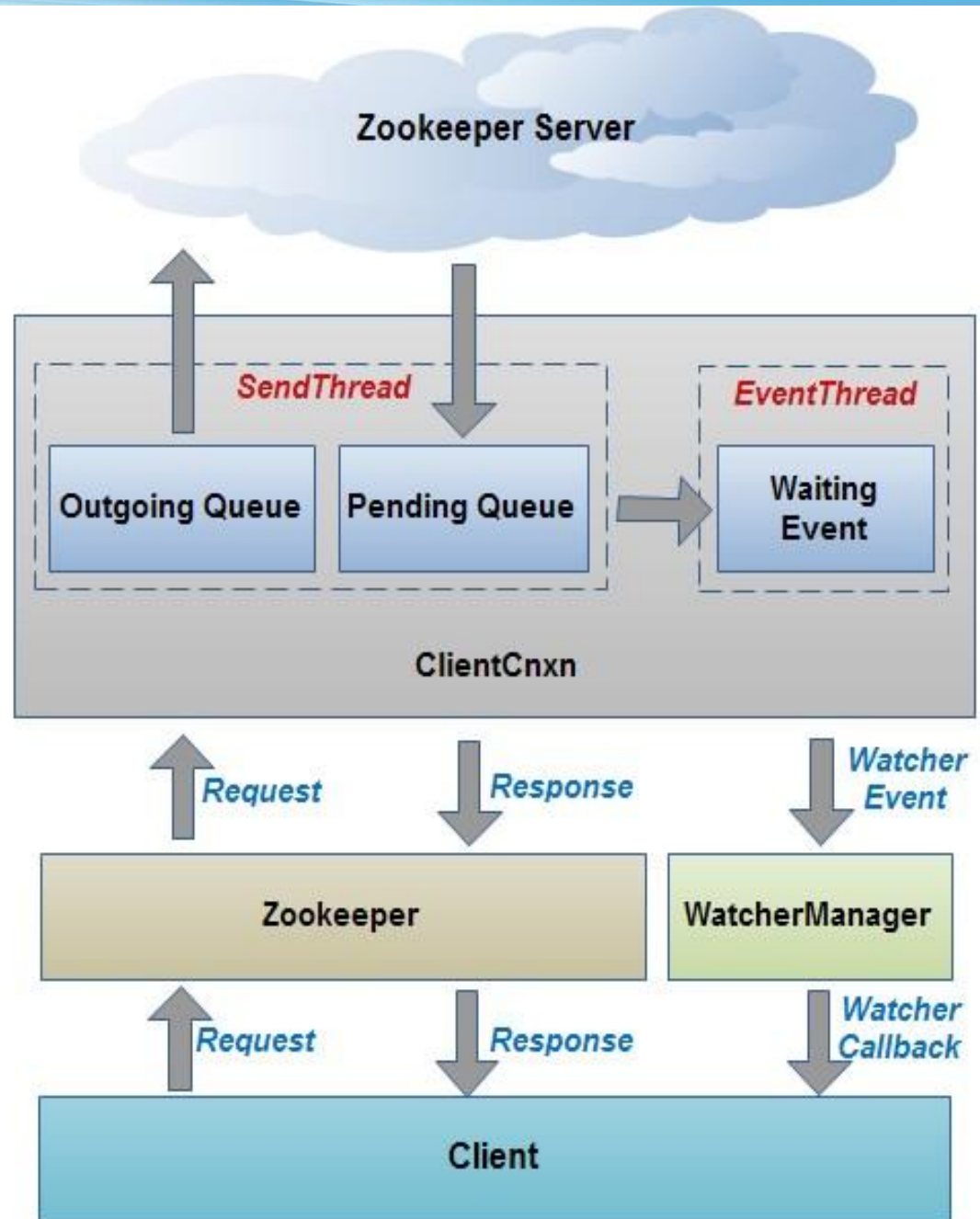


# HBase 0.20 特色

- 解決單點失效問題（single point of failure）
  - ◆ Ex: Hadoop NameNode failure
- 設定檔改變或小版本更新會重新啟動
- 隨機讀寫（Random access）效能如同 MySQL

# Zookeeper ?

- Hadoop的正式子項目
- 針對大型分散式系統的可靠協調系統
- Google的Chubby
- 存儲一些配置信息，確保文件寫入的一致性
- Master / Client 架構，Master 可由選舉而得





# 雲端入侵偵測日誌分析平台

Part-7 : Building IDS-log Cloud Analysis System (ICAS)

**Yao-Tsung Wang**

[jazz@nchc.org.tw](mailto:jazz@nchc.org.tw)

**Wei-Yu Chen**

[waue@nchc.org.tw](mailto:waue@nchc.org.tw)



# 專家說：雲端每個環節都有安全問題

ZDNet Taiwan - 專家談雲端：每個環節都有安全問題 - 新聞

2010/08/10 19:50:02

專家談雲端：每個環節都有安全問題

ZDNet記者曠文濤／台北報導 雲端的安全問題不是無解，只是不管是雲端服務供應商或者想要建立私有雲的企業用戶，都必須考量到每個環節。

微軟亞太區全球技術支援中心專案經理、同時也是ZDNet專欄作家林宏嘉今（10）日在ZDNet舉行的IT Priorities圓桌論壇中表示，**雲端的安全議題涉及了IaaS、PaaS乃至於SaaS的每個層面**，當然有些問題是原本就存在：例如在討論到IaaS時，就涉及到了**機房的管理**和**硬體設備的可用性**等；但是講到PaaS時，企業用戶倘若要選擇開原碼的作業系統，必須考量到後續的**安全維護**；在SaaS的層次，企業用戶必須確保每一個分區（partition）的安全更新和**資料安全**。

目前正如火如荼建立台灣第一個校園私有雲的台大計算機及資訊網路中心主任孫雅麗則呼應道，Amazon的雲端服務證實了在Hypervisor層有駭客入侵，也就是意味著過去大家在討論如何防範**虛擬機器的資料安全**，但是威脅已經深化到了更下一層。這些問題都有待解決。

「有些問題甚至是來自於內部，舉例而言，MIS可能會把存在記憶體裡的資料倒出來，或者在Hypervisor層就植入了可以蒐集資料的程式，」孫雅麗說。

安全議題是目前台灣企業對雲端持保留態度的最大主因，這也是何以台灣的大型企業對於雲端的想法，還是  
傾好建立私有雲。畢竟對用戶而言，資料放在別人家，還和其他企業，甚至具競爭對手「共處」，

