Grid and Cloud Computing for Bioinformatics education and research

Tan Tin Wee, Mark de Silva, KS Lim
10 March 2010
NUS BioGrid and APBioGrid 2001-2002

http://www.bic.nus.edu.sg/biogrid/
http://www.startap.net/starlight/igrid2002/apbiogrido2.html
Won Supercomputing SC’2003 Most Geographically Distributed Project

TECHNICAL PROGRAM

OVERVIEW TECHNICAL PAPERS TUTORIALS PLENARIES PANELS POSTERS MASTERWORKS
BOF GRID 2003 AWARDS & PRIZES WORKSHOPS SC GLOBAL HPC CHALLENGE
MSI PROGRAM BANDWIDTH CHALLENGE

HPC Challenge

The HPC Challenge Awards will honor participants in two categories for innovative uses of high performance computing resources. This will provide opportunities for contestants to showcase

Most Geographically Distributed Application: Global Analysis of Arthropod Evolution

A global team of researchers dynamically link a large number of high performance computers and networks across five continents to run fastDNAml, a parallel program for inference of phylogenetic relationships based on DNA sequences. They simultaneously demonstrate a global-scale distributed computing application and at the same time perform meaningful analysis of an important biological problem.
Integrated Biolaboratory Manufacturing and Services System 2004-2005

A*STAR SERC Thematic Strategic Research Programme on “Integrated Manufacturing & Services Systems”

APBioGrid, APBioBox, APBioKnoppix, Laboratory Workflow Integration and BioManufacturing Workflow in Life Sciences

PI: Tan Tin Wee, National University of Singapore
CO-PIs: I A Karimi (ChemEng), Lim Teck Sin (KooPrime), Lim Beng Siong (SIMTech)

For acceleration of specific areas of manufacturing, in particular biomanufacturing.

www2.jp.apan.net/meetings/bangkok2005/presentation/tantinwee-APAN19-eScienceBKKJan05.ppt
Workflow Integration in Laboratory Automation

Exemplar 2: Large scale Bioinstrument/Device Integration possible
Introduction

We started exploring Cycle Harvesting Grid Computing since 2003. In that year, a testbed with 120 computers was set up using the Platform Computing ActiveCluster middleware. At the end of 2004, a collaboration was established between NUS and Singapore Computer Systems (SCS) to expand the Cycle Harvesting Grid (code named TCG@NUS, which stands for Tera-Scale Campus Grid at NUS) to 800 computers in Phase 1 of the project and to 3000 PCs/servers subsequently.
Grid Apps on TCG@NUS Grid

Applications Enabled on TCG@NUS Grid

The following applications have been enabled on TCG@NUS:

- **AutoDock**: Tool for automated docking of flexible Ligands to Macromolecules
- **BLAST**: Tool for comparing nucleotide or protein sequences to databases & calculating the statistical significance of matches
- **HMMER**: A sequence analysis tool using profile hidden Markov models
- **MODELLER**: A program for comparative protein structure modeling by satisfaction of spatial restraints
- **Parametric**: A computer program to run
- **PovRay**: A high-quality tool for creating images

2. **Comparative Modelling of Protein Structures**
   - Application software: Modeller
   - Run 9999 jobs consisting of protein sequences with about 200 amino acid residues
     - Took 42 days to complete on a 600MHz PC.
     - Took about 3 days to complete on 100 computers in TCG@NUS.
   - **Speedup**: About 14 times faster.

3. **Protein Level Sequence Matching**
   - Application software: tblastx
   - 22K sequences have to be searched against 3,461,799 sequences
     - Estimated to take about a month to complete on a 16-CPU server.
     - Took 20.5 hrs compute time on TCG@NUS and 13 hrs reassembly time on a standalone PC.
   - **Speedup**: About 20 times (320 times with reference to a single CPU).
AeroGenome – Crunching on a PC Grid

- Aim: To study bacterial microorganisms in the air in Singapore.
- Motivation: Air plays significant role in public & environmental health.
- Initial dataset of 20,000 DNA sequences were matched against database of 3 million known sequences. Single run on isolated compute resources took an entire month.
- Same process took less than 2 days on TCG@NUS.
- Future: Scale up study using larger datasets.

