運用自由軟體打造生物資訊雲端服務

Building Bioinformatics Cloud Service using Open Source

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

Powered by DRBL

PART 1 :

淺談雲端運算的定義與驅動力

Part 1 : Overview of Cloud Computing

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

Powered by DRBL

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

1. On-demand self-service.
隨需自助服務

2. Broad network access
隨時隨地用任何網路裝置存取

3. Resource pooling
多人共享資源池

4. Rapid elasticity
快速重新佈署靈活度

5. Measured Service
可被監控與量測的服務

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

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

Public Cloud
公用雲端

amazon web services™
Microsoft
Google

Target Market is S.M.B.
主要客户爲中小企業

Hybrid Cloud

以大型企業爲主要客户
Enterprise is key market

IBM

私有雲端
Private Cloud

Community Cloud
社群雲端

Academia 學術爲主

3 Service Models of Cloud Computing
雲端運算的三種服務模式（市場區隔）

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

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

集中，大廠

Centerized，
Enterprise

端

symbian
OS

多元，中小廠

Diversify，
SMB

1 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

With Any Devices 使用任何裝置

Accessing Services 存取各種服務

Key spirit of Cloud ~

形成服務才是重點！！

Everything as a Service !!

Cloud Computing =~ Network Computing

雲端運算 =~ 網路運算

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

Source: http://www.cnet.co.uk/i/c/blg/cat/software/cloudcomputing/clouds1.jpg

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

e-Taiwan 成果系列報導 → 行動台灣 應用推動計畫 → 發展優質網路社會 (U-Taiwan) 計畫 → 智慧台灣 領航系列報導

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 Bullet Borad	電子佈告欄 BBS	部落格 Blog	微網誌 Twitter

Brief History of Computing (1/5)

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

1960 PDP-1

.

.

.

1965 PDP-7

.

.

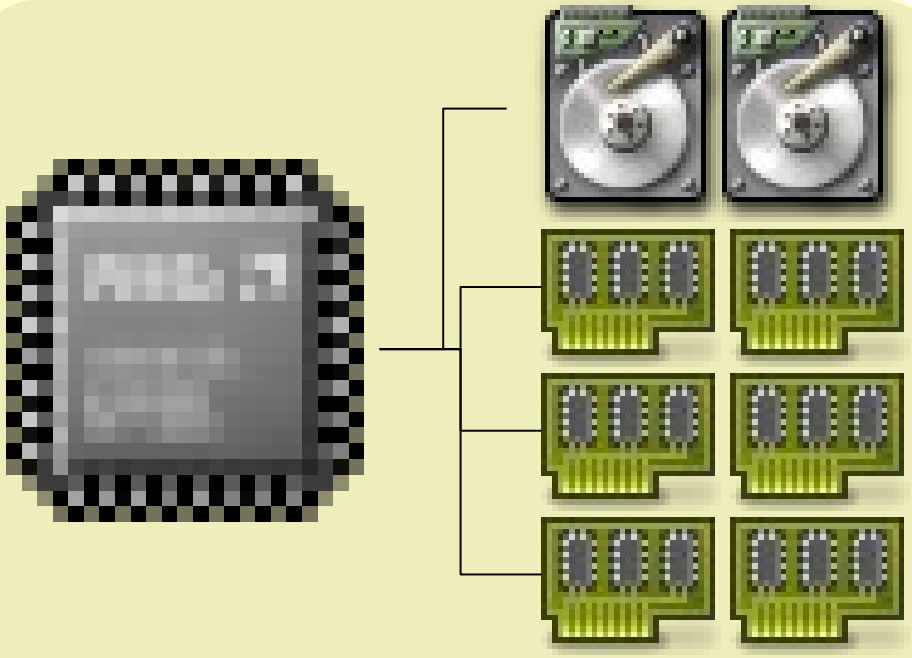
.

1969 1st Unix

Mainframe

Super Computer

Evolution of Computing Architecture (1/5)



Multiple Users

Single CPU

Shared Memory

One Admin.

Single Super Computer

Mainframe Super Computer

使用者心裡的『謎之聲』(1/5)

等執行程式，要排隊排好久喔~

可惡，程式又死掉了，又得重排一次

眞希望自己有一台電腦可以跑!!

超級電腦是有錢人才玩得起的玩具~

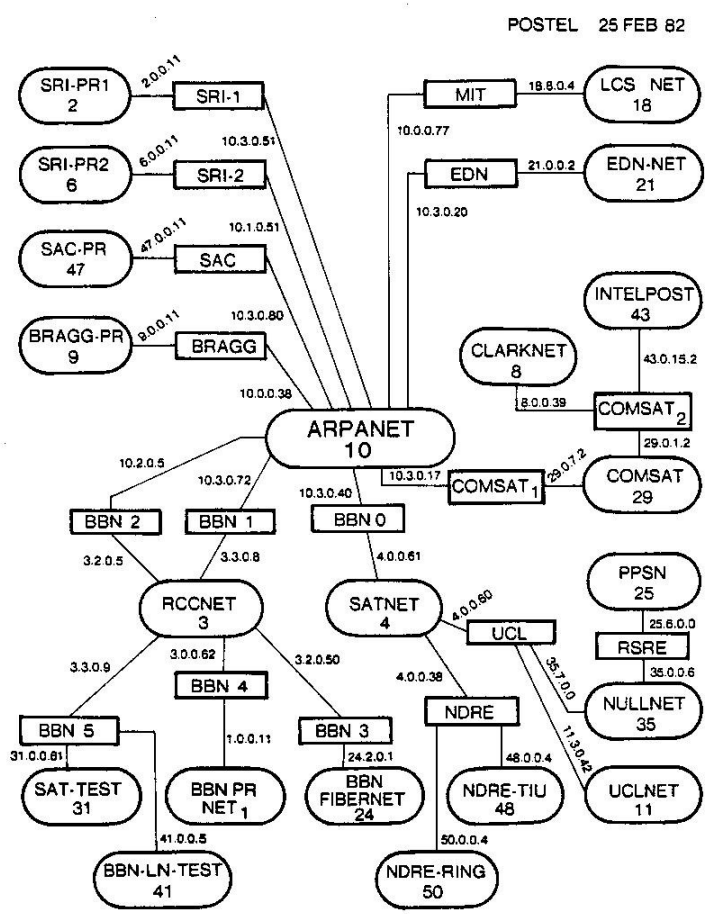
1977 Apple II

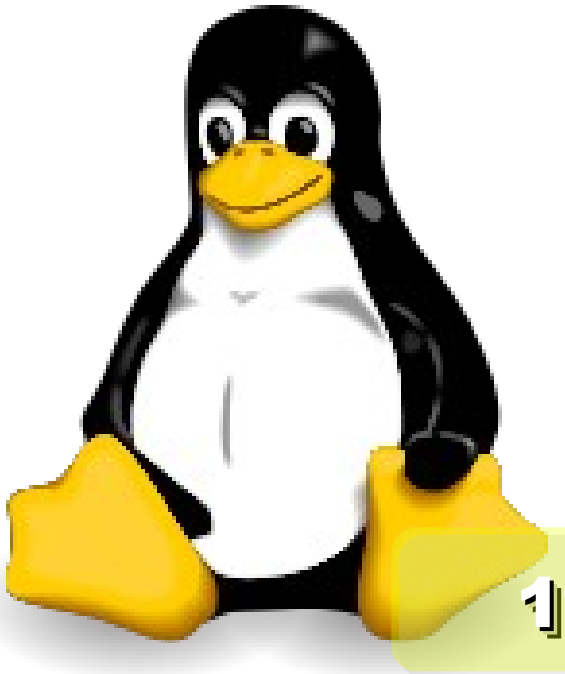
1981 IBM 1st PC 5150

Back to Year 1970s ...

1982 TCP/IP

1983 GNU





1991 Linux

Back to Year 1980s ...

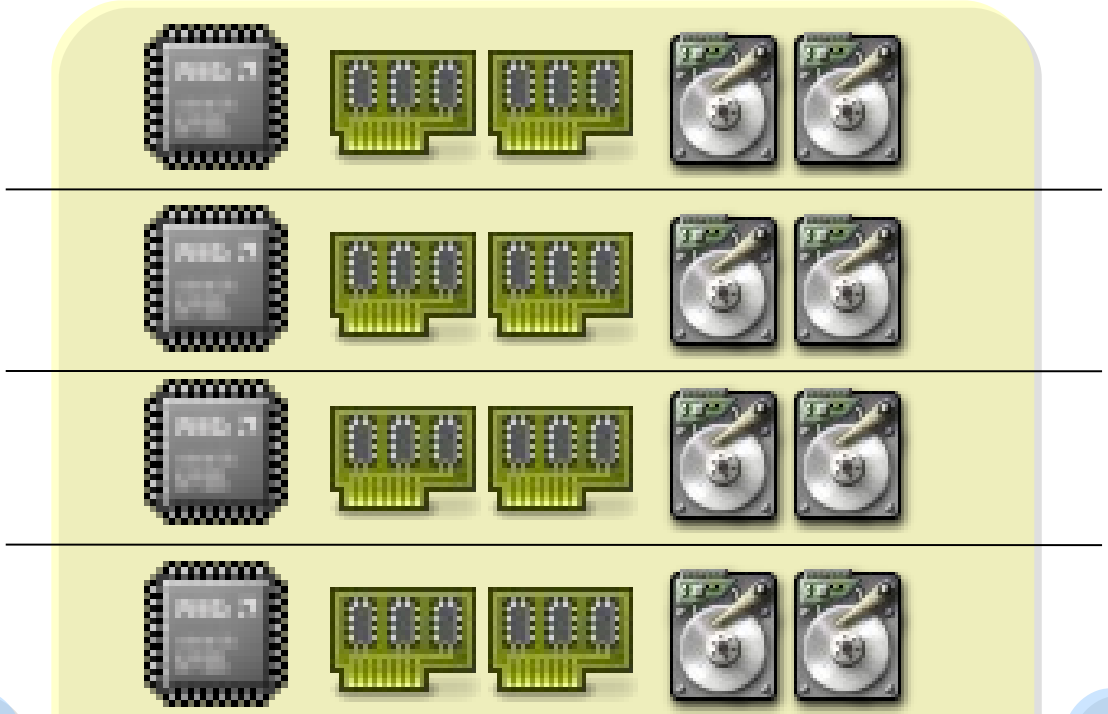
Brief History of Computing (2/5)

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

Mainframe
Super
Computer

PC / Linux
Cluster
Parallel

Evolution of Computing Architecture (2/5)



Multiple Users

Separate CPU

Separate Memory

One Admin.

Multiple PC in One Location

ame
er
uter

PC / Linux
Cluster
Parallel

使用者心裡的『謎之聲』(2/5)

可惡，記憶體不夠大，程式又死掉了

奇怪，我的程式為什麼不能跑？

眞希望自己有一組叢集可以跑!!

管理員老大，可以幫我裝 LiBT 嗎？

1990 World Wide Web
by CERN

....

....

