加值服務：多人雲端運算實驗叢集

Building Multiple Users Cloud Infrastructure with Open Source

Jazz Wang

Yao-Tsung Wang

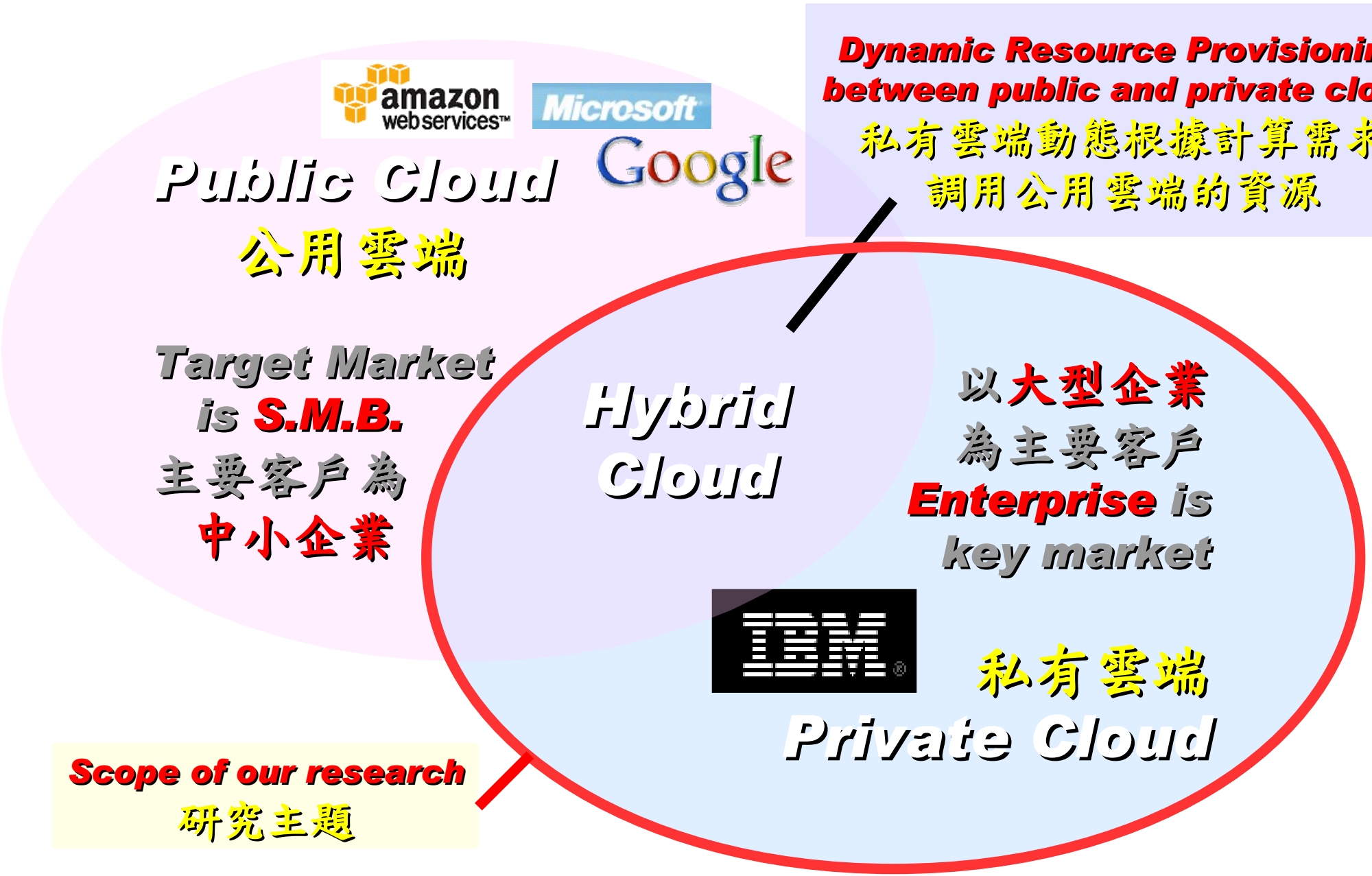
jazz@nchc.org.tw

Powered by DRBL

NCHC Cloud Computing Research Group
團隊小檔案：國網中心雲端運算研究小組

- 主要研究雲端運算的基礎架構組成元件
- 團隊成員：6 名
  - 王耀聰 - drbl-xen / drbl-hadoop (~6 Years) 架構
  - 陳威宇 - Hadoop / NutchEz / ICAS (~3 Years) 應用
  - 郭文傑 - Xen / OpenNebula / Eucalyptus (~3 Years) 元件
  - 涂哲源 - Xen GPU / OpenMP / VirtualGL (~3 Years) 元件
  - 鄭宗碩 - Google App Engine (~2 Years) 新技術
  - 鄧偉華 - AMQP / OpenID (~2 Years) 新技術
- 定位：
  - 研發快速佈建軟體，提供實驗平台服務，開辦訓練課程育才
- 獨特性：
  - 基於企鵝龍 (DRBL)，可快速佈署雲端運算的實驗叢集環境

Types of Cloud Computing
雲端運算的三種型態



Open Source for Private Cloud

建構私有雲端運算架構的自由軟體

應用
Social Computing, Enterprise, ISV,…

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

程式語言
Web 2.0 介面, Mashups, Workflows, …

Hadoop (MapReduce),
Sector/Sphere, AppScale

控制
Qos Negotiation, Ddmission Control,
Pricing, SLA Management, Metering…

OpenNebula, Enomaly,
Eucalyptus , OpenQRM, ...

