雲端運算於生物資訊之應用

Cloud Computing for Bioinformatics

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

Powered by DRBL

雲端運算於生物資訊之應用

Cloud Computing for Bioinformatics

PART 1 :

(60 %)

Introduction to Hadoop

How can I solve big data problem ?

PART 2 :

(30%)

Cloud & Bioinformatics Application

PART 3 :

(10%)

Open Source for Bioinformatics

PART 1 :

Introduction to Hadoop and its Ecosystem

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

Powered by DRBL

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
  - **Lucene create an inverse index of every word in 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.**

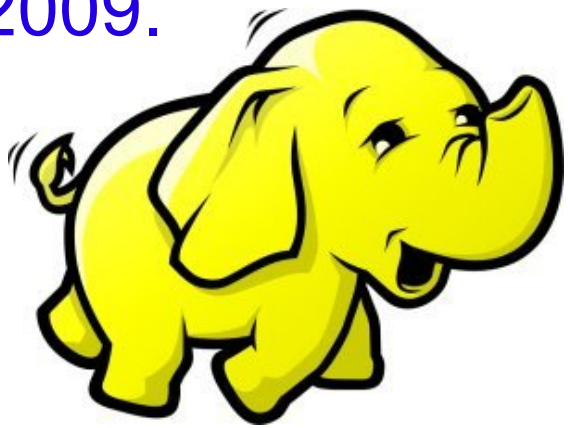
Three Gifts from Google ....
來自 Google 的三個禮物 ....

- 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.



Who Use Hadoop ??
有哪些公司在用 Hadoop 這套軟體 ??

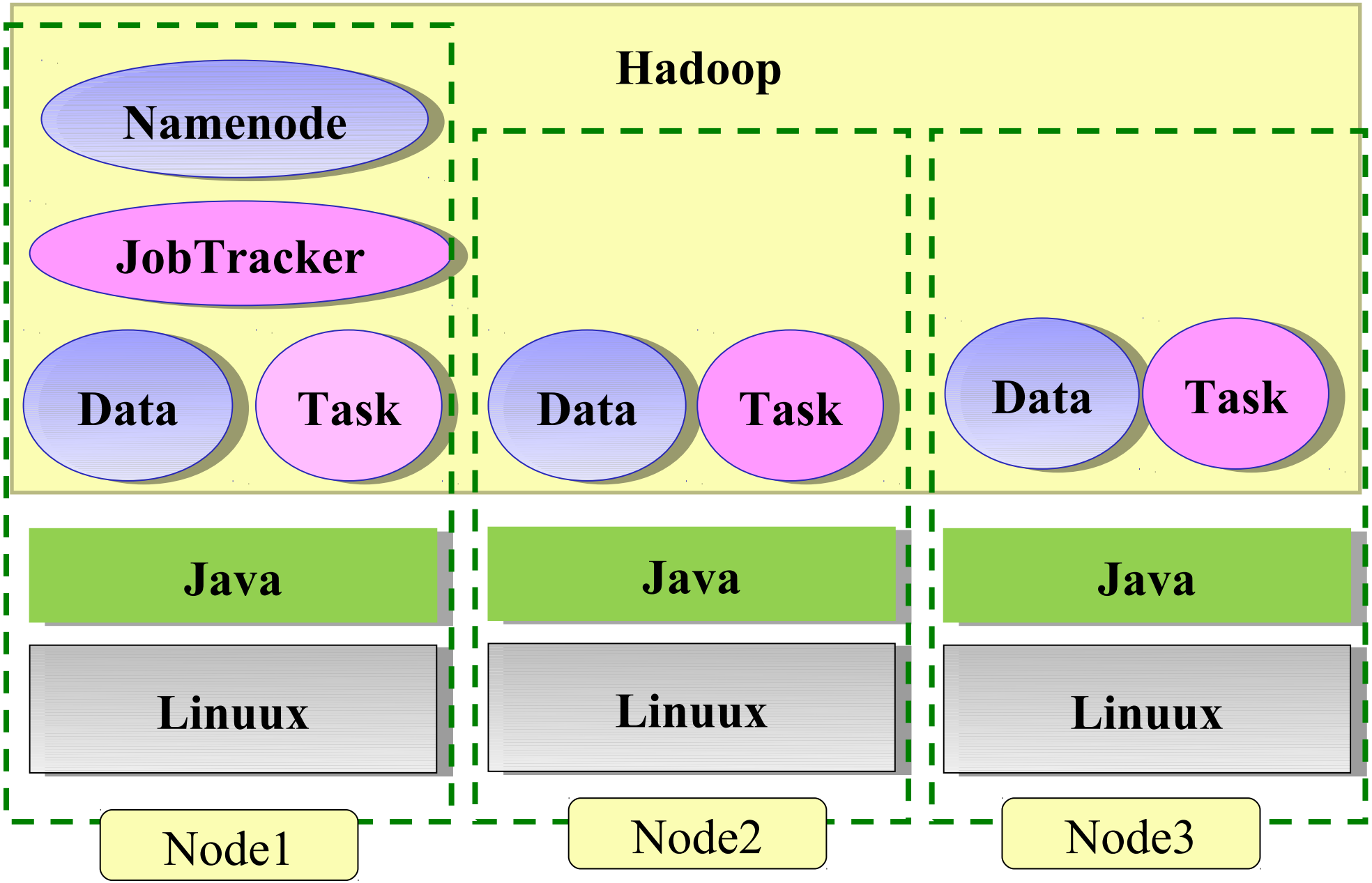
- Yahoo is the key contributor currently.
- IBM and Google teach Hadoop in universities …
- http://www.google.com/intl/en/press/pressrel/20071008_ibm_univ.html
- The New York Times used 100 Amazon EC2 instances and a Hadoop application to process 4TB of raw image TIFF data (stored in S3) into 11 million 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