# 雲端資安的範疇

用雲端  
處理資安

**Dealing Security  
issues using Cloud**

**Data Security  
In the Cloud**

雲內部  
的資安管制

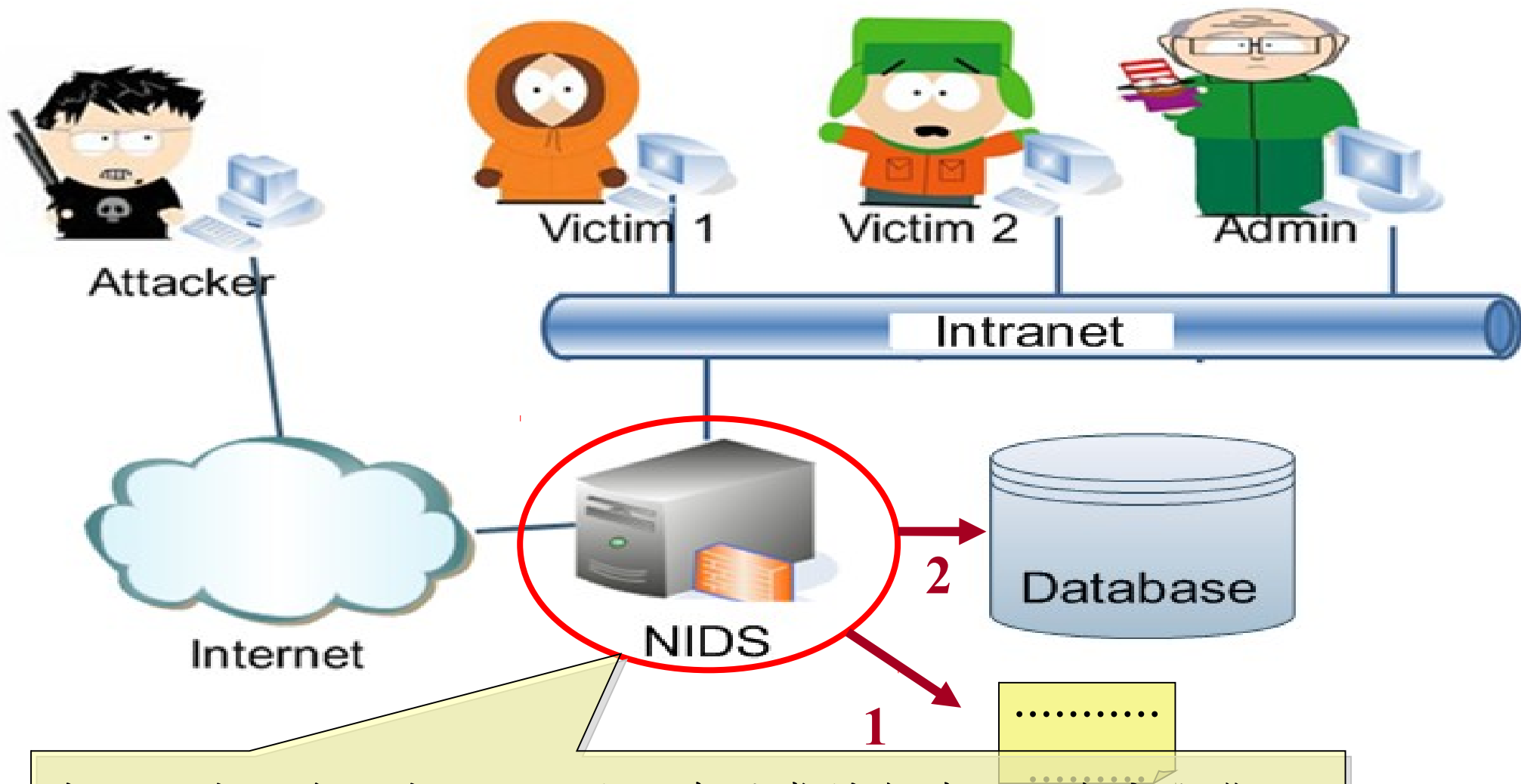
**Security Issues  
Inside the Cloud**

雲端資料  
安全性

端本身  
的資安威脅

**Security Threats  
to Internet of Things**

# 使用入侵偵測系統 (NIDS) 來找出入侵訊息



當入侵偵測系統偵測到網路上有異常封包時，就會產生警訊以告知有攻擊發生。警訊通常有兩種形式：  
1. 紀錄成 log 檔 2. 紀錄到資料庫



# 傳統 NIDS 的警訊型態 (1) 紀錄在日誌檔內

## 入侵偵測系統所產生警訊日誌檔內一小段內容

```
[**] [1:538:15] NETBIOS SMB IPC$ unicode share access [**]  
[Classification: Generic Protocol Command Decode] [Priority: 3]  
09/04-17:53:56.363811 168.150.177.165:1051 -> 168.150.177.166:139  
TCP TTL:128 TOS:0x0 ID:4000 IpLen:20 DgmLen:138 DF  
***AP*** Seq: 0x2E589B8 Ack: 0x642D47F9 Win: 0x4241 TcpLen: 20
```

```
[**] [1:1917:6] SCAN UPnP service discover attempt [**]  
[Classification: Detection of a Network Scan] [Priority: 3]  
09/04-17:53:56.385573 168.150.177.164:1032 -> 239.255.255.250:1900  
UDP TTL:1 TOS:0x0 ID:80 IpLen:20 DgmLen:161  
Len: 133
```

```
[**] [1:1917:6] SCAN UPnP service discover attempt [**]  
[Classification: Detection of a Network Scan] [Priority: 3]  
09/04-17:53:56.386910 168.150.177.164:1032 -> 239.255.255.250:1900  
UDP TTL:1 TOS:0x0 ID:82 IpLen:20 DgmLen:161  
Len: 133
```

```
[**] [1:1917:6] SCAN UPnP service discover attempt [**]  
[Classification: Detection of a Network Scan] [Priority: 3]  
09/04-17:53:56.388244 168.150.177.164:1032 -> 239.255.255.250:1900  
UDP TTL:1 TOS:0x0 ID:84 IpLen:20 DgmLen:161  
Len: 133
```

```
[**] [1:538:15] NETBIOS SMB IPC$ unicode share access [**]  
[Classification: Generic Protocol Command Decode] [Priority: 3]  
09/04-17:53:56.405923 168.150.177.164:1035 -> 168.150.177.166:139  
TCP TTL:128 TOS:0x0 ID:94 IpLen:20 DgmLen:138 DF  
***AP*** Seq: 0x82073DFF Ack: 0x2468EB82 Win: 0x4241 TcpLen: 20
```

```
[**] [1:1917:6] SCAN UPnP service discover attempt [**]  
[Classification: Detection of a Network Scan] [Priority: 3]  
09/04-17:53:56.417045 168.150.177.164:45461 -> 168.150.177.1:1900  
UDP TTL:1 TOS:0x0 ID:105 IpLen:20 DgmLen:161  
Len: 133
```

```
[**] [1:1917:6] SCAN UPnP service discover attempt [**]  
[Classification: Detection of a Network Scan] [Priority: 3]  
09/04-17:53:56.420759 168.150.177.164:45461 -> 168.150.177.1:1900  
UDP TTL:1 TOS:0x0 ID:117 IpLen:20 DgmLen:160  
Len: 132
```

```
[**] [1:1917:6] SCAN UPnP service discover attempt [**]  
[Classification: Detection of a Network Scan] [Priority: 3]  
09/04-17:53:56.422095 168.150.177.164:45461 -> 168.150.177.1:1900  
UDP TTL:1 TOS:0x0 ID:118 IpLen:20 DgmLen:161  
Len: 133
```

```
[**] [1:2351:10] NETBIOS DCERPC ISystemActivator path overflow attempt little endian  
unicode [**]  
[Classification: Attempted Administrator Privilege Gain] [Priority: 1]  
09/04-17:53:56.442445 198.8.16.1:10179 -> 168.150.177.164:135  
TCP TTL:105 TOS:0x0 ID:49809 IpLen:20 DgmLen:1420 DF  
***A**** Seq: 0xF9589BBF Ack: 0x82CCF5B7 Win: 0xFFFF TcpLen: 20  
[Xref => http://www.microsoft.com/technet/security/bulletin/MS03-026.msp][Xref =>  
http://cgi.nessus.org/plugins/dump.php3?id=11808][Xref => http://cve.mitre.org/cgi-bin/cvename.cgi?name=2003-0352][Xref => http://www.securityfocus.com/bid/8205]
```

```
[**] [122:3:0] (portscan) TCP Portsweep [**]  
[Priority: 3]  
09/04-17:53:56.499016 198.8.16.1 -> 168.150.177.166  
PROTO:255 TTL:0 TOS:0x0 ID:1750 IpLen:20 DgmLen:168
```

# 傳統 NIDS 的警訊型態 (2) 紀錄在資料庫內

以下為利用瀏覽器透過網頁方式呈現警訊資料庫的內容

The screenshot shows a Mozilla browser window displaying the 'Basic Analysis and Security Engine (BASE)' interface. The page title is 'Basic Analysis and Security Engine (BASE)'. The navigation menu includes 'Home', 'Search', and 'AG Maintenance'. A status message indicates 'Added 0 alert(s) to the Alert cache' and the query time is 'Thu October 14, 2004 22:04:44'. A table on the left shows search criteria: Meta Criteria, IP Criteria, TCP Criteria, and Payload Criteria, all set to 'any'. A 'Summary Statistics' box on the right lists: Sensors, Unique Alerts (classifications), Unique addresses: source | destination, Unique IP links, Source Port: TCP | UDP, Destination Port: TCP | UDP, and Time profile of alerts. Below this, it says 'Displaying alerts 1-50 of 81 total'. The main content is a table of alerts with columns for ID, Signature, Timestamp, Source Address, Dest. Address, and Layer 4 Proto.

<input type="checkbox"/>	ID	< Signature >	< Timestamp >	< Source Address >	< Dest. Address >	< Layer 4 Proto >
<input type="checkbox"/>	#0-(1-84)	[snort] NETBIOS SMB IPC\$ share unicode access	2004-10-08 11:25:41	192.168.1.100:1613	192.168.1.4:139	TCP
<input type="checkbox"/>	#1-(1-83)	[snort] NETBIOS SMB IPC\$ share unicode access	2004-10-08 11:25:31	192.168.1.100:1608	192.168.1.4:139	TCP
<input type="checkbox"/>	#2-(1-82)	[snort] NETBIOS SMB IPC\$ share unicode access	2004-10-08 11:25:05	192.168.1.100:1601	192.168.1.4:139	TCP
<input type="checkbox"/>	#3-(1-80)	[snort] (http_inspect) OVERSIZE CHUNK ENCODING	2004-10-04 22:25:41	192.168.1.4:42164	67.19.245.228:80	TCP
<input type="checkbox"/>	#4-(1-81)	[snort] (http_inspect) OVERSIZE CHUNK ENCODING	2004-10-04 22:25:41	192.168.1.4:42163	67.19.245.228:80	TCP

