高速運算於生物資訊之應用

HPC for Bioinformatics

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

Powered by DRBL

高速運算於生物資訊之應用

HPC for Bioinformatics

PART 1 : (60 %)

HPC = High Performance Computing

What is HPC? Types of HPC ?

Can I solve my problem with HPC ?

PART 2 : (30%)

HPC & Bioinformatics Application

PART 3 : (10%)

Open Source for Bioinformatics

PART 1 :

HPC 101

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

What is HPC ?

&

Why HPC ?

High performance computing brings together computers, software, and expertise to solve problems too difficult to solve effectively by other means.

www.insiceHPC.com/ HPCcan

Source: http://insidehpc.com/whatishpc/WhatIsHPC.pdf

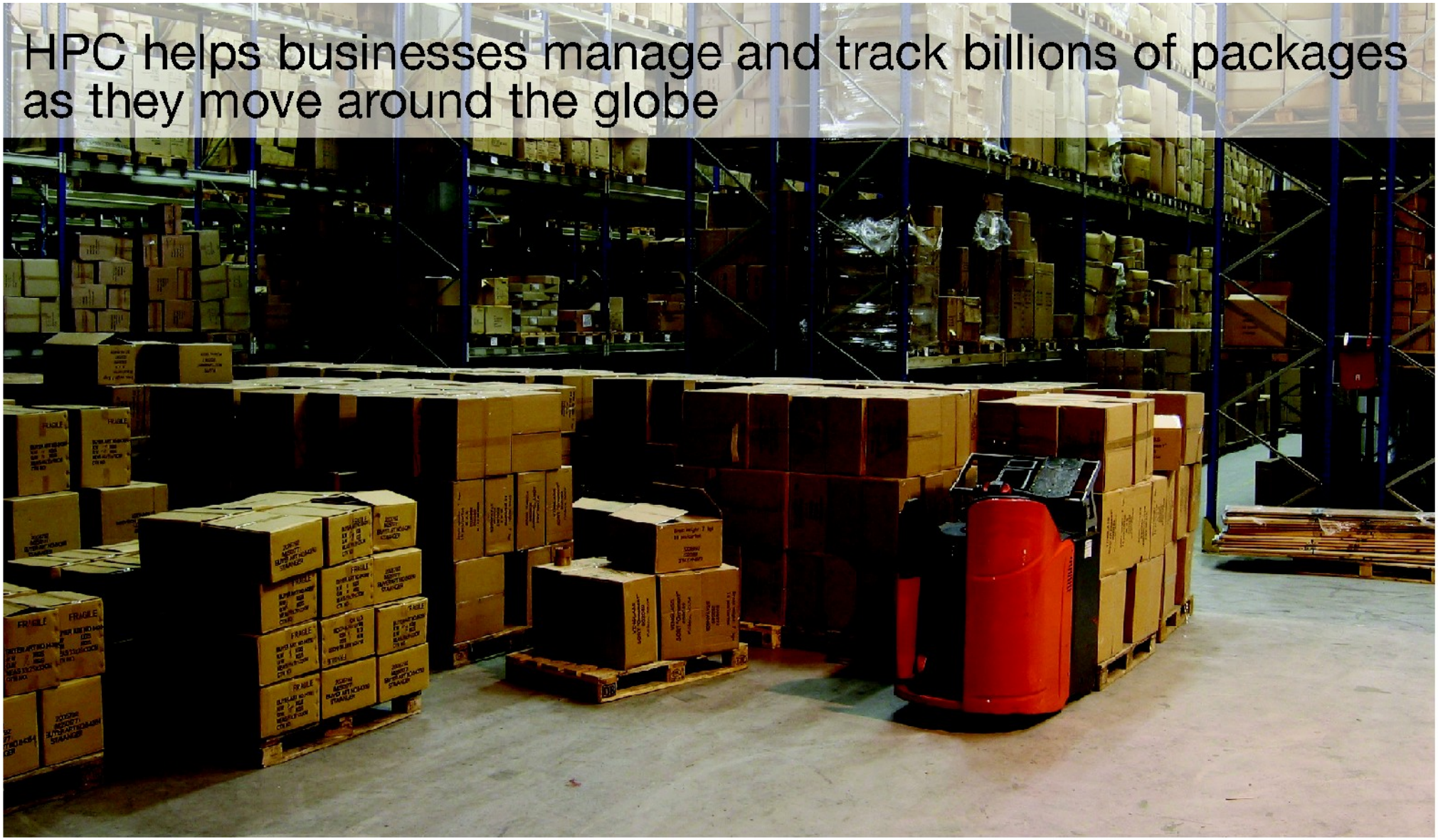
Simulated crash tests improve auto safety in more scenarios than could be tested in the real world