1993 Web Browser
Mosaic by NCSA

1991 CORBA

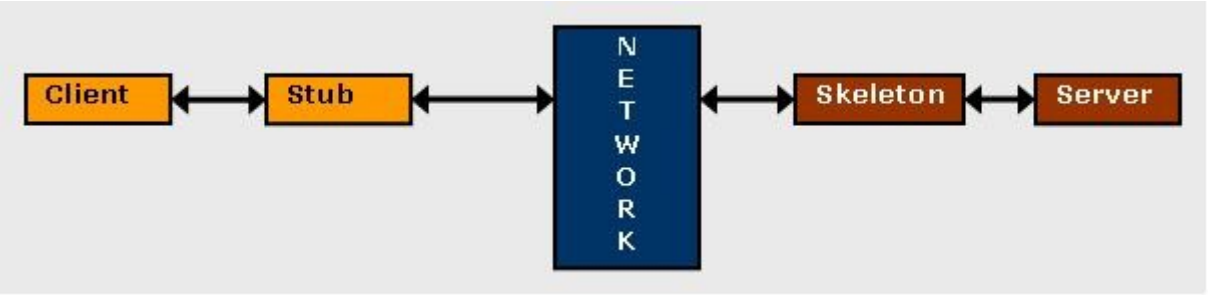
...

Java RMI

Microsoft DCOM

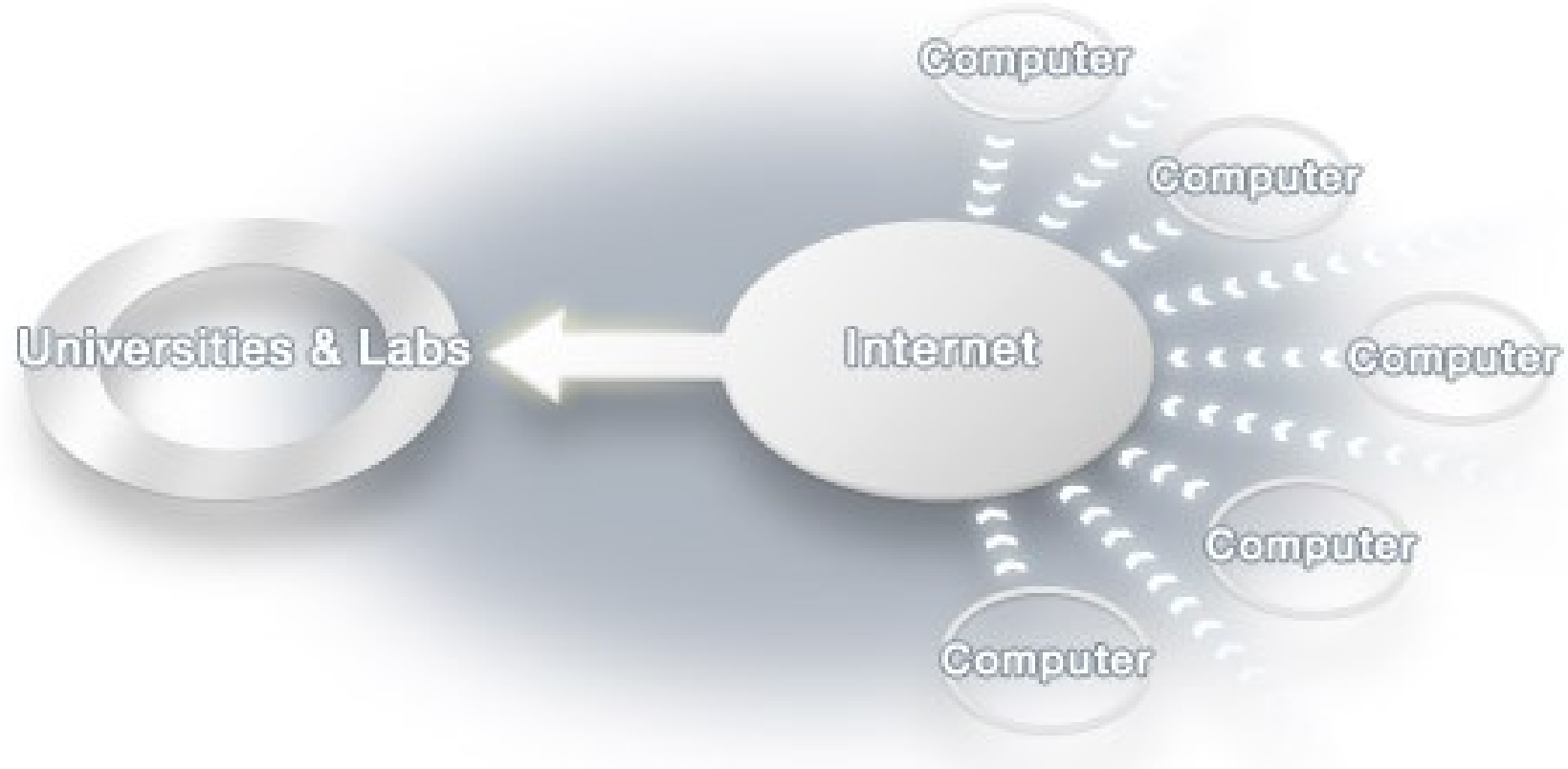
...

Distributed Objects



Back to Year 1990s ...

Brief History of Computing (3/5)



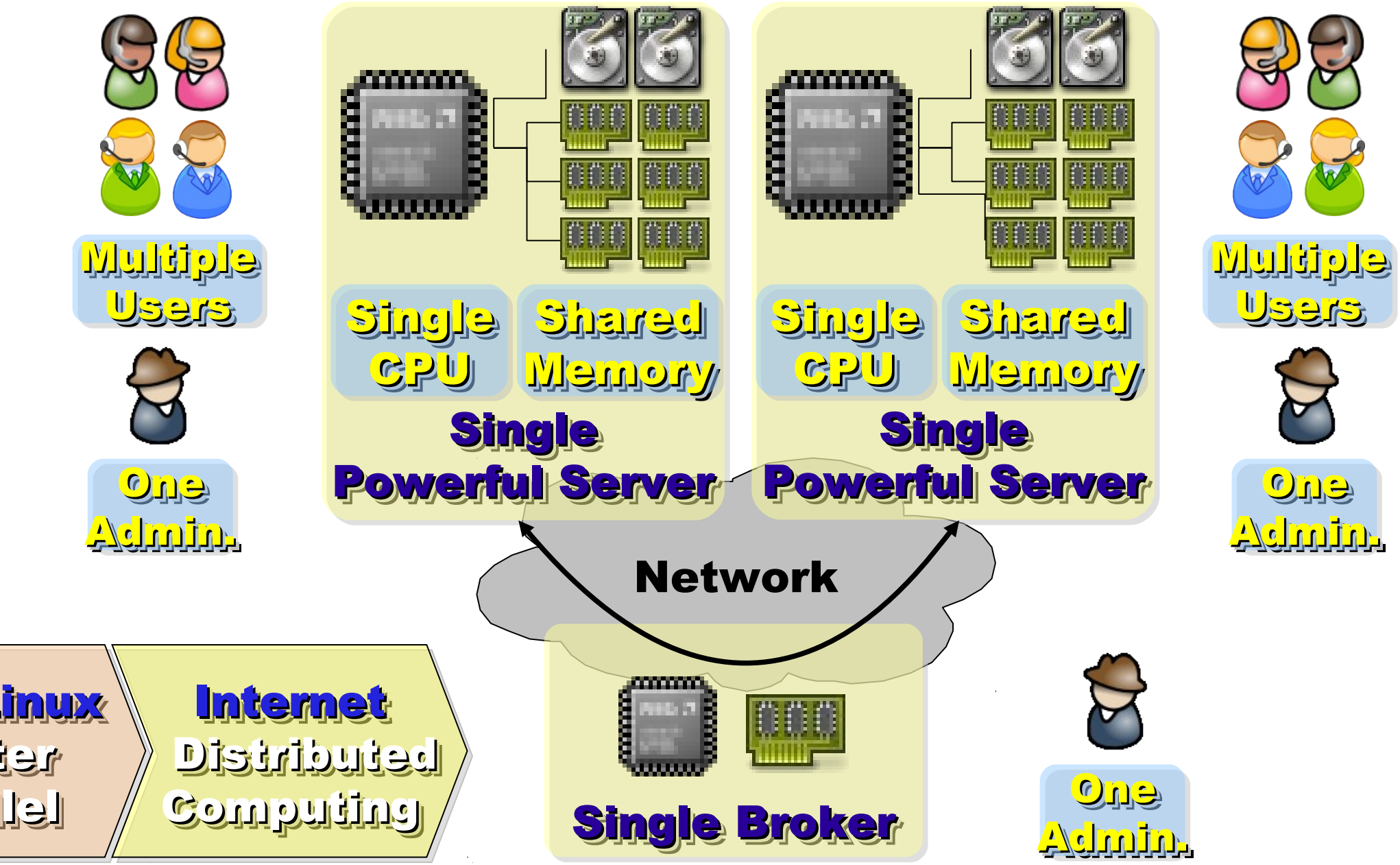
Source: http://www.scei.co.jp/folding/en/dc.html

Mainframe
Super
Computer

PC / Linux
Cluster
Parallel

Internet
Distributed
Computing

Evolution of Computing Architecture (3/5)



使用者心裡的『謎之聲』(3/5)

分散式物件怎麼這麼抽象啊～XD

啊！網路斷線了～不能動了～

大家把閒置電腦都貢獻出來吧!!

給我網路遊戲，其餘免談！

1997 Volunteer Computing
1999 SETI@HOME

2003 Globus Toolkit 2

SETI HOME

the globus toolkit®
www.globustoolkit.org

2002 Berkley BOINC

2004 EGEE gLite

BOINC!

gLite
Lightweight Middleware for Grid Computing

Back to Year 2000s ...

