iGridPortal

and

Biological Problems

Authors:
Hoang Le Minh, Vo Duc Cam Hai, **Nguyen Thi Thanh Nhan**
Tran Linh Thuoc, Do Anh Tuan, Vo Cam Quy

Viet Nam National University
Hochiminh City, University of Natural Sciences

BioGrid Workshop

Overview

- **Introduction and Motivation**
- **Demo**
- **iGridPortal**
  - **Grid Information Channel**
  - **GridBlast Channel**
- **Conclusions**

Introduction and Motivation

- ***Grid provides us a lot of computing services and resources. However, Grid service interfaces maybe good for programmers, but not for users.***
- ***Portal is a suitable framework to integrate Grid resources and help users communicate easily with the Grid infrastructure via web browsers***

Motivation

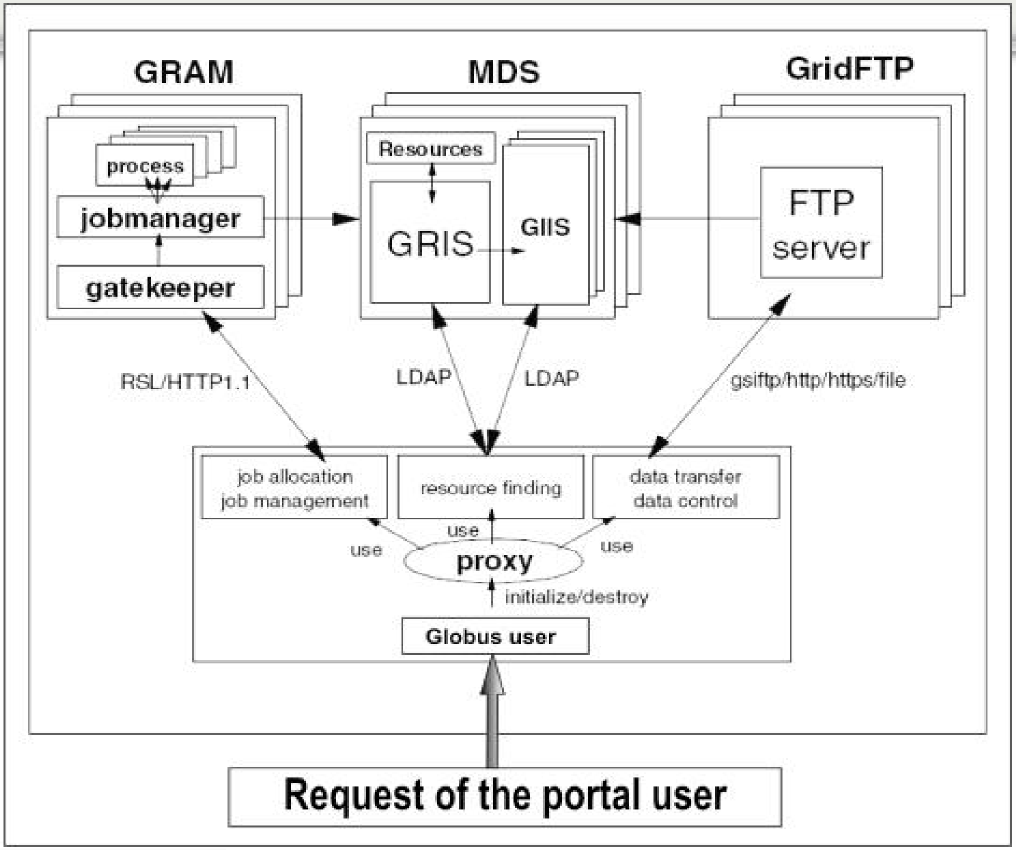
- *To demonstrate and develop best practice in using Grid technologies to support and promote collaborative research projects,*
- *To work with the bioinformatics to demonstrate the power of the Grid*
- *To create biogrid applications with an easy-to-use interface presented to users in terms of application science, not in terms of complex distributed computing protocols*