Too dangerous

Source: http://insidehpc.com/whatishpc/WhatIsHPC.pdf

Too time consuming

Source: http://insidehpc.com/whatishpc/WhatIsHPC.pdf

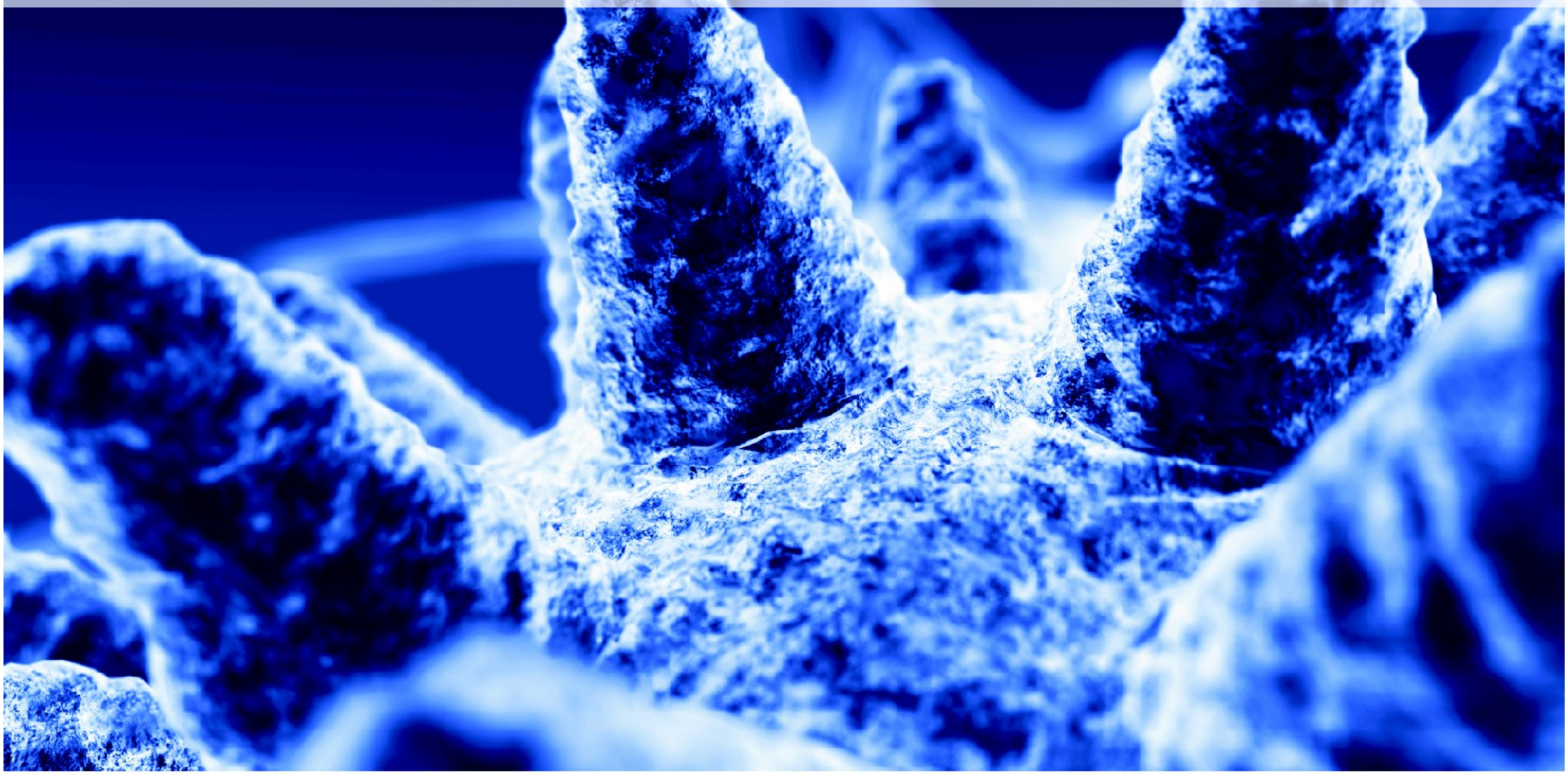


Too complex | Why HPC?

www.insideHPC.com/HPCcan

Source: http://insidehpc.com/whatishpc/WhatIsHPC.pdf

HPC gives doctors and scientists a window into the chemistry of our bodies to help develop new drugs and treatments



Too far beyond other tools

Why HPC?

www.insiceHPC.com/HPCcan

Source: http://insidehpc.com/whatishpc/WhatIsHPC.pdf

Types of HPC ?

Source: http://blog.tice.de/a_icons/icons/512%20Time%20Machine.png

Back to Year 1960s ...

Brief History of Computing (1/5)

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

1960 PDP-1

.

.

.

1965 PDP-7

.

.

.

1969 1st Unix

Mainframe

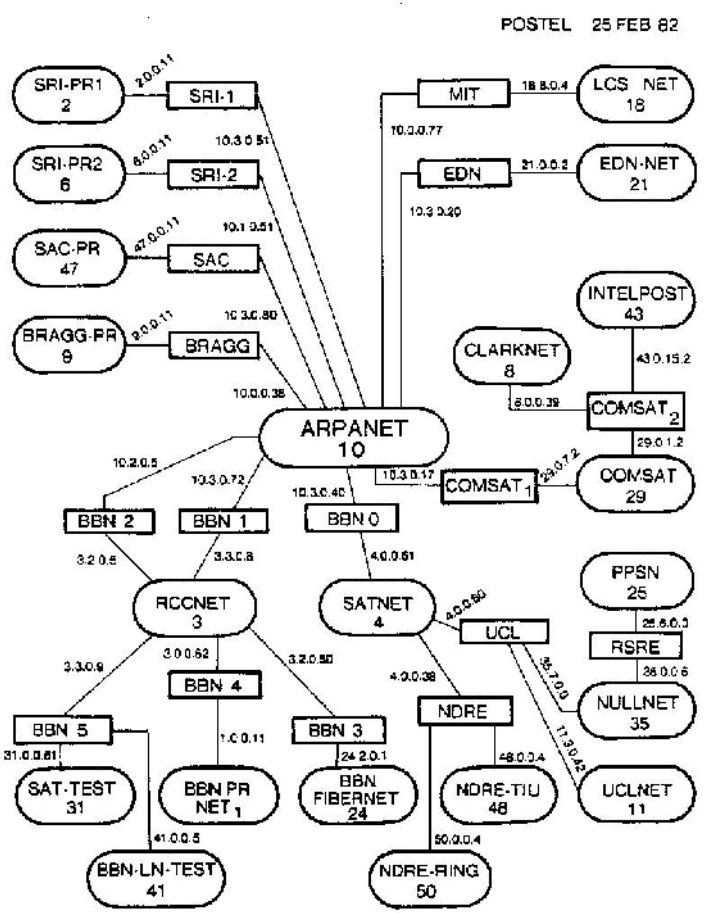
Super Computer

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)

