Course Information 課程資訊

- 講師介紹 About Me ：
  - 國網中心 王耀聰 副研究員 / 交大電控碩士
  - Yao-Tsung Wang (Jazz) / Associate Researcher, NCHC
  - Master, Electrical and Control Engineering, NCTU, Taiwan
  - jazz@nchc.org.tw
- 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。
  It changes too fast, don't waste your paper !!
- 更多資訊 All Info. will be updated :
  - http://trac.nchc.org.tw/cloud
  - http://www.classcloud.org/media - training course video archives
  - http://www.screentoaster.com/user?username=jazzwang
- 若需要實驗環境 If you need an environment of Hadoop :
  - http://hadoop.nchc.org.tw
- 相關問題討論 If you have questions about Hadoop ：
  - http://forum.hadoop.tw

Let's have a QUICK REVIEWS about Cloud

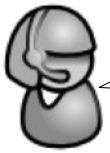
運用自由軟體打造私有雲端
Build Your Own Private Cloud using Open Source

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

Powered by DRBL

什麼是雲端運算啊？可以個簡單的定義嗎？
What is Cloud Computing ?

雲端運算怎麼聽起來要買一些新硬體、新軟體啊？
Is it about buying NEW Hardware and Software?



雲端運算可能只是拿來振興經濟的幌子吧？
Is it a trap to another bubble economy ?

我聽你們在那裡講五四三.....
Cloud Computing is as simple as 5..4..3..2..1...

National Definition of Cloud Computing
美國國家標準局 NIST 給雲端運算所下的定義

5 Characteristics 五大基礎特徵

4 Deployment Models 四個佈署模型

3 Service Models 三個服務模式

On-demand self-service.
隨需自助服務

Broad network access
隨時隨地用任何網路裝置存取

Resource pooling
多人共享資源池

Rapid elasticity
快速重新佈署靈活度

Measured Service
可被監控與量測的服務

3 Service Models of Cloud Computing
雲端運算的三種服務模式

IaaS
Infrastructure as a Service
架構即服務

PaaS
Platform as a Service
平台即服務

SaaS
Software as a Service
軟體即服務

4 Deployment Models of Cloud Computing
雲端運算的四種佈署模型

Public Cloud
公用雲端

amazon web services
Microsoft
Google

Target Market is S.M.B.
主要客戶為中小企業

Dynamic Resource Provisioning between public and private cloud
私有雲端動態根據計算需求調用公用雲端的資源

Hybrid Cloud

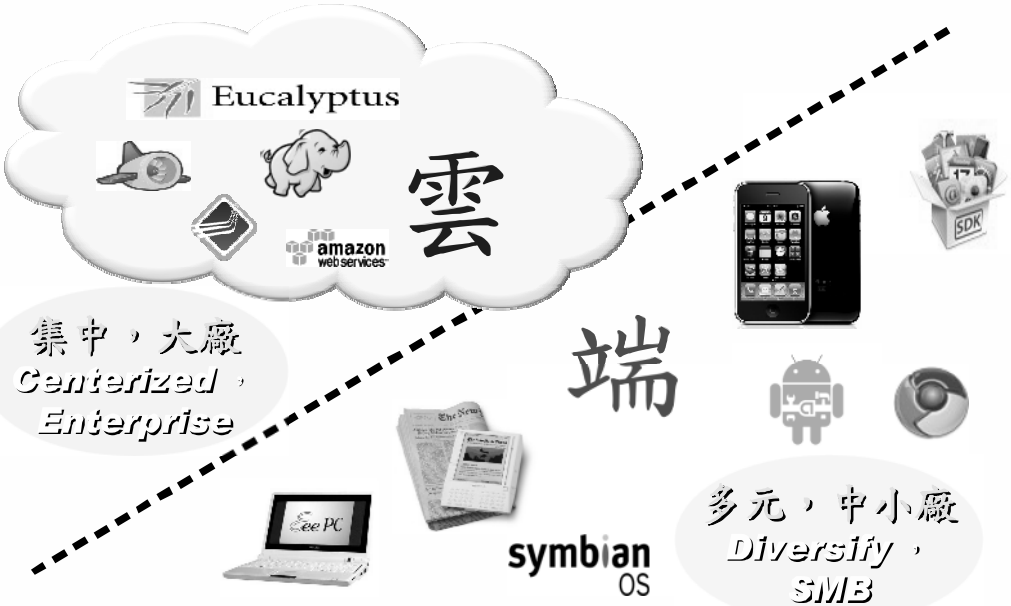
以大型企業為主要客戶
Enterprise is key market

IBM
私有雲端
Private Cloud

Community Cloud
社群雲端

Academia 學術為主

2 R&D directions : Cloud or Device
兩大研究方向：你該選「雲」還是「端」？



One key spirit of Cloud Computing
用一句話說明雲端運算！服務才是王道！

Anytime 隨時

Anywhere 隨地

Key spirit of Cloud ~

形成服務才是重點!!

Everything as a Service !!