Distributed Operating System of Hadoop

Hadoop 建構成一個分散式作業系統



Is Hadoop only support Java ?

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

Hadoop Pipes (C++, Python)

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

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

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

Hadoop Streaming

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

There are serveral Hadoop subprojects

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

Other Hadoop related projects

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

Hadoop Ecosystem

<table>
<tr><td>Pig</td><td>Chukwa</td><td>Hive</td><td colspan="2">HBase</td></tr>
<tr><td colspan="2">MapReduce</td><td colspan="2">HDFS</td><td>ZooKeeper</td></tr>
<tr><td colspan="3">Hadoop Core
(Hadoop Common)</td><td colspan="2">Avro</td></tr>
</table>

Source: *Hadoop: The Definitive Guide*

Avro

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

Zoo Keeper

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

Pig

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

Hive

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

Chukwa

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

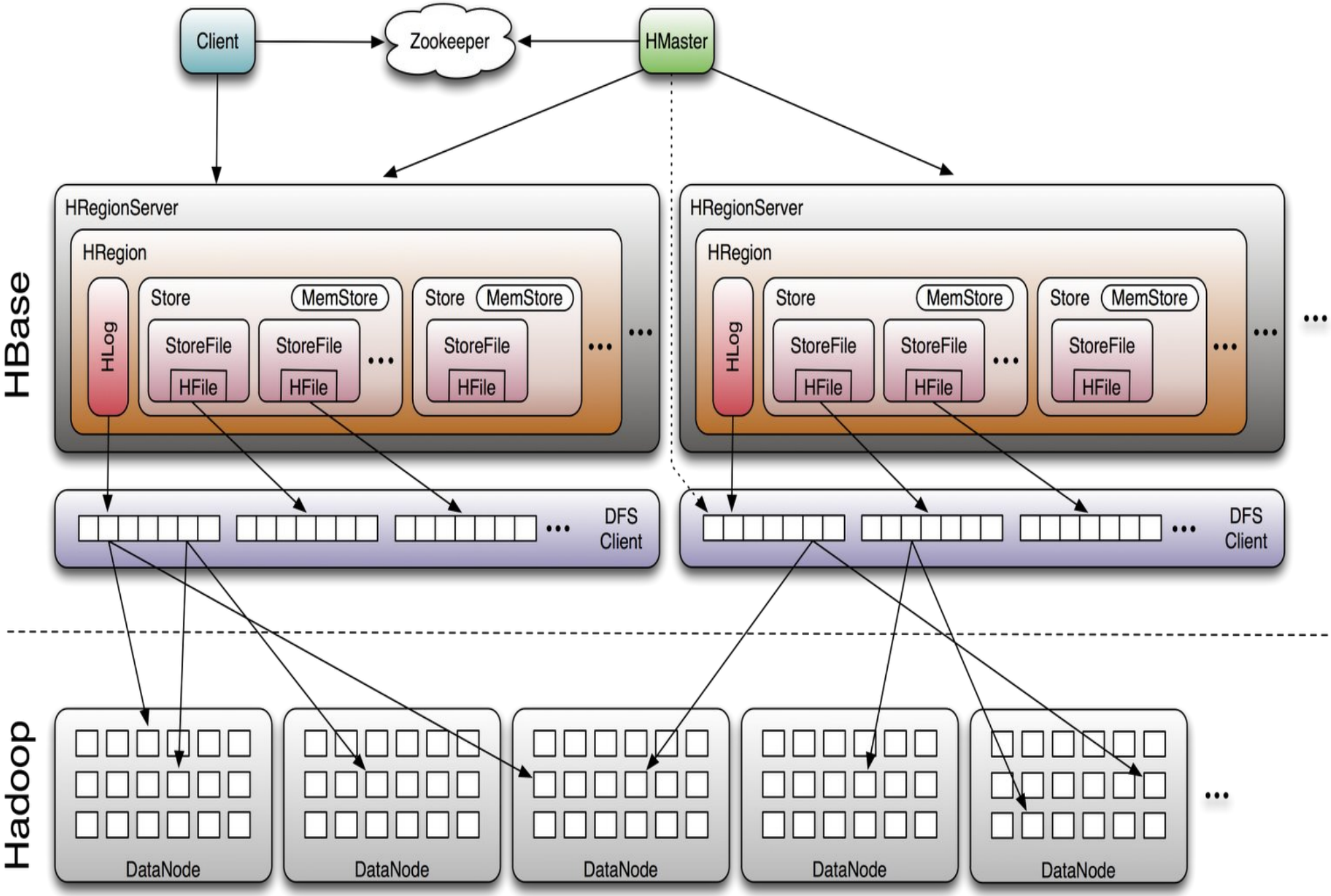
Mahout

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

HBase is ..

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

Architecture



PART 2 :