虛擬化
VM, VM management and Deployment

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

Open Cloud #1: Eucalyptus

- http://open.eucalyptus.com/
- 原是加州大學聖塔芭芭拉分校(UCSB)的研究專案
- 目前已轉由Eucalyptus System這間公司負責維護
- 創立目的是讓使用者可以**打造自己的EC2**
- 特色是相容於 Amazon EC2 既有的用戶端介面
- 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
- 缺點：目前仍有部分操作需透過指令模式

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

Open Cloud #2: OpenNebula

- http://www.opennebula.org

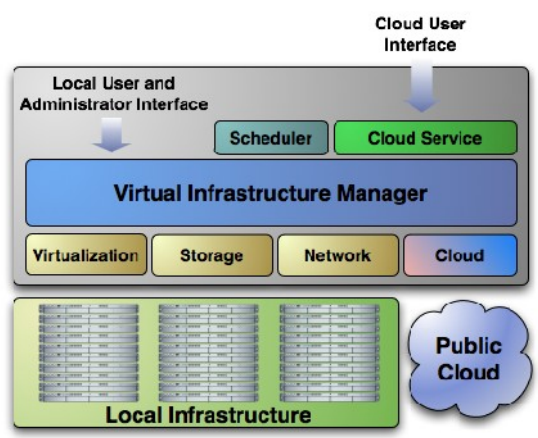
OpenNebula.org

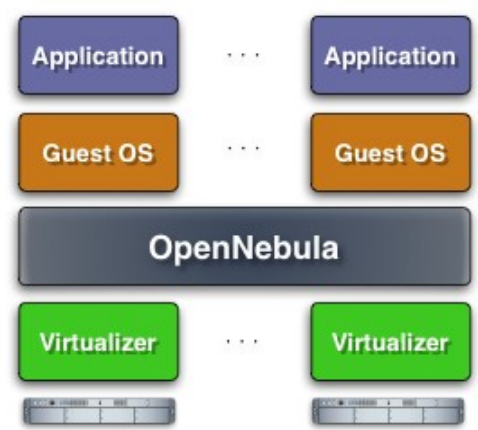
- 由歐洲研究學會(European Union FP7)贊助
- 將實體叢集轉換成具管理彈性的虛擬基礎設備
- 可管理虛擬叢集的狀態、排程、遷徙(migration)
- 優勢是Ubuntu 9.04 已經收錄 OpenNebula 的套件
- 缺點：需下指令來進行虛擬機器的遷徙(migration)。

SEVENTH FRAMEWORK PROGRAMME

RESERVOIR

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
- 目前主要由 Yahoo! 資助、開發與運用
- 創始者是Doug Cutting，參考Google Filesystem，以Java開發，提供HDFS與MapReduce API。
- 2006年使用在Yahoo內部服務中
- 已佈署於上千個節點。
- 處理Petabyte等級資料量。
- Facebook、Last.fm、Joost … 等著名網路服務均有採用Hadoop。

站在巨人的肩膀—國網中心自由軟體開發

多元化資訊教學的新選擇！

以個人叢集電腦 (PC Cluster) 經驗發展 DRBL&Clonezilla

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

是自由！不是免費…

分送、修改、存取、使用軟體的自由。免費是附加價值。

何謂企鵝龍 DRBL ??

- ***Diskless Remote Boot in Linux***
- 網路是便宜的，人的時間才是昂貴的。
- 企鵝龍簡單來說就是 .....
  - 用網路線取代硬碟排線
  - 所有學生的電腦都透過網路連接到一台伺服器主機