With Any Devices 使用任何裝置

Accessing Services 存取各種服務

Cloud Computing =~ Network Computing

雲端運算 =~ 網路運算

Evolution of Cloud Services
雲端服務只是軟體演化史的必然趨勢

數位化 → e-Taiwan 成果系列報導 → 行動台灣三 應用推動計畫 → 發展優質網路社會 (U-Taiwan) 計畫

實體 Physical	單機版 個人使用 Personal Software	網路版 多人共享 Share Service Software	行動版 隨時存取 Mobile Cloud Service
信箱 Mailbox	E-Mail	Web Mail	Mobile Mail
電視 TV	電視盒 Setop Box	Web TV Ex. Youtube	Mobile TV
打字機 Typer Writer	Office	Google Docs	M-Office
電話 Telephone	數位電話 PBX	Skype	Flash Wengo
佈告欄 Bulletin Board	電子佈告欄 BBS	部落格 Blog	微網誌 Twitter

Everything as a Service 啥米鬼都是一種服務

- AaaS Architecture as a Service
- BaaS Business as a Service
- CaaS Computing as a Service
- DaaS Data as a Service
- DBaaS Database as a Service
- EaaS Ethernet as a Service
- FaaS Frameworks as a Service
- GaaS Globalization or Governance as a Service
- HaaS Hardware as a Service
- IMaaS Information as a Service
- **IaaS Infrastructure or Integration as a Service**
- IDaaS Identity as a Service
- LaaS Lending as a Service
- MaaS Mashups as a Service
- OaaS Organization or Operations as a Service
- **SaaS Software or Storage as a Service**
- **PaaS Platform as a Service**
- TaaS Technology or Testing as a Service
- VaaS Voice as a Service

Customer-Oriented

客戶導向，服務至上

能把 AAA 做好就很強了

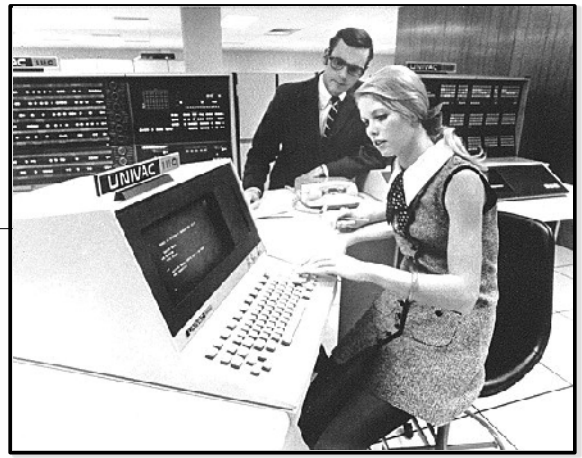
Authentication
Authorization
Accounting
as
a
Service

引用自： https://www.ibm.com/developerworks/mydeveloperworks/blogs/sbose/entry/gathering_clouds_of_xaas

When did the Cloud come ?!

這朵雲幾時飄過來的?!

Brief History of Computing (1/5)



Source: http://pinedakrch.files.wordpress.com/2007/07/

1960 PDP-1

,

,

,

1965 PDP-7

,

,

,

1969 1st Unix

Mainframe
Super
Computer

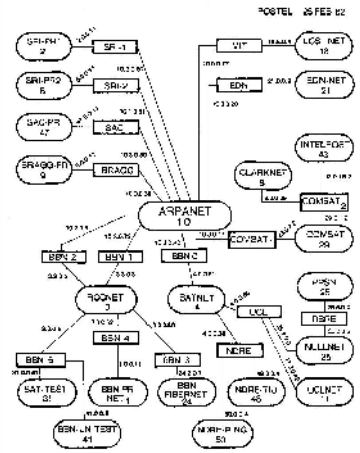
1977 Apple II

1981 IBM 1st PC 5150

Back to Year 1970s ...

1982 TCP/IP

1983 GNU



1991 Linux

Back to Year 1980s ...

Brief History of Computing (2/5)

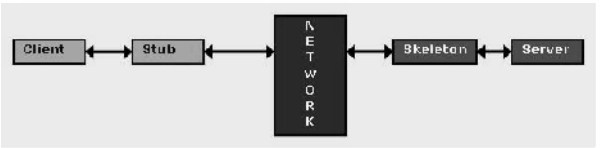
Source: http://www.nchc.org.tw

Mainframe
Super
Computer

PC / Linux
Cluster
Parallel

1990 World Wide Web by CERN
...
...
1993 Web Browser Mosaic by NCSA

1991 CORBA
...
Java RMI
Microsoft DCOM
...
Distributed Objects



Back to Year 1990s ...