Cloud & Bioinformatics Application

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

Powered by DRBL

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

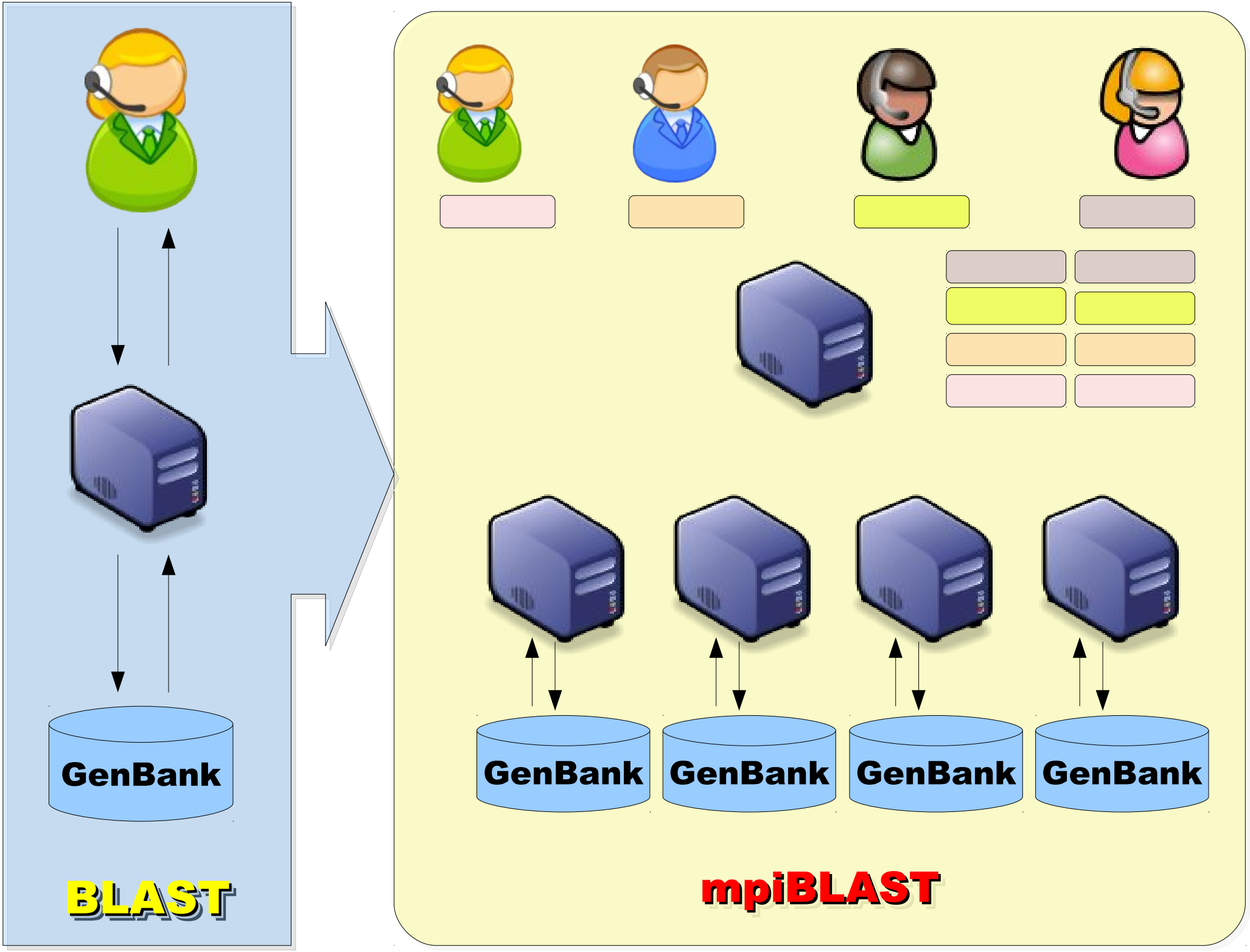
    (例如：不同蛋白質的氨基酸序列 DNA 序列的核甘酸)
- 用途：搜尋其他物種 (如：老鼠) 未知基因，是否也存在人類基因中
- 優點：使用啓發式搜索來找出相關的序列，比動態規劃快上 50 倍。
- 缺點：不能夠保證搜尋到的序列和所要找的序列之間的相關性。
- 技術問題：巨大的序列資料庫需要進行比對，怎樣計算才快？