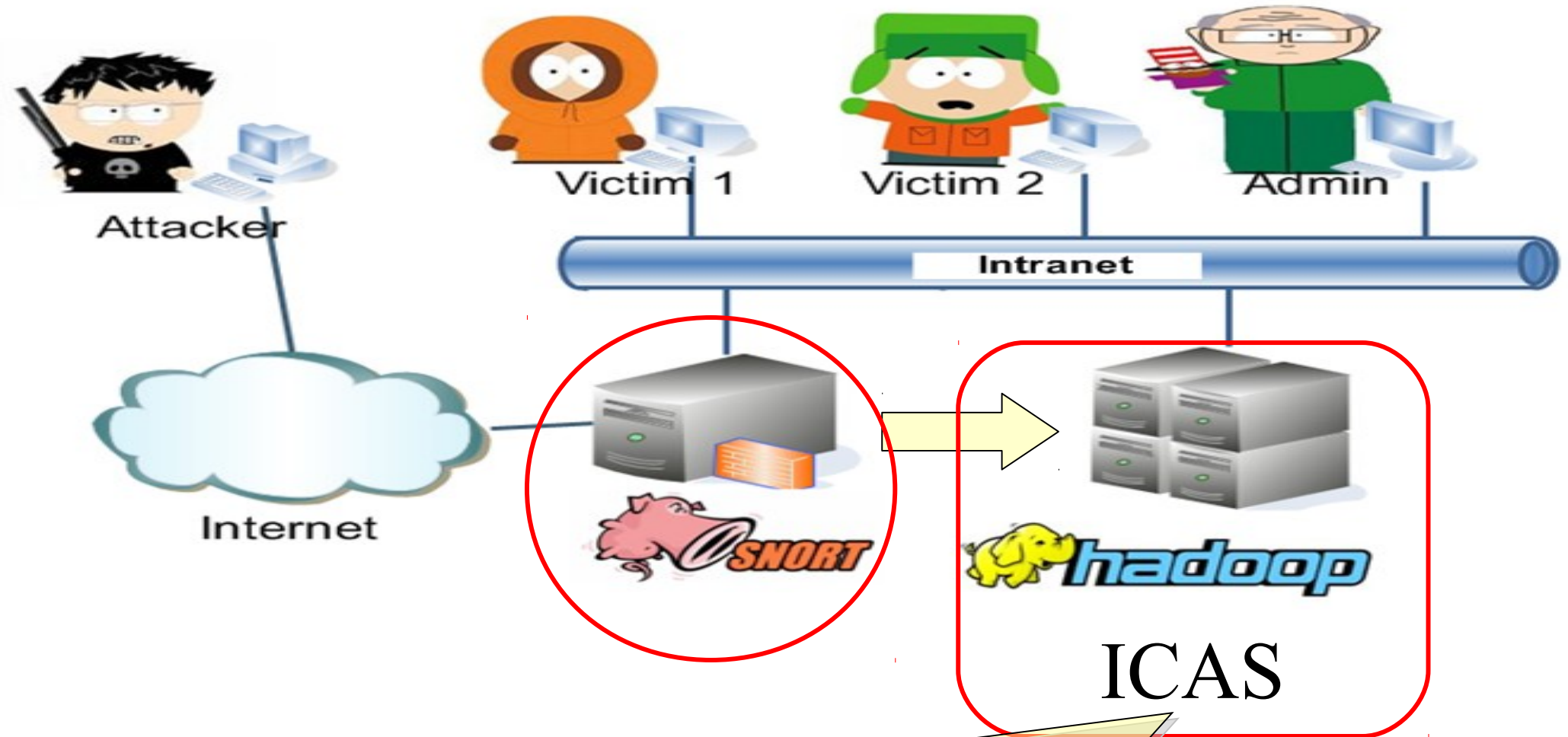
## 以上作法的缺點

- 警訊僅被『忠實』地被記錄下來，無法顯示彼此間的關聯性，因此系統管理者難以瞭解全部攻擊情形
- 過多的警訊，使得容易忽略重要內容
- 完全依賴單一資料庫，當資料量一大，該台主機的讀寫效率將成爲瓶頸

# 使用雲端運算的解決方案：ICAS

- ICAS, *IDS Cloud Analysis System*
- 利用雲端運算的特性提供以下好處
  - 對大量資料有高效率
  - 一般主機的叢集
  - 有錯誤容忍
- 分析演算法
  - 整合
  - 關聯

# 透過 ICAS 協助分析 IDS 的警訊

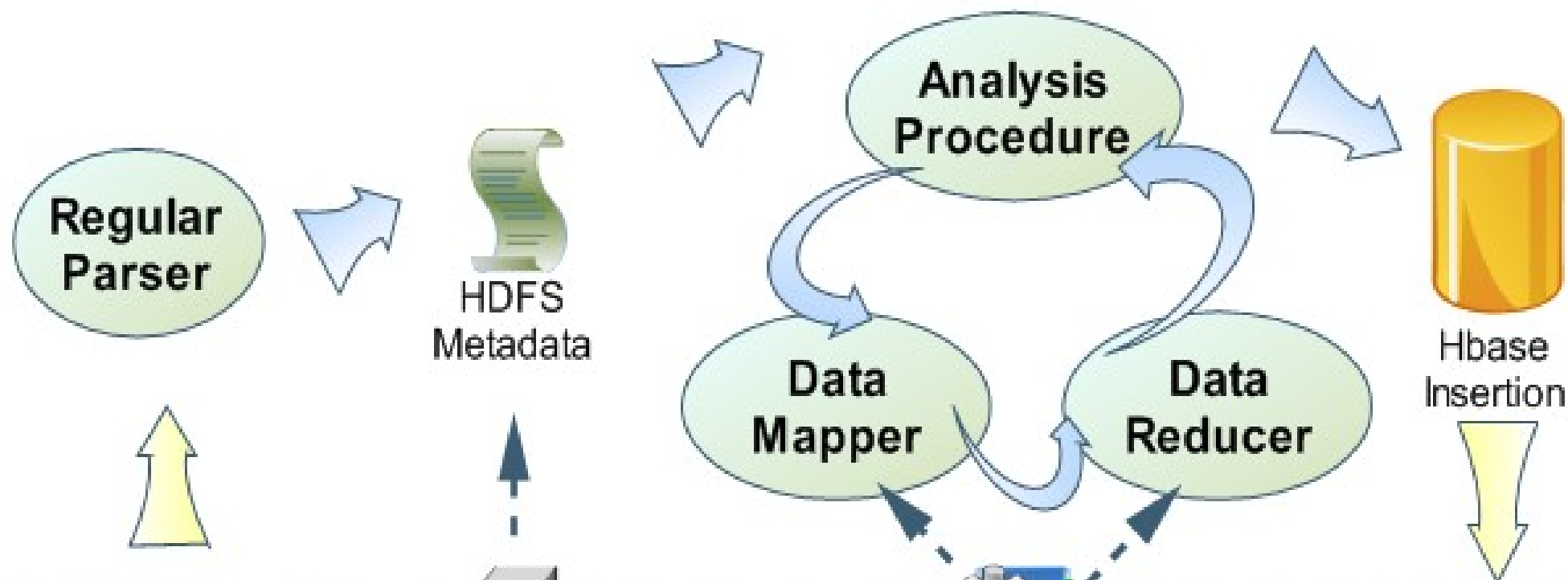


可多個 NIDS 共同產生警訊後，傳送至 ICAS，分析演算法  
目前有 ICAS-I 及 ICAS-II

# ICAS-I

- 將任意個原始警訊檔上傳到運行 ICAS-I 演算法的 Hadoop 檔案系統空間 ( HDFS )
- 利用 Hadoop 的 MapReduce 平台架構所設計的演算法來分析資料
- 分析完後的資料塞入分散式資料庫 HBase 內

# ICAS-I 流程圖



**Intrusion  
Detectoin  
System**



**HDFS**



**JobTracker**



**hadoop**

**Cloud Platform**



**HBASE**

**Database**

# ICAS-I 整合後的警訊結果

Destination IP	Attack Signature	Source IP	Destination Port	Source Port	Packet Protocol	Timestamp
Host_1	Trojan	Sip1	80	4077	tcp	T1
Host_1	Trojan	Sip2	80	4077	tcp	T2
Host_1	Trojan	Sip1	443	5002	tcp	T3
Host_2	Trojan	Sip1	443	5002	tcp	T4
Host_3	D.D.O.S	Sip3	53	6007	udp	T5
Host_3	D.D.O.S	Sip4	53	6008	tcp	T5
Host_3	D.D.O.S	Sip5	53	6007	udp	T5
Destination IP	Attack Signature	Source IP	Destination Port	Source Port	Packet Protocol	Timestamp



Key		Values				
Host_1	Trojan	Sip1,Sip2	80,443	4077,5002	tcp	T1,T2,T3
Host_2	Trojan	Sip1	443	5002	tcp	T4
Key		Values				

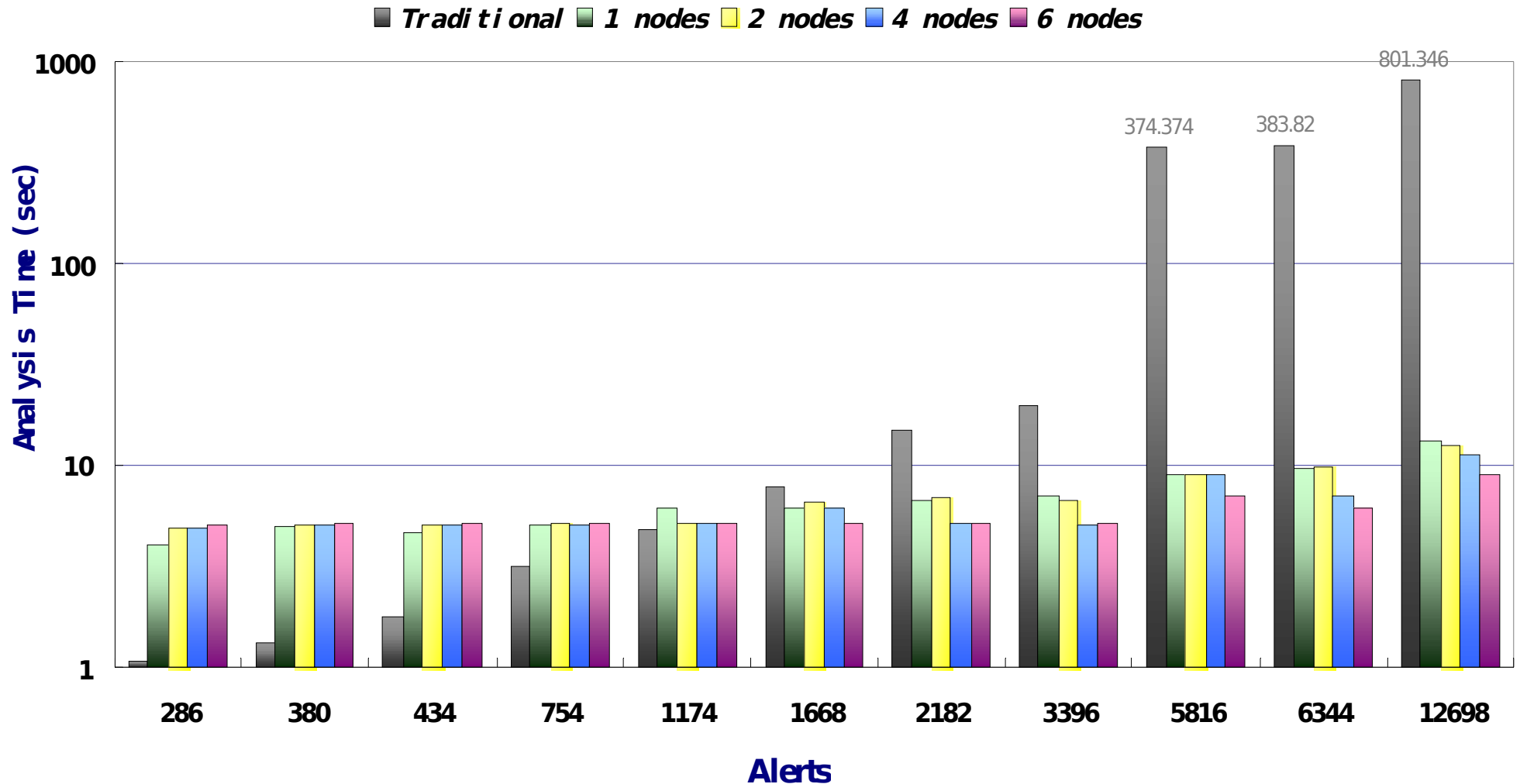


# ICAS-I 效能數據的環境

- Machine:
  - CPU : Intel quad-core, Memory : 2 GB,
- OS : Linux : Ubuntu 8.04 server
- Software : version
  - Hadoop : 0.16.4
  - Hbase : 0.1.3
  - Java : 6
- Alerts Data Sets
  - MIT Lincoln Laboratory, Lincoln Lab Data Sets
  - Computer Security group at UC Davis, tcpdump file

