Building the Bioscience Gateway

Lavanya Ramakrishnan
lavanya@renci.org

Renaissance Computing Institute
Duke University
North Carolina State University
University of North Carolina - Chapel Hill

NORTH CAROLINA
BIOPORTAL

www.ncbioportal.org

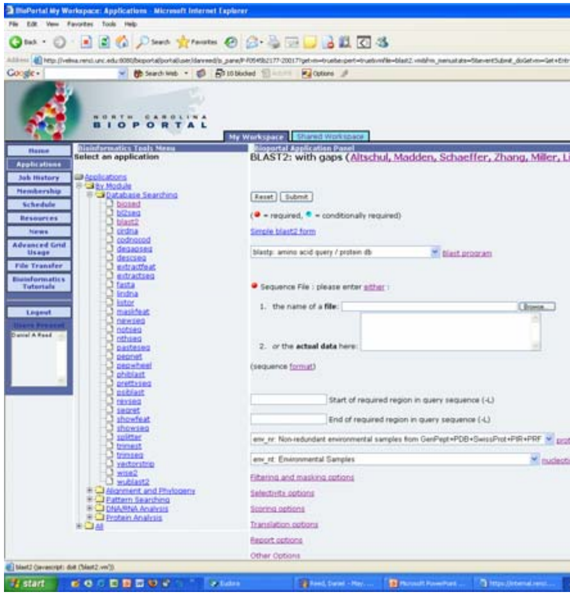
Building Science Communities

- **National Evolutionary Synthesis Center**
  - $15M, five year project, Duke, NCSU and UNC-CH
  - national center and resource
    - research, data federation and outreach
    - sabbaticals and teaching release
  - RENCI support
    - *data models, portals and Grid infrastructure*
- **The Carolina Center for Exploratory Genetic Analysis**
  - develop collaborative experiences and plans
    - preliminary data to apply for a P50 grant
  - develop a prototype informatics infrastructure
    - data models, methods, tools and portals
  - facilitate use of best practices for existing projects

- **North Carolina Bioportal**
  - leverage state-wide investment in bioinformatics and grid
  - undergraduate education, graduate education, faculty research

North Carolina Bioportal

- **Features**
  - access to common bioinformatics tools
  - extensible toolkit and infrastructure
    - OGCE and National Middleware Initiative (NMI)
    - leverages emerging international standards
  - remotely accessible or locally deployable
  - packaged and distributed with documentation
- **National reach and community**
  - TeraGrid deployment
    - scheduled for summer 2005
- **Education and training**
  - hands-on workshops across North Carolina
    - clusters, Grids, portals and bioinformatics



Bioinformatics Tools Menu

Select an application

- Applications
  - By Module
    - Database Searching
      - biosed
      - bl2seq
      - blast2
      - cirdna
      - codnocod
      - degapseq
      - descseq
      - extractfeat
      - extractseq
      - fasta
      - lindna
      - listor
      - maskfeat
      - newseq
      - notseq
      - nthseq
      - pasteseq
      - pepnet
      - pepwheel
      - phiblast
      - prettyseq
      - psiblast
      - revseq
      - seqret
      - showfeat
      - showseq
      - splitter
      - trimest
      - trimseq
      - vectorstrip
      - wise2
      - wublast2
    - Alignment and Phylogeny
    - Pattern Searching
    - DNA/RNA Analysis
    - Protein Analysis
  - All

Bioportal Application Panel

BLAST2: with gaps (Altschul, Madden, Schaeffer, Zhang, Miller, Li

Reset Submit

(● = required, ● = conditionally required)

Simple blast2 form

blastp: amino acid query / protein db Blast program

● Sequence File : please enter either :

1. the name of a **file**: Browse...
2. or the **actual data** here:

(sequence format)

Start of required region in query sequence (-L)