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

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

mpiBLAST-G2

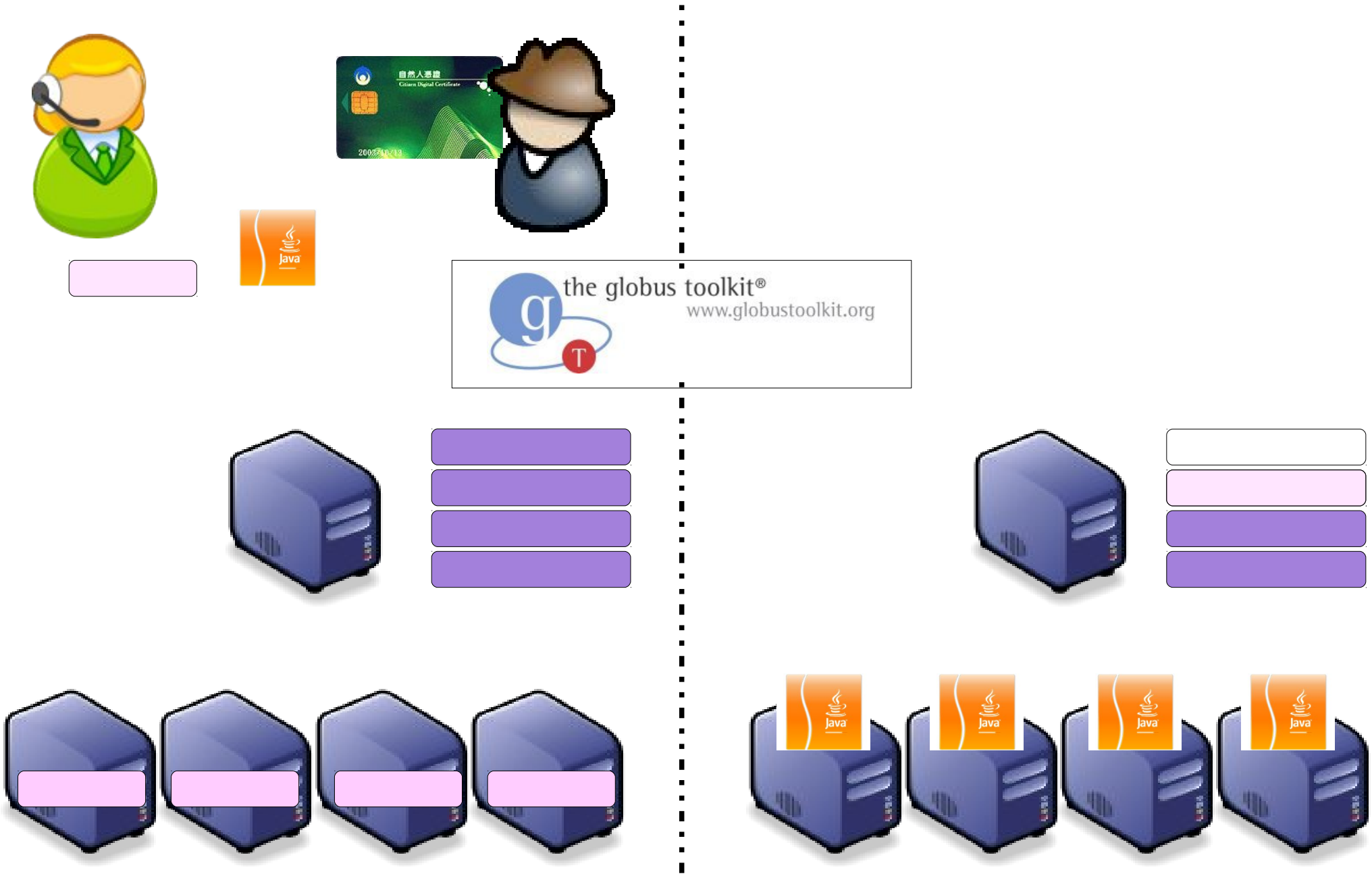
- **mpiBLAST-G2 is an enhanced parallel program of LANL's mpiBLAST. It is based on Globus Toolkit 2.x and MPICH-g2.**
- **Bioinformatics Technology and Service (BITS) team of Academia Sinica Computing Centre (ASCC), Taiwan**
- 參考：
  - **The MPIBLAST-g2 Introduction**
  - **MPIBLAST-g2 Example**
  - **mpiBlast-G2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