# ICAS-I 效能分析時間圖

## The Consuming Time of Each Number of Data Sets



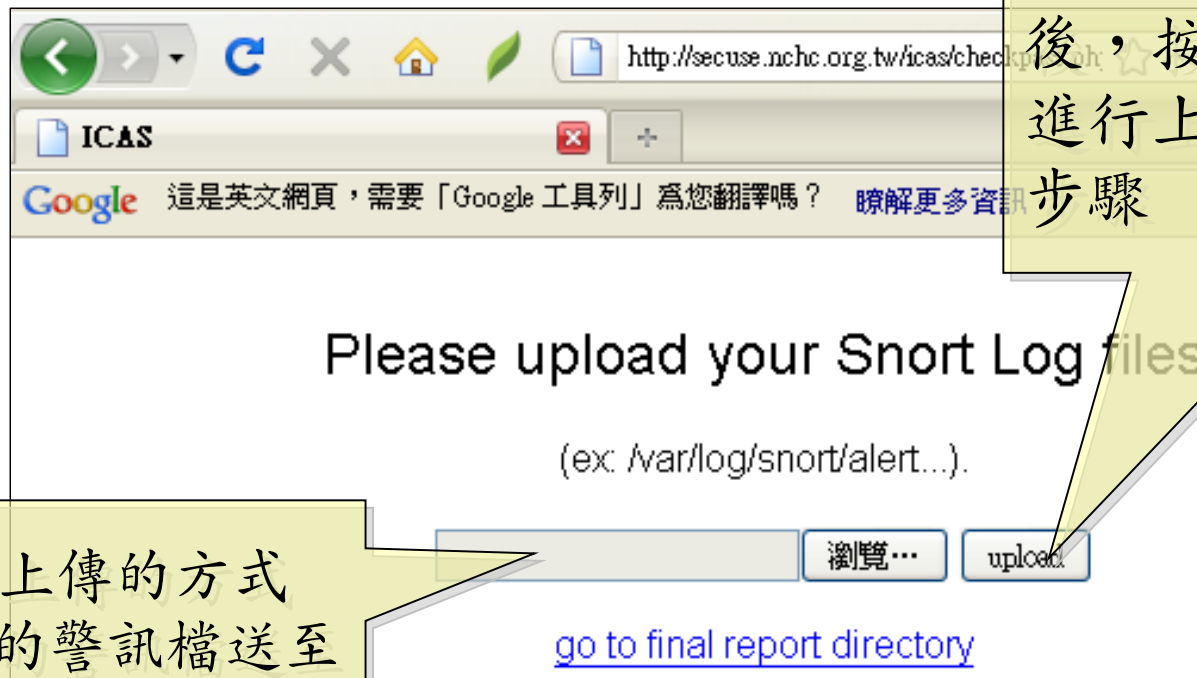
# ICAS-I 效能數據表

## Throughput Data Overall

Original Alerts	Analysis Time (sec)					Results	Reduction Rate
	Traditional	1 nodes	2 nodes	4 nodes	6 nodes		
286	1.068	4.087	4.869	4.864	5.077	30	89.51%
380	1.333	4.94	5.069	5.067	5.097	11	97.11%
434	1.76	4.61	5.066	5.068	5.09	9	97.93%
754	3.145	5.066	5.079	5.038	5.096	16	97.88%
1174	4.73	6.066	5.093	5.089	5.097	33	97.19%
1668	7.909	6.07	6.56	6.071	5.082	16	99.04%
2182	14.949	6.671	6.95	5.166	5.088	16	99.27%
3396	19.901	7.053	6.654	5.076	5.091	68	98.00%
5816	374.374	9.081	9.076	9.07	7.076	66	98.87%
6344	383.82	9.68	9.872	7.069	6.069	72	98.87%
12698	801.346	13.096	12.367	11.367	9.083	36	99.72%

# ICAS-II

- ICAS-I 僅將資料塞入資料庫，然而還是文字的敘述
- ICAS-II 將輸入的任意多個警訊整合成一張警訊關聯圖
- 資料的來源可以透過以下兩種方式上傳到分析平台
  - 系統自動設定以 SCP 傳送到 ICAS 工作目錄
  - 管理者透過 ICAS 網頁上傳