1997 Volunteer Computing
1999 SETI@HOME

2003 Globus Toolkit 2

the globus toolkit®
www.globustoolkit.org

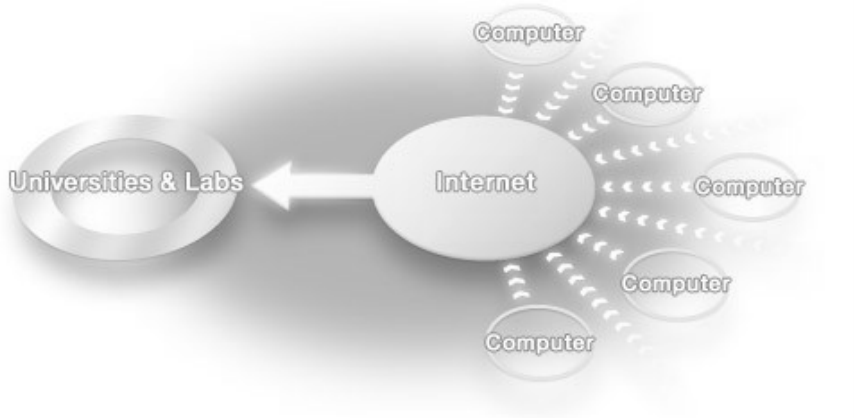
2002 Berkley BOINC

2004 EGEE gLite

Lightweight Middleware for Grid Computing

Back to Year 2000s ...

Brief History of Computing (3/5)



Source: http://www.scei.co.jp/folding/en/dc.html

Mainframe Super Computer → **PC / Linux** Cluster Parallel → **Internet** Distributed Computing

Brief History of Computing (4/5)

Source: http://gridcafe.web.cern.ch/gridcafe/whatisgrid/whatis.html

Mainframe Super Computer → **PC / Linux** Cluster Parallel → **Internet** Distributed Computing → **Virtual Org.** Grid Computing

2001 Autonomic Computing
IBM

2006 Apache Hadoop

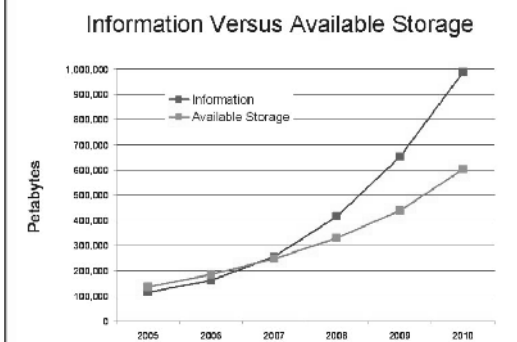
2005 Utility Computing
Amazon EC2 / S3

2007 Cloud Computing
Google + IBM

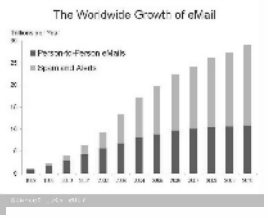
Back to Year 2007 ...

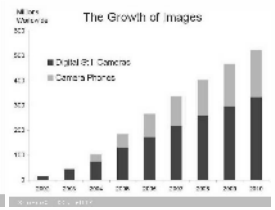
2007 Data Explore

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

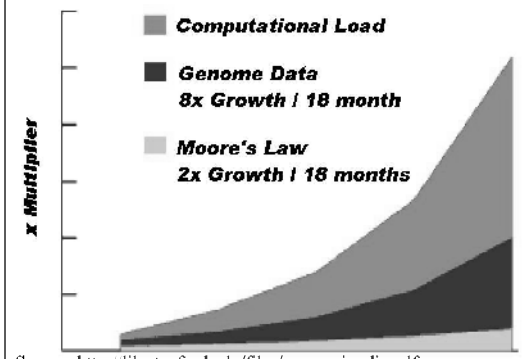


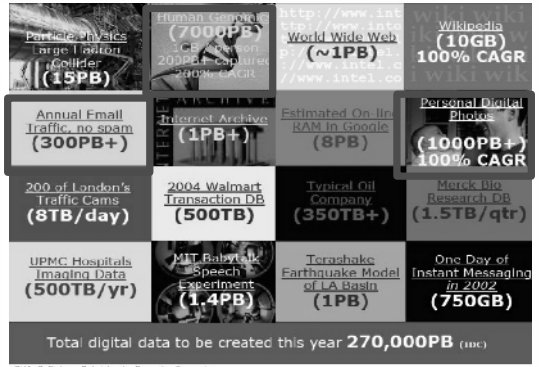
Source: http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf





Source: IDC, 2007

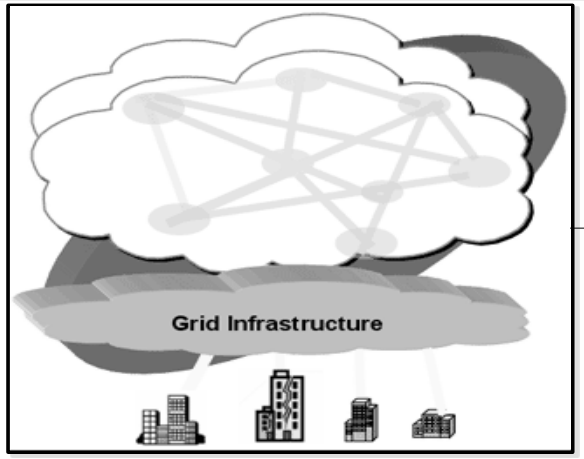




Source: http://lib.stanford.edu/files/see_pasig_die.pdf

Philip B. Gibbons, Data-Intensive Computing Symposium

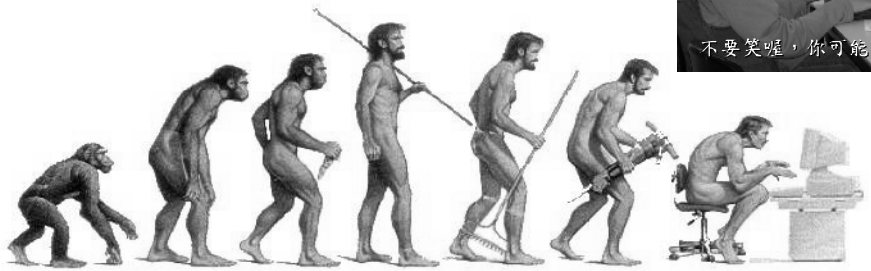
Brief History of Computing (5/5)