mpiBLAST

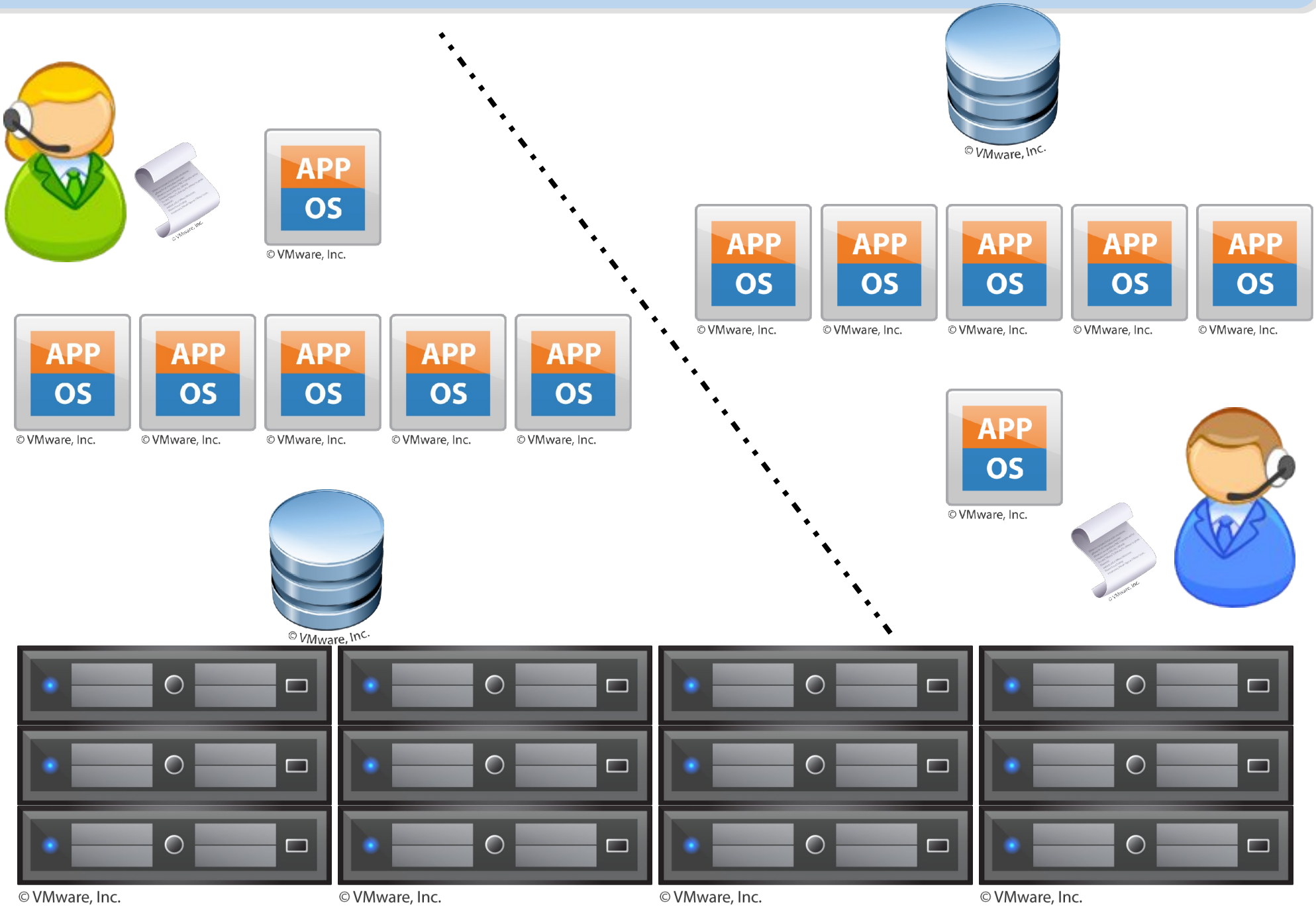
the globus toolkit®
www.globustoolkit.org