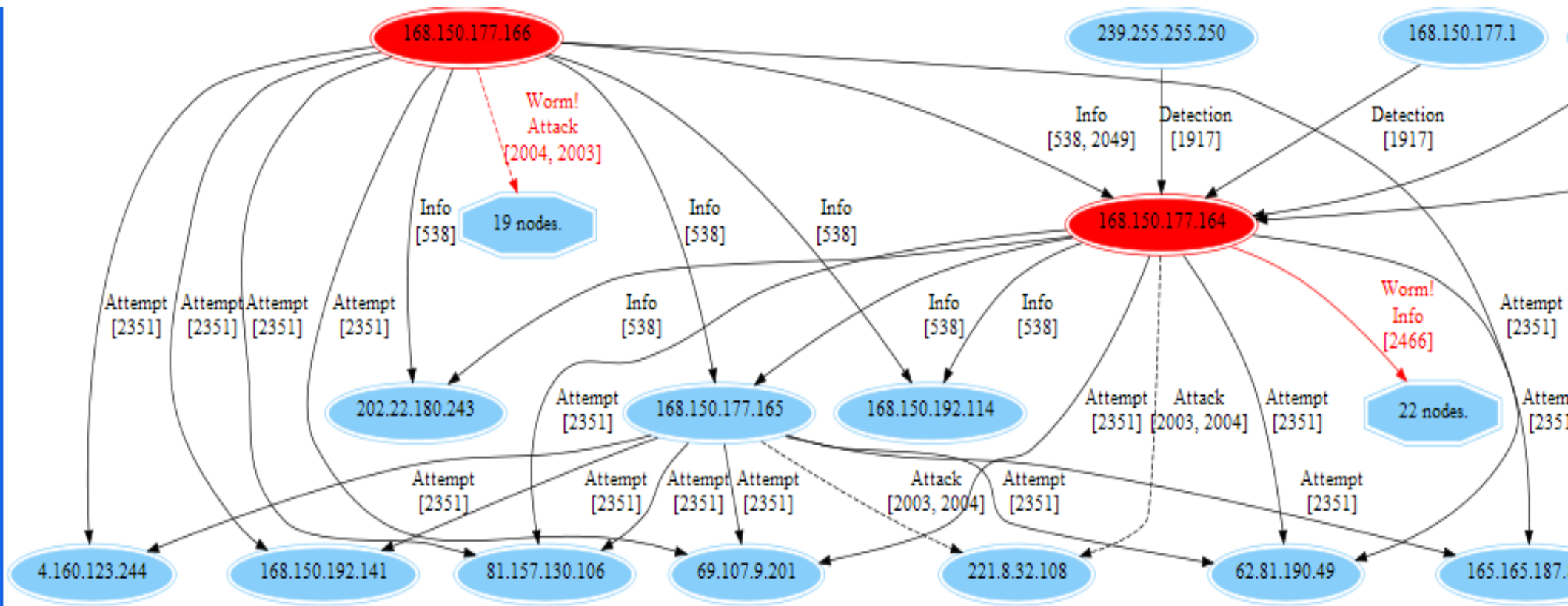


一旦選定需分析的日誌檔後，按下『上傳』，系統進行上傳→分析→繪圖等步驟

透過網頁上傳的方式將 Snort 的警訊檔送至 ICAS 分析

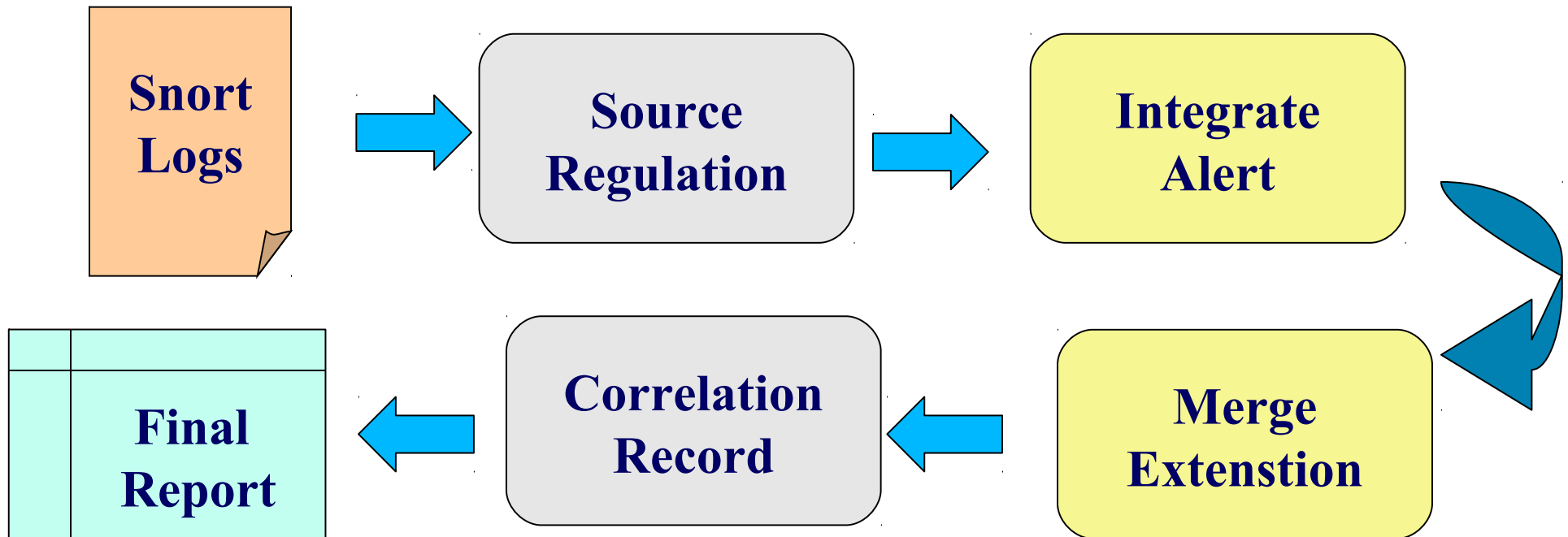
# ICAS-II 所產生的報表：警訊關聯圖

- 經過 ICAS-II 分析後，可以得到此警訊關聯圖。
- 圖中橢圓形代表節點，箭頭及線上文字代表攻擊方向與攻擊方法。
- 標為紅色則是經過系統分析之後，被判定有攻擊行為的節點與方法。
- 此圖說明 IP 168.150.177.166 與 168.150.177.164 有進行蠕蟲的攻擊行為



# ICAS-II 的分析流程

- Hadoop v 0.20



# ICAS-II 結論

- ICAS-II 可經過警訊的來源、目的、攻擊事件綜合分析
  - 提供巨觀攻擊關聯圖來瞭解攻擊事件的始末
  - 自動透過標記顏色的方法將較高危險的事件呈現出來。
- ICAS-II 尚在整合關聯式資料庫，因此未進行數據量測

# ICAS 總結

- 雲端運算處理資料格式相似且資料量大的情況下，能展現其效益
- 提供高容錯率、低獨占系統資源、多工作同時執行等能力
- 可搭配其他軟體作即時的警訊資料呈現，ICAS 可補充分析後資料的部份
- 未來工作
  - 整合多種資料來源平台
  - 產生更詳細與人性化的分析資料





# 運用抓抓龍製作個人化書籤搜尋引擎

Build Your Personal Bookmark Search Engine using Crawlzilla

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



Powered by DRBL

# Search is everywhere in our daily life !!

## 「搜尋」已經成爲我們生活中的一部分

搜尋結果

### 檔案搜尋

網址(D) 搜尋結果


搜尋小幫手

您想要搜尋什麼?

- 圖片、音樂，或視訊(P)
- 文件(文字處理、試算表，等等)(O)
- 所有檔案和資料夾(L)
- 電腦或人員(C)
- 說明和支援中心裡的資訊(I)

您也可能想要...

- 搜尋網際網路(S)
- 變更喜好(G)



0 個物件

Gmail Calendar Documents Photos Sites Web More -

Search

All Mail

From

To

Subject

Has the words

Doesn't have

Has attachment

Date within 1 day of

Examples: f

Search

### 信件搜尋

發的交談

jarwin.nchc.org.tw 於 2011年12月02日 (週五) 10時53分46秒 的交談

(10時53分48秒) Shunfa 楊順發

(10時53分51秒) Jazz Yao-Tsung

(10時54分08秒) Shunfa 楊順發

(10時54分42秒) Jazz Yao-Tsung

(10時54分49秒) Jazz Yao-Tsung

(10時54分51秒) Jazz Yao-Tsung

(10時55分02秒) Shunfa 楊順發

(10時55分04秒) Shunfa 楊順發

(10時55分39秒) Jazz Yao-Tsung

尋找(F)

關閉(C)

### 即時通訊搜尋

IEEE Xplore DIGITAL LIBRARY

BROWSE

- Journals & Magazines
- Conference Proceedings
- Standards
- Books
- Educational Courses

SIGN IN

Search 3,076,887 documents

SEARCH

Advanced Search | Preferences | Search Tips

### 資料庫搜尋

## 今天要談的是「網頁搜尋」

設Yahoo!奇摩為首頁 資訊展PK線上搶先

# YAHOO! 奇摩

網頁 | 知識+ | 圖片 | 影片 | 部落格 | 字典 | 新聞 | 購物 BETA

網頁搜尋

熱門: 第一美腿 12歲父親 嫩模女神 幼稚病 51區 花心星座 解夢 知識: 傷口癢竟是 電鍋料理

2011 資訊月 ONLINE 3G特展搶先看!!

# To speed up search, We need “Index”

爲了加速搜尋的效率，我們需要「索引」

## Index

出現頁碼

關鍵字

### Symbols

! (exclamation mark) command prefix, 368

### A

ack queue in HDFS, 66

ACLs (access control lists)  
for Hadoop services, 283  
ZooKeeper, 446, 456

ActiveKeyValueStore class (example), 464

ad hoc analysis and product feedback  
(hypothetical use case), 511

adjacency list, 560

installation, 565

prerequisites, 565

TeraByte sort on, 553

Apache Hadoop project, 10, 12

Apache Lucene project, 9

Apache Nutch, 9

Apache Thrift services, 49

Apache ZooKeeper, 442

(see also ZooKeeper)

APIs in ZooKeeper, 453

archive files, copying to tasks, 253

archive tool, 72

archives, 72

# History of Hadoop ... 2001~2005

## Hadoop 這套軟體的歷史源起 ... 2001~2005



- 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 ... 2005~2006

## Hadoop 這套軟體的歷史源起 ... 2005~2006

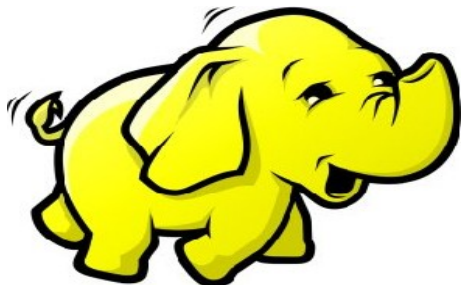
- Nutch
  - <http://nutch.apache.org/>
  - Nutch 是基於開放原始碼所開發的網站搜尋引擎
  - Nutch is open source **web-search** software.
  - 利用 Lucene 函式庫開發
  - It builds on **Lucene and Solr**, adding web-specifics, such as a **crawler**, a **link-graph database**, parsers for HTML and other document formats, etc.



# History of Hadoop ... 2006 ~ Now

## *Hadoop* 這套軟體的歷史源起 ... 2006 ~ Now

- Nutch 後來遇到儲存大量網站資料的瓶頸，剛好看到 Google 在一些會議分享他們的三大關鍵技術 ...
- 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.)
- Dong Cutting joined Cloudera at year 2009.



**YAHOO!**

**cloudera**



# lock files don't work in JDK 1.1

Log In

Views

### Details

Type:	Bug	Status:	Closed
Priority:	Major	Resolution:	Fixed
Affects Version/s:	1.2	Fix Version/s:	None
Component/s:	core/store		
Labels:	None		
Environment:	Operating System: All		

### People

Assignee: [Lucene Developers](#)  
 Reporter: [Doug Cutting](#)  
 Vote (0) Watch (0)

### Dates

Created: 09/Oct/01 16:19



# test

Log In

Views

### Details

Type:	Bug	Status:	Closed
Priority:	Trivial	Resolution:	Fixed
Affects Version/s:	None	Fix Version/s:	None
Component/s:	None		
Labels:	None		

### People

Assignee: Unassigned  
 Reporter: [Doug Cutting](#)  
 Vote (0) Watch (0)

### Dates

Created: 11/Feb/05 04:23



# initial import of code from Nutch

Log In

Views

### Details

Type:	Task	Status:	Closed
Priority:	Major	Resolution:	Fixed
Affects Version/s:	None	Fix Version/s:	0.1.0
Component/s:	None		
Labels:	None		

### People

Assignee: [Doug Cutting](#)  
 Reporter: [Doug Cutting](#)  
 Vote (0) Watch (0)

### Dates

Created: 01/Feb/06 02:54

# Do you like to write notes?

## 你有寫筆記的習慣嘛？



**EVERNOTE** 想知道更多 | 升級至專業版 | 檔案下載 | 部落格 | 百寶箱

• 登入  
• 立即註冊

### 什麼都記得住

**擷取每件事。**  
儲存點子、喜愛事物以及所見所聞。

**隨處存取。**  
Evernote 幾乎能在所有電腦、手機以及其他行動裝置上使用。

**快速尋找事物。**  
根據標題、標籤或甚至是影像內的印刷與手寫文字進行搜尋。

取得 Evernote [立即下載](#) 免費使用。

大腦記憶力有限，只好靠筆記啦！



# Tools that I used to write notes

## 我作筆記的工具 (1) 維基 **Oddmuse Wiki**

軟體下載：<http://www.oddmuse.org/>

[關於本站](#) [更新紀錄](#) [[年曆\(c\)](#)] [[登入\(l\)](#)] [[上一頁](#)] [[首頁\(h\)](#)] [[12-02\(9\)](#)] [[12-03\(0\)](#)] [[12-04\(1\)](#)] [[12-05\(2\)](#)] [[12-06\(3\)](#)]

December 2011						
Su	Mo	Tu	We	Th	Fr	Sa
				1	2	
4	5	6	7	8	9	1
11	12	13	14	15	16	1
18	19	20	21	22	23	2
25	26	27	28	29	30	3