Brief History of Computing (4/5)

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

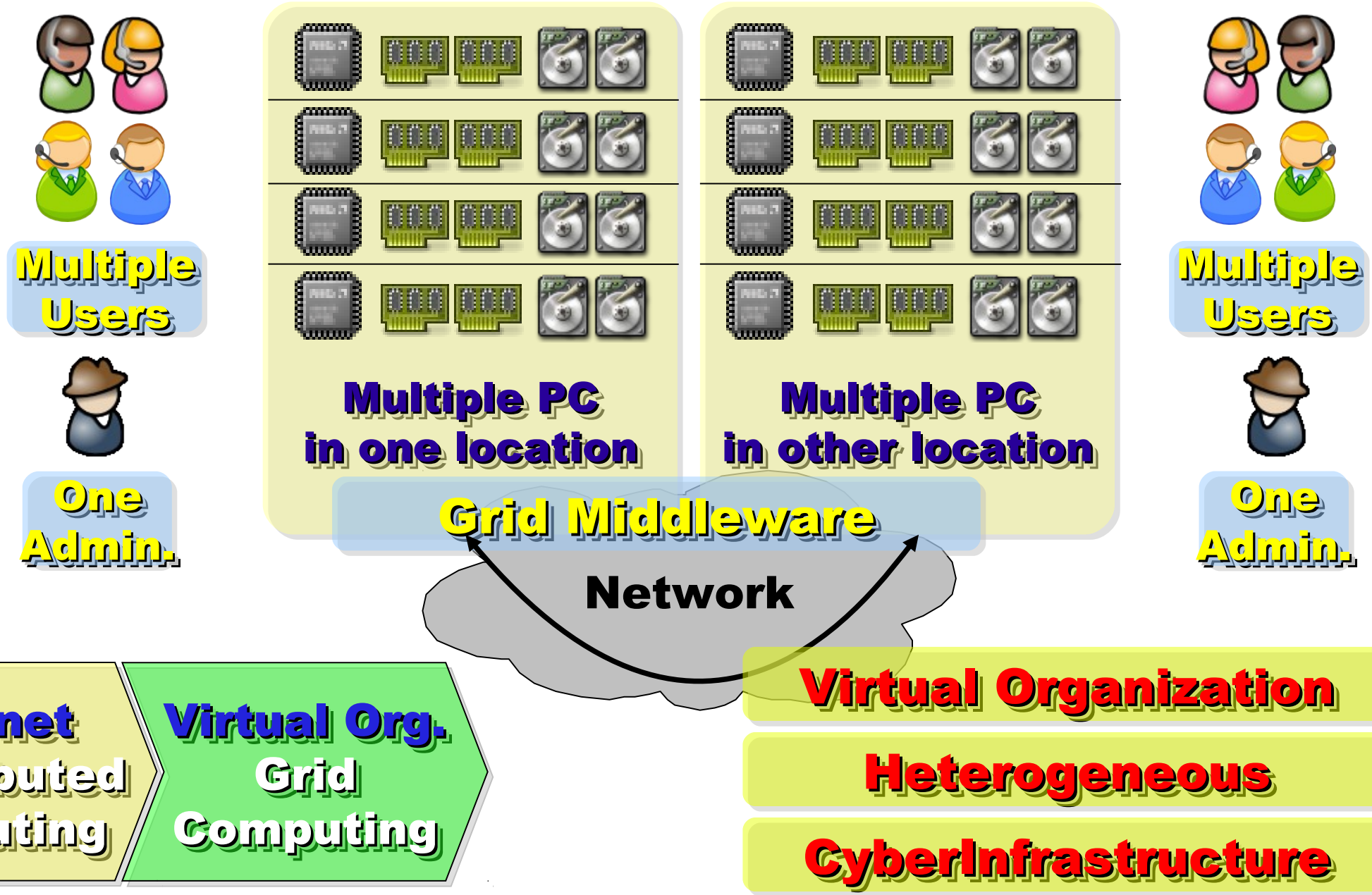
Mainframe
Super
Computer

PC / Linux
Cluster
Parallel

Internet
Distributed
Computing

Virtual Org.
Grid
Computing

Evolution of Computing Architecture (4/5)



使用者心裡的『謎之聲』(4/5)

已給我認證了，爲什麼要不到資源？

啥？可用資源在美國，慢慢搬檔案吧！

爲什麼人家Google那麼會算?!

長官，請幫我們去談好資源共享政策吧！

2001 Autonomic Computing
IBM

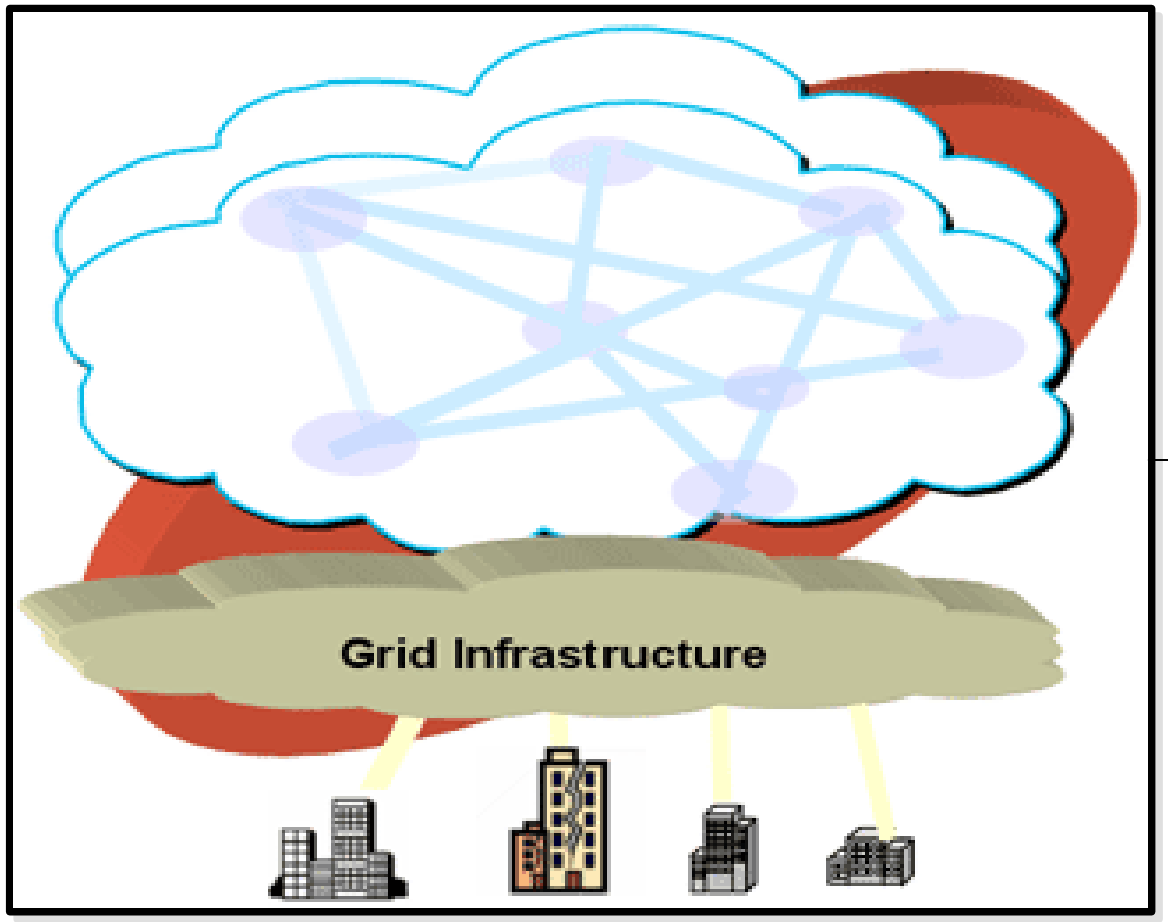
2006 Apache Hadoop

2005 Utility Computing
Amazon EC2 / S3

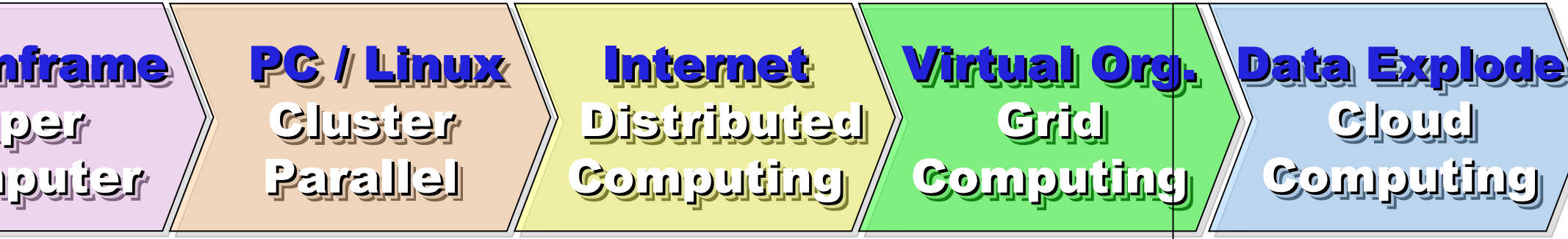
2007 Cloud Computing
Google + IBM

Back to Year 2007 ...

Brief History of Computing (5/5)



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



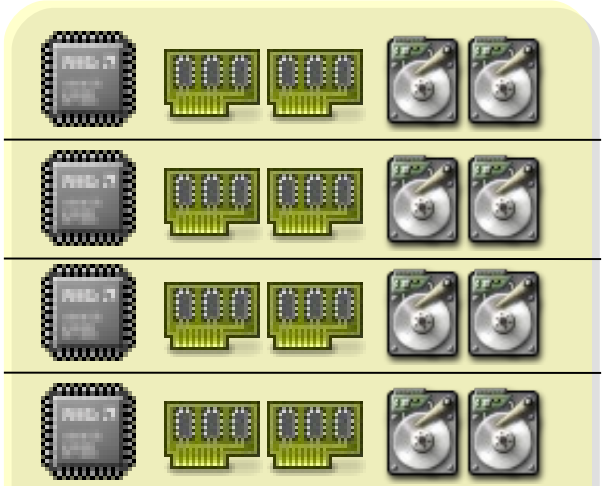
Evolution of Computing Architecture (5/5)

Each User

||

Virtual Admin.

Access any time, any where with mobile device



Multiple PC in different locations

Multiple Admin.

Virtual World

Physical World

tual Org.
Grid
mputing

Data Explode
Cloud
Computing

What is NEXT ?!

Mobile Computing ?!

使用者心裡的『謎之聲』(5/5)

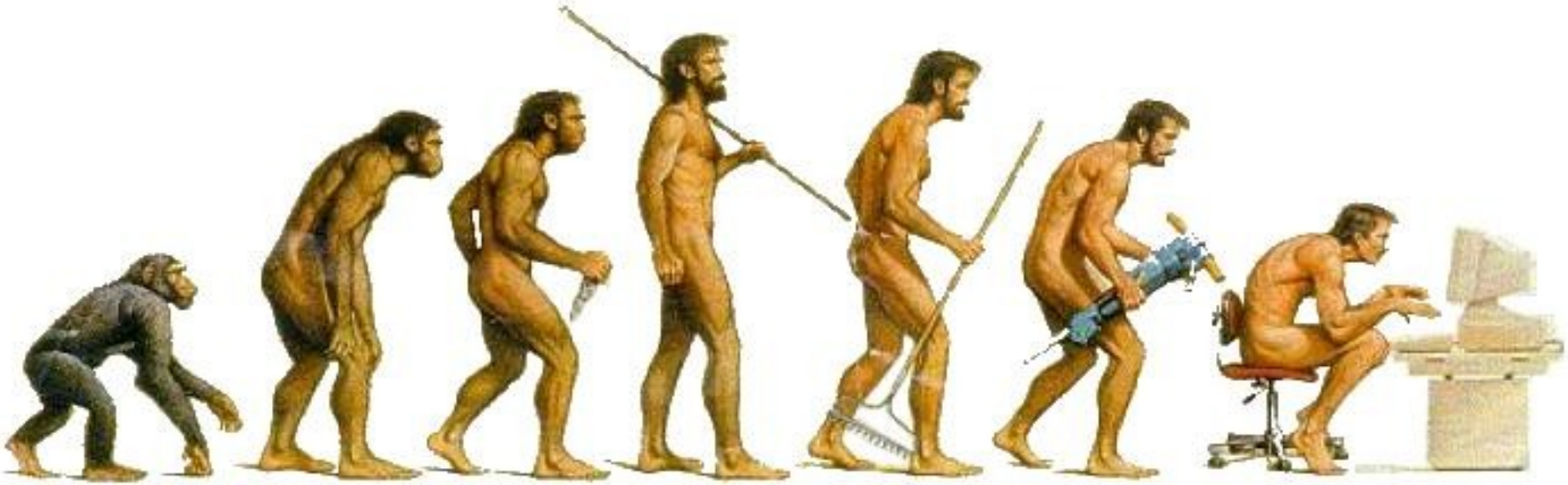
雲端運算合適我用嗎？

按使用時間計費，真的比較省？

Google 到底有沒有偷窺我的信？!