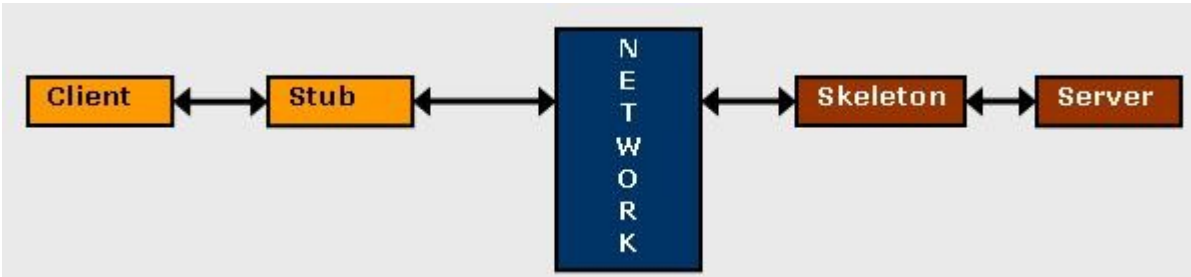
Source: http://www.nchc.org.tw

Mainframe
Super Computer

PC / Linux
Cluster Parallel

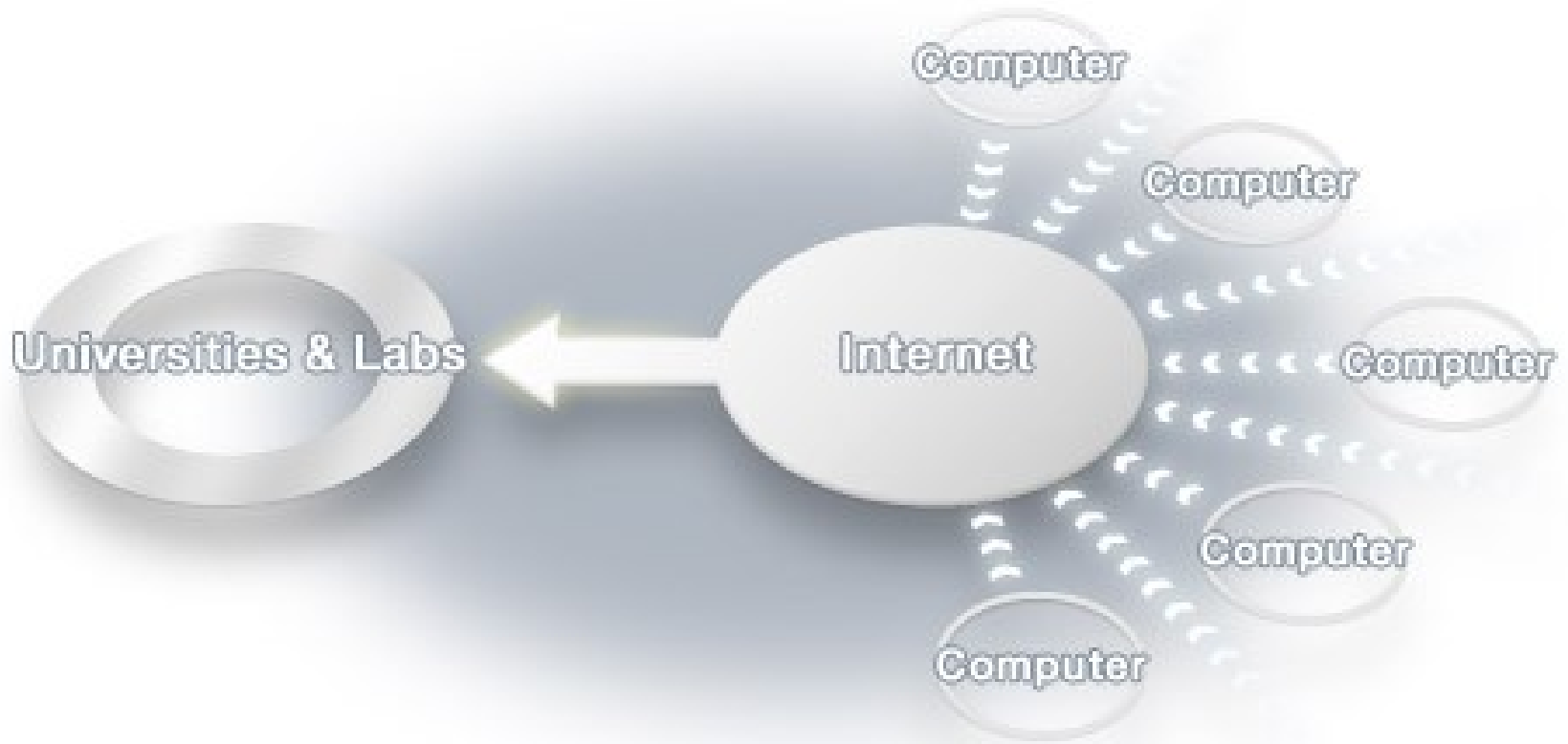
1990 World Wide Web
by CERN
...
...
1993 Web Browser
Mosaic by NCSA

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



Back to Year 1990s ...

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

1997 Volunteer Computing

1999 SETI@HOME

2003 Globus Toolkit 2

2002 Berkley BOINC

2004 EGEE gLite

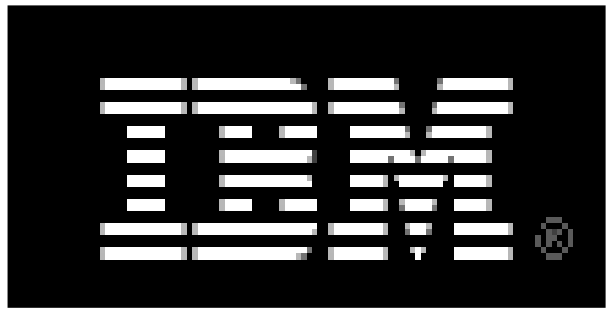
Back to Year 2000s ...

Brief History of Computing (4|5)

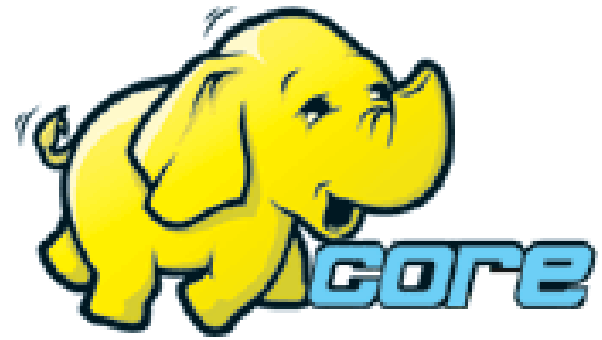
Source: http://gridcafe.web.cern.ch/gridcafe/whatisgrid/whatis.html