Diskfull PC

=

+

+

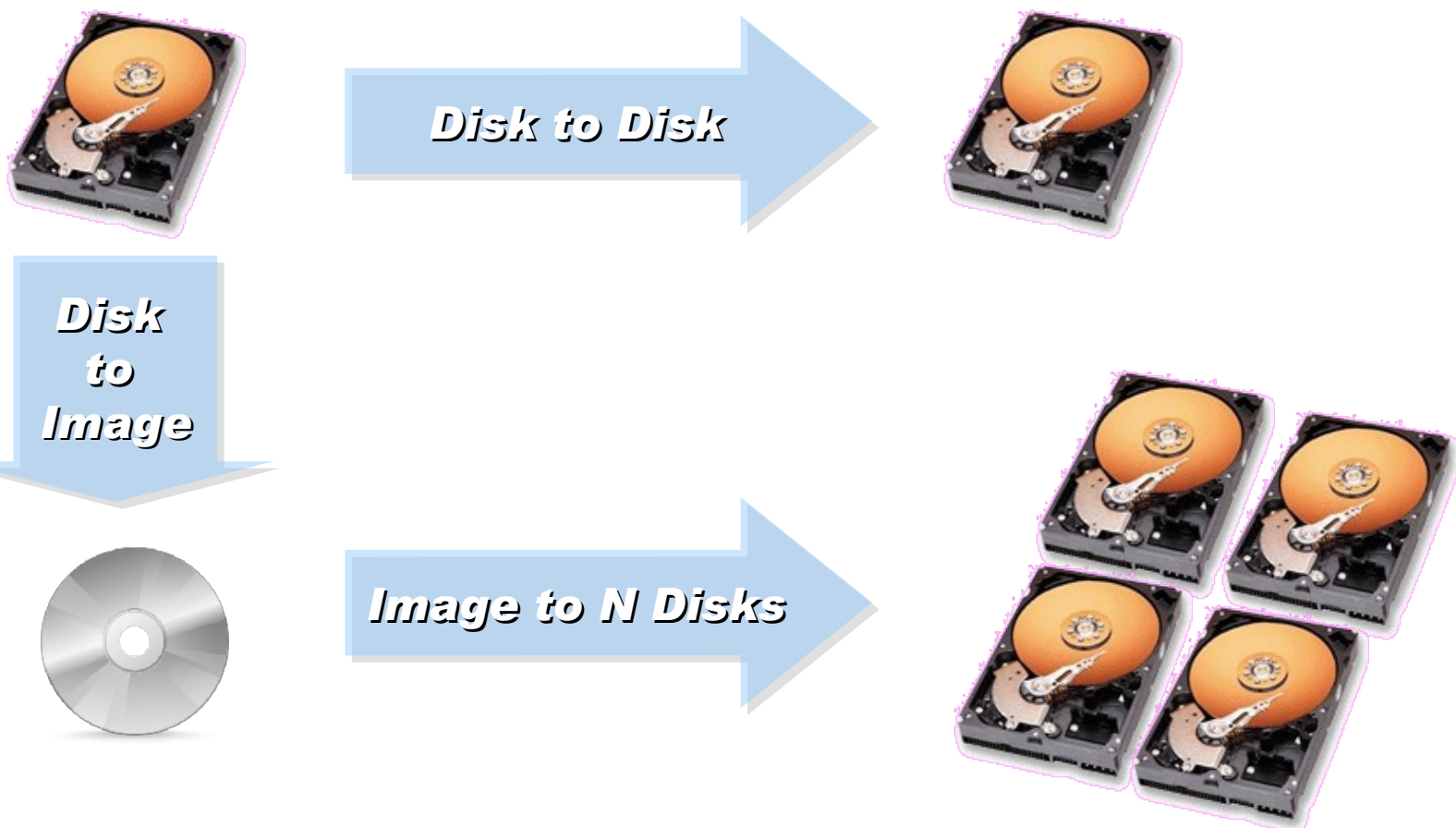
Diskless PC

Server

source: http://www.mren.com.tw

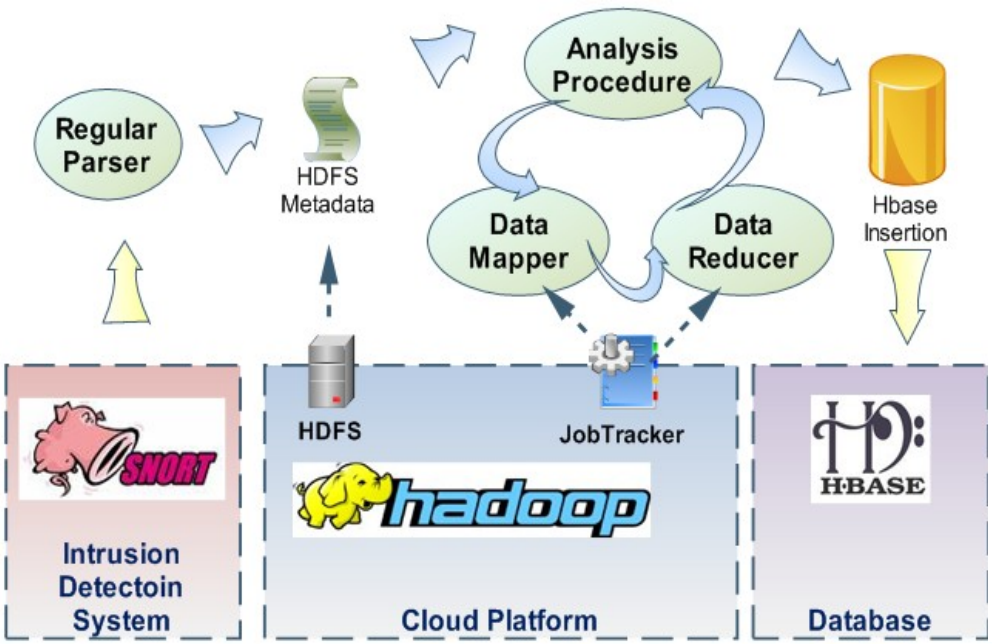
何謂再生龍 Clonezilla ??

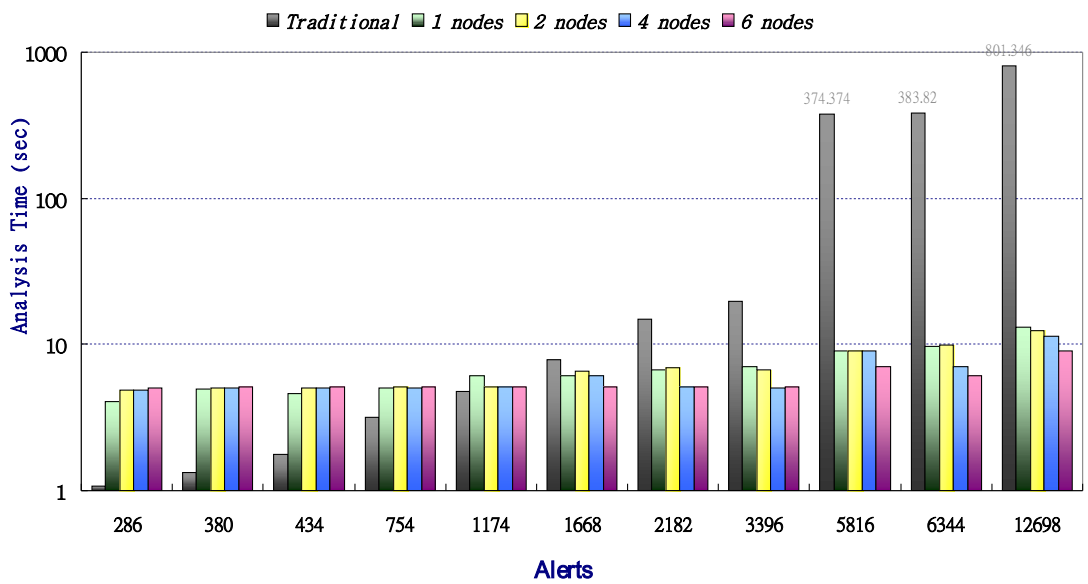
- **Clone (複製) + zilla = Clonezilla (再生龍)**
- 裸機備分還原工具
- **Norton Ghost** 的自由軟體版替代方案