我們自己可以架雲端運算的環境嗎？

Evolution



(or is it?)

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

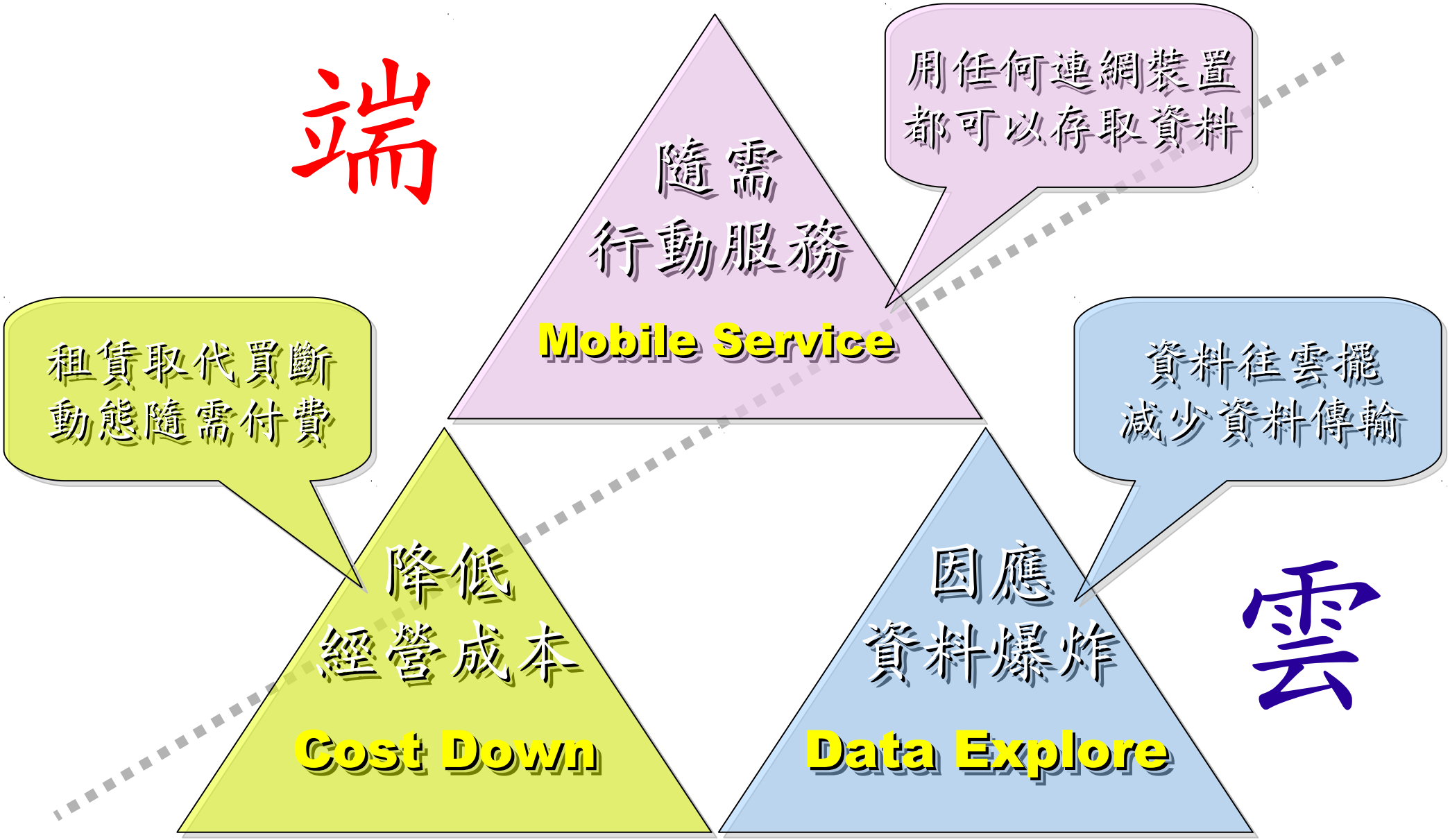
Lesson #3: Extra cost to move data to Grid !
教訓三：資料搬運的網路與時間成本！

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

Cost Down is the Key Drive !!

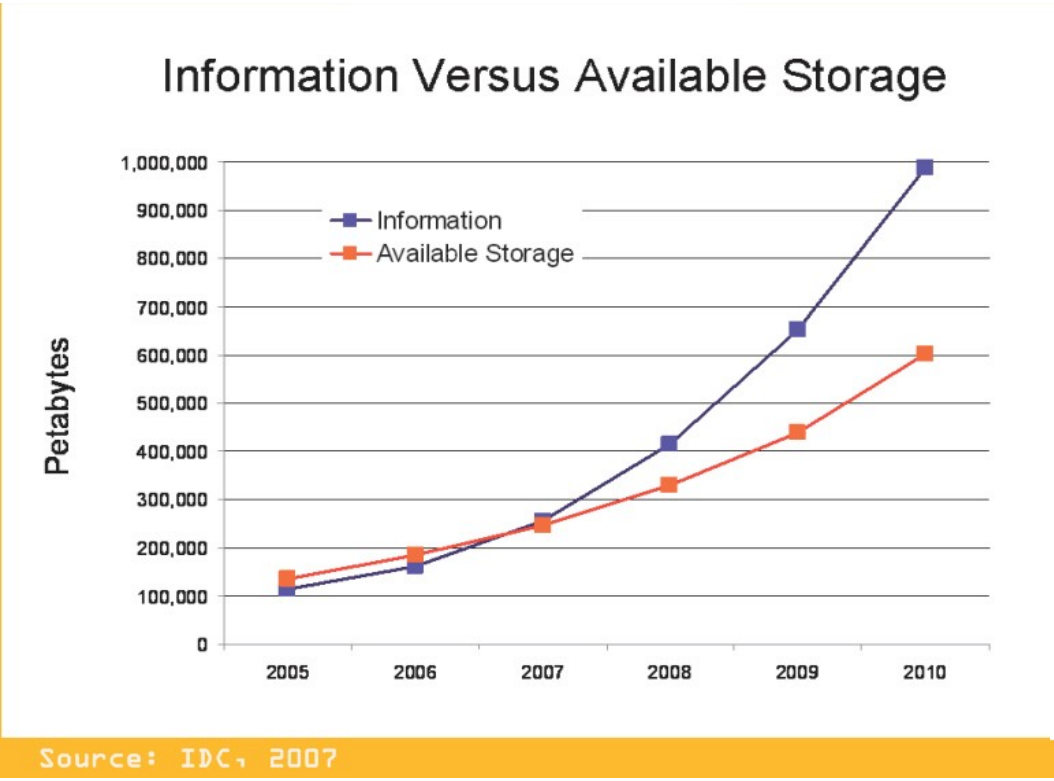
降低營運成本才是企業導入雲端運算的關鍵考量！！

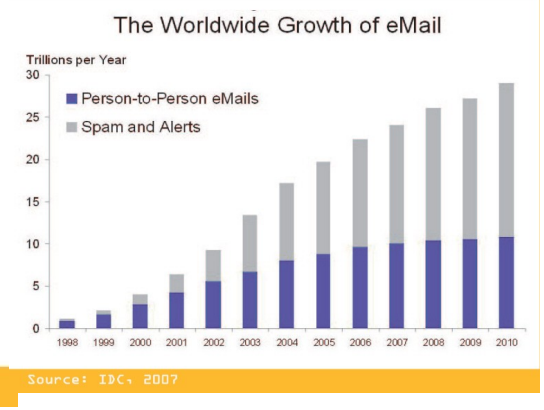
Key Driving Forces of Cloud Computing
雲端運算的關鍵驅動力