End of required region in query sequence (-L)

env_nr: Non-redundant environmental samples from GenPept+PDB+SwissProt+PIR+PRF

env_nt: Environmental Samples

Filtering and masking options

Selectivity options

Scoring options

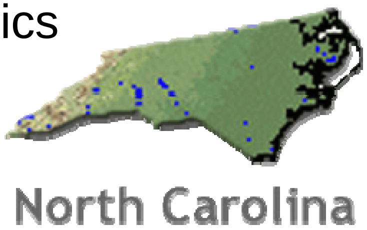
Translation options

Report options

Other Options

blast2 (javascript: doit ('blast2.vm'))

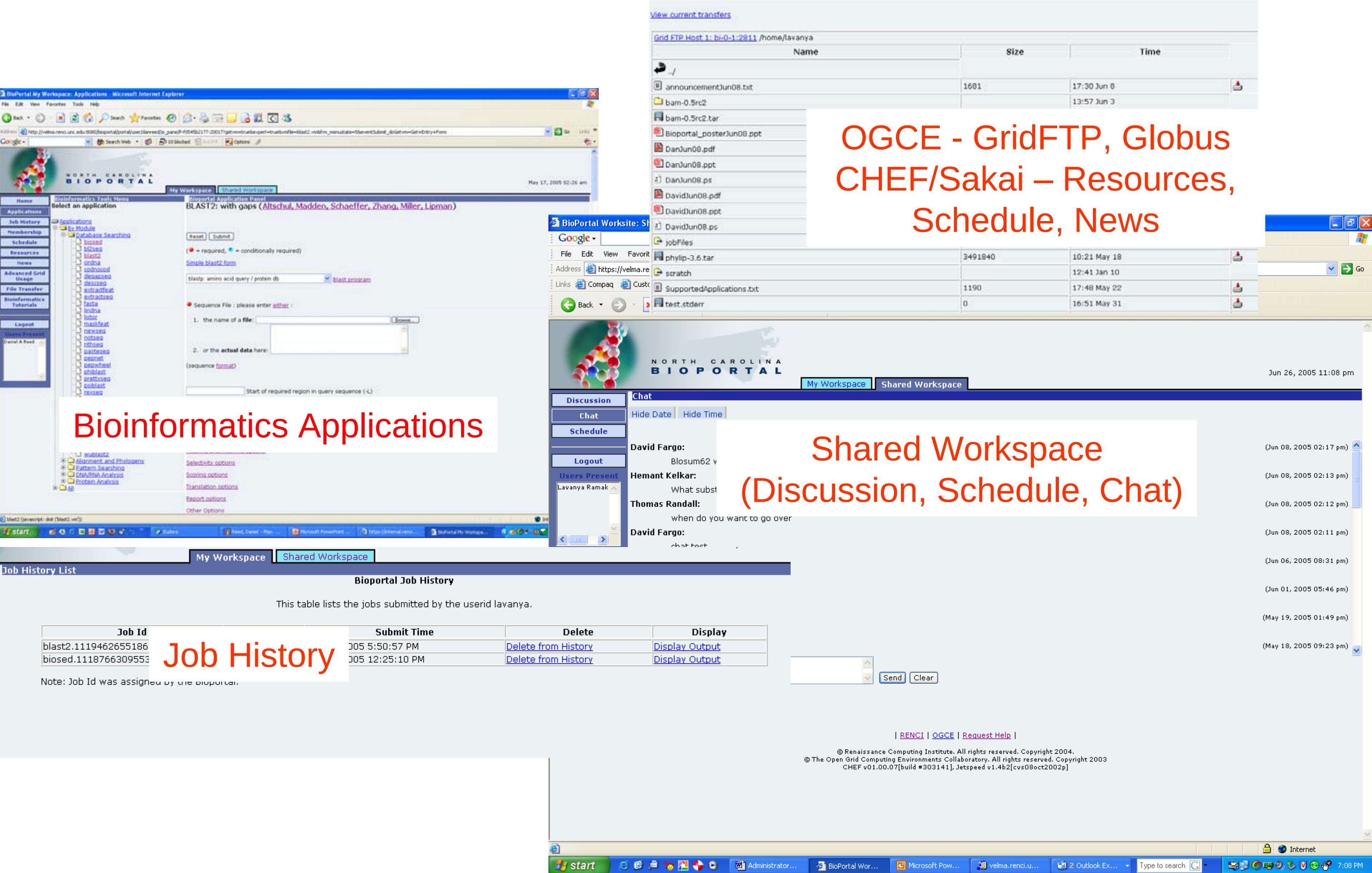
NORTH CAROLINA
BIOPORTAL



Bioportal Computing Infrastructure

- **34 node Linux cluster**
  - one head node
  - 32 compute nodes
  - one storage node
- **Configuration**
  - 3.06 GHz dual Xeon processors
  - 4 GB memory/node
  - 8 GB memory on storage node
- **1.73 TB storage array**
  - 14 x 146GB U320 SCSI Drives
  - RAID 5 partitioned
- **Software stack**
  - ROCKS cluster software
  - Globus toolkit + Torque/Maui + MyProxy
  - OGCE portal software

Portlets



OGCE - GridFTP, Globus
CHEF/Sakai – Resources, Schedule, News

Bioinformatics Applications

Shared Workspace
(Discussion, Schedule, Chat)

Job History

View current transfers

Grid FTP Host 1: bi-0-1:2811 /home/lavanya

Name	Size	Time
../		
announcementJun08.txt	1681	17:30 Jun 8
bam-0.5rc2		13:57 Jun 3
bam-0.5rc2.tar		
Bioportal_posterJun08.ppt		
DanJun08.pdf		
DanJun08.ppt		
DanJun08.ps		
DavidJun08.pdf		
DavidJun08.ppt		
DavidJun08.ps		
jobFiles		
phylip-3.6.tar	3491840	10:21 May 18
scratch		12:41 Jan 10
SupportedApplications.txt	1190	17:48 May 22
test.stderr	0	16:51 May 31

May 17, 2005 02:26 am

Select an application

- Applications
  - By Module
    - Database Searching
      - biosed
      - bl2seq