Source: http:/ mmdays.com/2008/02/14 cloud-computing

...nframe ...puter	**PC / Linux** Cluster Parallel	**Internet** Distributed Computing	**Virtual Org.** Grid Computing	**Data Explode** Cloud Computing

Evolution



(or is it?)

What can we learn from the past ?!

在這漫長的演化中，我們到底學到些什麼 ?!

Source: http://cyberpingui.free.fr/humour/evolution-white.jpg

Lesson #1: One cluster can't fit all !
教訓一：叢集的單一設定無法滿足所有需求！

Answer #1: Virtual Cluster 新服務：虛擬化叢集

Lesson #2: Grid for Heterogeneous Enterprise !
教訓二：格網運算該用在異業結盟的資源共享！

Answer #2: Peak Usage Time 尖峰用量發生時間點

Lesson #3: Extra cost to move data to Grid !
教訓三：資料搬運的網路與時間成本！

Answer #3: Total Cost of Ownership 總擁有成本

This is why Cloud Computing matters ?!
這就是為什麼雲端運算變得熱門?!

CIO 2010 : Virtualization, Cloud and Web 2.0

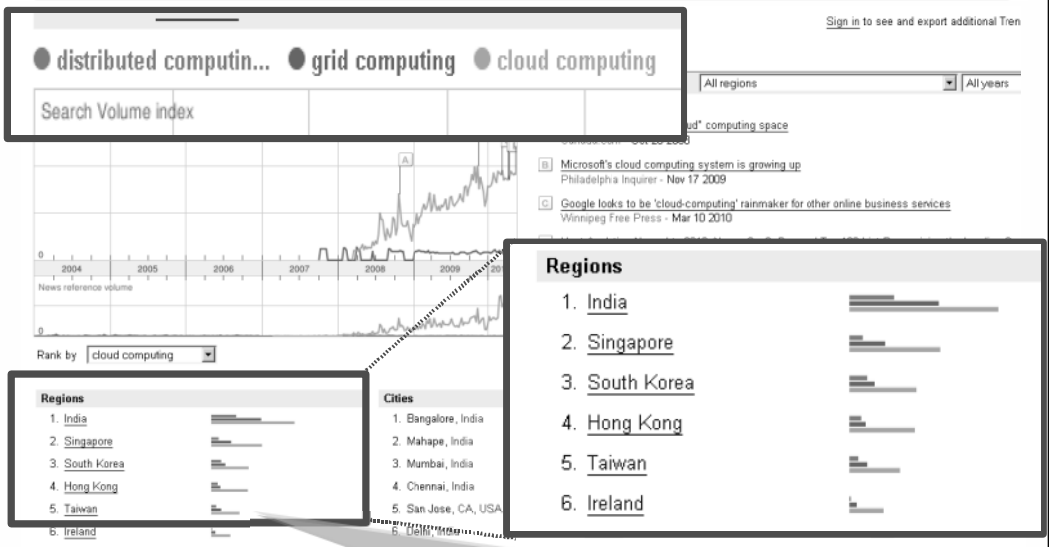
CIO strategic technologies reflect increased interest in "lighter-weight" solutions

CIO technologies	Ranking of technologies CIOs selected as one of their top 5 priorities in 2010				
Ranking	2010		2009	2008	2007
Virtualization	1	↑	3	3	5
Cloud computing	2	↑	16	*	*
Web 2.0	3	↑	15	15	*
Networking, voice and data communications	4	↑	6	7	4
Business intelligence (BI)	5	↓	1	1	1
Mobile technologies	6	↑	12	12	11
Data/document management and storage	7	↑	10	9	9
Service-oriented applications and architecture	8	↑	9	10	7
Security technologies	9	↓	8	5	6
IT management	10		*	*	*
Enterprise applications	11	↓	2	2	2

* New question for that year

Source: *Gartner Executive Programs : " Leading in Times of Transition: The 2010 CIO Agenda "*

What are the trend of next 10 years ?
什麼是下個十年的熱門技能？



似乎亞洲國家特別熱愛雲端?! **Too Hot in Asia ?!**

Trend #1: Data are moving to the Cloud
趨勢一：資料開始回歸集中管理

Access data anywhere anytime 為了隨時存取

Reduce the risk of data lost 降低資料遺失風險

Reduce data transfer cost 減少資料傳輸成本

Enhance team collaboration 促進團隊協同合作

How to store huge data ?!
如何儲存大量資料呢?!