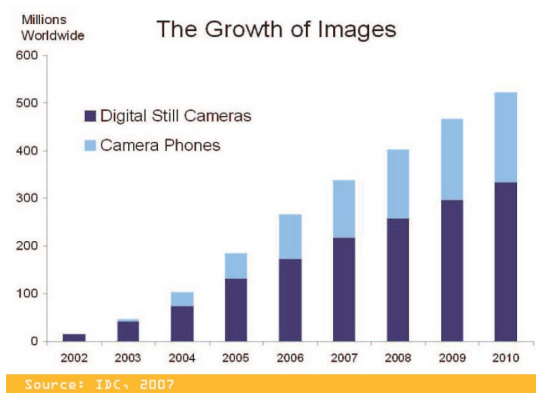


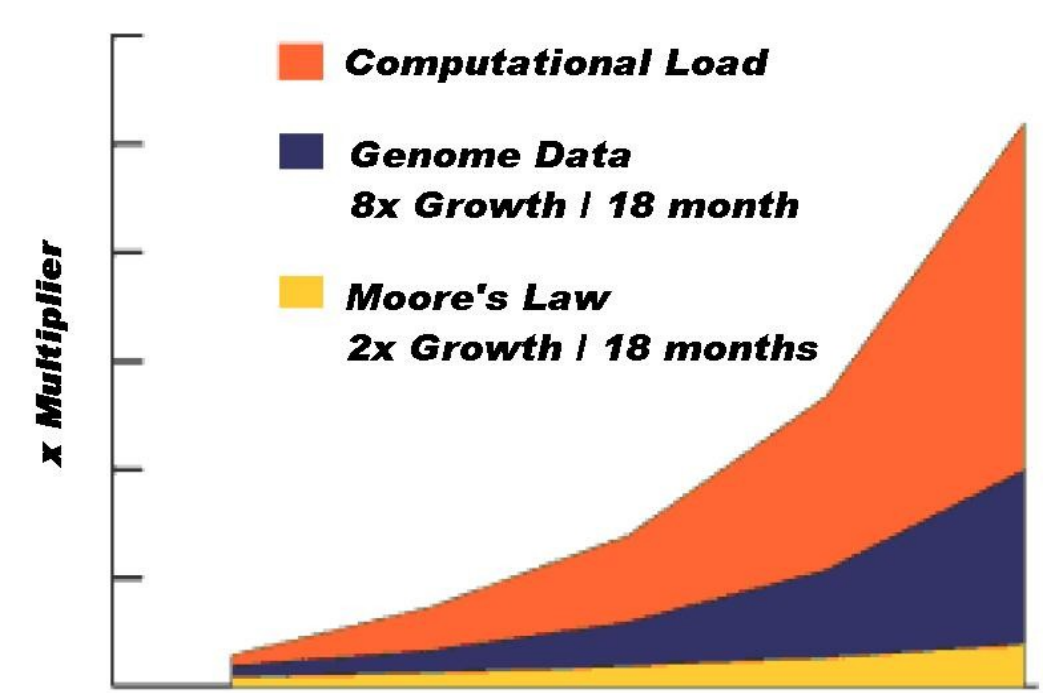
2007 Data Explore

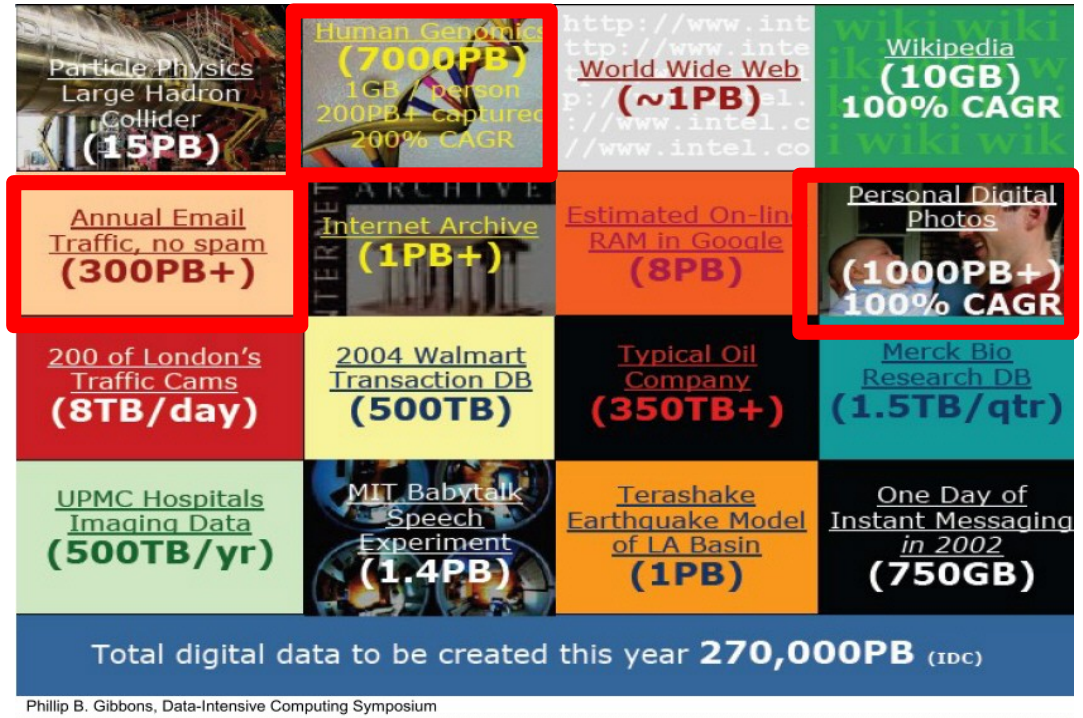
Top 1 : Human Genomics – 7000 PB / Year
Top 2 : Digital Photos – 1000 PB+/ Year
Top 3 : E-mail (no Spam) – 300 PB+ / Year







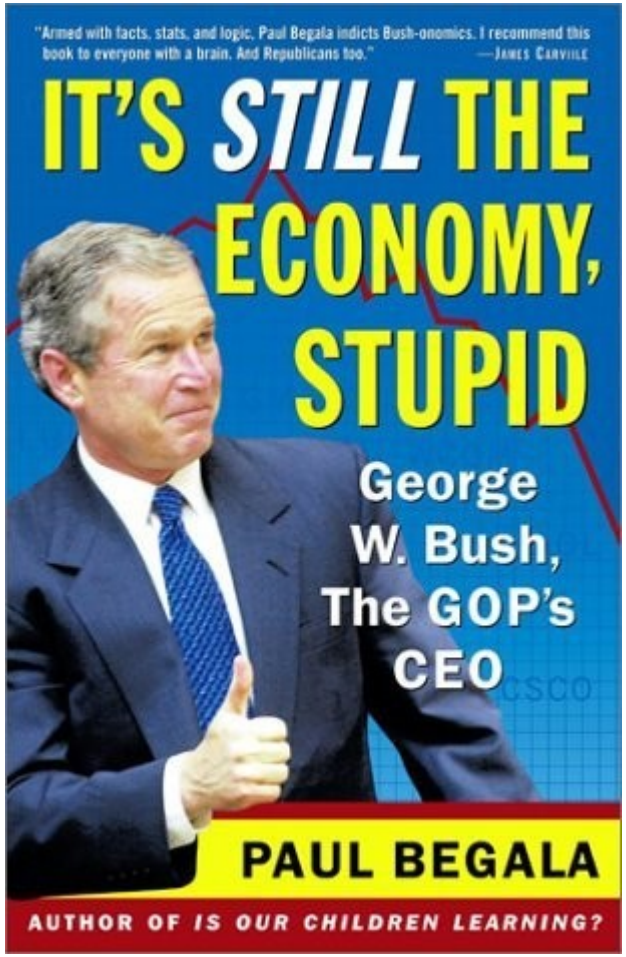




Phillip B. Gibbons, Data-Intensive Computing Symposium

Source: http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf
Source: http://lib.stanford.edu/files/see_pasig_dic.pdf

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

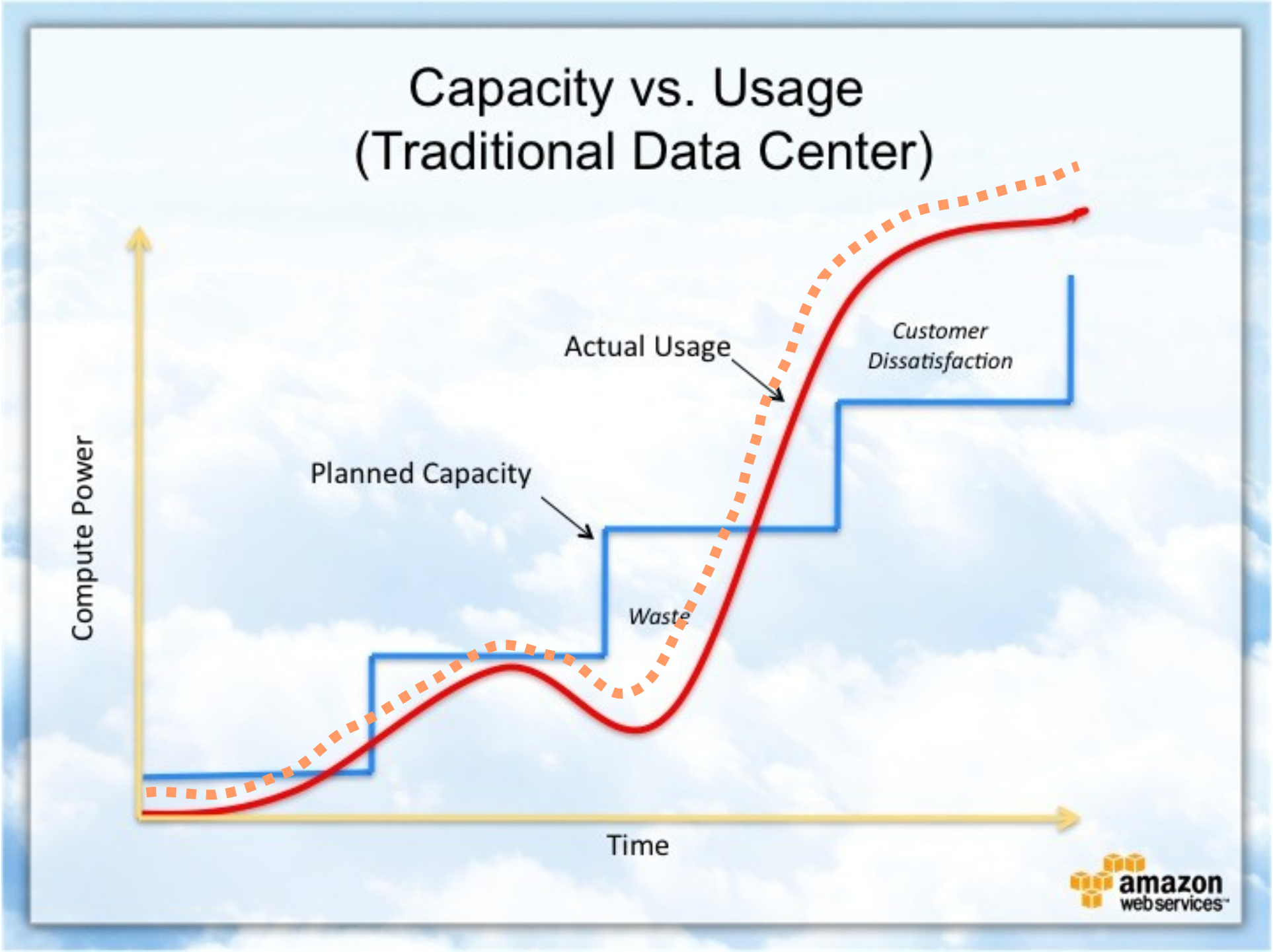
雲端時代，谷歌會說：「笨蛋！重點在資料」

（**"It's the data, stupid"**）

誰掌握了你的資料，就有機會掌握你的荷包

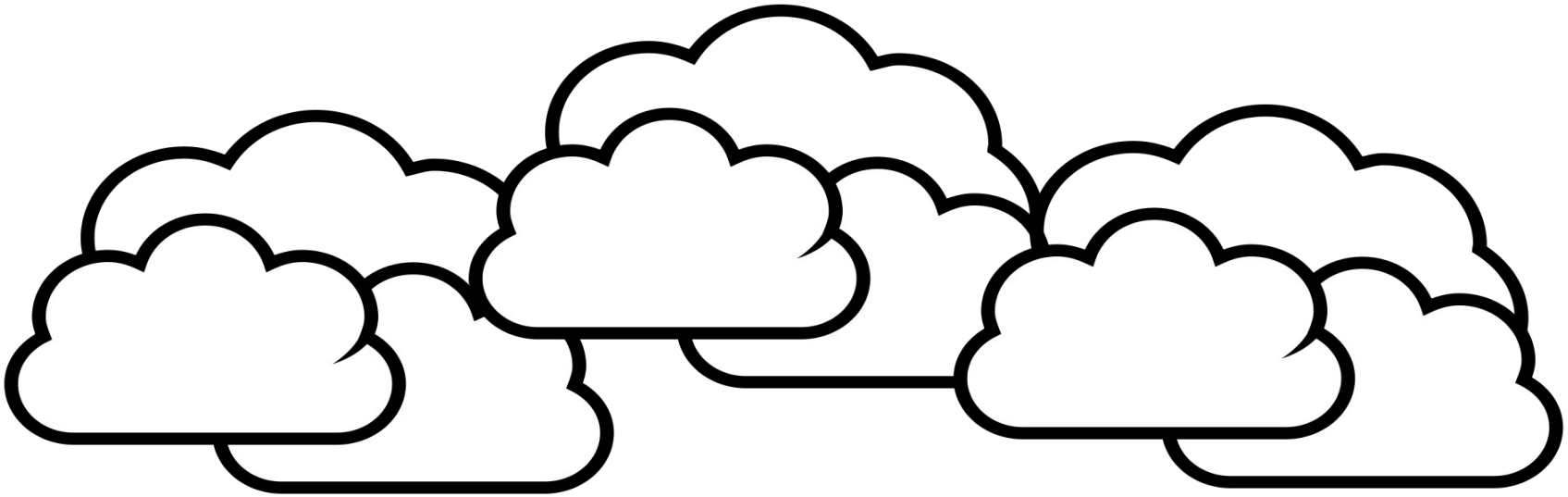
想想看，電腦、手機掉了，您心疼的是甚麼呢？

- 2007 年



Source : http://awsmedia.s3.amazonaws.com/chart01_traditional_720x540.jpg

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

PART 2 :

從叢集到雲端的生物資訊應用範例

Part 2 : Open Source Application for Bioinformatics

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

Powered by DRBL

自由軟體 (Free Software)

站在巨人的肩膀上，是自由軟體發展的理念。其靈活、可自由複製、分享的價值，將有效解決資訊教育的管理成本及商業軟體高成本負擔的問題。

Open Source is your Friend !!