      - blast2
      - cirdna
      - codnocod
      - degapseq
      - descseq
      - extractfeat
      - extractseq
      - fasta
      - lindna
      - listor
      - maskfeat
      - newseq
      - notseq
      - nthseq
      - pasteseq
      - pepnet
      - pepwheel
      - phiblast
      - prettyseq
      - psiblast
      - revseq
      - wublast2
    - Alignment and Phylogeny
    - Pattern Searching
    - DNA/RNA Analysis
    - Protein Analysis
  - All

BLAST2: with gaps (Altschul, Madden, Schaeffer, Zhang, Miller, Lipman)

(● = required, ● = conditionally required)

Simple blast2 form

blastp: amino acid query / protein db — blast program

● Sequence File : please enter either :

1. the name of a file:
2. or the actual data here:

(sequence format)

Start of required region in query sequence (-L)

Selectivity options
Scoring options
Translation options
Report options
Other Options

BioPortal Worksite: Sh

NORTH CAROLINA BIOPORTAL

Jun 26, 2005 11:08 pm

My Workspace | Shared Workspace

Discussion
Chat
Schedule
Logout
Users Present
Lavanya Ramak

Chat

Hide Date | Hide Time

David Fargo: (Jun 08, 2005 02:17 pm)
Blosum62 v

Hemant Kelkar: (Jun 08, 2005 02:13 pm)
What subst

Thomas Randall: (Jun 08, 2005 02:12 pm)
when do you want to go over

David Fargo: (Jun 08, 2005 02:11 pm)
chat test

(Jun 06, 2005 08:31 pm)
(Jun 01, 2005 05:46 pm)
(May 19, 2005 01:49 pm)
(May 18, 2005 09:23 pm)

| RENCI | OGCE | Request Help |

My Workspace | Shared Workspace

Job History List

Bioportal Job History

This table lists the jobs submitted by the userid lavanya.