➔ *Networked Computing Lab (NCLab) and Cooperation Bioinformatics Lab (CBLab) have worked together to set up the grid environment to be used to deploy biological applications such as sequence alignment, gene prediction, and so more with portal interface - iGridPortal*

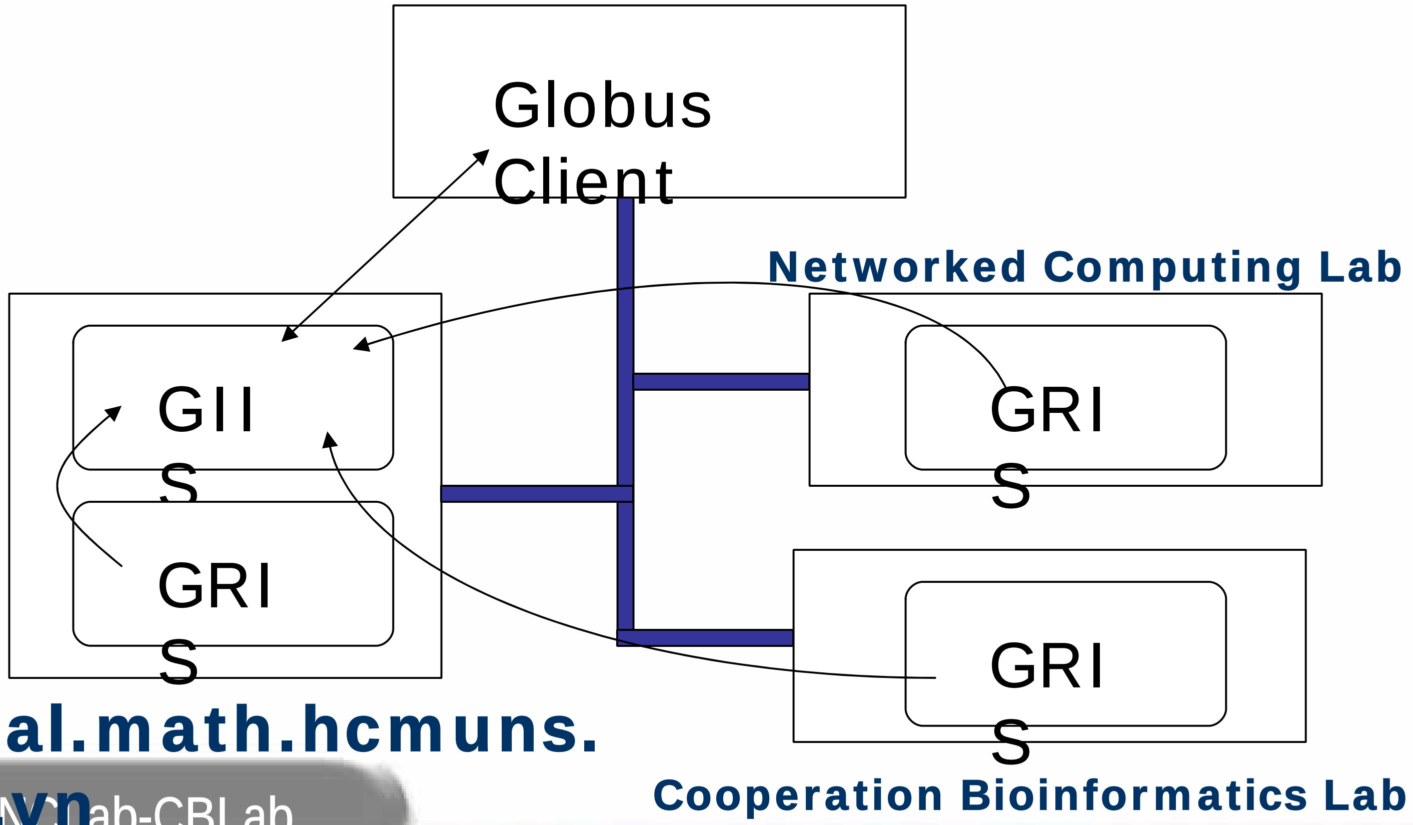
iGridPortal

- Built at the computer laboratory of Department of Mathematics and Computer Science, University of Natural Sciences, HCMC, Vietnam with the support of following software:
  - Globus Toolkit (www.globus.org)
  - uPortal (www.uportal.org)
  - JavaCog (http://www-unix.globus.org/cog/java/)
  - MyProxy (http://grid.ncsa.uiuc.edu/myproxy/)
  - Resin (www.caucho.com)
- Be a web based application framework with necessary software to interact with the grid services and resources
- Provide single point of access to distributed information and services