- Open Bioinformatics Foundation - http://www.bioinfomatics.org
  - BioPerl - http://bio.perl.org
  - BioPython - http://biopython.org
  - BioPHP - http://biophp.org
  - BioJava - http://biojava.org
- C++ Bio Sequence Library
  - http://libseq.sourceforge.net/
  - C++ 版本的序列分析函式庫
- Bio-SPICE - http://biospice.sourceforge.net/
- BioEra - http://bioera.net/
  - 跟腦科學有蠻強的關聯性，主要功能是在做訊號處理。
- NCBI Viewer - http://ncbiviewer.bravehost.com/

BLAST (Basic Local Alignment Search Tool)

- http://blast.ncbi.nlm.nih.gov/
- National Center for Biotechnology Information
- BLAST is an algorithm for comparing primary biological sequence information. (BLAST 用來比對生物序列的主要結構)
  - the amino-acid sequences of different proteins
  - the nucleotides of DNA sequences

  (例如：不同蛋白質的氨基酸序列 DNA 序列的核甘酸)
- 用途：搜尋其他物種 (如：老鼠) 未知基因，是否也存在人類基因中
- 優點：使用啓發式搜索來找出相關的序列，比動態規劃快上 50 倍。
- 缺點：不能夠保證搜尋到的序列和所要找的序列之間的相關性。
- 技術問題：巨大的序列資料庫需要進行比對，怎樣計算才快？