Job Id			Submit Time	Delete	Display
blast2.1119462655186			005 5:50:57 PM	Delete from History	Display Output
biosed.1118766309553			005 12:25:10 PM	Delete from History	Display Output

Note: Job Id was assigned by the Bioportal.

Current Bioportal Applications

- **Applications**
  - ~140 distinct codes
- **Application Suites**
  - EMBOSS
    - European Molecular Biology Open Software Suite
  - GLIMMER
    - gene identification in microbial DNA
  - HMMER
    - Hidden Markov Model program for profile-based sequence analysis
  - NCBI
    - diverse set of tools
  - PHYLIP
    - PHYLogeny Inference Package for inferring phylogenies
- **Others (incomplete list)**
  - ClustalW, FASTA
- **Standard bioinformatics databases**
  - NCBI Aggregate (95 GB)
    - three formats: native, BLAST and WUBLAST
  - GenBank (206 GB)
  - GenPept (3 GB)
  - PDB (6.3 GB)
  - Prints (72 MB)
  - RepBase (8.6 MB)
  - UniProt (12 GB)
  - PFam (8.7 GB)
  - ProSite (16 MB)
  - TransFac (36 MB)
- **Database update mechanism**
  - follows the schedule of the distribution source
  - currently NCBI Aggregate is the only one updated nightly

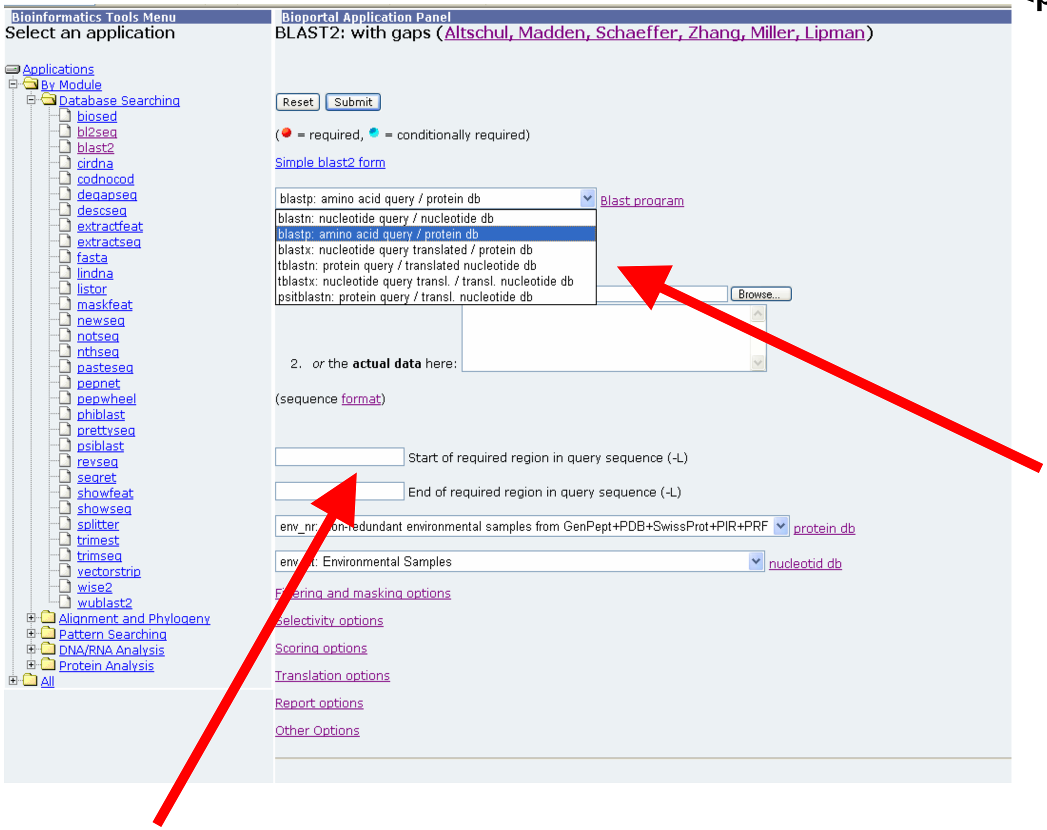
PISE

- **Pasteur Institute Software Environment (PISE)**
  - generates web interfaces for molecular biology tools
    - XML specification for command line interfaces
  - see www.pasteur.fr/recherche/unites/sis/Pise