軟體研發（1）：雲端入侵偵測分析系統
（IDS-log Cloud Analysis System，ICAS）

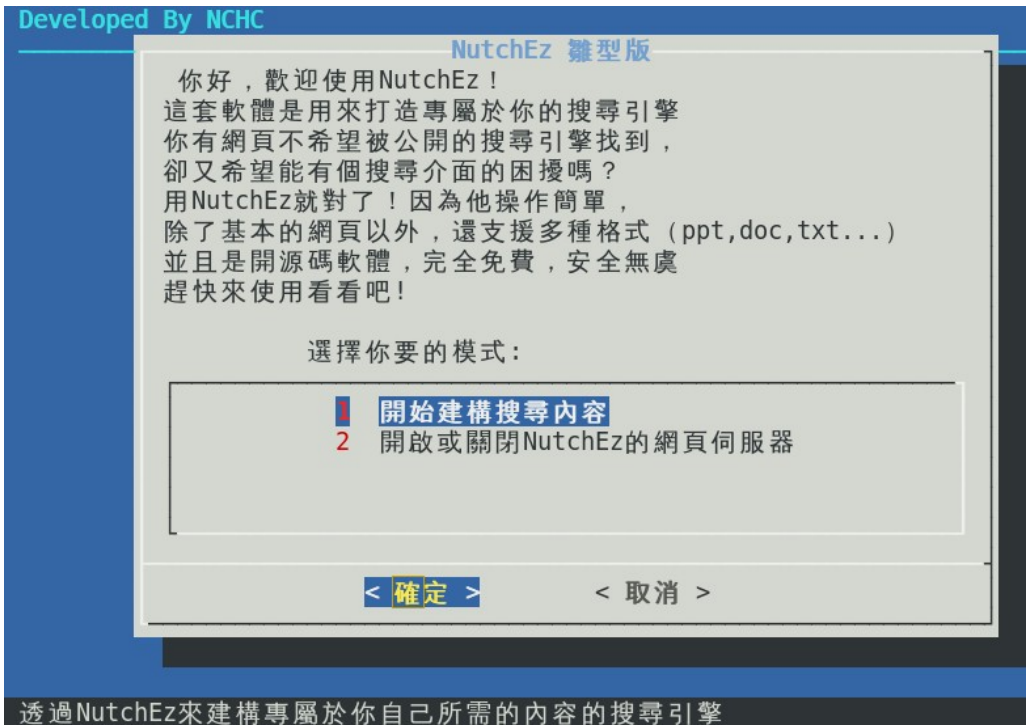
- 持續開發中，待整理套件
- 結合 **Hadoop** 與 **HBase** 來處裡 **SNORT** 產生的網路入侵報告。
- 雲端運算處理資料格式相似且資料量大的情況下，能展現其效益，並提供高容錯率、低獨占系統資源、多工作同時執行等能力
- **Key-Value** 資料庫寫入慢，讀取效能相對快，但缺乏其他語言支援。關聯式資料庫對小量資料的讀寫的效率較好，且支援的語言也較多。

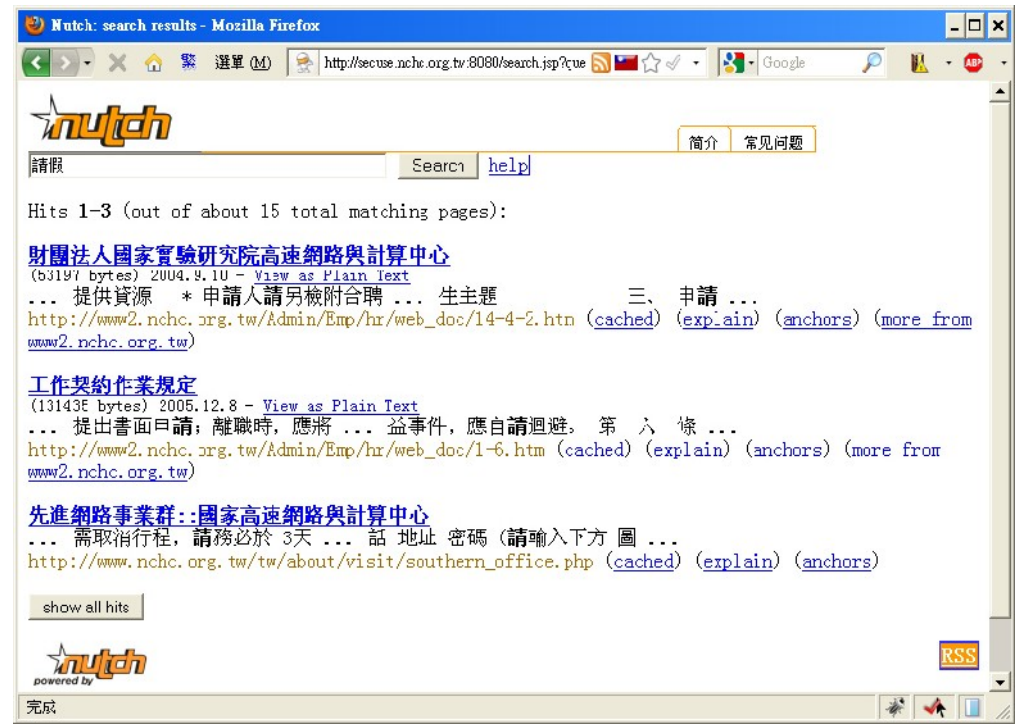




軟體研發 (2)：簡易架設個人搜尋引擎（NutchEz）

- 已釋出中文版套件 **http://trac.nchc.org.tw/cloud/wiki/NutchEz**
- 合適用來建立屬於組織內部的網頁搜尋引擎
- 核心為強大的 **Nutch** ，建於 **Hadoop** 上，貢獻：簡化安裝步驟
- 效能數據：搜尋 **699 doc, 322 pdf, 9 ppt, 13 odt.** 費時 **11** 分
- 系統負載： **CPU Quad 4 2.4G (19%)/ 4GB RAM (20%)**