Trend #2: Web become default Platform!
趨勢二：網頁變成預設開發平台

Open Standard 網頁是開放標準

Open Implementation 實作不受壟斷

Cross Platform 瀏覽器成為跨平台載具

Web Application 網頁程式設計成為顯學

Browser difference become entry barrier ?!

瀏覽器的差異造成新的技術門檻?!

該使用別人打造的雲端，還是自己打造專屬雲端呢？

Trend #3: HPC become a new industry
趨勢三：高速計算已悄悄變成新興產業

Parallel Computing 平行運算的技能

Distributed Computing 分散運算的技能

Multi-Core Programming 多核心程式設計

Processing Big Data 處理大資料的技能

Education and Training are needed !!

為了讓這些技能與產業接軌，亟需教育訓練！！

Let's SKIP Public Cloud
公用雲端服務，講過了就跳過啦！！

Public Cloud
公用雲端

amazon web services
Microsoft
Google

Target Market is S.M.B.
主要客戶為中小企業

Hybrid Cloud

以大型企業為主要客戶
Enterprise is key market

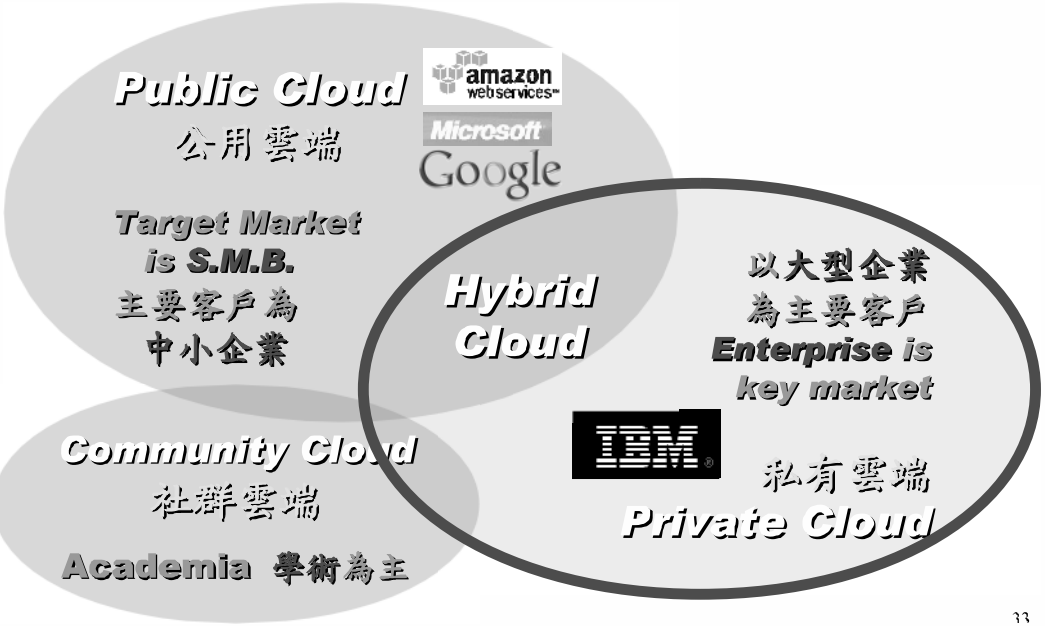
IBM

私有雲端
Private Cloud

Community Cloud
社群雲端

Academia 學術為主

How can we build our Private Cloud ??
那我們如何打造私有雲端呢??



Open Source for Private Cloud
建構私有雲端運算架構的自由軟體

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	

Reference Cloud Architecture
雲端運算的參考架構

應用軟體 Application Social Computing, Enterprise, ISV,...	User-Level
程式語言 Programming Web 2.0 介面, Mashups, Workflows, ...	User-Level Middleware
控制管理 Control Qos Negotiation, Admission Control, Pricing, SLA Management, Metering...	Core Middleware
虛擬化 Virtualization VM, VM management and Deployment	
硬體設施 Hardware Infrastructure: Computer, Storage, Network	System Level