iGridPortal

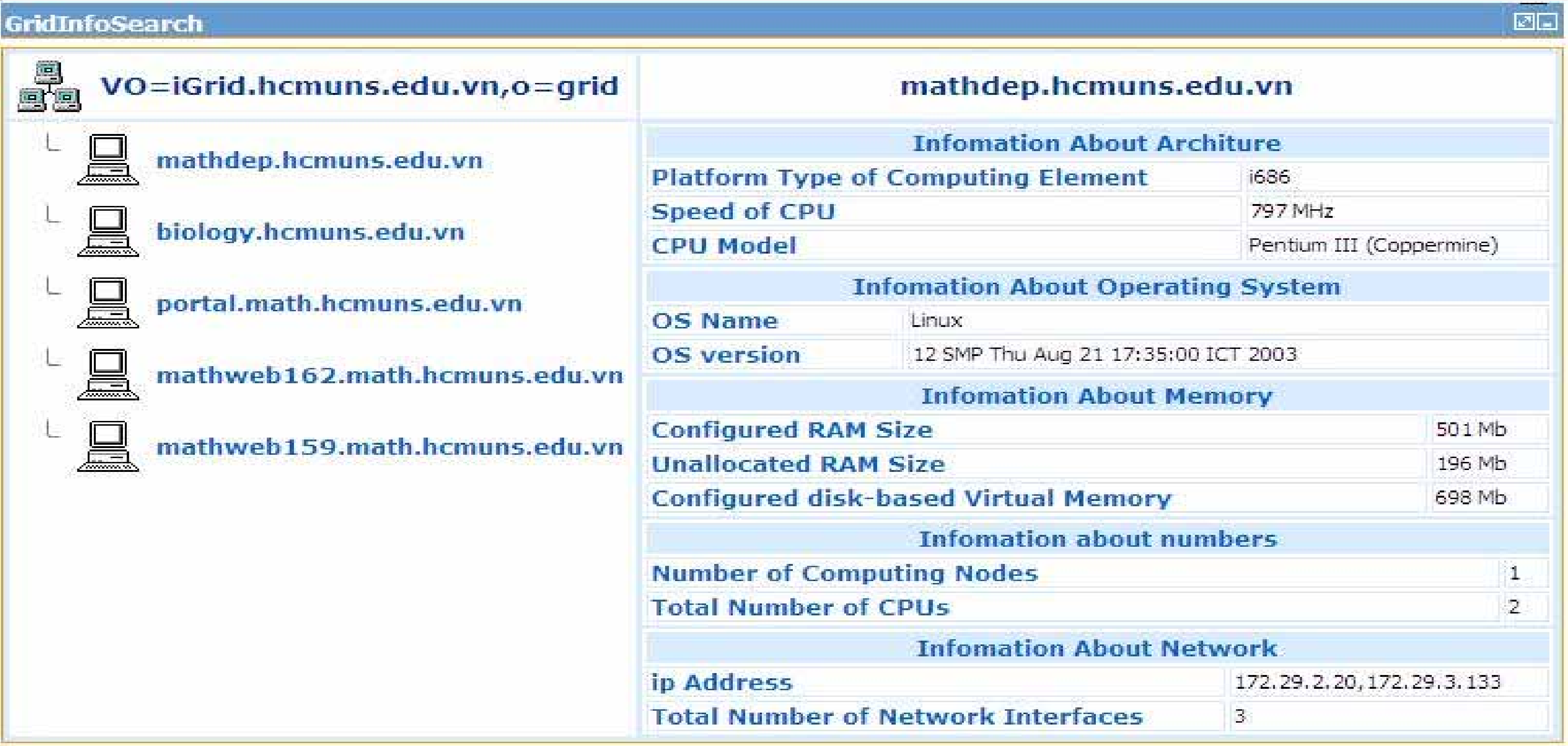


Monitoring and Discovery Service (MDS) configuration



portal.math.hcmuns.edu.vn

Grid Information Channel



GridInfoSearch

VO=iGrid.hcmuns.edu.vn,o=grid

- mathdep.hcmuns.edu.vn
- biology.hcmuns.edu.vn
- portal.math.hcmuns.edu.vn
- mathweb162.math.hcmuns.edu.vn
- mathweb159.math.hcmuns.edu.vn

mathdep.hcmuns.edu.vn

Infomation About Architure	
Platform Type of Computing Element	i686
Speed of CPU	797 MHz
CPU Model	Pentium III (Coppermine)

Infomation About Operating System	
OS Name	Linux
OS version	12 SMP Thu Aug 21 17:35:00 ICT 2003

Infomation About Memory	
Configured RAM Size	501 Mb
Unallocated RAM Size	196 Mb
Configured disk-based Virtual Memory	698 Mb

Infomation about numbers	
Number of Computing Nodes	1
Total Number of CPUs	2

Infomation About Network	
ip Address	172.29.2.20,172.29.3.133
Total Number of Network Interfaces	3