- **Rationale and objectives**
  - simplify specification of program interfaces
    - homogeneous specification mechanisms
  - reuse of existing software interfaces
    - independent development and integration
  - extension for integration with graphical interfaces
    - complexity hiding and commonality

- **Bioportal program described in PISE**
  - semi-automated GUI synthesis from XML via Perl
- **Output is a generated command line, for example**
  - blastall -p blastp -d env_nr -i query.dat.blast2.1116248106513 -a 2

An Example PISE XML



BLAST2: with gaps (Altschul, Madden, Schaeffer, Zhang, Miller, Lipman)

(● = required, ● = conditionally required)

Simple blast2 form

blastp: amino acid query / protein db — Blast program

- blastn: nucleotide query / nucleotide db
- blastp: amino acid query / protein db
- blastx: nucleotide query translated / protein db
- tblastn: protein query / translated nucleotide db
- tblastx: nucleotide query transl. / transl. nucleotide db
- psitblastn: protein query / transl. nucleotide db

2. *or* the **actual data** here:

(sequence format)

Start of required region in query sequence (-L)

End of required region in query sequence (-L)

env_nr: non-redundant environmental samples from GenPept+PDB+SwissProt+PIR+PRF — protein db

env_nt: Environmental Samples — nucleotid db

Filtering and masking options

Selectivity options

Scoring options

Translation options

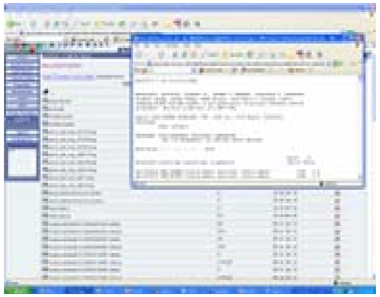
Report options

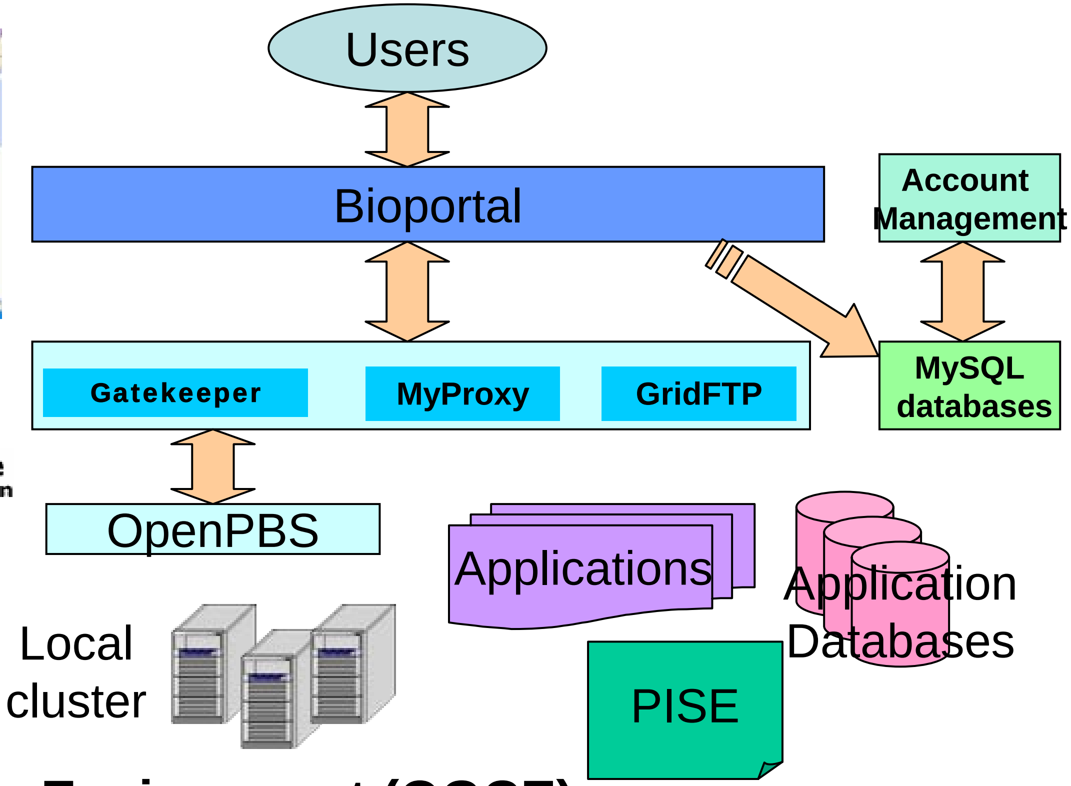
Other Options

```
<parameter ismandatory="1" iscommand="1"
    issimple="1" type="Excl">
  <name>blast2</name>
  <attributes>
    <prompt>Blast program</prompt>
    <format>
      <language>perl</language>
      <code>"blastall -p $value"</code>
    </format>
    <vdef><value>blastp</value></vdef>
    <group>1</group>
    <vlist>
      <value>blastn</value>
      <label>blastn: nucleotide query / nucleotide
    db</label>
      <value>blastp</value>
      <label>blastp: amino acid query / protein
    db</label>
      <value>blastx</value>
      <label>blastx: nucleotide query translated / protein
    db</label>
      <value>tblastn</value>
      <label>tblastn: protein query / translated
    nucleotide db</label>
      <value>tblastx</value>
      <label>tblastx: nucleotide query transl. / transl.
    nucleotide db</label>
      <value>psitblastn</value>
      <label>psitblastn: protein query / transl. nucleotide
    db</label>
    </vlist>
</parameter>
```

```
<parameter type="Integer">
   <name>start_region</name>
   <attributes>
     <prompt>
       Start of required region in query sequence (-L)
     </prompt>
   </attributes>
  </parameter>
```