IaaS PaaS SaaS

Building IaaS with Open Source
用自由軟體打造 IaaS 服務

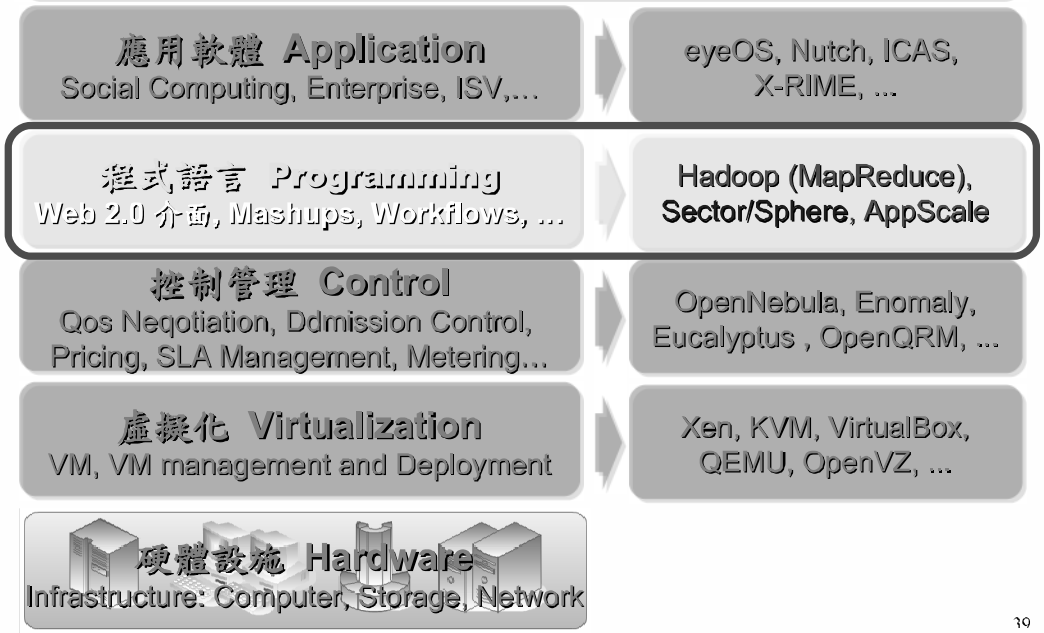
應用軟體 Application Social Computing, Enterprise, ISV,...	eyeOS, Nutch, ICAS, X-RIME, ...
程式語言 Programming Web 2.0 介面, Mashups, Workflows, ...	Hadoop (MapReduce), Sector/Sphere, AppScale
控制管理 Control Qos Negotiation, Admission Control, Pricing, SLA Management, Metering...	OpenNebula, Enomaly, Eucalyptus , OpenQRM, ...
虛擬化 Virtualization VM, VM management and Deployment	Xen, KVM, VirtualBox, QEMU, OpenVZ, ...
硬體設施 Hardware Infrastructure: Computer, Storage, Network	

Open Cloud #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**

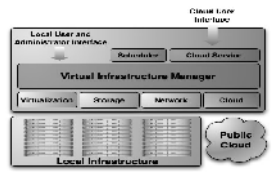
關於 Eucalyptus 的更多資訊，請參考
http://trac.nchc.org.tw grid/wiki/Eucalyptus

Building PaaS with Open Source
用自由軟體打造 PaaS 雲端服務



Open Cloud #2: OpenNebula

- 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



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

Open Cloud #3: 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

Open Cloud #4: 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MalStone效能評比軟體
- **Open Cloud Consortium provide Open Cloud Testbed and develop MalStone toolkit for benchmark**

Sector / Sphere

National Center for Data Mining
University of Illinois at Chicago

UIC

open data

Open Data Group
http://www.opendatagroup.com/

Questions?

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

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

Powered by DRBL

What we learn today ?

WHAT	隨時隨地用任何裝置存取各種服務 !! Accessing services with any device anytime anywhere!!
WHO	亞馬遜、谷歌、微軟等！ 什麼都可以是服務 ~ Amazon, Google, Microsoft and more! Everything as a Service!
WHEN	雲端運算是 2007 年繼格網運算之後的新趨勢 !! Cloud Computing become new trend since year 2007 !!
WHY	資料集中、虛擬化、異業資源共享 Data-intensive, Virtualization, Heterogeneous
HOW	採用自由軟體也能打造私有雲端 Hadoop, Sector/Sphere, Eucalyptus, and more ....

Hadoop 簡介：源起與術語
Introduction to Hadoop : History and Terminology

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

What is Hadoop ?
用一句話解釋 Hadoop 是什麼 ??

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

Hadoop 是一個讓使用者簡易撰寫並執行處理海量資料應用程式的軟體平台。

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

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

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

Founder of Hadoop – Doug Cutting
Hadoop 這套軟體的創辦人 – Doug Cutting

Doug Cutting Talks About The Founding Of Hadoop
http://www.youtube.com/watch?v=qxC4urJOchs

History of Hadoop ... 2002~2004

Hadoop 這套軟體的歷史源起 ... 2002~2004

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

History of Hadoop ... 2002~2004

Hadoop 這套軟體的歷史源起 ... 2002~2004

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

Three Gifts from Google ....

來自 Google 的三個禮物 ....

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

History of Hadoop ... 2004 ~ Now

Hadoop 這套軟體的歷史源起 ... 2004 ~ Now

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

YAHOO! cloudera

Who Use Hadoop ??

有哪些公司在用 Hadoop 這套軟體 ??

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

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

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

Performance improvement of Hadoop

Hadoop 過去幾年的效能改進 (from Yahoo)

年份	日期	節點數	耗時（小時）
2006	四月	188	47.9
2006	五月	500	42
2006	十一月	20	1.8
2006	十一月	100	3.3
2006	十一月	500	5.2
2006	十一月	900	7.8
2007	七月	20	1.2
2007	七月	100	1.3
2007	七月	500	2
2007	七月	900	2.5

Sort benchmark, every nodes with terabytes data.

Hadoop in production run ….

商業運轉中的 Hadoop 應用 ….

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

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

Hadoop in production run ….

商業運轉中的 Hadoop 應用 ….