- Source: http://zh.wikipedia.org/w/index.php?title=BLAST_(生物資訊學)&variant=zh-tw

At First, We have " 4 + 1 " PC Cluster

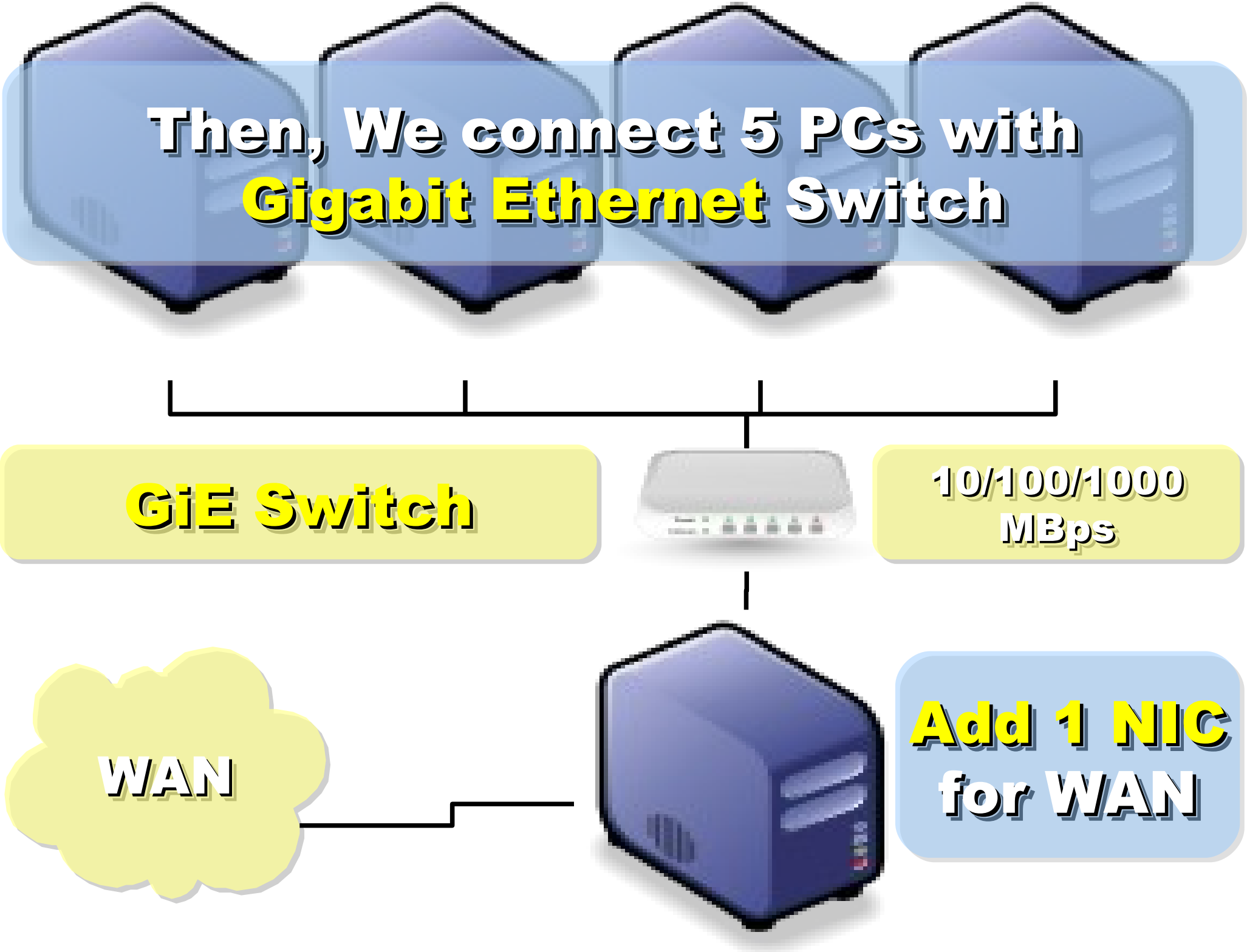
It'd better be

2^n

Manage

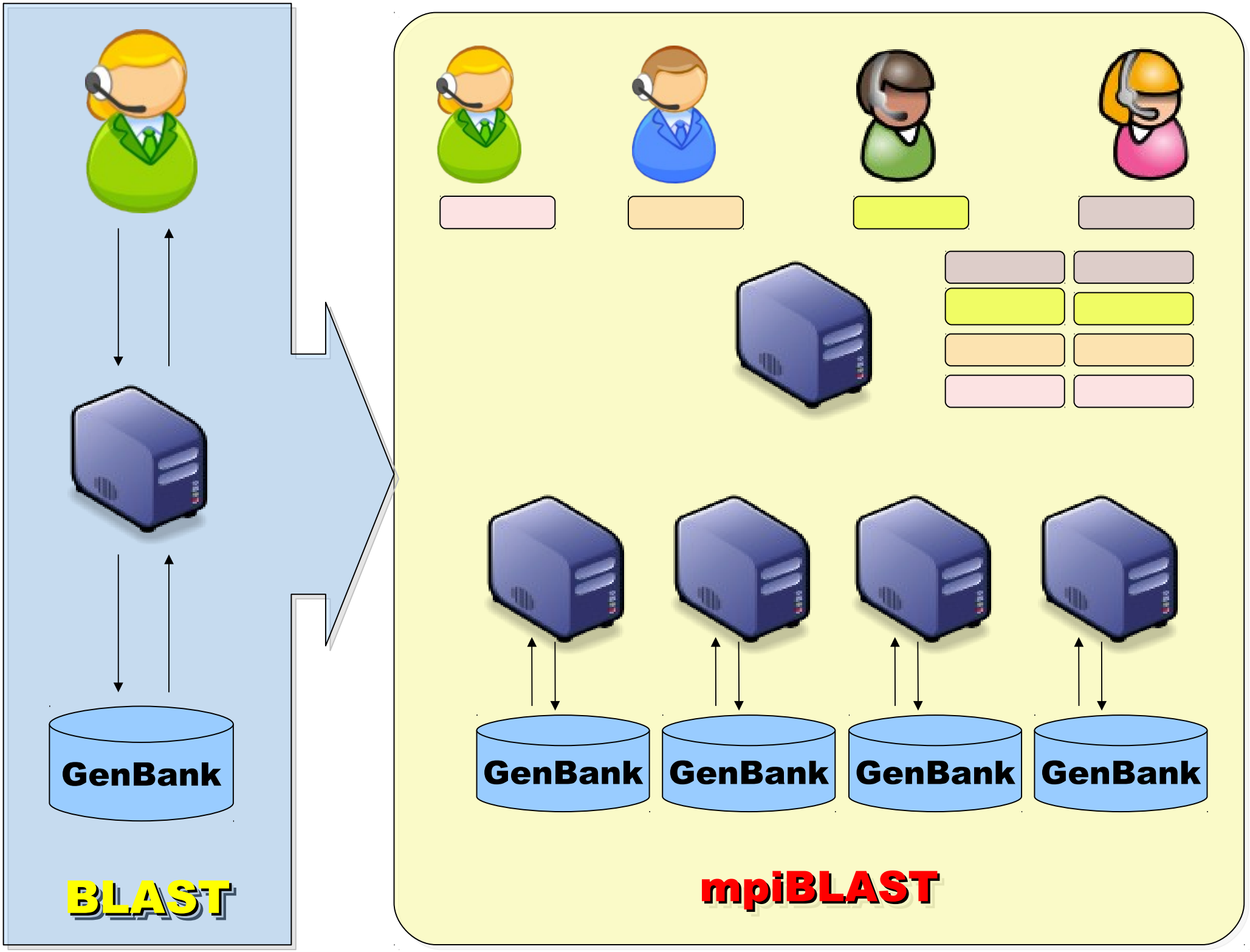
Scheduler

Then, We connect 5 PCs with Gigabit Ethernet Switch

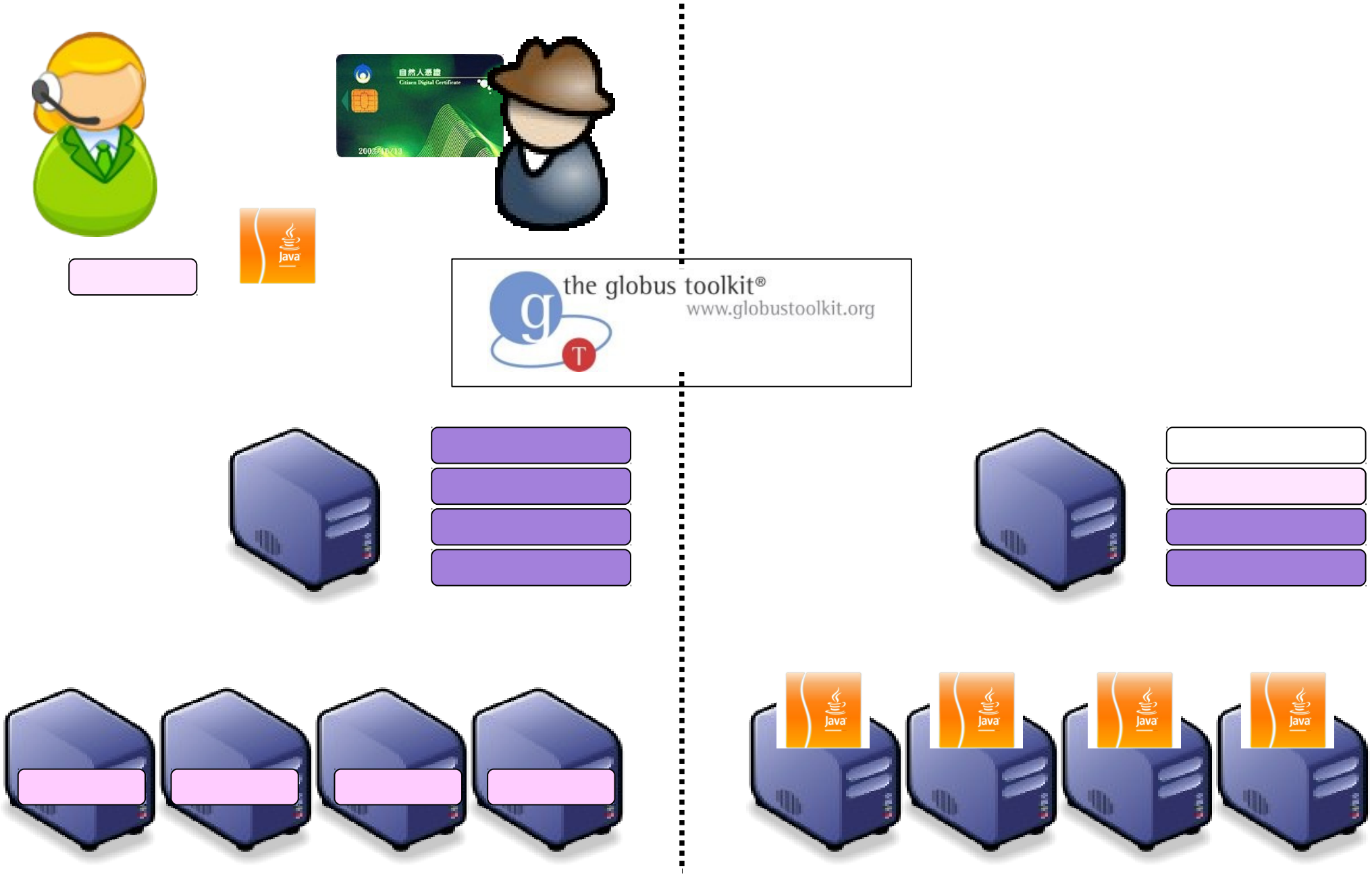


mpiBLAST

- http://www.mpiblast.org/
- An open-source, parallel implementation of NCBI BLAST
- 特點：
  - Database fragmentation
  - Query segmentation
  - Parallel input/output
- 設計理念：
  - The Design, Implementation, and Evaluation of mpiBLAST.
  - http://www.mpiblast.org/downloads/pubs/cwce03.pdf
- 類似工具：
  - TurboWorx TurboBLAST
  - Parallel BLAST by Caltech


GenBank
BLAST
GenBank
GenBank
GenBank
GenBank
mpiBLAST

Grid =~ Cluster of Cluster



mpiBLAST-G2

- **mpiBLAST-G2 is an enhanced parallel program of LANL's mpiBLAST. It is based on Globus Toolkit 2.x and MPICH-g2.**
- **Bioinformatics Technology and Service (BITS) team of Academia Sinica Computing Centre (ASCC), Taiwan**
- 參考：
  - **The MPIBLAST-g2 Introduction**
  - **MPIBLAST-g2 Example**
  - **mpiBlast-G2 with GT4**

mpiBLAST

the globus toolkit®
www.globustoolkit.org

ASCC
ACADEMIA SINICA
Computing Center
中央研究院計算中心

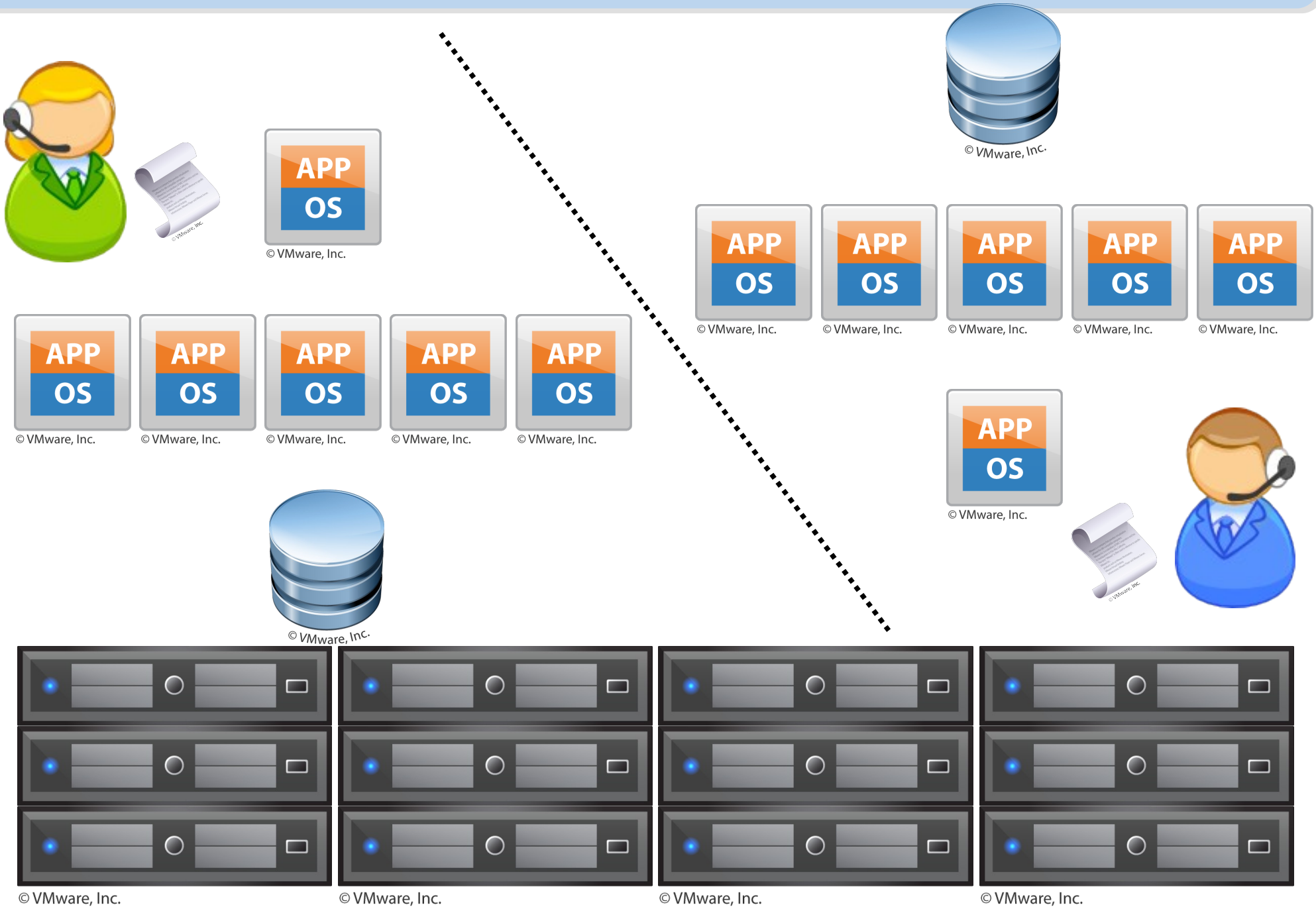
PART 2.3 :

Cloud 101 & CloudBLAST

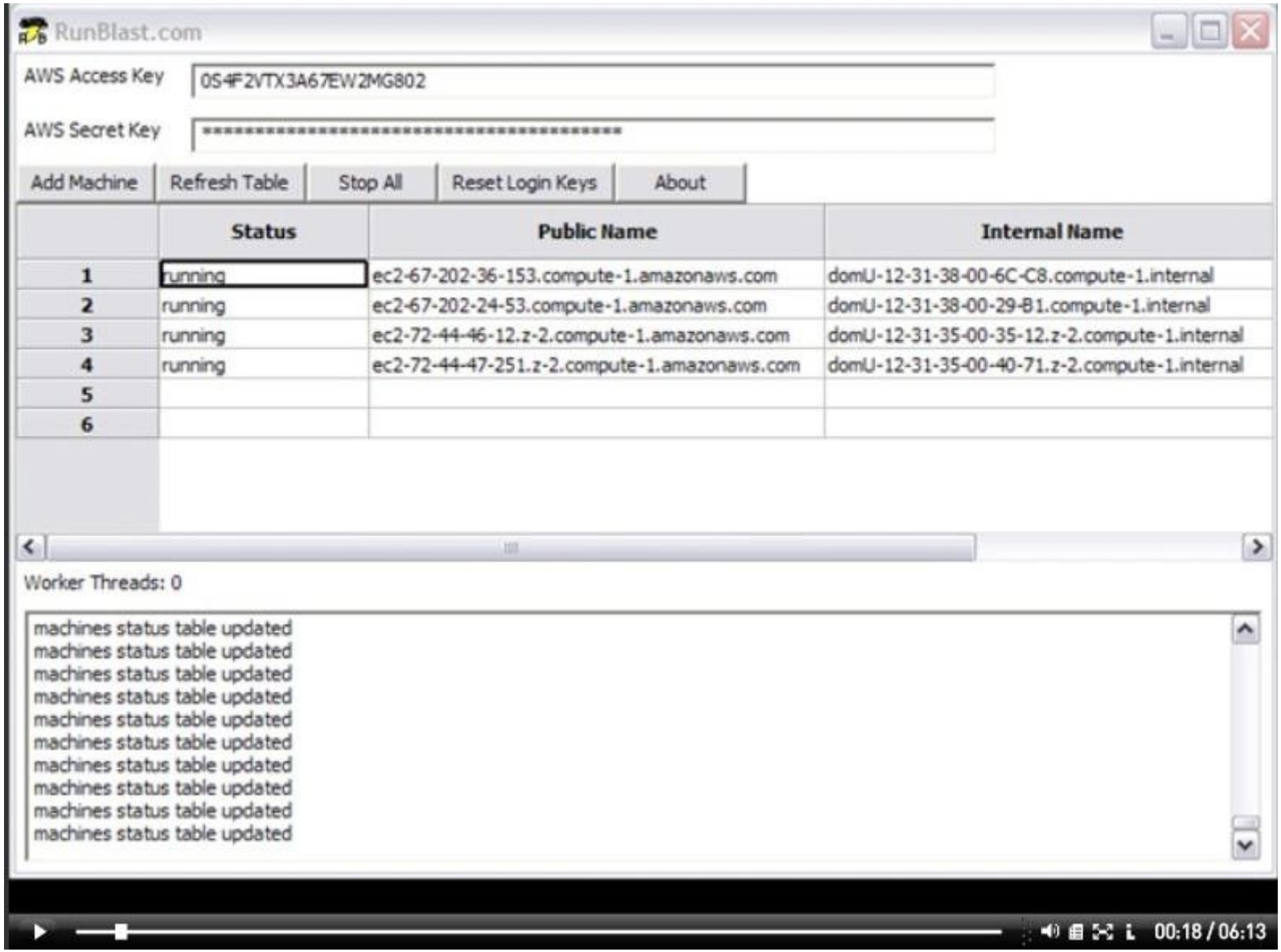
Jazz Wang
Yao-Tsung Wang
jazz@nchc.org.tw

