Deployment of BioXSD-enabled services on a Cloud
Outline

- IBCP, provider of BioXSD-enabled services
- Cloud Computing
- RENABI GRISBI, French infrastructure
Bioinformatics Integrated Tools

The IBCP have integrated several tools for protein sequence analysis with the Web services technology. These Bioinformatics Web services provide scientists and developers with programmatic access to these tools. Our Web services are build upon standards from the W3C like SOAP, WSRF and HTTP. These tools can be use remotely through a graphical and integrated SOAP client like Taverna or Triana. You can also write your own SOAP client with languages such as Python & ZSI, C/C++ gSOAP, perl SOAP::Lite or Java.

### Bioinformatics Tools available

<table>
<thead>
<tr>
<th>Tool</th>
<th>Type of analysis</th>
<th>Description</th>
<th>Documentation</th>
<th>Examples of clients</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClustalW</td>
<td>multiple alignment</td>
<td>wsdl</td>
<td>usage</td>
<td>PyZSI Tav2 ( pict)</td>
</tr>
<tr>
<td>Multalin</td>
<td>multiple alignment</td>
<td>wsdl</td>
<td>usage</td>
<td>PyZSI Tav2 ( pict)</td>
</tr>
<tr>
<td>BLAST</td>
<td>sequence similarity</td>
<td>wsdl</td>
<td>usage</td>
<td>PyZSI Tav2 ( pict)</td>
</tr>
<tr>
<td>FastA</td>
<td>sequence similarity</td>
<td>wsdl</td>
<td>usage</td>
<td>PyZSI Tav2 ( pict)</td>
</tr>
<tr>
<td>SSearch</td>
<td>sequence similarity</td>
<td>wsdl</td>
<td>usage</td>
<td>PyZSI Tav2 ( pict)</td>
</tr>
<tr>
<td>Dsc</td>
<td>secondary structure of protein</td>
<td>wsdl</td>
<td>usage</td>
<td>PyZSI Tav2 ( pict)</td>
</tr>
<tr>
<td>Gor I</td>
<td>secondary structure of protein</td>
<td>wsdl</td>
<td>usage</td>
<td>PyZSI Tav2 ( pict)</td>
</tr>
<tr>
<td>Gor III</td>
<td>secondary structure of protein</td>
<td>wsdl</td>
<td>usage</td>
<td>PyZSI Tav2 ( pict)</td>
</tr>
<tr>
<td>Gor IV</td>
<td>secondary structure of protein</td>
<td>wsdl</td>
<td>usage</td>
<td>PyZSI Tav2 ( pict)</td>
</tr>
<tr>
<td>Predator</td>
<td>secondary structure of protein</td>
<td>wsdl</td>
<td>usage</td>
<td>PyZSI Tav2 ( pict)</td>
</tr>
<tr>
<td>Simpa96</td>
<td>secondary structure of protein</td>
<td>wsdl</td>
<td>usage</td>
<td>PyZSI Tav2 ( pict)</td>
</tr>
<tr>
<td>Protein Secondary</td>
<td>secondary structure of protein</td>
<td>wsdl</td>
<td>usage</td>
<td>PyZSI Tav2 ( pict)</td>
</tr>
</tbody>
</table>

gbio-pbil.ibcp.fr/ws

christophe.blanchet@ibcp.fr
GBIO WS resources

Virtual Machines

Virtualization
VMware ESX/Xen

Physical Machines

Central Storage

NFS / S3

Web Services HOST

Jobs

Computing Nodes

Virtual Machines

Physical Machines

Central Storage

NFS / S3

NFS / S3

HTTP
SOAP

Biologist 's Machine
Web Services clients
What is seekda?

seekda™ is the free search engine for Web API (Web Services at the moment) and their providers.

We run a focused crawled that gathers information about services available on the Web and we monitor these services daily. We also allow our users to edit certain data regarding providers or their services.

Web Services

seekda’s Web Services portal provides a direct access to a robust technology platform enabling search for public Web Services, and in the near future a one-stop-shopping marketplace for services offered by providers from all around the world.