## 關於本站

架站日期：2005/01/01

基本架構：<http://www.oddmuse.org/> Oddmuse Wiki

## 更新紀錄

2005-01-01		
chinese-utf8.pl	<b>Chinese Translations</b>	<a href="http://www.oddmuse.org/cgi-bin/wiki/Chinese">http://www.oddmuse.org/cgi-bin/wiki/Chinese</a>
download.pl	<b>Download Extension</b>	<a href="http://www.oddmuse.org/cgi-bin/wiki/Download_Extension">http://www.oddmuse.org/cgi-bin/wiki/Download_Extension</a>
headers.pl	<b>Header Markup Extension</b>	<a href="http://www.oddmuse.org/cgi-bin/wiki/Header_Markup_Extension">http://www.oddmuse.org/cgi-bin/wiki/Header_Markup_Extension</a>
tables-long.pl	<b>Long Table Markup Extension</b>	<a href="http://www.oddmuse.org/cgi-bin/wiki/Long_Table_Markup_Extension">http://www.oddmuse.org/cgi-bin/wiki/Long_Table_Markup_Extension</a>
tables.pl	<b>Table Markup Extension</b>	<a href="http://www.oddmuse.org/cgi-bin/wiki/Table_Markup_Extension">http://www.oddmuse.org/cgi-bin/wiki/Table_Markup_Extension</a>
usemod.pl	<b>Usemod Markup Extension</b>	<a href="http://www.oddmuse.org/cgi-bin/wiki/Usemod_Markup_Extension">http://www.oddmuse.org/cgi-bin/wiki/Usemod_Markup_Extension</a>
2005-02-16		
calendar.pl	<b>Calendar Extension</b>	<a href="http://www.oddmuse.org/cgi-bin/wiki/Calendar_Extension">http://www.oddmuse.org/cgi-bin/wiki/Calendar_Extension</a>

2005~2008

# Tools that I used to write notes

## 我作筆記的工具 (2) 維基 PmWiki

軟體下載：<http://www.pmwiki.org/>

### 導覽列 (編輯)

首頁  
民視專訪  
福山參訪  
超級視訊  
阿聰的首頁  
穎燦的首頁  
測試沙箱  
近期更新

### PmWiki 官方網站

程式安裝  
常見問題  
外掛程式  
布景主題  
臭蟲管理

### PmWiki 使用手冊

頁面編輯入門  
頁面編輯技巧  
建立新的頁面

Main » [Home Page](#)

[本文](#) · [編輯](#) · [附檔](#) · [列印](#) · [歷史](#)

Search

縮小 正常 放大

### 歡迎使用 PmWiki:

以下是一些 [PmWiki](#) 安裝後的預設頁面，你可以從這些頁面開始瞭解 PmWiki:

- [Pm Wiki](#) 的 [相關文件目錄](#)。
- 什麼是 [WikiWikiWeb](#) ?
- [頁面編輯技巧](#) 和 [寫作語法](#) 這兩份完文件描述如何建立一個 Wiki 頁面。
- 你可以用 [測試沙箱](#) 來練習 Wiki 的語法。

請定期拜訪 [PmWiki 官方網站](#) 以便獲得最新的資訊。

« [October 2011 period](#) · [February 2012 period](#) »

### December 2011

Mon	Tue	Wed	Thu	Fri	Sat	Sun
			<a href="#">01</a>	<a href="#">02</a>	<a href="#">03</a>	<a href="#">04</a>
<a href="#">05</a>	<a href="#">06</a>	<a href="#">07</a>	<a href="#">08</a>	<a href="#">09</a>	<a href="#">10</a>	<a href="#">11</a>
<a href="#">12</a>	<a href="#">13</a>	<a href="#">14</a>	<a href="#">15</a>	<a href="#">16</a>	<a href="#">17</a>	<a href="#">18</a>
<a href="#">19</a>	<a href="#">20</a>	<a href="#">21</a>	<a href="#">22</a>	<a href="#">23</a>	<a href="#">24</a>	<a href="#">25</a>
<a href="#">26</a>	<a href="#">27</a>	<a href="#">28</a>	<a href="#">29</a>	<a href="#">30</a>	<a href="#">31</a>	

### January 2012

Mon	Tue	Wed	Thu	Fri	Sat	Sun
						<a href="#">01</a>
<a href="#">02</a>	<a href="#">03</a>	<a href="#">04</a>	<a href="#">05</a>	<a href="#">06</a>	<a href="#">07</a>	<a href="#">08</a>
<a href="#">09</a>	<a href="#">10</a>	<a href="#">11</a>	<a href="#">12</a>	<a href="#">13</a>	<a href="#">14</a>	<a href="#">15</a>
<a href="#">16</a>	<a href="#">17</a>	<a href="#">18</a>	<a href="#">19</a>	<a href="#">20</a>	<a href="#">21</a>	<a href="#">22</a>
<a href="#">23</a>	<a href="#">24</a>	<a href="#">25</a>	<a href="#">26</a>	<a href="#">27</a>	<a href="#">28</a>	<a href="#">29</a>
<a href="#">30</a>	<a href="#">31</a>					

# Tools that I used to write notes

## 我作筆記的工具 (3) 離線網頁 ScrapBook

安裝：<https://addons.mozilla.org/zh-TW/firefox/addon/scrapbook/>

The screenshot shows the Mozilla Add-ons website for the ScrapBook extension. At the top right, there are links for "註冊 或 登入" and "其他應用程式", and the Mozilla logo. The main heading is "附加元件" (Add-ons) with sub-links for "擴充套件", "個性面板", "佈景主題", "收藏集", and "更多...". A search bar contains "搜尋附加元件". Below the search bar, a breadcrumb trail reads "家 >> 擴充套件 >> ScrapBook". The main card for "ScrapBook 1.4.8" by "Gomita" features a green "精選" (Featured) banner, a description "協助您擷取網頁，方便整理擷取後的資料。", and a green "+ 新增至 Firefox" button. To the right, it shows a 4.5-star rating, "564 位使用者評論", and "448,868 位使用者". Below the card, there is a "捐款" (Donate) button with "建議 US\$2.99" and a "覺得這個附加元件很棒嗎?" section with a "G" icon and text "此附加元件的開發者希望您小額捐款支持附加元件的永續開發。". A yellow box on the right contains the text "2005~NOW".



# Tools that I used to write notes

## 我作筆記的工具 (4) 維基 + 版本控制 Trac

軟體下載：<http://trac.edgewall.org/>

 Search

[Login](#) | [Preferences](#) | [Help/Guide](#) | [About Trac](#)

	<a href="#">Wiki</a>	<a href="#">Timeline</a>	<a href="#">Roadmap</a>	<a href="#">Browse Source</a>	<a href="#">View Tickets</a>	<a href="#">Search</a>
--	----------------------	--------------------------	-------------------------	-------------------------------	------------------------------	------------------------

[Start Page](#) | [Index](#) | [History](#) | [Last Change](#)

~ Welcome to NCHC Grid Architecture Research Group ~

### 【Project News】

- 2011-01-28: [2011 Project Deliverable Results](#) had updated~
- 2010-07-01: [2010 Project Deliverable Results](#) had updated~
- 2009-03-18: [2009 Project Deliverable Results](#) had updated~
- 2008-10-08: [2008 Project Deliverable Results](#) had updated~
- [雲端平台維護日誌](#)
- [Paper Reading Schedule \(讀書會行程\)](#)
- [Submission Deadline List \(投稿截止日期列表\)](#)
- [2008 Related NEWS](#)

### 【Project News】

#### 【1: Distributed / Parallel Computing】

- 1.1: [Kerrighed](#)
- 1.2: [BOINC](#)
- 1.3: [WebOS](#)
- 1.4: [Parallel Computing](#)

#### 【2: File System & Data Grid】

- 2.1 [Distributed & Parallel File System](#)
- 2.2 [FS Realition](#)

#### 【3: Virtualization】

#### 【Projects】

#### 【Related Conference Paper / Poster / Demo Submission】

#### 【Possible Technical Whitepaper Submission】

#### 【Technical Tutorials】

[Visitor](#)

### 【1: Distributed / Parallel Computing】

#### 1.1: [Kerrighed](#)

#### 1.2: [BOINC](#)

- 2008-05-27: [How to Setup BOINC example project](#)
- 2008-02-13: [How to Build BONIC Server](#)