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

2001 Autonomic Computing
IBM



2006 Apache Hadoop

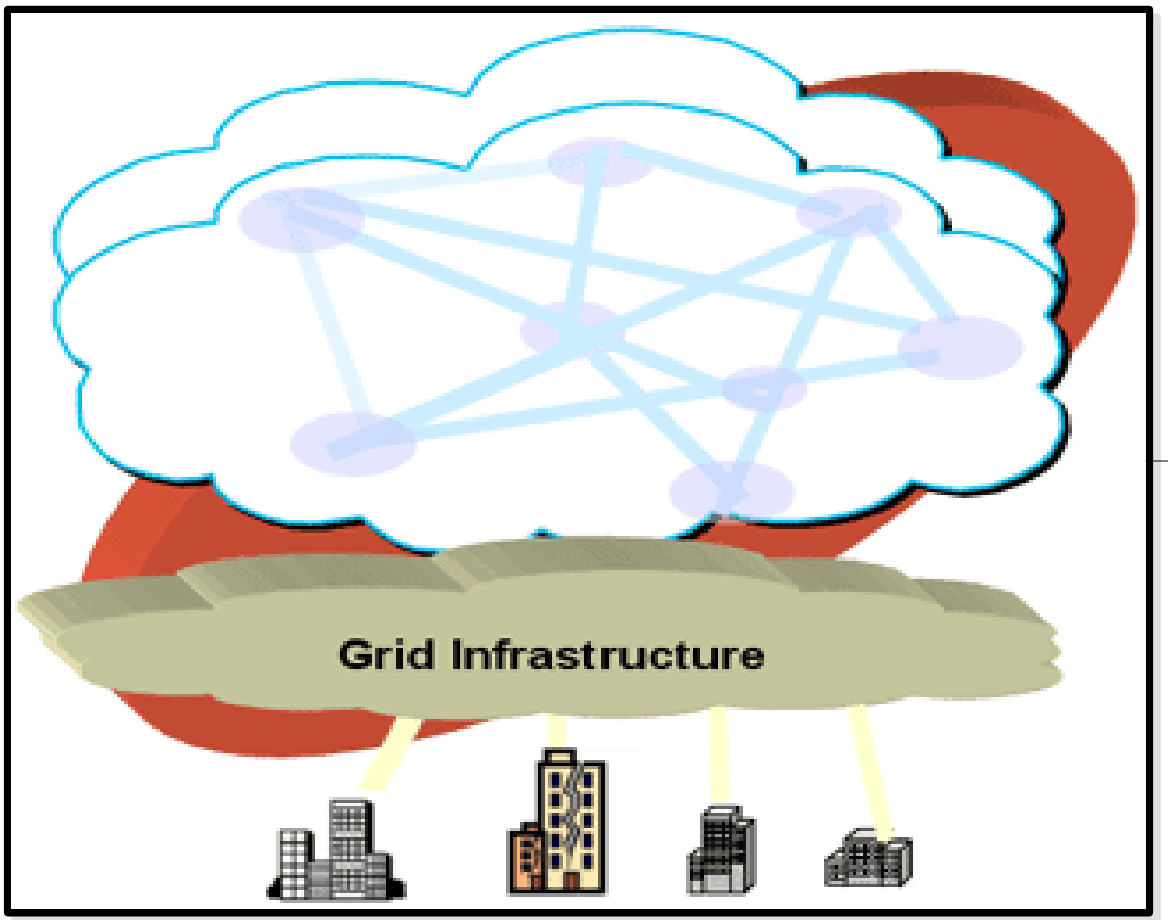


2005 Utility Computing
Amazon EC2 / S3

2007 Cloud Computing
Google + IBM

Back to Year 2007 ...

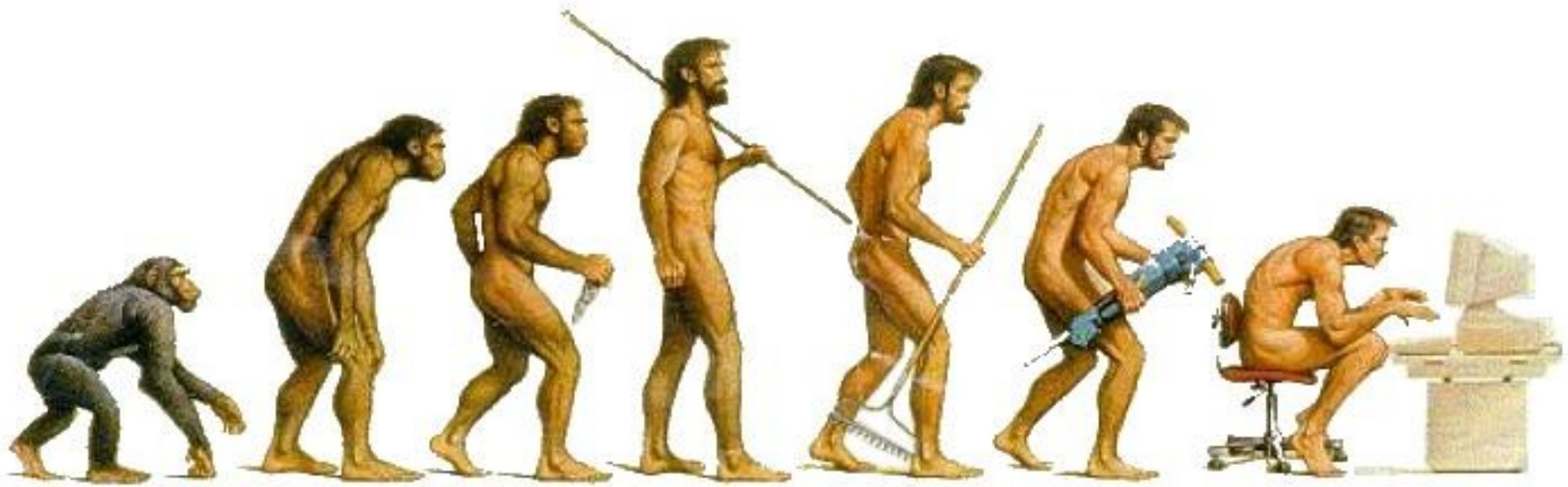
Brief History of Computing (5/5)



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

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

Evolution



(or is it?)

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

Source: http://media.photobucket.com/image/falling%20ground/preeto_f10/falling.jpg

Which Type of HPC is the Right ONE to solve My Problem ?