Powered by DRBL

Cloud =~ Virtualization + Cluster

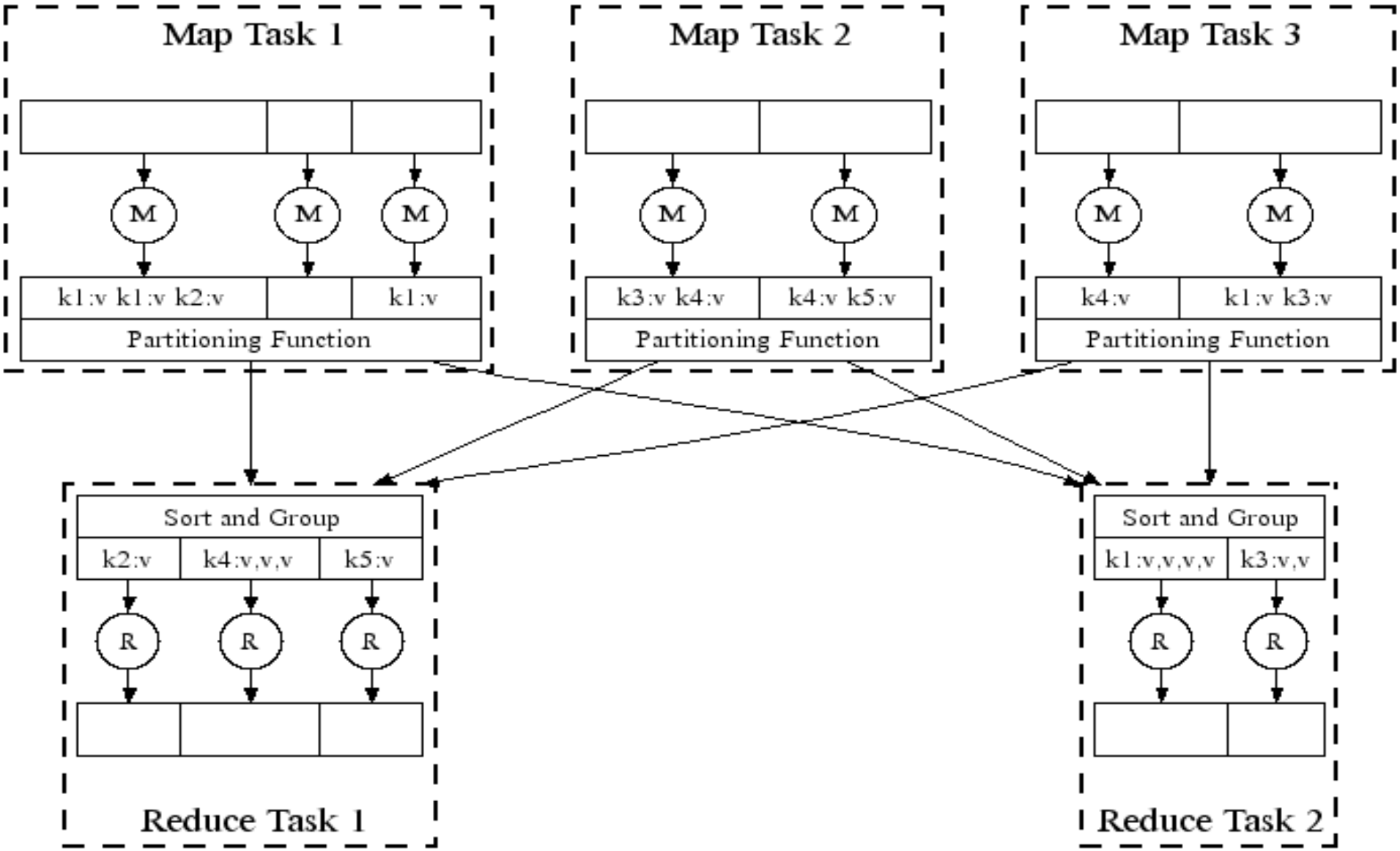


RunBLAST : mpiBLAST in Amazon EC2



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

Map/Reduce



Ref. MapReduce: Simplified Data Processing on Large Clusters, Google

CloudBLAST

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

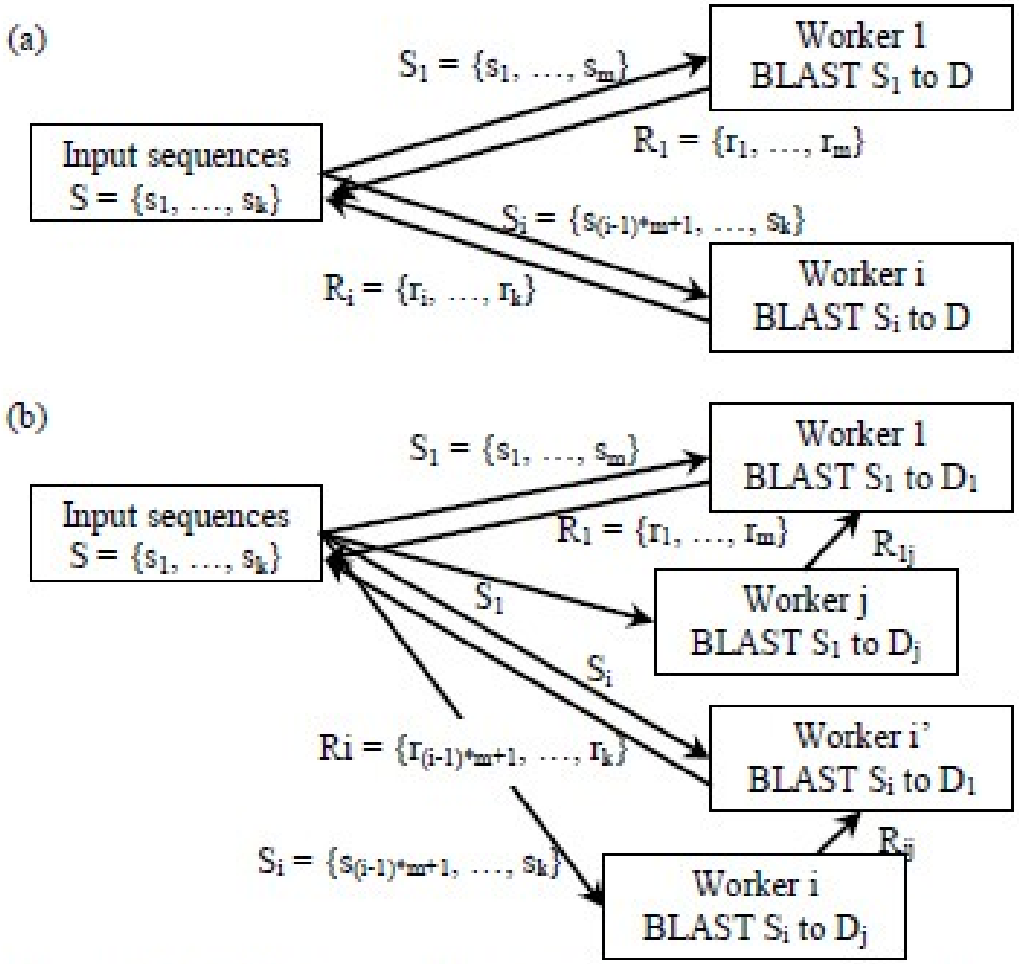


Figure 1. Parallelization of BLAST: (a) input sequences are partitioned into i subsets, each of which is processed by a worker, and results are

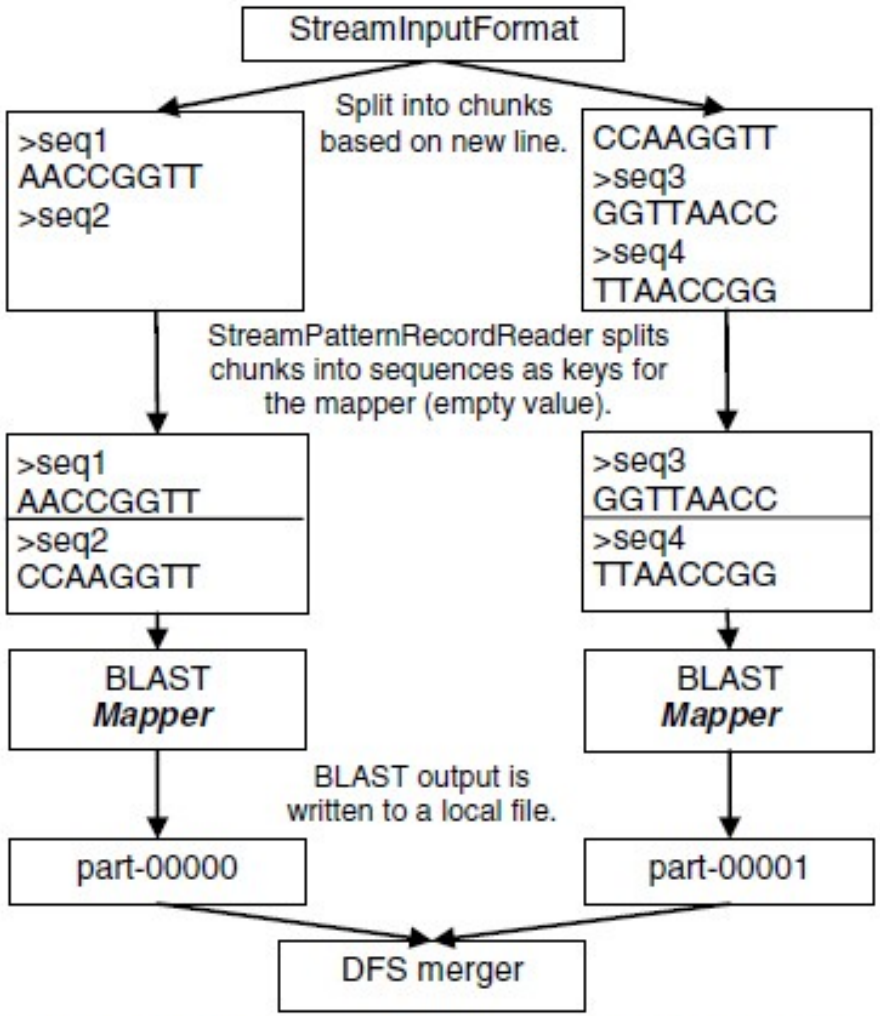


Figure 2. BLASTing with MapReduce. Given a set of input sequences

Questions?

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

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

Powered by DRBL