[Courtesy of Dr. Patrick Tan, GIS]
BioSLAX is a new live CD/DVD suite of bioinformatics tools that has been released by the resource team of the BioInformatics Center (BIC), National University of Singapore (NUS). Bootable from any PC, this CD/DVD runs the compressed SLACKWARE flavour of the LINUX operating system also known as SLAX.

SLAX is becoming the live CD/DVD of choice because of its ability to modularize almost any application and plug it into the system on the fly. The system can also be installed to USB thumbdrives or directly to the PC as a regular Linux using the BioSLAX installer provided.

How is it built?

There are two parts to the SLAX build, the core system and the individual modules. The core system is the OS itself and basic tools, usually referred to as the 'base', while the modules are the individual utilities that a user wants to have on their system. Since these modules can be put in or removed prior to creating the CD/DVD/USB, the system is fully modular and easily customisable.

All credit and thanks go to Tomas Matejicek, the developer of the SLAX and Linux Live Scripts.
BioSlax with VMplayer

Bioinformatics software in a virtual machine

| Packages                                      | Applications                                   | Programming languages – C, perl, python etc
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Packages</td>
<td>Applications</td>
<td>GUI interfaces</td>
</tr>
<tr>
<td>Applications</td>
<td>Programming languages – C, perl, python etc</td>
<td>Servers – apache, mysql</td>
</tr>
<tr>
<td>Programming languages – C, perl, python etc</td>
<td>GUI interfaces</td>
<td>No need to reboot, just install VMplayer, run</td>
</tr>
<tr>
<td>GUI interfaces</td>
<td>Servers – apache, mysql</td>
<td>BioSlax vmx file and create</td>
</tr>
<tr>
<td>Servers – apache, mysql</td>
<td></td>
<td>a new instance of a BioSlax virtual machine</td>
</tr>
</tbody>
</table>

| For comments or assistance, please e-mail bioslax@bioslax.com. |

The BioSLAX BioInformatics LiveCD Suite was developed at the BioInformatics Center, National University of Singapore, April 2006 by Mark De Silva, Lim Kuan Siong and Tan Tin Wee.
Desktop-as-a-Service  BioSlax+VM+MokaFive
BioSlax Addable Modules

Compile $\rightarrow$ dir2lzm $\rightarrow$ copy into virtual HDD /mnt/sda1/module
Grid-enabling BioSlax: UD MP agent in BioSlax

- History – TCG@NUS
- UDBlast and UDHMMer
- Rapid recruitment of machines
- No need for user installation
- Push or pull instances
- Enduser joining of Grid
- In the process of embedding gLite inside BioSlax
- Standard Linux with Bio applications
Grid-Cloud initiatives

- GoGrid.com
- Amazon EC2
- BIC BioCloud – Citrix Xen Hypervisor
- Participation in EUAsiaGrid through Asian Bioinformation Centers (ABC) initiative
GoGrid.com

- Used to run a windows server running LAMS
- 250 students per semester accessing this platform
Amazon EC2

• Collaboration as beta tester for latest Amazon technologies: NDA in progress
• BioSlax 7.5 to be configured to run as an Amazon image for as a cloud service
• Creation of Slax 5.1.8.1 Amazon Machine Image
• Amazon Machine Image, S3 storage of the image, Instance and ssh session into Slax 5.1.8.1 works