不負責解析

	執行程式 記憶體需求	執行程式 邏輯可分割	輸入資料 數量 / 大小	應用特性 計算特性
超級電腦 **Mainframe**	**非常大 (比 PC 大)**	**不易分割**	單一大檔 **GB**	即時性高 **共享記憶體**
叢集運算 **Cluster Parallel**	**小於單一 計算節點 配置量**	**邏輯相近 可分割**	一般數量 **MB**	**即時性一般 共享檔案**
分散式運算 **Distributed**	**小於單一 計算節點 配置量**	**邏輯不同 偕同運作**	一般數量 **MB**	即時性高 **訊息傳遞**
格網運算 **Grid**	**小於單一 計算節點 配置量**	**邏輯相近 可分割**	一般數量 **MB**	**即時性較低 資料就計算**
雲端運算 **Cloud**	**小於單一 計算節點 配置量**	**邏輯相近 各自獨立**	海量檔案 **MB**	**即時性低 計算就資料**

PART 2 :

HPC & Bioinformatics Application

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

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：使用啟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

PART 2.1 :

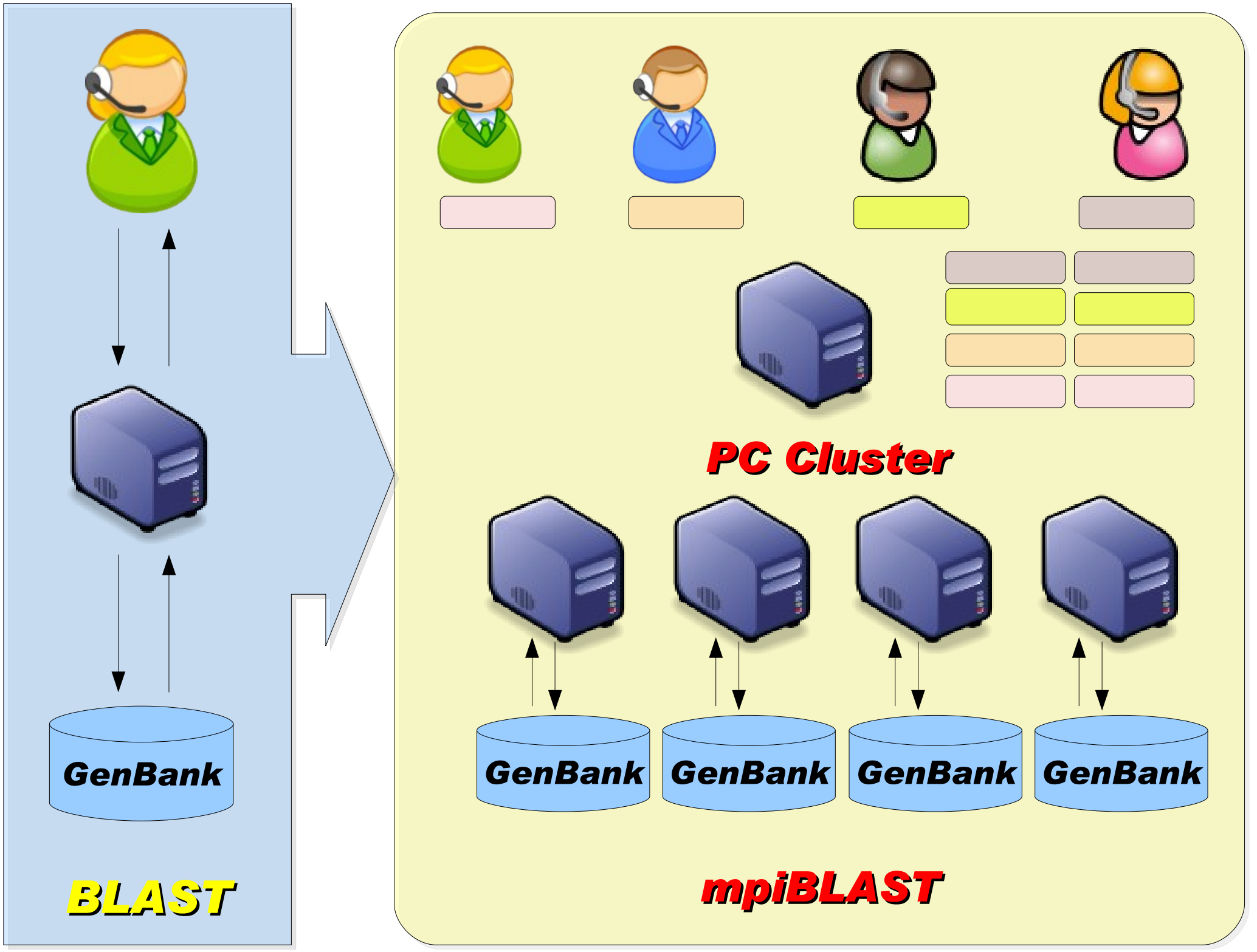
Cluster 101 & mpiBLAST

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

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
PC Cluster
GenBank
GenBank
GenBank
GenBank
mpiBLAST

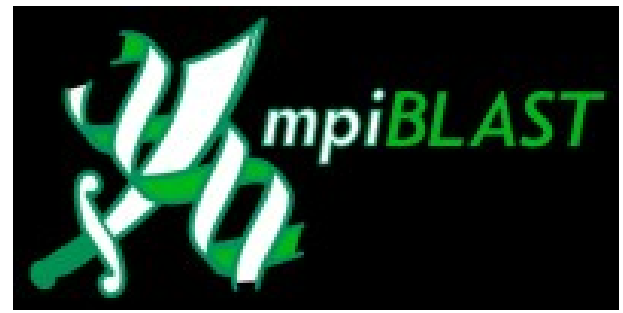
PART 2.2 :

Grid 101 & mpiBLAST-G2

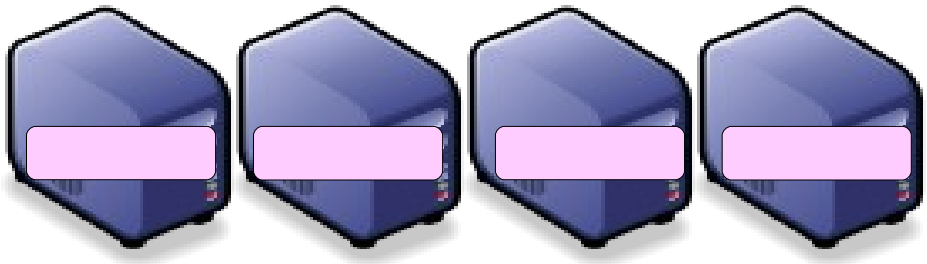
Jazz Wang
Yao-Tsung Wang
jazz@nchc.org.tw