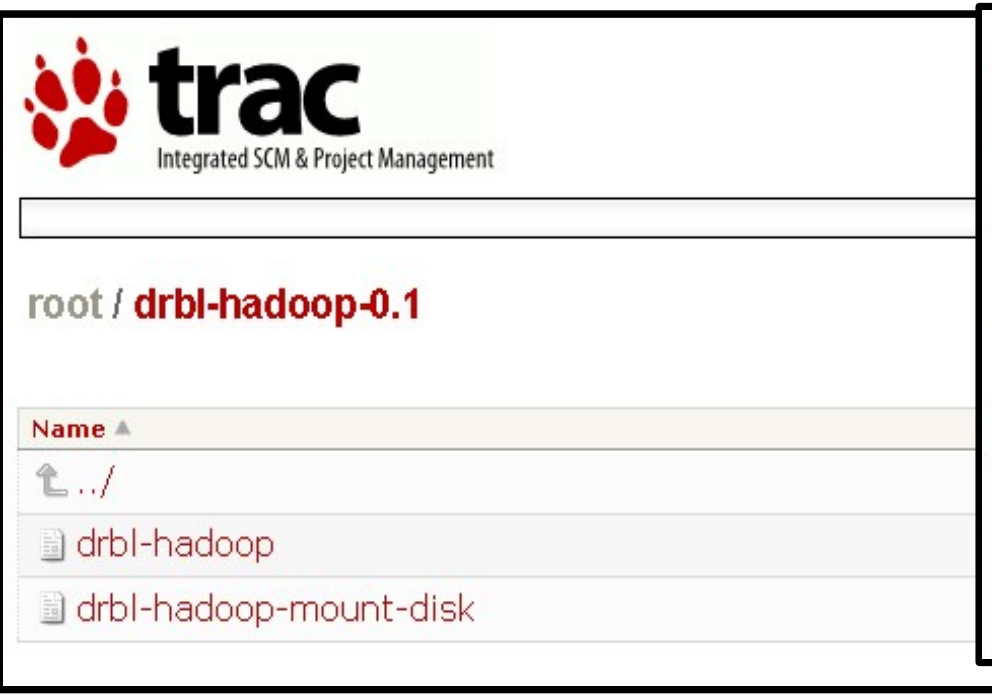
軟體研發 (3)：用企鵝龍佈署 Hadoop 雲端實驗環境

- 持續開發中，待整理套件
- **drbl-hadoop –** 掛載本機硬碟給 **HDFS** 用

```
svn co http://trac.nchc.org.tw/pub/grid/drbl-hadoop
```

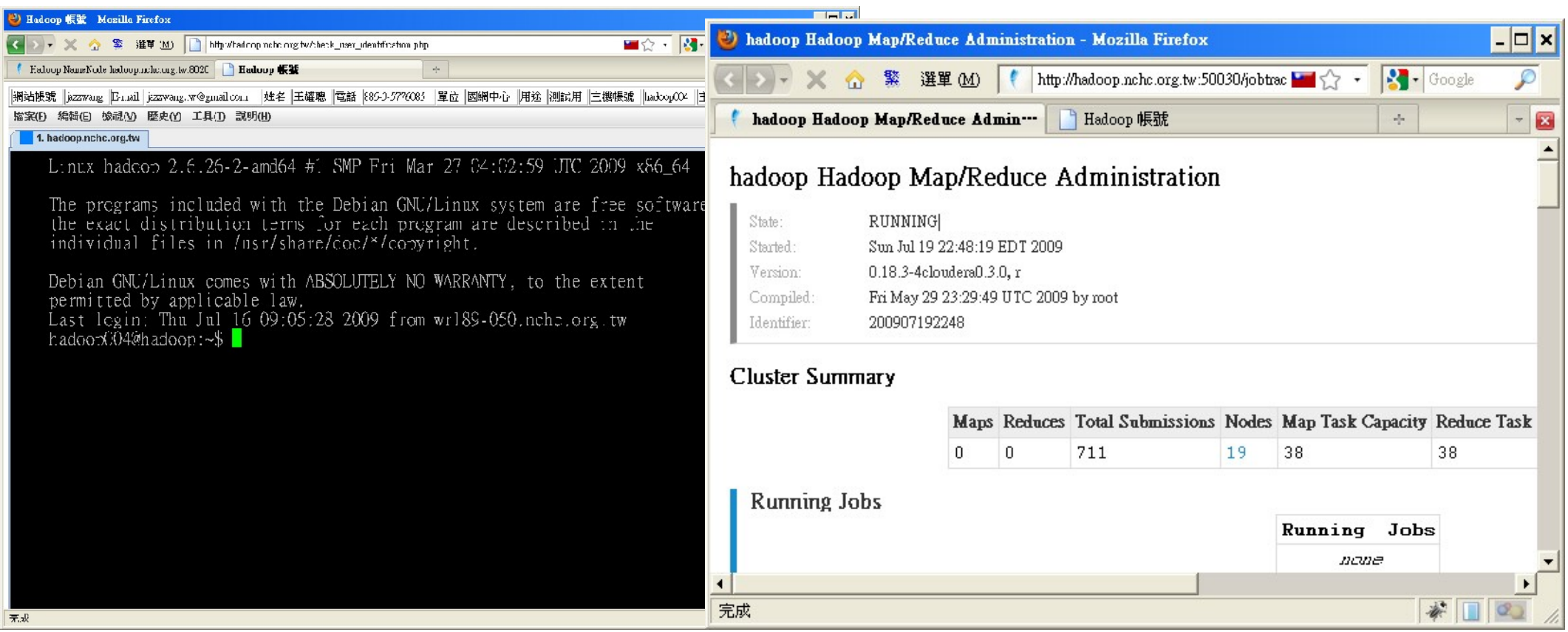
- **hadoop-register –** 註冊網站與 **ssh applet**

```
svn co http://trac.nchc.org.tw/pub/cloud/hadoop-register
```



實驗服務：hadoop.nchc.org.tw 多人雲端實驗叢集

- **DRBL Server - 1** 台 **(hadoop)** ，加大 **/home** 與 **/tftpboot** 空間。
- **DRBL Client - 19** 台 **(hadoop101~hadoop119)**
- 使用 **Cloudera** 的 **Debian** 套件，並針對多人環境進行讀寫權限加強。
- 使用 **drbl-hadoop** 的設定跟 **init.d script** 來協助部署
- 使用 **hadoop-register** 來提供使用者註冊與 **ssh applet** 介面