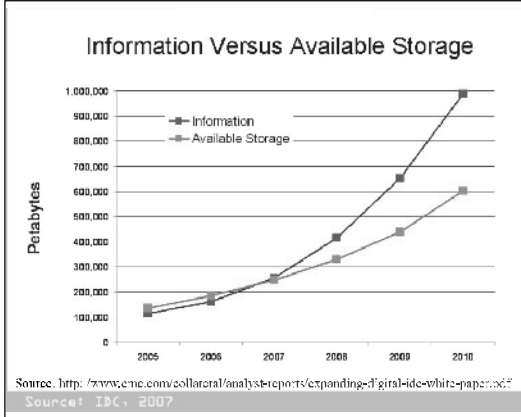
- September 30, 2008
- Scaling Hadoop to 4000 nodes at Yahoo!
- http://developer.yahoo.net/blogs/hadoop/2008/09/scaling_hadoop_to_4000_nodes_a.html

Total Nodes	**4000**
Total cores	**30000**
Data	**16PB**

	500-node cluster		**4000-node cluster**	
	write	**read**	**write**	**read**
number of files	990	990	14,000	14,000
file size (MB)	320	320	360	360
total MB processes	316,800	316,800	5,040,000	5,040,000
tasks per node	2	2	4	4
avg. throughput (MB/s)	**5.8**	**18**	**40**	**66**

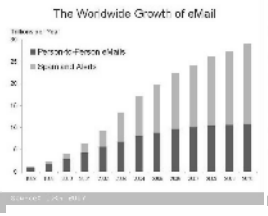
Comparison between Google and Hadoop
Google 與 Hadoop 的比較表

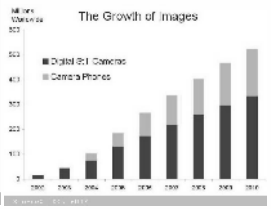
Develop Group	Google	Apache
Sponsor	Google	Yahoo, Amazon
Algorithm Method	MapReduce	MapReduce
Resource	open document	open source
File System (MapReduce)	GFS	HDFS
Storage System (for structure data)	big-table	HBase
Search Engine	Google	Nutch
OS	Linux	Linux / GPL