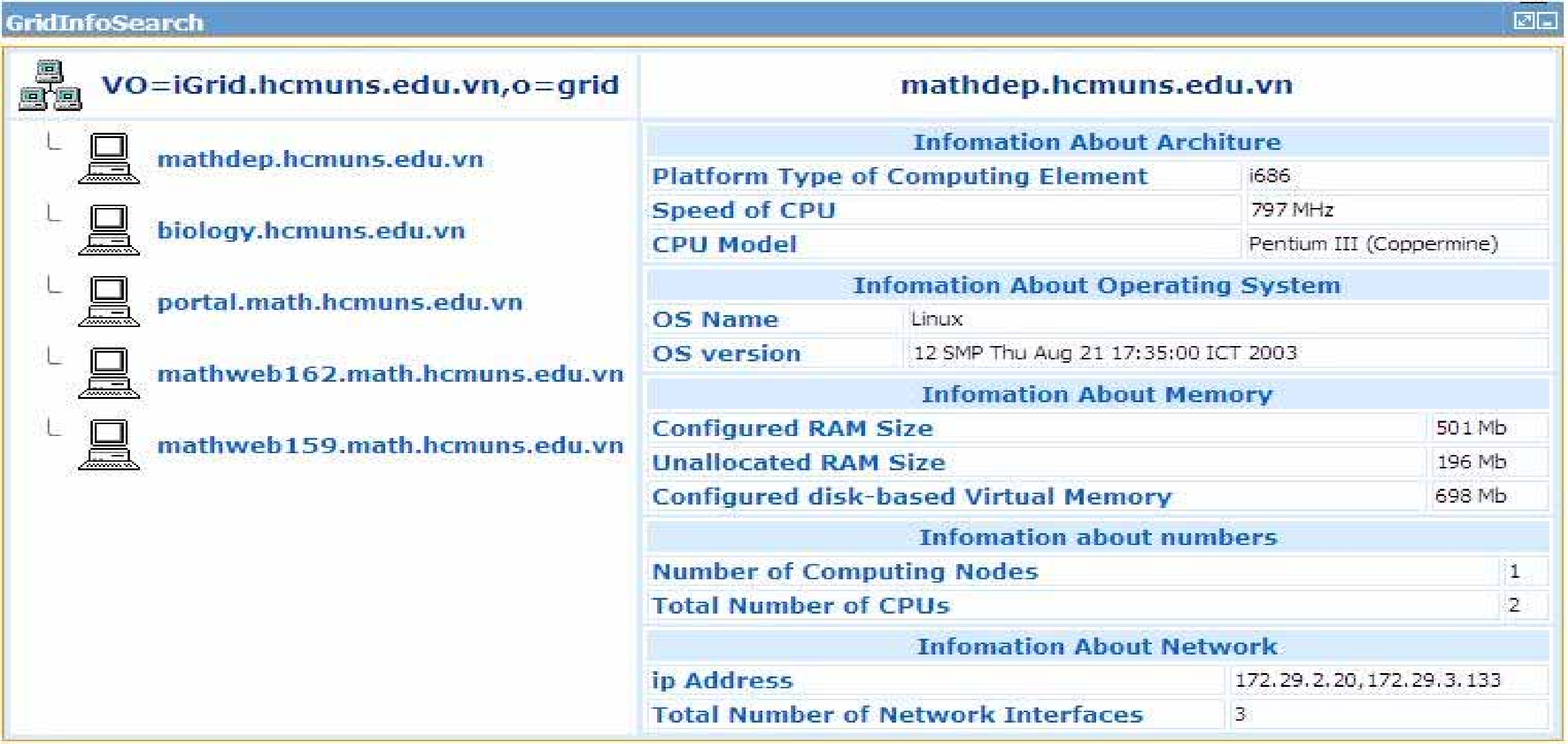
GridInforSearch

- At the first time when GridInformation channel is selected , it connects to MyProxy server to get proxy of the globus user *iportal*. We can assume that portal user and globus user are the same. If GridInformation has not received the proxy, GridInformation channel can not be used.
- We have used classes of Java Commodity Kit (JavaCoG) to build GridProxy class to get an valid proxy or create a proxy if it is necessary.
- After having a valid Globus proxy, an instance of GridInforSearch is created to obtain vital information about the grid and grid resource in XML format.

GridInforSearch result

```
<computer hn="mathweb162.math.hcmuns.edu.vn" show="false">
        <architure>
                <Mds-Computer-platform>i686</Mds-Computer-platform>
                <Mds-Cpu-speedMHz>1002</Mds-Cpu-speedMHz>
                <Mds-Cpu-model>Pentium III (Coppermine)</Mds-Cpu-model>
        </architure>
        <os>
                <Mds-Os-name>Linux</Mds-Os-name>
                <Mds-Os-version>1 Thu Mar 13 17:54:28 EST 2003</Mds-Os-version>
        </os>