人才培育：雲端運算基礎課程(一~三)開放課程

- 雲端運算基礎課程(一)：Hadoop 簡介、安裝與實作
- 雲端運算基礎課程(二)：Xen 虛擬化叢集建置、管理與應用
- 雲端運算基礎課程(三)：Google App Engine 體驗課程
- 最新課程訊息與課程錄影詳見 http://trac.nchc.org.tw/cloud/


教育訓練網

新消息 | 會員專區 | 常見問題 FAQ | 住宿資訊 | 交通導引 | 電子報

課程總覽 » 最近六個月的課程 » **課程介紹**

課程編號： NE-2009-TH06
課程名稱： 雲端運算基礎課程(一)：Hadoop簡介、安裝與範例實作
課程領域： 電腦及網路
相關領域： 無
上課方式： 實體教室
上課地點： 竹科 B教室 竹
上課時間： 2009/11/24 (二) ~ 2009/11/25 (三) 09:30 ~ 16:30
上課總天數： 2天，共計 12 個小時
期截止(含)： 2009/11/22 (日) 17:00
費截止(含)： 2009/11/23 (一) 05:00
費截止(含)： 2009/11/23 (一) 17:00
提供午餐： 是
招生人數： 8 ~ 20 人
講師： 國家高速網路與計算中心 王耀聰 先生
國家高速網路與計算中心 陳威宇 先生


教育訓練網

新消息 | 會員專區 | 常見問題 FAQ | 住宿資訊 | 交通導引 | 電子報

課程總覽 » **課程介紹**

課程編號： NE-2009-CH05
課程名稱： 雲端運算基礎課程(二)：Xen虛擬化叢集建置、管理與應用
課程領域： 電腦及網路
相關領域： 無
上課方式： 實體教室
上課地點： 台中 電腦教室 A 中
上課時間： 2009/10/27 (二) ~ 2009/10/28 (三) 09:30 ~ 16:30
上課總天數： 2天，共計 12 個小時
期截止(含)： 2009/10/25 (日) 17:00
費截止(含)： 2009/10/26 (一) 05:00
費截止(含)： 2009/10/26 (一) 17:00
提供午餐： 是
招生人數： 8 ~ 20 人
講師： 國家高速網路與計算中心 涂哲源 先生
國家高速網路與計算中心 郭文傑 先生


教育訓練網

新消息 | 會員專區 | 常見問題 FAQ | 住宿資訊 | 交通導引 | 電子報

課程總覽 » 課程搜尋結果 » **課程介紹**

課程編號： NE-2009-CH06
課程名稱： 雲端運算基礎課程(三)：Google App Engine體驗課程
課程領域： 電腦及網路
相關領域： 無
上課方式： 實體教室
上課地點： 台中 電腦教室 A 中
上課時間： 2009/11/30 (一) 09:30 ~ 16:30
上課總天數： 1天，共計 6 個小時
期截止(含)： 2009/11/27 (五) 17:00
費截止(含)： 2009/11/27 (五) 17:00
費截止(含)： 2009/11/27 (五) 17:00
提供午餐： 是
招生人數： 8 ~ 20 人
講師： 國家高速網路與計算中心 鄭宗碩 先生
報名費用： 一般人士 1000 元
學生 500 元

對學界的幫助 (1): 實驗叢集間接促成研究成果

- 促成台大資工系資訊網路與多媒體研究所發表論文至 **ACM Multimedia 2009**
- 自 **2009** 年四月至 **2009** 年九月，雲端實驗叢集共註冊 **238** 人，服務 **37** 個學術單位 (計 **154** 人) ，、 **5** 個研究單位 (計 **21** 人) 、**19** 間業界公司 (計 **20** 人) 、**2** 所醫院 (計 **3** 人) 及不願提供單位的一般民眾計 **30** 人。累計於五個月內執行 **3341** 個 **Job** 。
- ***註冊人數排行前五大依序為交通大學、台灣大學、成功大學、中央大學與陽明大學***

Canonical Image Selection and Efficient Image Graph Construction for Large-Scale Flickr Photos

Liang-Chi Hsieh, Kuan-Ting Chen, Chien-Hsing Chiang, Yi-Hsuan Yang, Guan-Long Wu
Chun-Sung Ferng, Hsiu-Wen Hsueh, Charng-Rurng Tsai, Winston H. Hsu
National Taiwan University, Taipei, Taiwan
viirya@gmail.com, ktchen@cmlab.csie.ntu.edu.tw, {pacifistboy, affige}@gmail.com, {b95109, b95108, b95057, b95093, winston}@csie.ntu.edu.tw

ABSTRACT

Efficient image search clustering is prominent for image search engines for exponentially growing photo collections. In this work, we propose an image search clustering approach which selects multiple canonical images from image search results and constructs image clusters in real time on an image subgraph for the search results. The efficiency is achieved with the help of offline-computed image context graphs by distributed computing methods. Extending our prior works, we demonstrate the results of the proposed canonical image selection and preliminary outcomes of large-scale image graph construction in this proposal. We experiment in Flickr550 dataset, containing 540,321 Flickr photos.