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

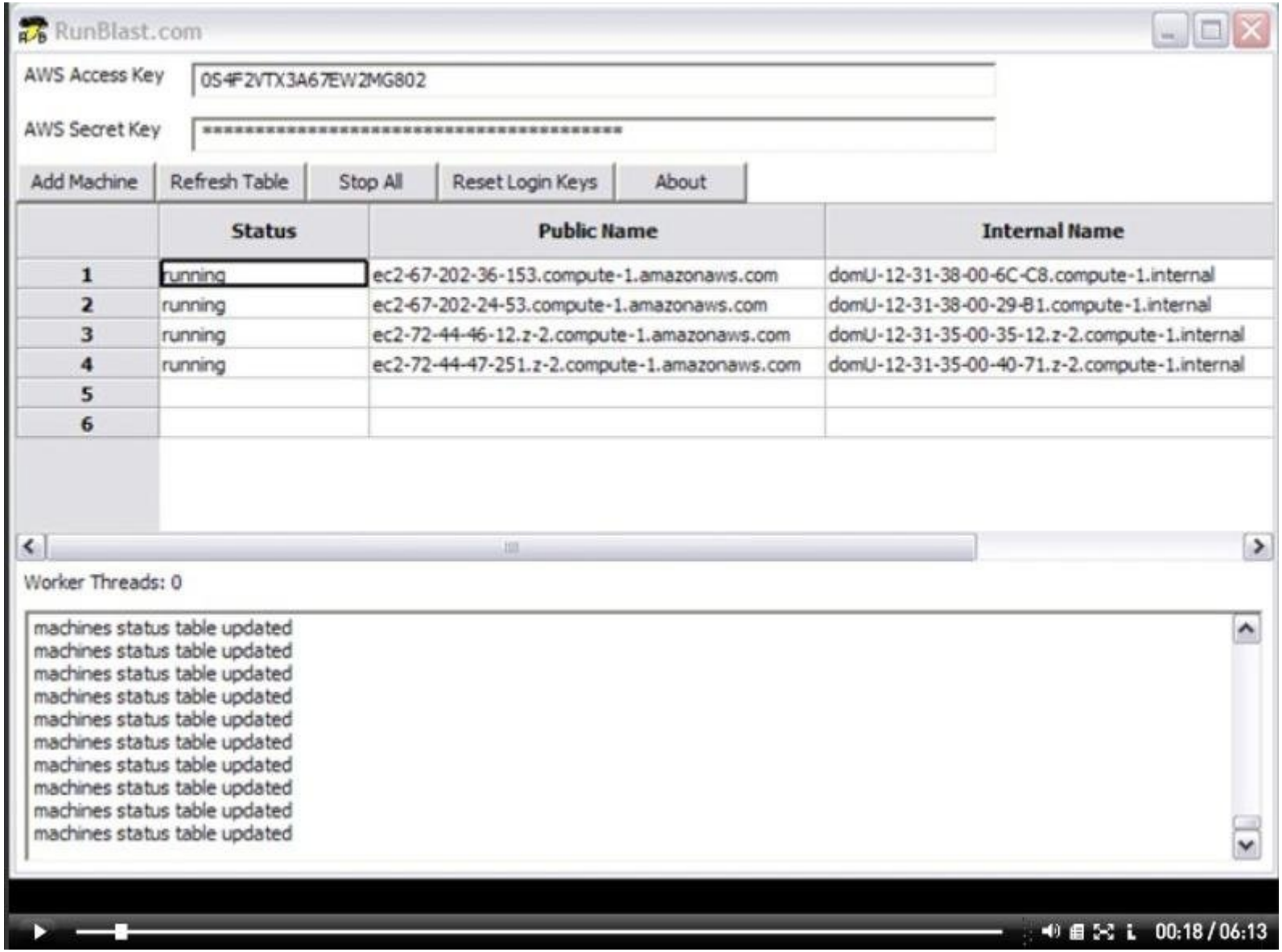
Grid =~ Cluster of Cluster



Cloud =~ Virtualization + Cluster



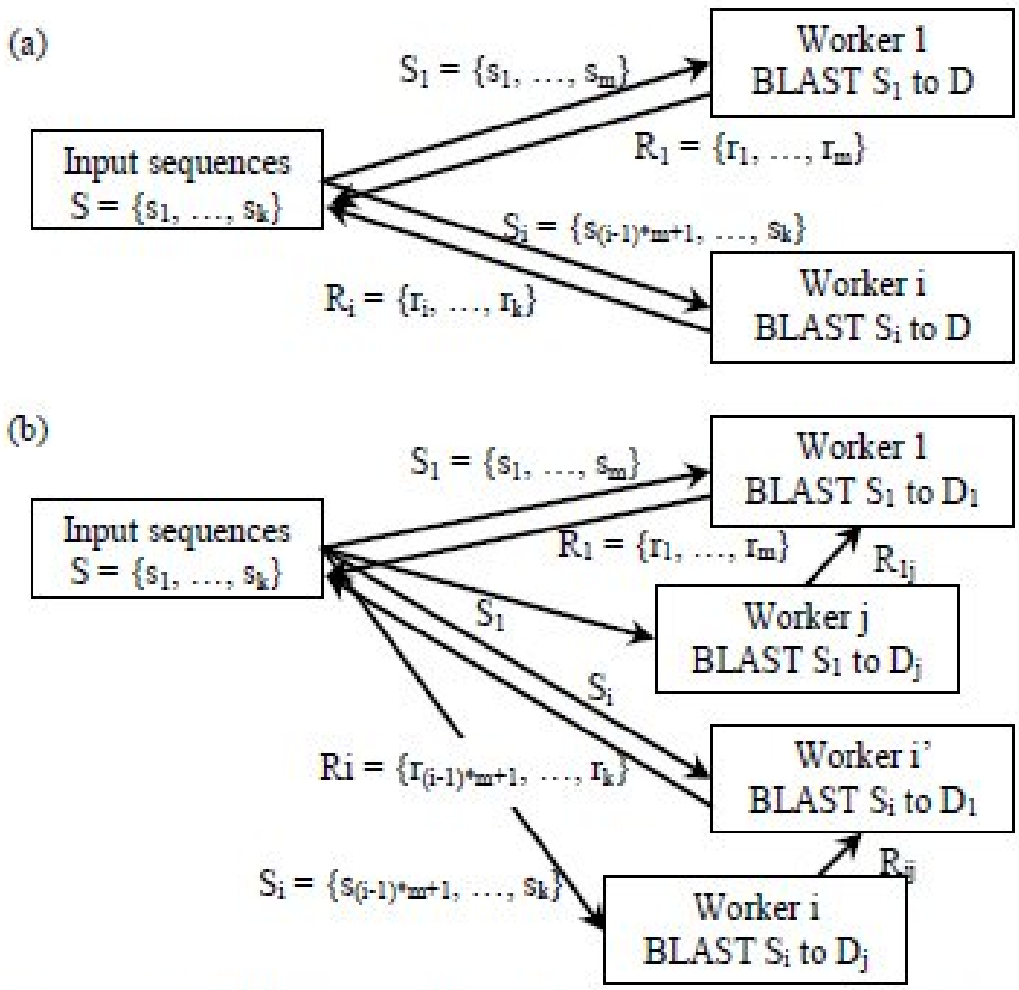
RunBLAST : mpiBLAST in Amazon EC2