        <memory>
                <Mds-Memory-Ram-Total-sizeMB>121</Mds-Memory-Ram-Total-sizeMB>
                <Mds-Memory-Ram-freeMB>33</Mds-Memory-Ram-freeMB>
                <Mds-Memory-Ram-Total-SizeMB>121</Mds-Memory-Ram-Total-SizeMB>
                <Mds-Memory-Vm-sizeMB>1513</Mds-Memory-Vm-sizeMB>
        </memory>
        <numbers>
                <Mds-Computer-Total-nodeCount>1</Mds-Computer-Total-nodeCount>
                <Mds-Cpu-Total-count>1</Mds-Cpu-Total-count>
        </numbers>
        <network>
                <Mds-Net-name>eth0</Mds-Net-name>
                <Mds-Net-name>lo</Mds-Net-name>
                <Mds-Net-addr>172.29.3.162</Mds-Net-addr>
                <Mds-Net-Total-count>2</Mds-Net-Total-count>
        </network>
</computer>
```

Grid Information Channel



VO=iGrid.hcmuns.edu.vn,o=grid

- mathdep.hcmuns.edu.vn
- biology.hcmuns.edu.vn
- portal.math.hcmuns.edu.vn
- mathweb162.math.hcmuns.edu.vn
- mathweb159.math.hcmuns.edu.vn

mathdep.hcmuns.edu.vn

Infomation About Architure	
Platform Type of Computing Element	i686
Speed of CPU	797 MHz
CPU Model	Pentium III (Coppermine)