With Web Services technologies you enable, convert and enhance your applications to truly Web-based solutions. By using Web Services, you can plug-in and immediately use new components as easily as you already use software libraries today. The only difference in comparison to the
EMBRACE Use case «Usage of Generic WS»

- Protein sequence analysis at large scale with generic services,
- User: get the most information possible regarding their protein(s)
- Bioinformatics tools
  - use case with common software BLAST, ClustalW, GOR4, CATH-Gene3D, HMMscan, ...
- Biological data
  - Analyzing large sets of proteins obtained for example from Next Generation Sequencing
  - Using international databases, i.e. UniProt

https://bioinformatics.bmc.uu.se/WP4/content/view/125/50/

christophe.blanchet@ibcp.fr

Workshop for Web Service Providers in Bioinformatics
2 June 2010, CBS, Lyngby
BioXSD Services

The IBCP have integrated several tools for protein sequence analysis with the Web services technology. These Bioinformatics Web services provide scientists and developers with programmatic access to these tools. Our Web services are build upon standards from the W3C like SOAP, WSRF and HTTP. These tools can be used remotely through a graphical and integrated SOAP client like Taverna or Triana. You can also write your own SOAP client with languages such as Python SUDS, C/C++ gSOAP, perl SOAP::Lite or Java.

This page is pre-showing the future release of the IBCP Web services, especially featuring some BioXSD-enabled Web services [bioxsd]. If you are looking for the original one, with more integrated bioinformatics services, please go there.

Bioinformatics Tools available

Multiple alignment
- ClustalW [WSDL]
- ClustalW BioXSD [WSDL]

Secondary structure of protein
- GorIV [WSDL]
- GorIV BioXSD [WSDL]

Sequence similarity
- Blast [WSDL]
Benefits of Controlled Vocabulary, Ontology and Schema

Providers

- verify the quality of the input/output data
  - expected data type: integer, string, ...
  - expected value: size, allowed range, ...
  - identify abnormal data
- verify the format of the input data
  - return detailed error message: what's wrong in the format
  - convert the data in a good format
- evaluate complexity/cost for the data
  - input data and its transfer/computation cost (time and €)
  - output data and its storage/transfer cost (time and €)
- workflow
  - group different process according to the data

Users

- make the Web services «user-friendly»
- Web services selection
  - BioCatalogue, Seekda
- Web services composition
  - Taverna, Triana, ...
- switch to other provider in case of failure
- Web services customization
  - to let users map/connect to their own ontology
- rich plugins to input data
  - specific forms, with data converters, an adequate help/doc about the data to input, on different hardware/support
- data conversion in workflow
- rich plugins to display data
Outline

• IBCP, provider of BioXSD-enabled services
• Cloud Computing
• RENABI GRISBI, French infrastructure
Cloud ?
Cloud: principles
Eucalyptus, Hipernet, OpenNebula cloud computing solution comparison

Abstract

This document compares the HIPCAL cloud middleware, Hipernet, with two other ones from the open source community, Eucalyptus and OpenNebula. We have compared their functionalities, from the provider point of view, easiness to install and operate them, and from the user point of view, easiness to use them to deploy scientific applications from Bioinformatics, and efficiency to compute the results.

Alexandre JOSEPH  CNRS IBCP  alexandre.joseph@ibcp.fr
Christophe BLANCHET  CNRS IBCP  christophe.blanchet@ibcp.fr
Cloud: middlewares

• Open source
  • hipernet (HIPCAL project),
  • eucalyptus (Amazon EC2/S3 like),
  • open nebula (OGF, StratusLab EU FP7)
  • others: nimbus, cloudStack, ...