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***



Grid =~ Cluster of Cluster

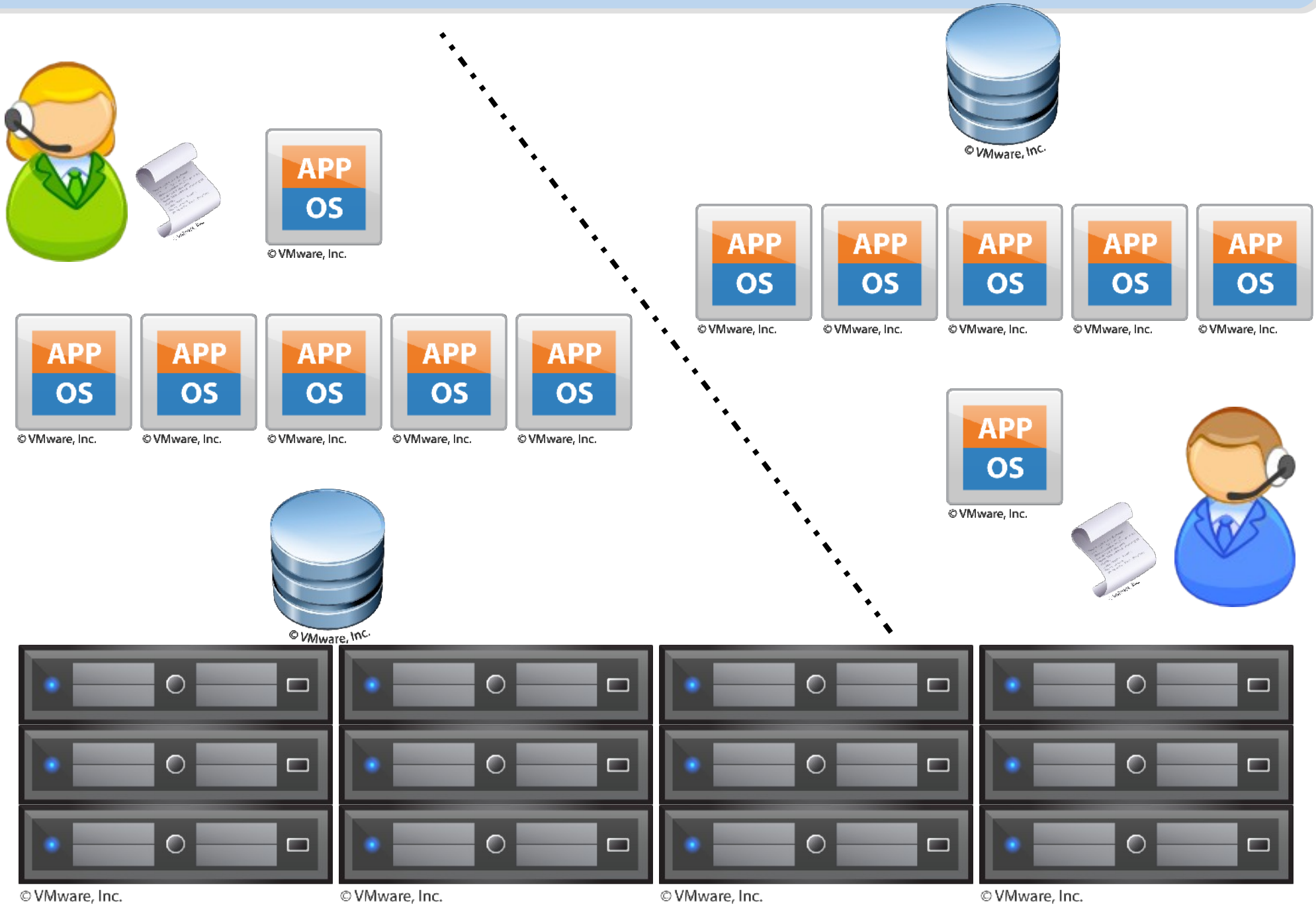


PART 2.3 :