Infomation About Operating System	
OS Name	Linux
OS version	12 SMP Thu Aug 21 17:35:00 ICT 2003

Infomation About Memory	
Configured RAM Size	501 Mb
Unallocated RAM Size	196 Mb
Configured disk-based Virtual Memory	698 Mb

Infomation about numbers	
Number of Computing Nodes	1
Total Number of CPUs	2

Infomation About Network	
ip Address	172.29.2.20,172.29.3.133
Total Number of Network Interfaces	3

GridBlast Channel

- **BLAST**
- **Databases**
- **Characteristics**
- **The process**
- **Interface**

Basic Local Alignment Search Tool (BLAST)

- In biology, a common application of sequence alignment is searching a database or sequences that are similar to a query sequence. By far, the most popular tool for searching sequence databases is a program called BLAST (Basic Local Alignment Search Tool). It performs pair wise comparisons of sequences, seeking regions of local similarity.
- GridBlast Channel in iGridPortal allows researchers to use BLAST software on Grid computing systems through a portal.
- It is hard to do alignment job on single machine if the databases is too large, such as human genomic. So, we have divided these databases, and storing them in different servers.

GridBlast channel

- NCBI allows one to perform BLAST searches online (http://www.ncbi.nlm.nih.gov/BLAST/).
- However, there is a need for customized interfaces rather than using the fixed interface provided by NCBI. In addition, we need to set up portals for running BLAST within virtual organizations, using dedicated computing resources, especially databases from our Biology Department.

Databases

- DNA and protein databases are downloaded from ftp://ftp.ncbi.nih.gov/blast/db/ and then formatted with BLAST Tool.
- There are databases:
  - human_genomic.tar.gz
  - swissprot.gz
  - alu.n.gz
  - alu.a.gz
  - nr.gz
  - en_nr.gz

The process for solving alignment problem

- **Step 1. At the first time when GridBlast channel is selected , it connects to MyProxy server to get proxy of globus user (*iportal*). We can assume that uPortal user and globus user are the same. If GridBlast channel has not received the proxy, GridBlast channel can not be used.**

The process for solving alignment problem (Cont.)

- **Step 2. GridBlast queries the gridblast database to find available databases. These databases will be presented in GridBlast channel for the user to select.**
- **Step 3. When the user submits job, GridBlast queries the distributed database to get information about child databases of selected databases.**
- **Step 4. With all information of each child database, a thread will be created containing these information, including: the name of test file, the name of child database, host name and gridblast directory. Every thread registers with a monitor that will log status of job. All threads are started.**

The process for solving alignment problem (Con.t)

- **Step 5. After that, the thread will do two following jobs: using GridFTP to send test sequence file on the host containing child database, and using GRAM to run the script runblast.sh.**

- **Step 6. Now, every thread listens the status of job on remote host, and monitors log status of all thread running. When job finish, the thread use GridFTP to copy the result file to portal server. These result files are collected and merged.**

The process for solving alignment problem (Cont.)

- The user will receive a notice email when the job complete
- Then, the user simply download output files from iGridPortal to local workstation.