• Commercial
  • Amazon EC2,
  • Google Apps,
  • IBM,
  • Microsoft,
  • Yahoo, ...
Cloud Comparison

<table>
<thead>
<tr>
<th>Feature</th>
<th>Eucalyptus</th>
<th>Hipernet</th>
<th>OpenNebula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine reservation</td>
<td>No</td>
<td>Yes¹</td>
<td>Yes</td>
</tr>
<tr>
<td>Administration interface</td>
<td>CLI, SOAP, REST, EC2 tools, libvirt</td>
<td>CLI</td>
<td>CLI, REST API (OGF), EC2 tools, libvirt</td>
</tr>
<tr>
<td>Virtualisation management</td>
<td>XEN, KVM</td>
<td>XEN</td>
<td>XEN, KVM, VMWare</td>
</tr>
<tr>
<td>Front-end installation type</td>
<td>System-wide, self-contained</td>
<td>System-wide²</td>
<td>System-wide, self-contained</td>
</tr>
<tr>
<td>Hybrid cloud</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Save VM</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Reboot VM</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Stop VM</td>
<td>Yes</td>
<td>Yes⁵</td>
<td>Yes</td>
</tr>
<tr>
<td>Multi-cluster</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Cluster VM migration</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Live cluster VM migration</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Group VM deployment</td>
<td>Yes</td>
<td>Yes</td>
<td>Possible⁶</td>
</tr>
<tr>
<td>VM user image</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Image transfer</td>
<td>HTTP</td>
<td>SSH</td>
<td>NFS, SFTP, HTTP</td>
</tr>
<tr>
<td>Network isolation</td>
<td>Optional⁷</td>
<td>Experimental⁸</td>
<td>Optional⁹</td>
</tr>
<tr>
<td>Throughput control</td>
<td>No</td>
<td>Experimental¹⁰</td>
<td>Possible¹¹</td>
</tr>
</tbody>
</table>
Cloud Console

```
[cblanchet@grive20 ~]$ onevm create chb.one
[cblanchet@grive20 ~]$ onevm list

<table>
<thead>
<tr>
<th>ID</th>
<th>USER</th>
<th>NAME</th>
<th>STAT</th>
<th>CPU</th>
<th>MEM</th>
<th>HOSTNAME</th>
<th>TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>74</td>
<td>cblanche</td>
<td>chb runn</td>
<td>0</td>
<td>524056</td>
<td>524056</td>
<td>griage16</td>
<td>20 02:55:08</td>
</tr>
<tr>
<td>85</td>
<td>cgauthey</td>
<td>xtremfs runn</td>
<td>0</td>
<td>524088</td>
<td>524056</td>
<td>griage16</td>
<td>15 02:50:48</td>
</tr>
<tr>
<td>86</td>
<td>aJoseph</td>
<td>devel runn</td>
<td>0</td>
<td>524056</td>
<td>524056</td>
<td>griage16</td>
<td>13 00:12:50</td>
</tr>
<tr>
<td>87</td>
<td>aJoseph</td>
<td>gbiows-1 runn</td>
<td>0</td>
<td>524124</td>
<td>524124</td>
<td>griage16</td>
<td>10 22:07:24</td>
</tr>
</tbody>
</table>
```

```
[root@grive09 ~]# euca-describe-instances

RESERVATION r-31CF0814 admin default
INSTANCES
- i-24FG6205 emi-613B12DC 172.16.2.10 172.16.2.10 pending 0 m1.small 2010-06-01T15:14:39Z Euca00
- i-59FG6705 emi-613B12DC 172.16.2.19 172.16.2.19 pending 0 m1.small 2010-06-01T15:14:39Z Euca00
- i-46FG6715 emi-613B12DC 172.16.2.24 172.16.2.24 pending 0 m1.small 2010-06-01T15:14:39Z Euca00
- i-35FG6725 emi-613B12DC 172.16.2.25 172.16.2.25 pending 0 m1.small 2010-06-01T15:14:39Z Euca00
- i-22FG6625 emi-613B12DC 172.16.2.26 172.16.2.26 pending 0 m1.small 2010-06-01T15:14:39Z Euca00
- i-11FG6635 emi-613B12DC 172.16.2.27 172.16.2.27 pending 0 m1.small 2010-06-01T15:14:39Z Euca00
```
Deploying on Cloud