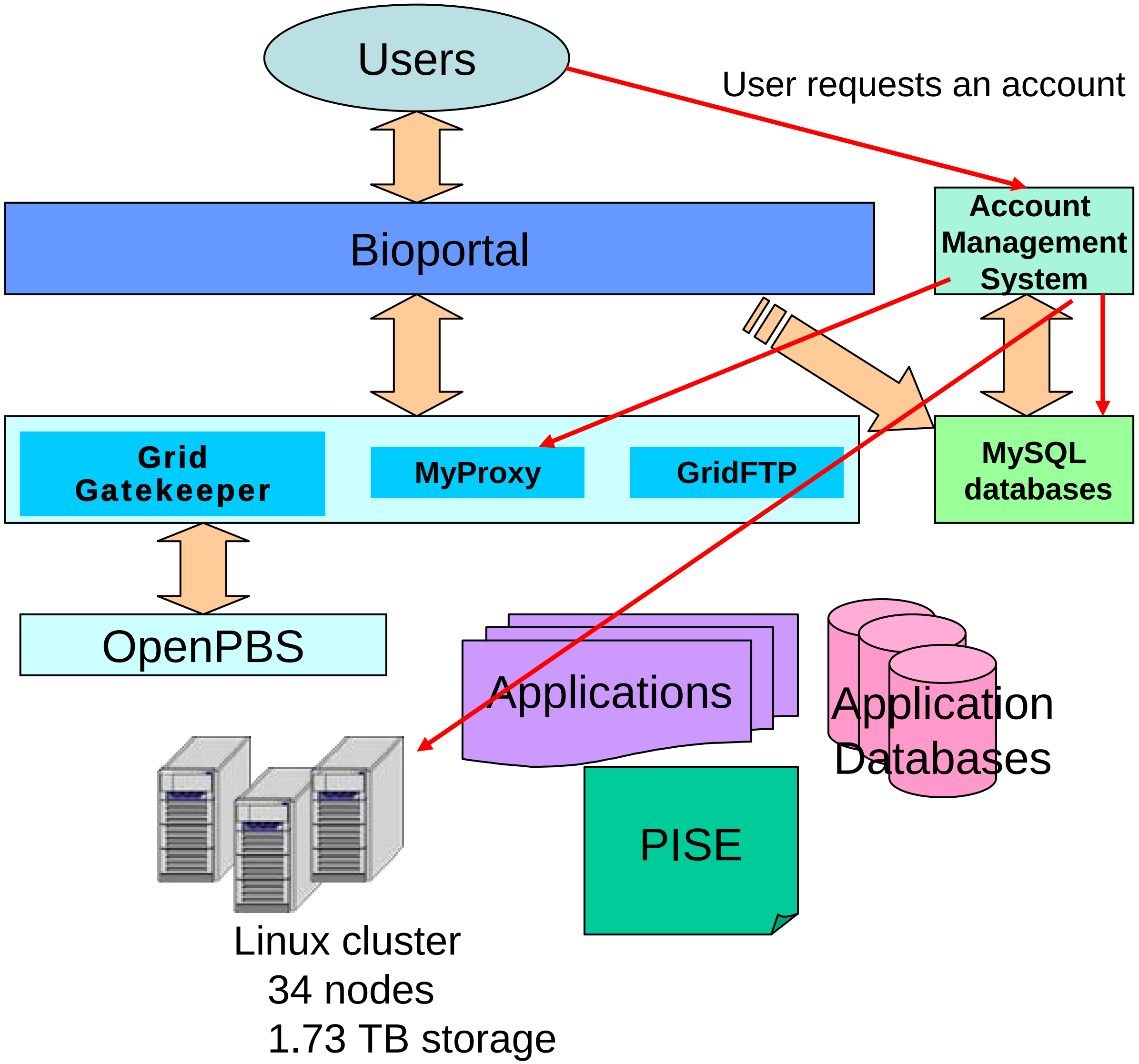
North Carolina Bioportal





- **Open Grid Computing Environment (OGCE)**
  - shared development based on NMI toolkit
  - standard web services
  - adopting portal standards (JSR168)
  - used by cyberinfrastructure projects
    - LEAD, NEES, PACI, DOE, TeraGrid …