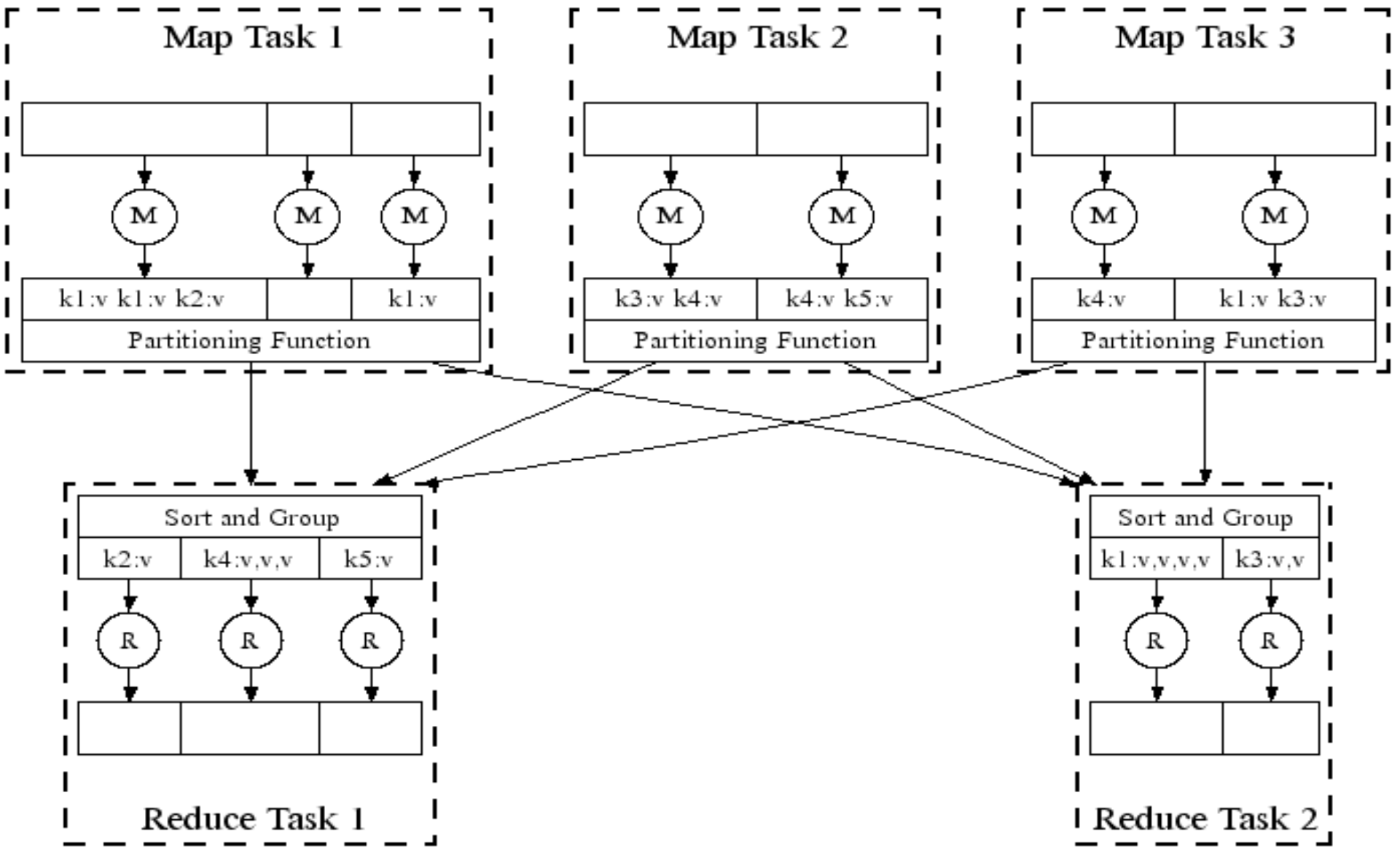
Cloud 101 & CloudBLAST

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

Cloud =~ Lots of Virtual Cluster



MapReduce



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

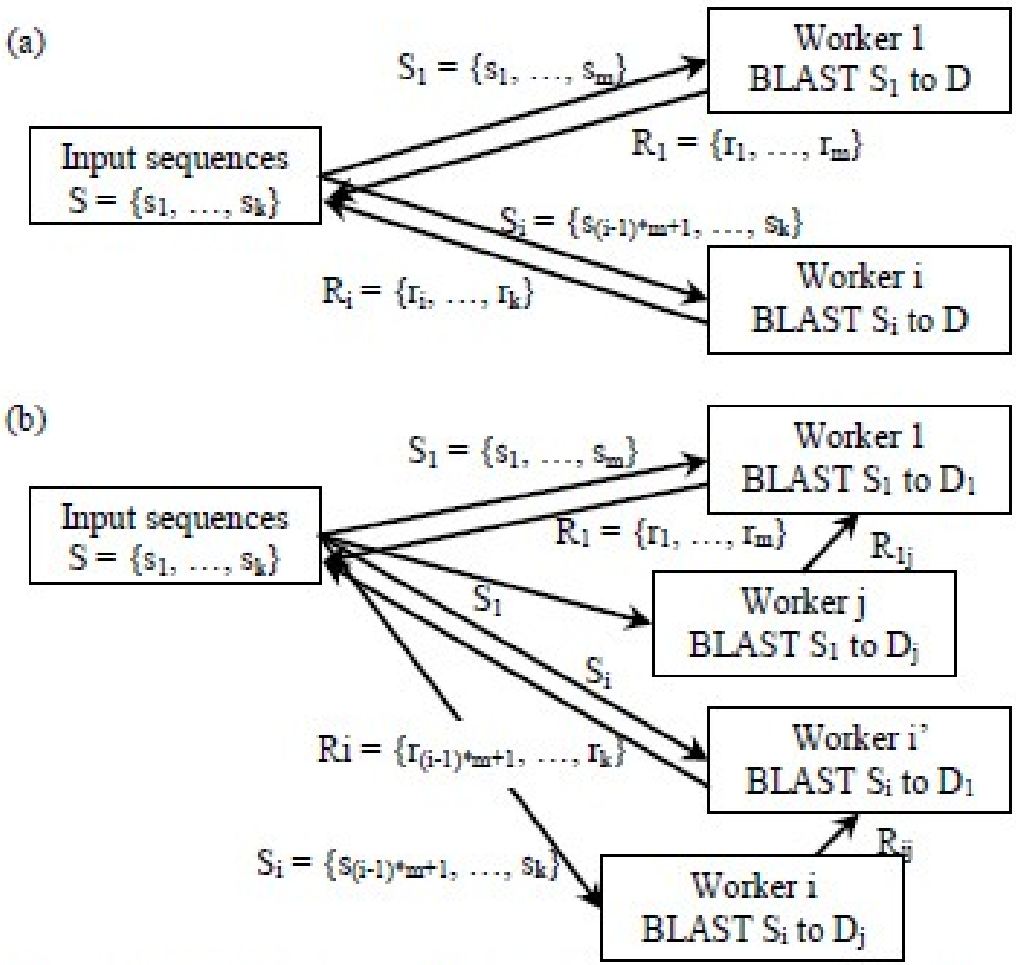
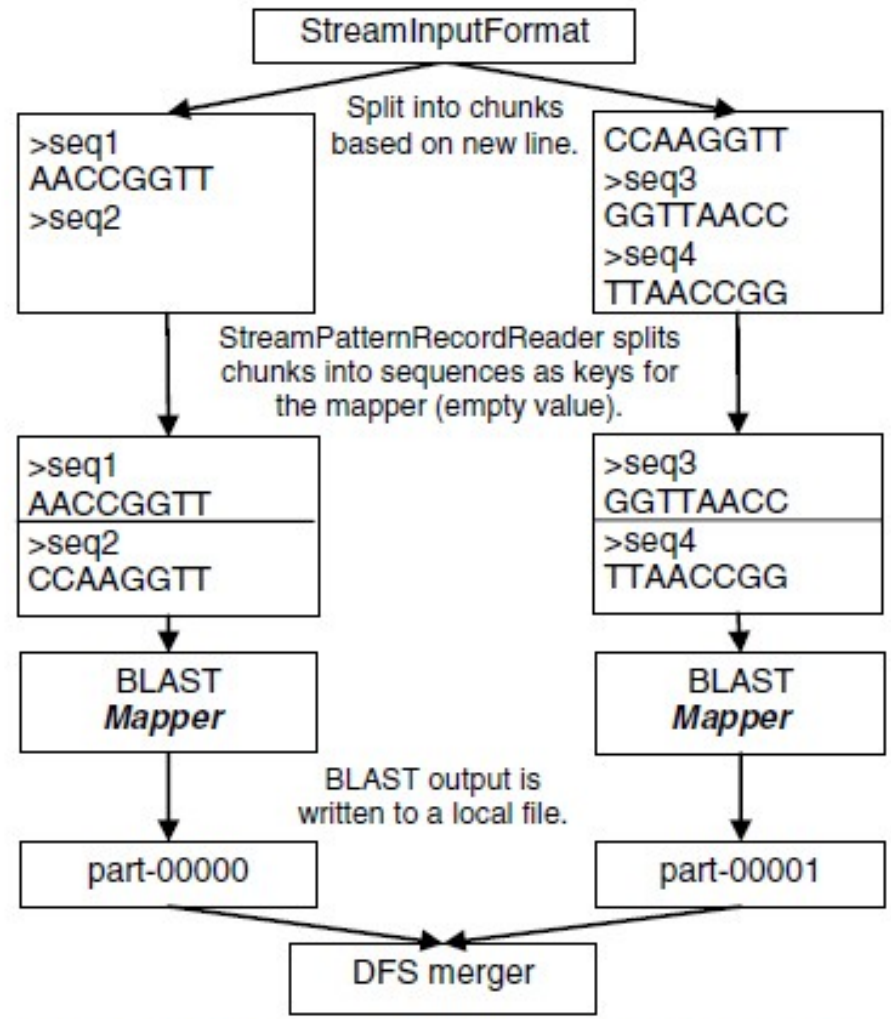
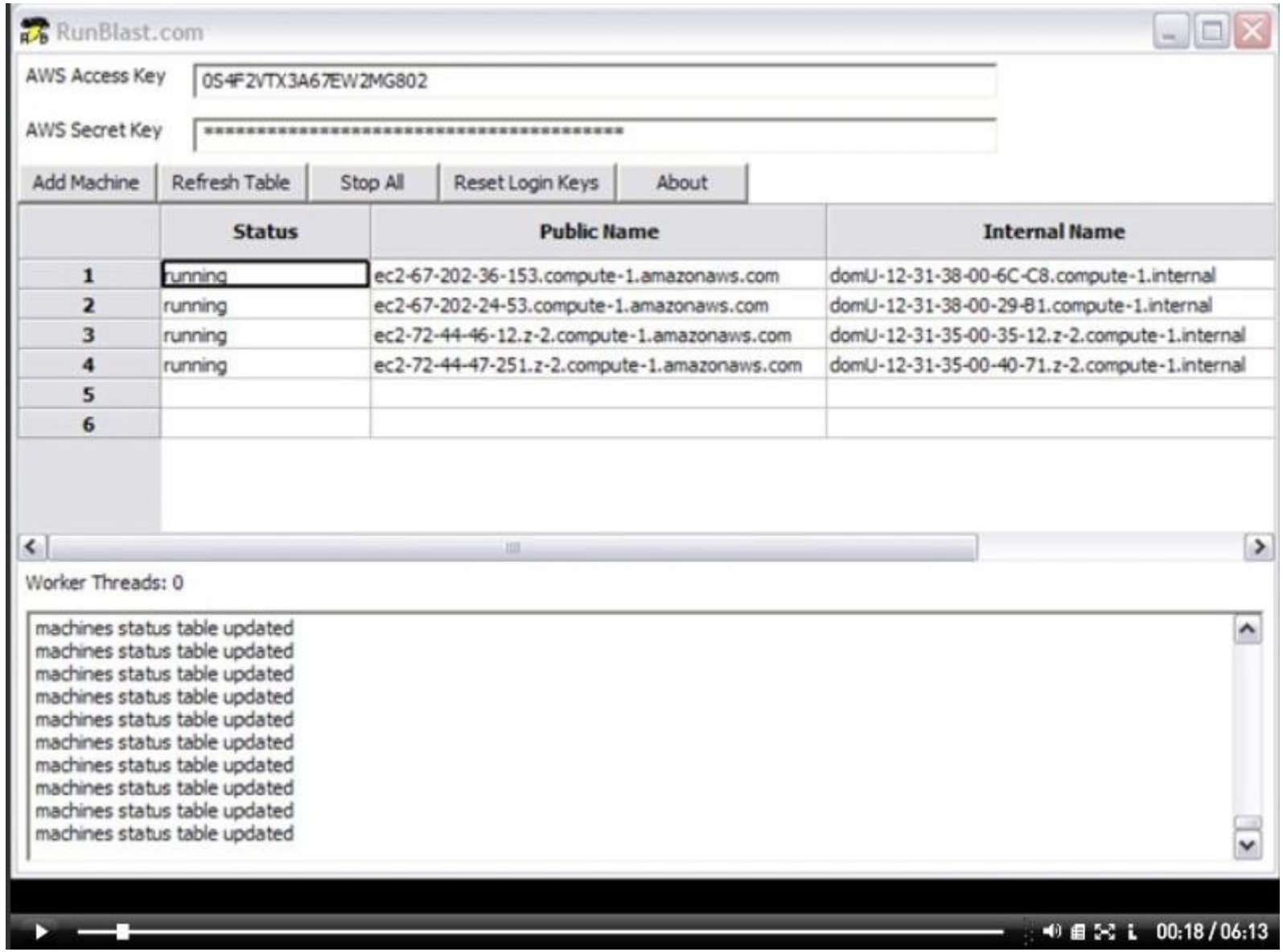