Categories and Subject Descriptors: H.3.5 [Informa-

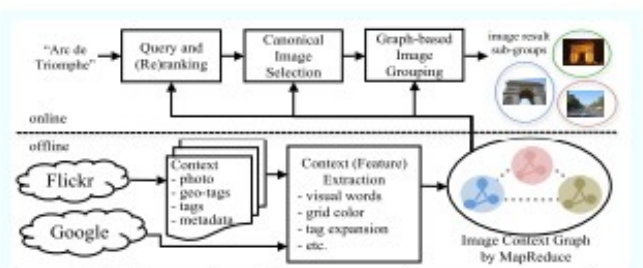
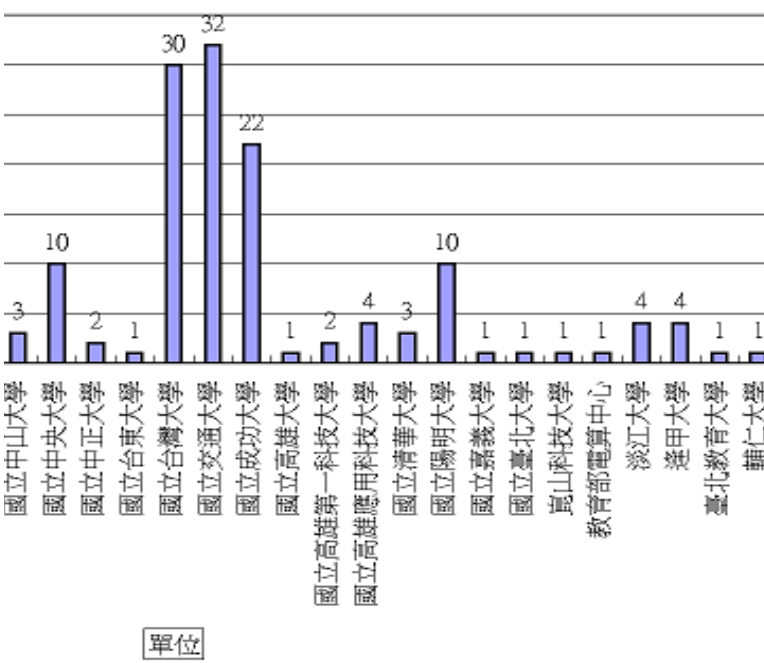


Figure 1: The system diagram: online query reranking, canonical image selection, and image result grouping are based on the image context graph, constructed offline by Hadoop MapReduce over rich context features in Flickr photos.