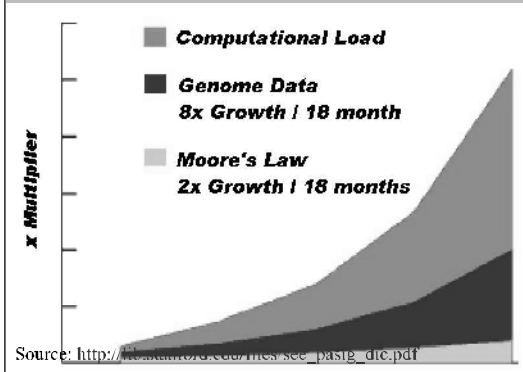


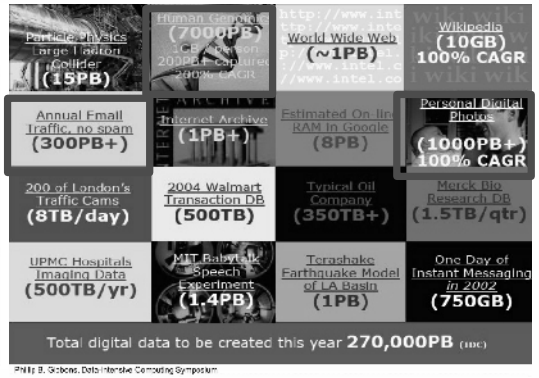
2007 Data Explore

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









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

1. Data Explore
資訊大爆炸

2. Data Mining Tool
方便作資料探勘的工作

3. Looking for Jobs
好找工作 !!

Hadoop 專業術語
Introduction to Hadoop Terminology

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

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

Scheduler
程序排程

File System
檔案系統

Terminologies of Hadoop
Hadoop 文件中的專業術語

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

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

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

名稱節點 NameNode

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

資料節點 DataNode

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

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

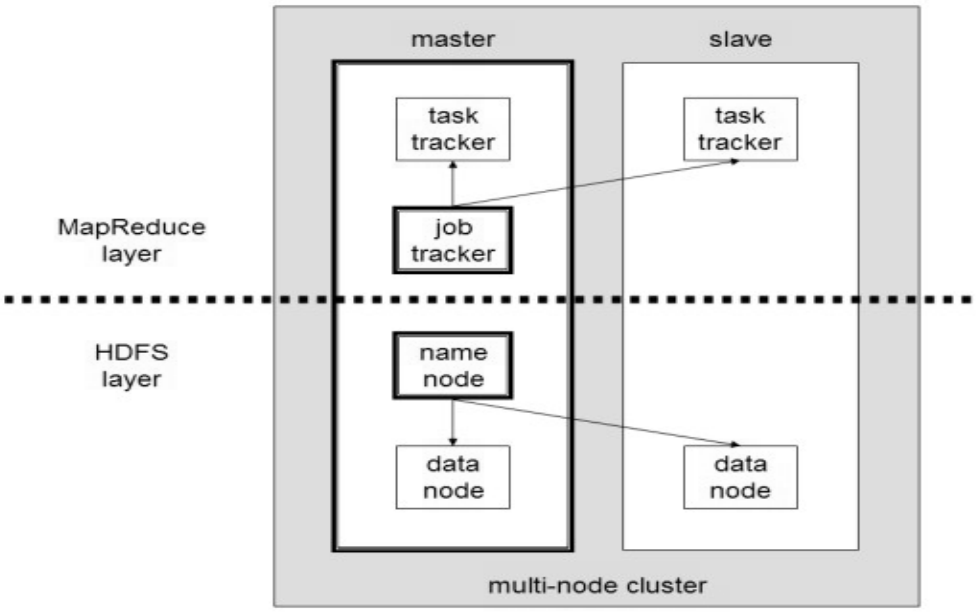
JobTracker

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

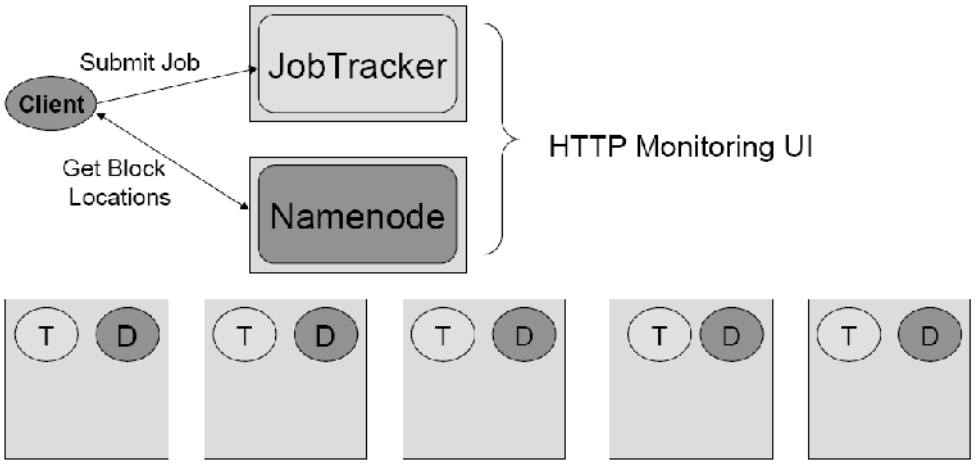
TaskTracker

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

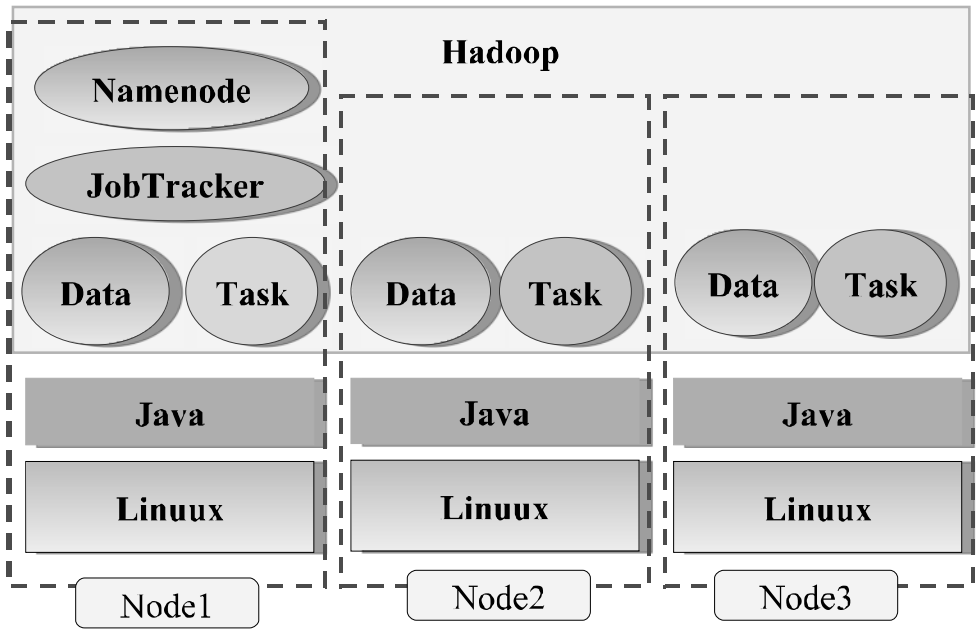
Different Roles of Hadoop Architecture
Hadoop 軟體架構中的不同角色



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



Distributed Operating System of Hadoop
Hadoop 建構成一個分散式作業系統



What we learn today ?

WHAT	Hadoop 是運算海量資料的軟體平台 !! hadoop is a software platform to process vast amount of data!!
WHO	始祖是 Doug Cutting，Apache 社群支持，Yahoo 贊助 From Doug Cutting to Apache Community, Yahoo and more !
WHEN	Hadoop 是 2004 年從 Nutch 分裂出來的專案 !! Hadoop became separate project since year 2004 !!
WHY	資料大爆炸、資料探勘、找工作 *Data Explore, Data Mining, Jobs !!*
HOW	採用自由軟體也能打造私有雲端 Install on large clusters built of commodity hardware !!

Questions?

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

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