Figure 1. Parallelization of BLAST: (a) input sequences are partitioned



RunBLAST : mpiBLAST in Amazon EC2



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

PART 3 :

Open Source for Bioinformatics

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

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/

What we learn today ?

Q1: What is HPC? 何謂高速運算？
A1: HPC 就是結合電腦硬體、軟體和一堆專家，
用各種方式解決困難的問題。

Q2: Types of HPC? 高速運算的種類有哪些？
A2: Mainframe, PC Cluster, Parallel, Distributed, Grid, Cloud
超級電腦、電腦叢集、平行、分散、格網、雲端運算

Q3: Can HPC solve all your problems? 高速運算可以解決所有問題？
A3: No. 高速運算無法解決所有問題，各種類別也各有所長。

Q4: What is PC Cluster? 何謂電腦叢集？
A4: Cluster = lots of PCs. 很多電腦用內部網路串起來，就是叢集。

Q5: What is Grid? 何謂格網運算？
A5: Grid = Cluster of Cluster. 把好幾座叢集視為一座抽象的叢集。

Q6: What is Cloud? 何謂雲端運算？
A6: Cloud = lots of Virtual Cluster. 在實體叢集中打造多座虛擬叢集。

Conclusion

神兵利器很多，挑一件來玩！
Open Source is Your Friend !!

學海無涯，團結合作力量大！
Team Work !! Do Cooperation !!

放大格局，創造差異！
Think Bigger than others !!

Crazy Idea + Right Tool + Good Team

Questions?

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

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