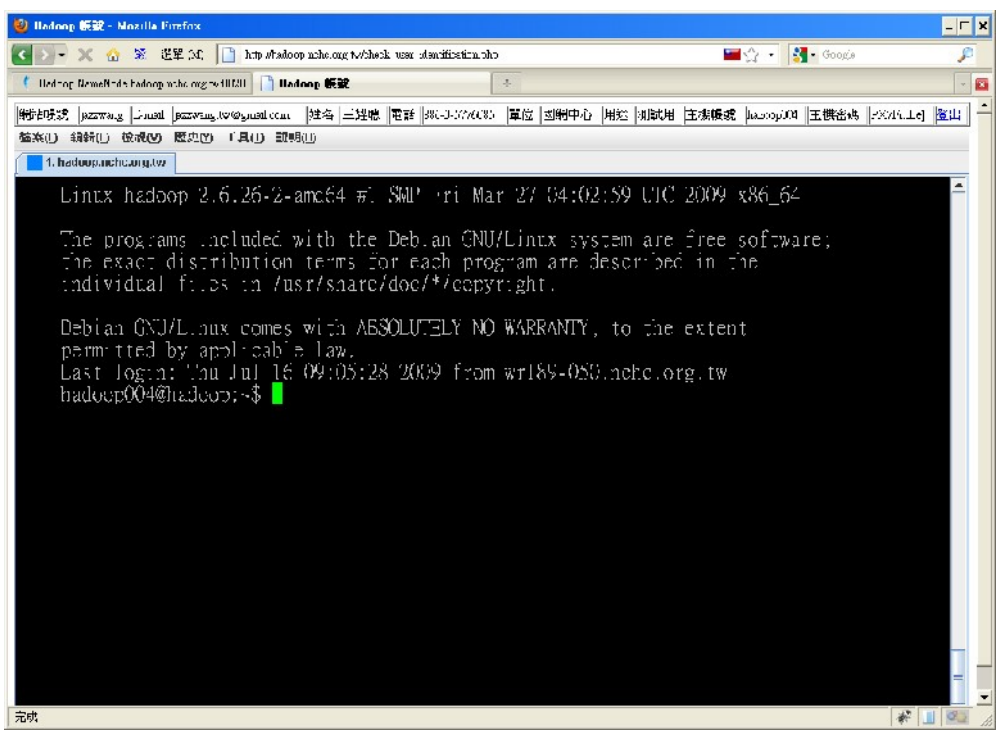


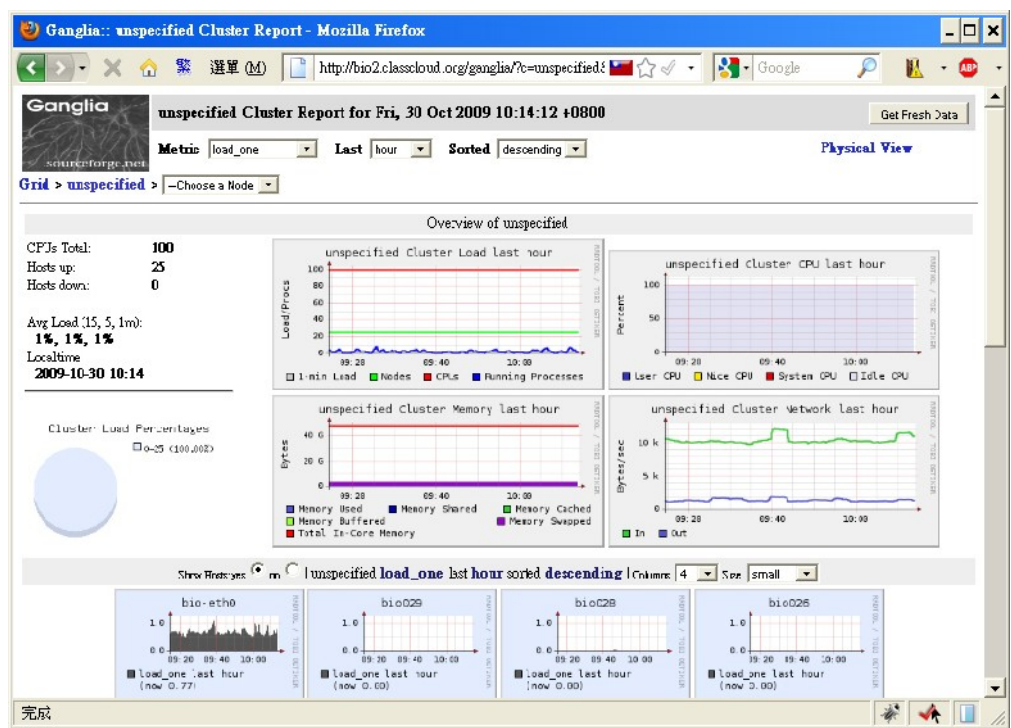
對學界的幫助 (2): 發展用企鵝龍佈署生物叢集的工具

- 持續整理中，待整理套件
- **drbl-biocluster –** 彙整安裝多人共用生物資訊叢集的批次檔

svn co http://trac.nchc.org.tw/pub/grid/drbl-biocluster

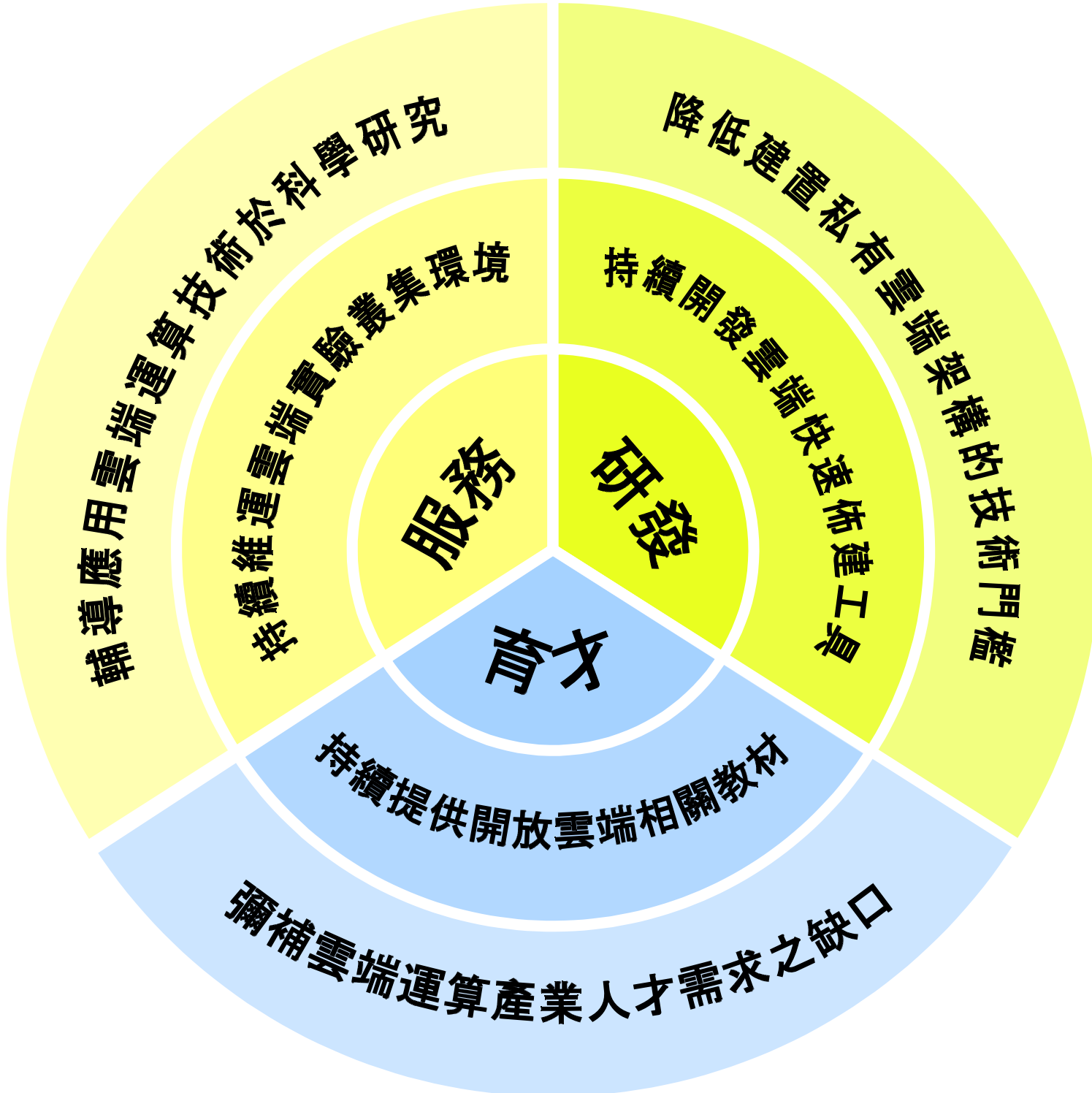
- 簡化安裝與測試生物資訊叢集常用軟體的程序： **DRBL** 、 **MPICH2** 、 **R** 、 **Rmpi** 、 **BioCondoctor** 、 **Ganglia** 、 **Nagios** 、 **AutoFACT** 、 **BLAST** 、 **SIM4** 、 **Clustal** 、 **PipMaker** 、 **Phylip** 、 **Eland** 、 **Velvet** 、 **Bowtie** 、 **SOAP**
- 成果： **http://bio2.classcloud.org**





對學界的幫助 (3): 更多開放教材－生物叢集、GAE...

- 陽明生資所 **97** 年度暑期學分班 格網及平行運算 (實驗課程) **http://trac.nchc.org.tw/course/**
- 陽明生資所 **98** 年度暑期學分班 格網及平行運算 (實驗課程) **http://bio.classcloud.org**
- 雲端運算基礎課程 (一) **Hadoop** 簡介、安裝與範例實作 **http://www.classcloud.org/media/**
- 「 **Ruby on Rails** 初學」電子書 **by** 鄭立竺 **http://nchcrails.blogspot.com**
- **Google App Engine** 電子書 **by** 鄭宗碩 **http://nchc-gae.blogspot.com/**
- **More to come .....**


服務
研發
育才
輔導應用雲端運算技術於科學研究
持續維運雲端實驗叢集環境
持續開發雲端快速佈建工具
降低建置私有雲端架構的技術門檻
持續提供開放雲端相關教材
彌補雲端運算產業人才需求之缺口

Questions?

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

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