Slax 5.1.8.1 Amazon Machine Image, S3, Instances
SSH into Slax 5.1.8.1 Instance

```
Using username "root".
root@ec2-67-202-37-165.compute-1.amazonaws.com:~$ ls
Desktop/  Set\ IP\ address
root@slax:~$ ls
Desktop/  Set\ IP\ address
root@slax:~$ ls
Desktop/  Set\ IP\ address
root@slax:~$ ls
Desktop/  Set\ IP\ address
root@slax:~$ ls
Desktop/  Set\ IP\ address
root@slax:~$ ls
Desktop/  Set\ IP\ address
```

```
root@slax:~$ ifconfig eth0
eth0  Link encap:Ethernet  HWaddr 12:31:3B:02:2C:66
     inet addr:10.243.47.148  Bcast:10.243.47.255  Mask:255.255.254.0
     UP BROADCAST NOTRAILERS RUNNING MULTICAST  MTU:1500  Metric:1
     RX packets:2374 errors:0 dropped:0 overruns:0 frame:0
     TX packets:2418 errors:0 dropped:0 overruns:0 carrier:0
     collisions:0 txqueuelen:1000
     RX bytes:266376 (260.1 KiB)  TX bytes:443248 (432.8 KiB)
```

```
root@slax:~$ nslookup 10.243.47.148
Server: 172.16.0.23
Address: 172.16.0.23#53

Non-authoritative answer:
148.47.243.10.in-addr.arpa name = ip-10-243-47-148.ec2.internal.
```
Citrix XenServer 5.5 Update 2 was tested and verified that it is able to boot BioSlax 7.5 in Hardware Virtual Machine (HVM) mode.

NX Client used to connect to the virtual machines.

XenServer host is running dual Intel Xeon Quad-core 3.0 GHz CPUs with 6 Gb of RAM.

Each XenServer can run 10 Virtual Machines at the moment – limitation due to memory as each VM is allocated 550 MB.
Bioinformatics Teaching

- 100 students doing year 2 miniprojects
- 20 student doing year 3 miniprojects
- 10 Graduate students
BIC Citrix XenServer

NxClient for Windows

Console Access to Virtual Machine
WebBased Virtual Machine Control Portal for easy access

- Similar basic functions to Citrix XenCenter
- Displays various Virtual Machine details in a single web page for quick overview
- Virtual Machines can be started, shutdown or rebooted using the portal
- Console access is also available via the portal
- Future plans: userbased login; owner control shutdown-reboot; resource control.
Next Generation BioMirror.net
“BioDataGrid”

- BioMirror 1998 – 12 countries in Asia Pacific; 1 Terabyte of compressed biodatabases
- APBioGrid 2002 - Grid computing installable in three steps; Funded by International Development Research Centre IDRC PAN-Asia Grant
- Azereus P2P database distribution system – Grant funded by IDRC in 2007-2009 Thailand, KOBIC, NUS collaboration.
- Asian Bioinformation Centers – APBioNet-AIMBN-eIMBL project
Asian Bioinformation Centers and EUAsiaGrid

- Collaborate with ASGC, Taiwan and National Yang Ming University
- Create next generation BioMirror2
- AuthorID system
- DocID Depository for Bioinformatics published datasets
- Minimum Information about a Bioinformatics Investigation MIABi standards
- Initiate Computationally intensive projects for complex diseases.
Welcome to CloudAsia 2010

The aggregation point in Asia for Grid and Cloud activities, conferences & test-beds

Welcome

This series, with the Grid and Cloud Computing themes, caters to everyone keen on Grid and Cloud, from interested researchers, practitioners, to industry players in Grid and Cloud Computing.

Program

CloudAsia 2010 will be held in the week of 3 - 7 May 2010. The program will cover industry & R&D presentations, discussions, and more information will be posted as program details are finalized.

Highlights

Case Studies on Grid and Cloud
Business in the Cloud
Users Training

Keynote Speakers

Keynote speakers will include thought leaders and Chief Technology Officers in the fields of Grid and Cloud Computing, from leading organizations in Singapore and overseas. More details will be posted in due course.

Past Conferences

- GridAsia 2005
- GridAsia 2006
- GridAsia 2007
- GridAsia 2008
- GridAsia 2009

Contact:
Tan Tin Wee tinwee@bic.nus.edu.sg
Mark De Silva mark@bic.nus.edu.sg
KS Lim kslim@bic.nus.edu.sg