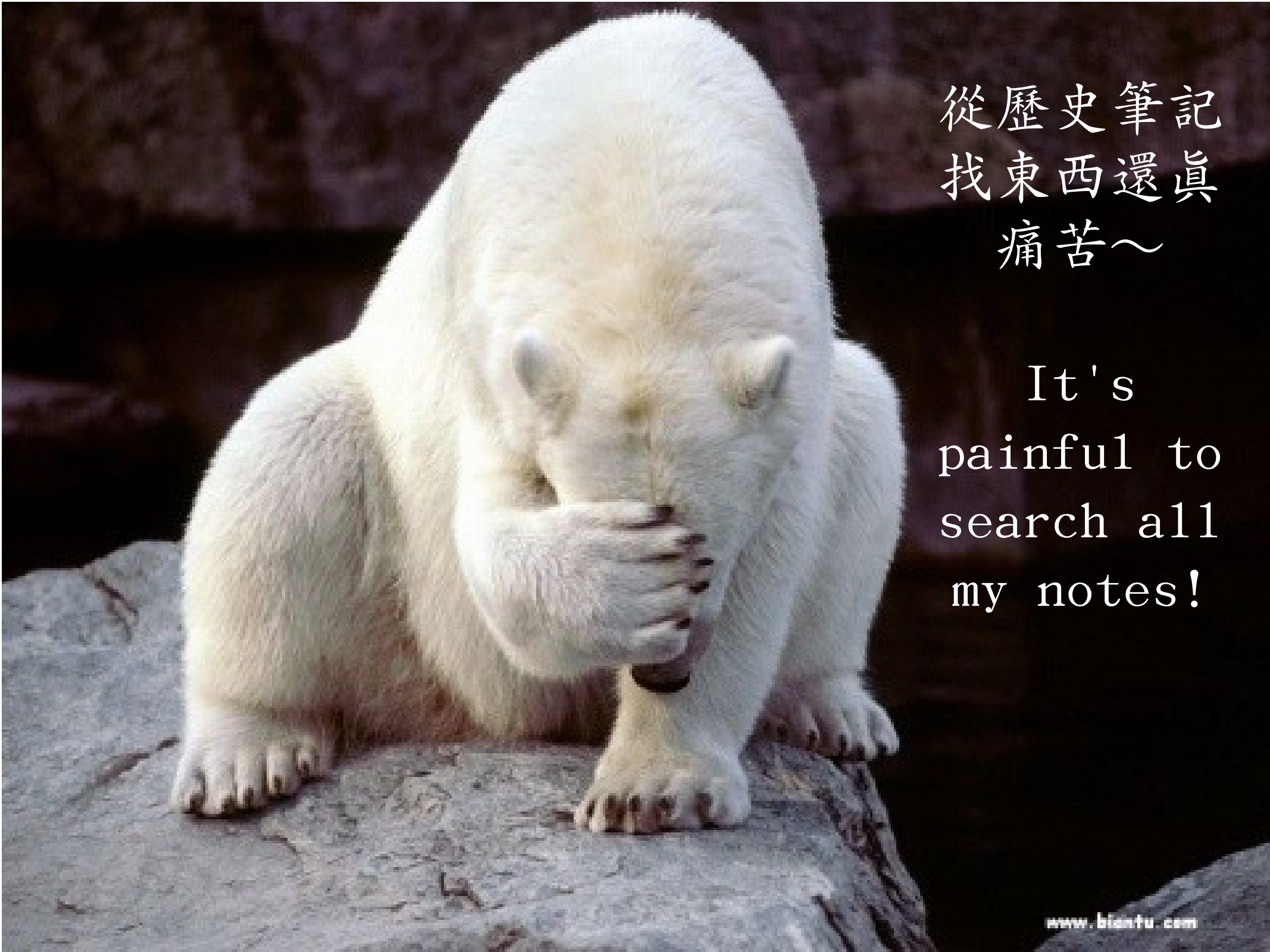
2006~NOW

# Tools that I used to write notes

## 我作筆記的工具 (5) 線上書籤 ReadItLater

安裝：<http://readitlaterlist.com/>

The screenshot shows the Read It Later website with a dark theme. At the top, there are navigation links for 'Login', 'Sign Up', 'API', and 'Help'. Below that, there are links for 'Firefox', 'iPhone/iPad', 'Android', 'Mobile', and 'All Browsers'. The main heading 'Read it Later' is prominently displayed. A yellow box highlights the text '2010~NOW'. The central message reads 'One reading list, everywhere you are.' Below this, there are two call-to-action buttons: 'Add to Firefox' and 'Add to iPhone/iPad'. Under 'Add to Firefox', it lists 'Internet Explorer', 'Chrome', 'Safari', and 'others'. Under 'Add to iPhone/iPad', it lists 'Android', 'Blackberry', 'WebOS', and 'others'. The bottom section features three icons with corresponding text: a computer and phone icon for 'Save' (Save pages from your computer or phone), an airplane and train icon for 'Read on or offline' (Read at home, work, on the plane, or during your commute; even without an internet connection), and a globe icon for 'Access Anywhere' (Read It Later is integrated into many popular applications and platforms you may already use).



從歷史筆記  
找東西還真  
痛苦～

It's  
painful to  
search all  
my notes!



既然我有許多筆記放在網頁上，  
何不試試自家研發的抓抓龍呢？

# Crawlzilla 系統功能

## Feature of Crawlzilla

- 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

如果您還不認識抓抓龍，

不妨看一下 2010 年 Hadoop 使用者會議的錄影

[http://cloud.nchc.org.tw/20101202/slides/01\\_Crawlzilla.wmv](http://cloud.nchc.org.tw/20101202/slides/01_Crawlzilla.wmv)



# System Architecture of Crawlzilla

## 抓抓龍系統架構

Web UI ( Crawlzilla Website + Search Engine)

JSP + Servlet + JavaBean

Nutch

Lucene

Crawlzilla System Management

Tomcat

Hadoop

PC1

PC2

PC3

# Comparison with other projects

## 抓抓龍與其他搜尋引擎專案的比較

	<b>Spidr</b>	<b>Larbin</b>	<b>Jcraw</b>	<b>Nutch</b>	<b>Crawlzilla</b>
Install	Rube Package Install	Gmake Compiler and Install	Java Compiler and Install	Deploy Configure Files	<b>Provide Auto Installation</b>
Crawl website pages	O	O	O	O	<b>O</b>
Parser Content	X	X	X	O	<b>O</b>
Cluster Computing	X	X	X	O	<b>O</b>
Interface	Command	Command	Command	Command	<b>Web-UI</b>
Support Chinese Segmentation	X	X	X	X	<b>O</b>

# New Feature of Crawlzilla 1.0

## 抓抓龍 1.0 的新功能

- 支援多重使用者
- 採用 jQuery Ui 打造的新網頁管理介面
- 支援重新爬取 ( Re-Crawl )
- 支援排程爬取 ( Schedule / Crontab )
- 支援雲端服務：
  - 懶得自己建？沒關係！這裡可以試用！
  - <http://demo.crawlzilla.info>

# Multi-user Web Search Cloud Service : Crawlzilla 1.0

## Crawlzilla 1.0 多人版雲端服務 (1)

首先連線到 <http://demo.crawlzilla.info>

(1)

Home	Crawl	索引庫管理	系統排程	Slave安裝	系統設定	登入/註冊
------	-------	-------	------	---------	------	-------

使用者登入

使用者註冊

使用者帳號	demo
使用者密碼	.....
密碼確認	.....
電子信箱	jazz@nchc.org.tw

Submit Reset

運算節點

工作排

空間管

(2)

(3)

▲ Step 1 : 新使用者註冊頁面

# Multi-user Web Search Cloud Service : Crawlzilla 1.0

## Crawlzilla 1.0 多人版雲端服務 (2)

接著等待管理者幫您開啓帳號！

尚未啓用會員列表

使用者	e-mail	申請時間	確認使用者
demo	jazz@nchc.org.tw	2011-12-04 21:18:37	Accept User

系統概況

(1)

項目	
會員人數	
搜尋引擎總數	



(2)

**Your account has been accepted from Crawlzilla System**

crawlzilla@gmail.com <crawlzilla@gmail.com>  
To: jazz@nchc.org.tw

We are pleased to inform you that your account has been accepted.

Please visit <http://140.110.134.197:8080> to build your search engine!

Thank you for use crawlzilla.

-This mail sent by the system automatically, do not reply to this mail.-

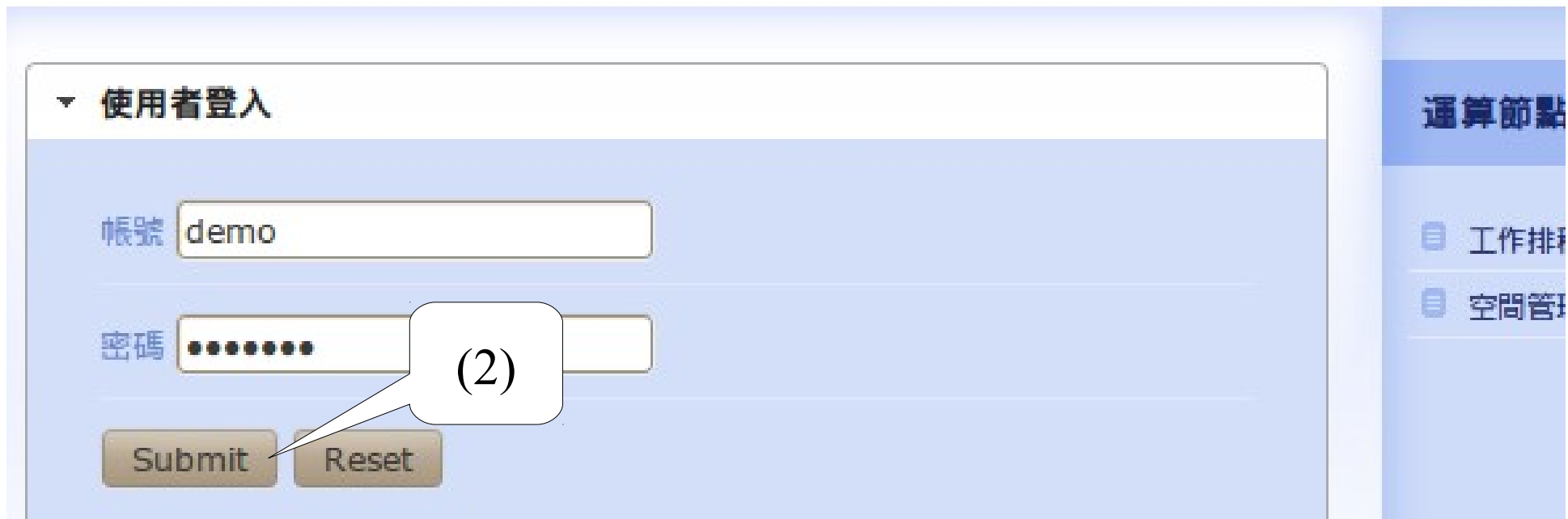
▲ Step 2 : 等待管理者啓用，您會收到啓用通知

# Multi-user Web Search Cloud Service : Crawlzilla 1.0

## Crawlzilla 1.0 多人版雲端服務 (3)

重新連線到 <http://demo.crawlzilla.info>

(1)



▲ Step 3 : 登入您的個人化管理頁面

## Crawlzilla 1.0 多人版雲端服務 (4)

建立新的搜尋索引庫

(1)