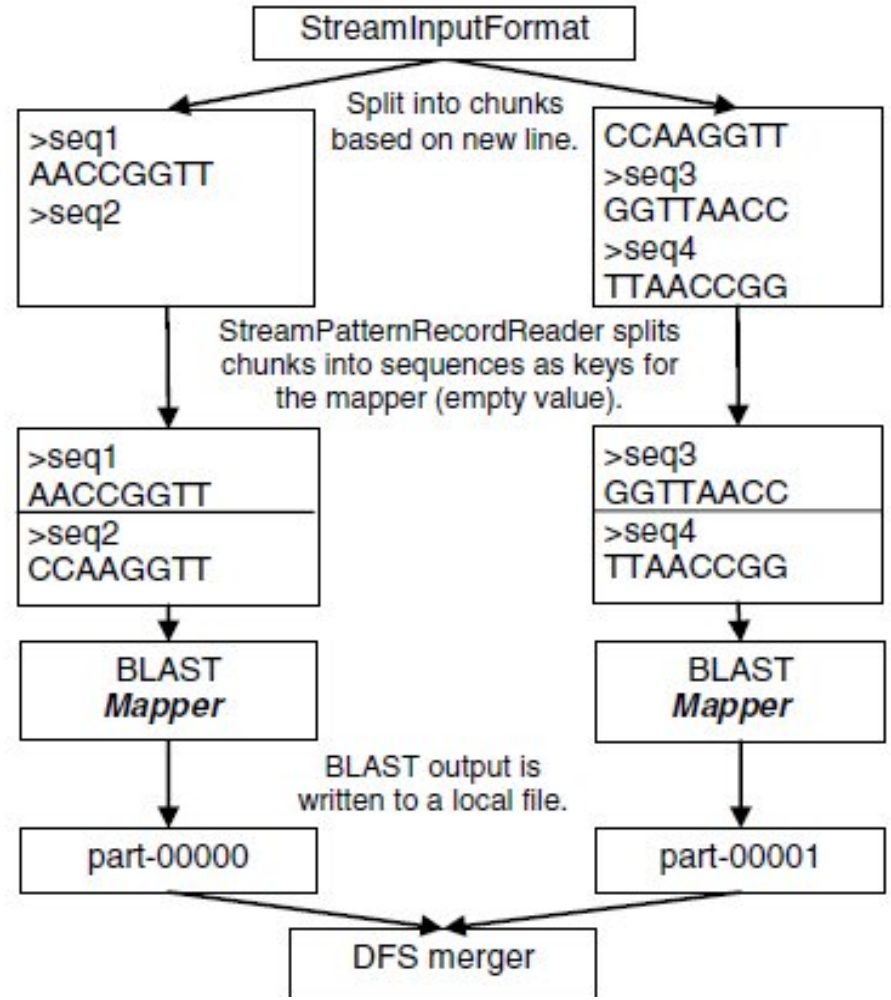


Video: **http://www.runblast.com/videos/runblast-blastwizard.swf**

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





PART 3 :

Open Source for Bioinformatics

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

Powered by DRBL

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/

Questions?

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

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

Powered by DRBL