IBCP’s cloud
5 servers
40 cores, 160GB

Master
User’s sequences splitting

Tools
- BLAST, FastA, SSearch

10s to 1,000s

Results gathering

Protein Similarity

Knowledge Bases
- MB to 100s GB
- Swiss-Prot, TrEMBL

Read only

Shared Storage

Tools

Virtual Infrastructure

IBCP's cloud
5 servers
40 cores, 160GB

Workshop for Web Service Providers in Bioinformatics
2 June 2010, CBS, Lyngby

christophe.blanchet@ibcp.fr
Bioinformatics Services on a CLOUD

1. & 5. HTTP SOAP

2. Register Biological Data

3. Jobs

4. Get Results

Physical Machines

Central Storage

NFS / S3

Virtualization
VMware ESX/Xen

Bioinformatics Virtual Apps

Cloud Console

Mgmt VMs

Computing Nodes

Virtual Machines

Connect to Cloud

Choose VMs

Deploy VMs

Submit SOAP

Process SOAP

Compute job

Results

Yet done by IBCP

Biologist's Machine

Cloud Storage (S3)

christophe.blanchet@ibcp.fr

Workshop for Web Service Providers in Bioinformatics
2 June 2010, CBS, Lyngby
Perspectives

1. Register Biological Data
2. HTTP SOAP
3. Get Results

Virtual Machines

Virtualization
VMware ESX/Xen

Central Storage

Physical Machines

Virtual Appliances Deployment

NFS / S3

Cloud Container (EC2, ONE, hipcal)

Mgmt VMs

Jobs

Computing Nodes

Biologist’s Machine

Biological Apps

Cloud Storage (S3)

0. HTTP/S SSH

1. Register Biological Data

Connect to Cloud

Choose VMs

Deploy VMs

Submit SOAP

Process SOAP

Compute job

Results

christophe.blanchet@ibcp.fr

Workshop for Web Service Providers in Bioinformatics
2 June 2010, CBS, Lyngby
What next?

- **Platform-as-a-Service**
  - Provide scientists with pre-defined VM to deploy
  - on the Research Infrastructure, e.g. RENABI, GRISBI
  - on their own computer/cloud
  - Bioinformatics centers switch to virtual appliances provider

- **Infrastructure-as-a-Service**
  - User can connect to the community cloud to deploy required VM
  - Deploy the required infrastructure according to the workflow:
    - VXDL/VPXI developed by HIPCAL project
    - Haizea with the OpenNebula.org
Outline

- IBCP, provider of BioXSD-enabled services
- Cloud Computing
- RENABI GRISBI, French infrastructure
GRISBI
- Grid Support to Bioinformatics -

Make possible challenging bioinformatics applications dealing with large scale biological systems

- National Research Infrastructure
  - RENABI, IBISA 2008-2010, Institut des Grilles 2009-2010
- 6 centers from RENABI
  - PRABI, MIGALE, GenOuest, CBiB Bordeaux, BIPS, CIB
- 8 sites, with 7 CNRS institutes
  - IBCP Lyon, SBR Roscoff, CBiB Bordeaux, CIB Lille, IRISA Rennes, LBBE Lyon, MIGALE Jouy-en-Josas, BIPS Strasbourg
- 40 participants
- Computing resources
  - 1200 cores, 220 TB storage

© RENABI GRISBI 2009 - www.grisbio.fr

Will be used tomorrow as hands-on material by LABRI people (RENABI GRISBI partner)
Conclusions

• Standard Integration of bioinformatics tools
  • Freely available to community
  • Annotated tools with BioXSD/EDAM

• Cloud perspective
  • Use standard infrastructures
  • Managed infrastructure with Web interface
  • Platform- and Infrastructure-as-a-Service

• Perspectives
  • Research Infrastructures based on public/private cloud
  • Provide the community with bioinformatics Virtual Appliances
Acknowledgment

CNRS - Centre National de la Recherche Scientifique
University of Lyon 1
The ANR through the HIPCAL project
GIS IBISA through the project GRISBI PF 2008
The European Commission through the EU FP7 EGEE III project, contract number INFSO-RI-222667.

HIPCAL: Pascale Vicat-Blanc and partners
EMBRACE partners
GRISBI partners

CNRS IBCP
Institute of Biology and Chemistry of Proteins
7 passage du Vercors, 69007 LYON, FRANCE

christophe.blanchet@ibcp.fr
Thanks

• Questions?