Home

Crawl

索引庫管理

系統排程

### Crawl-建立搜尋引擎

#### ▼ Crawl Setup

索引庫名稱: all

(2)

(5)

#### ▼ Crawl Setup

索引庫名稱: all

輸入欲爬取的網址(可多行):

http://cloud.nchc.org.tw/~jazz/rii\_export.html  
http://cloud.nchc.org.tw/~jazz/firefox\_bookmarks.html  
http://trac.nchc.org.tw/grid/wiki/jazz/Work\_2011  
http://trac.nchc.org.tw/grid/wiki/jazz/Work\_2010  
http://trac.nchc.org.tw/grid/wiki/jazz/Work\_2009  
http://trac.nchc.org.tw/grid/wiki/jazz/Work\_2008

(3)

爬取深度設定: 2

(4)

註: 因系統負載有限, 此一體驗網站 能建立3個索引庫, 不便之處請見諒

▶ 排程設定 (Option)

Submit

Reset

▲ Step 4 : 輸入索引庫名稱、起始網址與搜尋深度

## Crawlzilla 1.0 多人版雲端服務 (5)

接著只能靜候抓抓龍幫您建立搜尋索引庫

HRER to check the status of your jobs!' is provided. At the bottom of the main content area, a red warning message states: 'The crawling time depends on your system performance, URLs number, and depth.' On the right side of the interface is a sidebar with a blue background and white text. It contains a '搜尋引擎列表' section and a '運算節點即時狀態' section. The '運算節點即時狀態' section has two items: '工作排程器狀態' and '空間管理員狀態'. A white speech bubble with the number '(1)' is pointing to the '工作排程器狀態' item."/>

Crawlzilla Search Engine Hands-on Display

Home Crawl 索引庫管理 系統排程 Slave安裝 系統設定 系統登出

**Setup has been submit !!! But, it need time to crawl !!!**

ex. 4URLs with 1 depth -> 10~20 minute  
4URLs with 2 depth -> 40~80 minute  
100URLs with 10 depth -> very very long

click [HRER](#) to check the status of your jobs!

**The crawling time depends on your system performance, URLs number, and depth.**

搜尋引擎列表

運算節點即時狀態

- 工作排程器狀態
- 空間管理員狀態

(1)

▲ Step 5 : 等待抓抓龍幫您建立專屬的搜尋索引庫



## Crawlzilla 1.0 多人版雲端服務 (6)

### 索引庫管理

Home Crawl **索引庫管理** 系統排程 Slave安裝 系統設定

(1)

### 索引庫管理

系統爬取狀態

索引庫名稱	爬取狀態	爬取時間	刪除狀態
all	crawling	0h:3m:31s	

(2)

▲ 您可以在索引庫管理看到目前爬取已使用的時間

## Crawlzilla 1.0 多人版雲端服務 (7)

搜尋索引庫建立完成後，  
可以於「索引庫管理」處進行**手動**重新爬取（re-crawl）  
或刪除索引庫的動作

索引庫管理

系統爬取狀態

索引庫名稱	爬取狀態	爬取時間	刪除狀態
-------	------	------	------

索引庫列表

索引庫名稱	建立時間	爬取時間	爬取深度	索引庫操作	執行
ril	2011-12-04 16:15:24	0:18:13	2	Select Re-Crawl Delete IDB	Run

(1) 告訴您該索引庫花了多久時間爬取

(2) Re-Crawl

(3) Delete IDB

▲ 您可以在索引庫管理進行手動重新爬取或刪除索引庫

## Crawlzilla 1.0 多人版雲端服務 (8)

可以於「系統排程」處進行排程重新爬取 ( schedule )

The screenshot shows the '系統排程' (System Scheduling) page in the Crawlzilla 1.0 interface. The page has a navigation bar at the top with links for 'Home', 'Crawl', '索引庫管理', '系統排程', 'Slave安裝', and '系統設定'. The main content area is titled '索引庫爬取排程' (Index Library Crawl Scheduling) and is divided into two sections: '排程資訊' (Scheduling Information) and '新增排程' (Add Scheduling).

Callout (1) points to the '系統排程' menu item in the navigation bar. Callout (2) points to the '刪除排程' (Delete Scheduling) button in the table header. Callout (3) points to the '索引庫選擇' (Index Library Selection) dropdown menu. Callout (4) points to the 'Submit' button at the bottom of the form.

索引庫名稱	爬取排程時間	週期	刪除排程
-------	--------	----	------

新增排程

索引庫選擇

排程日期

排程時間  時  分

週期

▲ 您可以在索引庫管理看到目前爬取已使用的時間

## Crawlzilla 1.0 多人版雲端服務 (9)

可以於「索引庫管理」處進行即時讀取索引庫資訊

索引庫列表

索引庫名稱	建立時間	爬取時間	爬取深度	索引庫操作	執行
ril	2011-12-04 16:15:24	0:18:13	2	Select ▼	Run

(1)

(2)

### 關於搜尋引擎ril

▼ 基本資訊

索引庫名稱 ril

搜尋引擎連結位置 /home/crawler/crawlzilla/user/jazzwang/IDB/ril/index

搜尋引擎狀態 OK

爬取深度 2

建立時間 20111204-15:57:16

執行時間 0:18:13

起始連結 [http://cloud.nchc.org.tw/~jazz/ril\\_export.html](http://cloud.nchc.org.tw/~jazz/ril_export.html)

▶ 立即搜尋及網頁嵌入語法

▲ 您可以在索引庫管理取得即時搜尋索引庫的資訊

## Crawlzilla 1.0 多人版雲端服務 (10)

在搜尋索引庫資訊中，可以取得加入個人化搜尋引擎的語法



The screenshot displays a web page titled "關於搜尋引擎ril". It features a navigation menu with "基本資訊" and "立即搜尋及網頁嵌入語法". The "立即搜尋及網頁嵌入語法" section is expanded, showing a search engine logo for "CRAWLZILLA" and a search input field. Below this, the "語法" section displays the following HTML code:

```
  
  
<form name="search" action="http://demo.crawlzilla.info/jazzwang_ril/search.jsp" method="get">  
  
<input name="query" size=15></form>
```

Callout (1) points to the search engine logo and input field. Callout (2) points to the HTML code block.

▲ 您可以在索引庫管理取得嵌入網頁的語法

## Crawlzilla 1.0 多人版雲端服務 (11)

索引庫內容說明了共搜尋了多少個文件（網頁），  
並且會統計最常到訪的網址排行榜

索引庫內容 ril

資料總覽

總共文字數 213155

文件檔數量 3234

索引庫更新日期 Sun Dec 04 16:15:23 CST 2011

被搜尋分析到的網址

分析的文件型態

出現次數前五十分的字符

(1)

索引庫內容 ril

資料總覽

被搜尋分析到的網址

Order	Contents	Counts	Order	Contents
0	site:www.digitimes.com.tw	204	1	site:www.bnext.c
2	site:groups.google.com	74	3	site:highscalabilit
4	site:www.theregister.co.uk	49	5	site:www.ithome
6	site:www.cloudera.com	47	7	site:gigaom.com
8	site:en.wikipedia.org	39	9	site:www.networ
10	site:wiki.apache.org	37	11	site:www.zdnet.c
12	site:www.howtoforge.com	33	13	site:www.freegro
14	site:ajaxian.com	29	15	site:news.networ
16	site:www.linuxfordevices.com	25	17	site:ieeexplore.ie
18	site:www.linux-mag.com	24	19	site:insidehpc.cor
20	site:www.readwriteweb.com	22	21	site:only-percept
22	site:www.nosqldatabases.com	19	23	site:www.openfo
24	site:www.inside.com.tw	17	25	site:www.sys-cor

(2)

▲ 索引庫內容提供了許多統計資訊

## Crawlzilla 1.0 多人版雲端服務 (12)

此外，索引庫內容也說明了共搜尋了哪幾種文件，並且會統計最常出現的關鍵字排行榜

### 索引庫內容 ril

▶ 資料總覽

▶ 被搜尋分析到的網址

▼ 分析的文件型態

Order	Contents	Counts	Order	Contents	Counts
0	type:text	3232	1	type:text/html	3231
2	type:html	3231	3	type:xhtml+xml	2
4	type:application/xhtml+xml	2	5	type:application	2
6	type:text/plain	1	7	type:plain	1

(1)

▶ 出現次數前五十分的字符

### 索引庫內容 ril

▶ 資料總覽

▶ 被搜尋分析到的網址

▶ 分析的文件型態

▼ 出現次數前五十分的字符

Order	Contents	Counts	Order	Contents	Counts
0	content:a	2350	1	content:1	2346
2	content:2011	2240	3	content:2	2235
4	content:s	2218	5	content:3	2082
6	content:4	2030	7	content:all	1988
8	content:com	1903	9	content:about	1879
10	content:from	1861	11	content:you	1851
12	content:5	1804	13	content:10	1794
14	content:more	1786	15	content:can	1746
16	content:2010	1714	17	content:new	1653
18	content:your	1626	19	content:i	1620
20	content:use	1609	21	content:data	1602

(2)

▲ 索引庫內容提供了許多統計資訊

# 你也可以擁有自己的搜尋引擎 !!!

## Start from Here!

- **Crawlzilla** 示範多人自訂搜尋雲端服務
  - <http://demo.crawlzilla.info>
- **Crawlzilla @ Google Code Project Hosting ( 中文 )**
  - <http://code.google.com/p/crawlzilla/>
- **Crawlzilla @ Source Forge (Tutorial in English)**
  - <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>



# Authors of Crawlzilla

## 抓抓龍作者群

陳威宇 (左)

***waue@nchc.org.tw***

***waue0920@gmail.com***

郭文傑 (中)

***rock@nchc.org.tw***

***goldjay1231@gmail.com***

楊順發 (右)

***shunfa@nchc.org.tw***

***shunfa@gmail.com***





## Questions?

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

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



Powered by DRBL