Conclusions

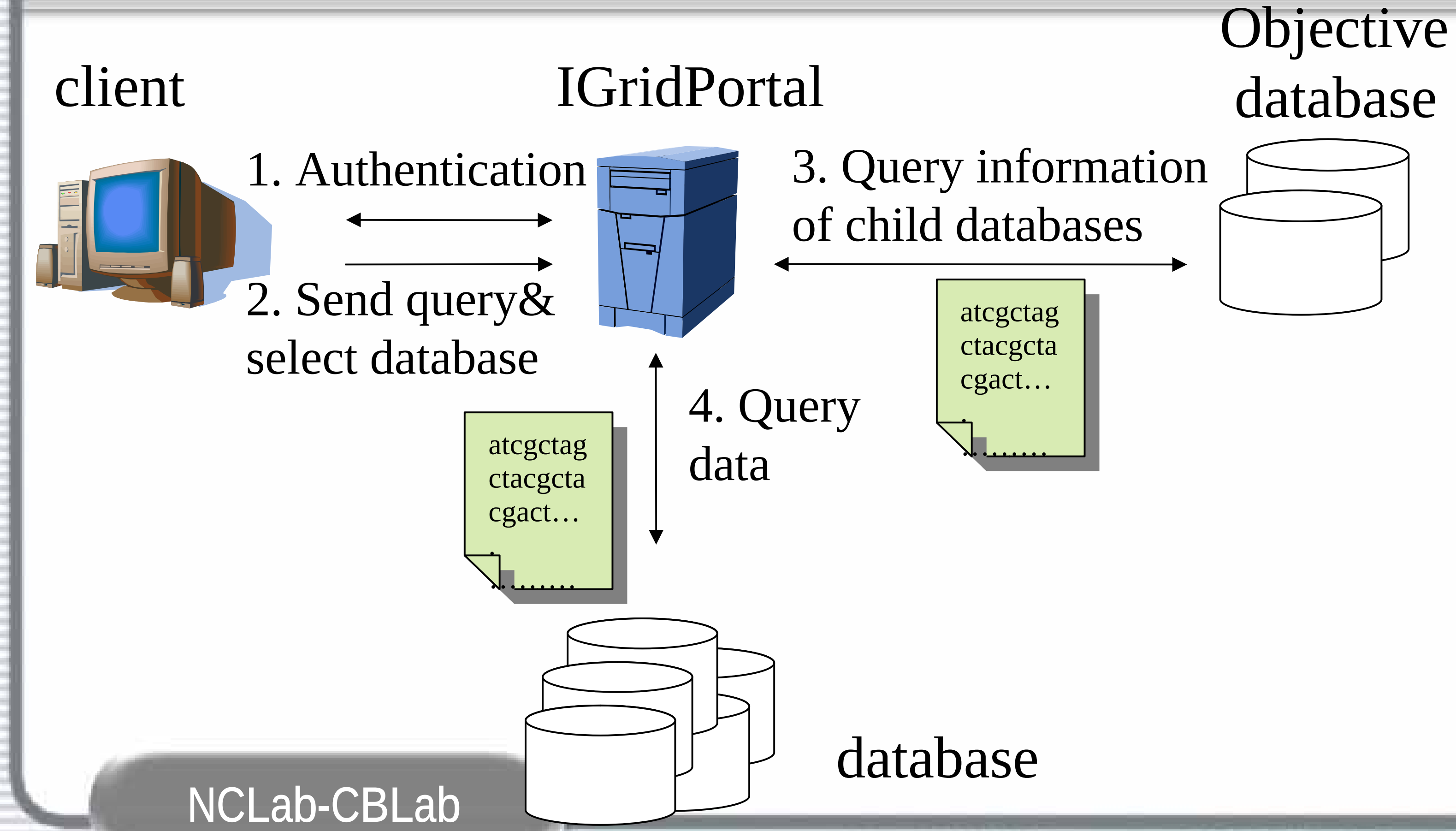
- **I have described iGridPortal, particularly Grid Information channel and GridBlast Channel. It has an easy-to-user interface, utilizes advantages of grid computing, especially reduces cost and time to solve problems.**
- **Now, we are planning to increase the number and types of biological applications.**

Acknowledgments

- **Thank to:**
  - **The organizers**
  - **All who sent us comments on our program**
  - **Pro. Peter Arzberger and Pro. Hoang Le Minh**
  - **Cybermedia Center, Osaka University**

iGridPortal

Thank you for your attention!



client

IGridPortal

Objective database

1. Authentication

2. Send query& select database

3. Query information of child databases

4. Query data

database

iGridPortal

and

Biological Problems

Authors:
Hoang Le Minh, Vo Duc Cam Hai, **Nguyen Thi Thanh Nhan**
Tran Linh Thuoc, Do Anh Tuan, Vo Cam Quy

Viet Nam National University
Hochiminh City, University of Natural Sciences

NCLab-CBLab