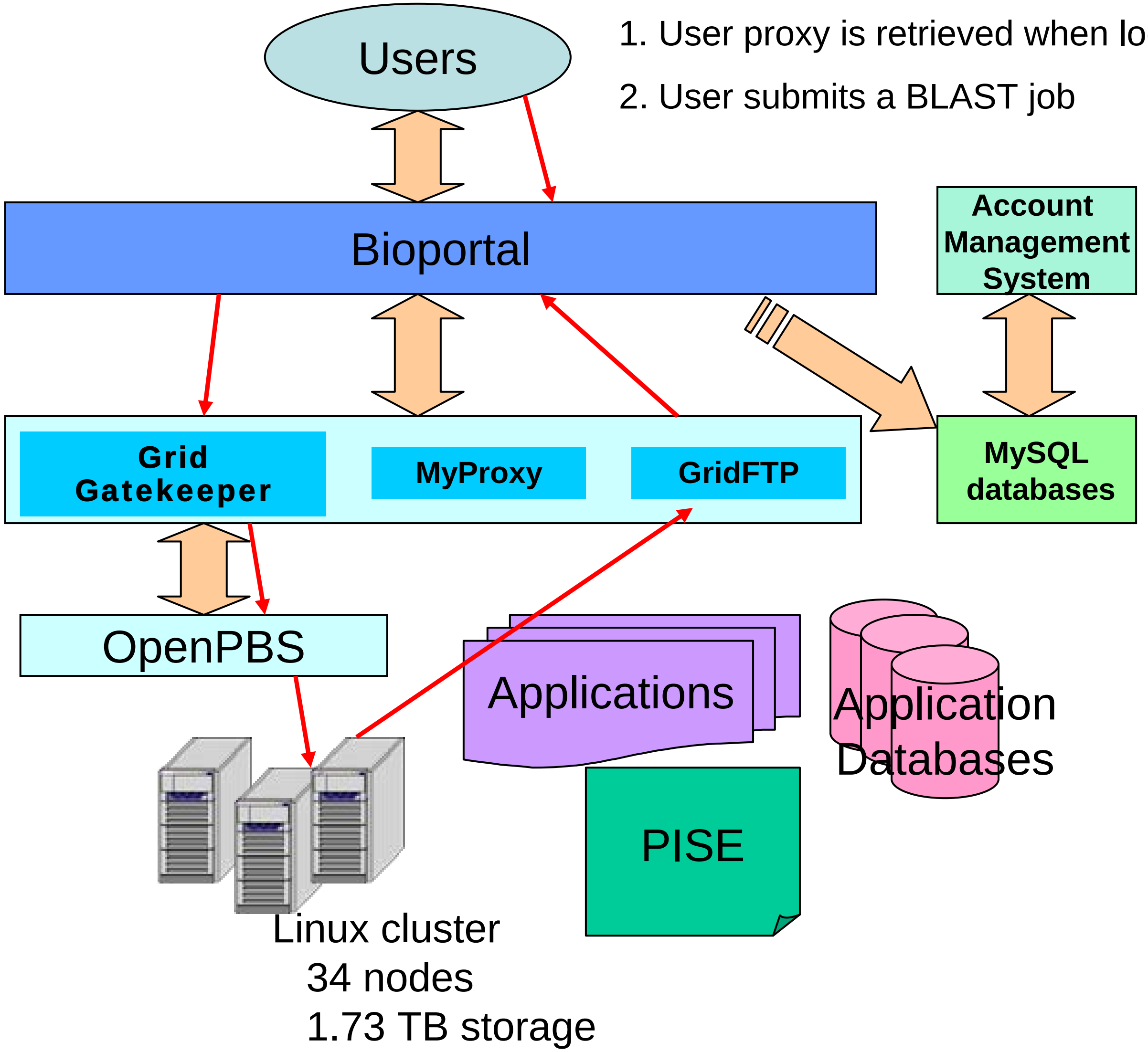
Bioportal User Interactions



Behind the scenes

1. Create Unix account
2. Create a certificate request
3. Sign the certificate request
4. Update MyProxy
5. Add entry to gridmap file
6. Create a portal account

Bioportal User Interactions



1. User proxy is retrieved when logged in
2. User submits a BLAST job
3. User can view job history
4. User can view job output

Behind the scenes

1. Job history database is updated
2. Job submitted to the gatekeeper
3. Enqueued in OpenPBS queue
4. Job executes
5. Output files viewed via GridFTP

Bioportal Experiences

- **Security**
  - account creation and management
  - Grid Security Infrastructure (GSI) ,SSL
- **Job Management**
  - unique job directory
  - manage job files -14 day policy
- **Application Domain Issues**
  - conflicts with Globus RSL
  - size and policy of database updates

Bioportal: What’s Next

- **Community Engagement**
  - workshops, experiences and deployments
  - Software and documentation
- **Infrastructure**
  - dynamic job scheduling across multiple sites
    - load driven based on community use
  - fully automated database updates, possible distributed replication
    - driven by user needs and available disk space
- **Portal tool suite**
  - expand application and databases based on user feedback
    - phylogeny, morphology, microarray analysis, …
  - different file format support
- **Leverage national presence**
  - TeraGrid/NCSA bioinformatics portal
  - NESCent evolutionary biology portal