ELIXIR Compute platform

Authors and contributors:
Alexander Agafonov (UIT – NO), Lars Ailo Bongo (UIT - NO), Mikael Borg (BILS - SE), Amelie Cornelis (EMBL-EBI), Rob Finn (EMBL-EBI), Montserrat Gonzalez (ELIXIR), Rob Hooft (DTL - NL), Ilkka Lappalainen (CSC - FI), Mikael Linden (CSC - FI), Ludek Matyska (MU - CZ), Conor McMenamin (EMBL-EBI), Steven Newhouse (EMBL-EBI), Tommi Nyrönen (CSC - FI), Michael Prochazka (MU - CZ), Inge Alexander Raknes (UIT – NO), Mirek Ruda (CESNET - CZ), Harri Salminen (CSC - FI), Maxim Scheremetjew (EMBL-EBI), Christine Staiger (SURFsara - NL), Dario Vianello (EMBL-EBI).

Acknowledgements:
Niklas Blomberg (ELIXIR), Rafael Jimenez (ELIXIR), Stephanie Suhr (ELIXIR).

European Life Sciences Infrastructure for Biological Information

www.elixir-europe.org
ELIXIR Compute platform for life sciences

ELIXIR Compute Platform supports bioinformatics: new technologies, tools, and data-analysis services and helps research to create scalable services for new data resources.
Data movement!

Transfer data (from 4 scientific cases) to platform. Some ELIXIR data resources are available on the platform.

Data from ENA, UniProt, TaraOceans, Ocean Sampling Day, Biosample DB, EURISCO, EGA
Sustainability

Cloud & Compute provided by ELIXIR-nodes from 20 countries

Scientific software environment

Virtual Services and cloud storage
Positioning

- **Relying Services** receive technical service capability from the ELIXIR Compute Platform. (e.g. ELIXIR Intranet, Biobank Data access portal, Marine metagenomics software service)

- **ELIXIR Compute platform’s** service capabilities consist of ELIXIR Nodes service delivery plans and European e-Infrastructure. Consumption by the Relying Services is supported by ELIXIR Compute experts

- **External Service Providers** allow ELIXIR to leverage investments made nationally, by EC (e-Infrastructures) or commercially by other Service Providers
1. ELIXIR COMPUTE SERVICES
Status

- The ELIXIR Compute Platform has established distributed Authentication & Authorisation Infrastructure (AAI), started Cloud & Compute, Storage and File Transfer Services that are delivered by the individual ELIXIR Nodes.
  - Demonstrated with Marine metagenomics analysis service
  - Contact contact-compute@elixir-europe.org
Authentication and Authorisation Infrastructure

- Reliable electronic identification of users (ELIXIR ID) is needed to access the most of the Compute platform and other services.

- **You can link existing user accounts to create your ELIXIR Identity today.** ELIXIR Authentication and Authorisation services allow Users to continue using their federated academic, corporate or social media identity by linking it to a personal ELIXIR ID.

- The ELIXIR service providers connected to ELIXIR AAI will benefit from a centralised user identity and access management services
Cloud & Compute

- A cloud and local compute infrastructure is needed to undertake life science data analysis.

- Cloud services from the nodes and external service providers need to be federated to provide uniform operation.

- Specialized (private) network solutions to access service are possible and ideal for users that require high performance, high security and certified environments for e.g. sensitive human data handling.
Data storage and transfer

Data transfers are important across life science and various data transport mechanisms have been investigated

- Data replication and data submission to or from ELIXIR Data Resources
- Services to pull relevant datasets from Data Resources or their replicas to cloud or compute services for detailed local analysis
- Data location services to manage and discover data replicas within ELIXIR to discourage *ad hoc* data transfer and storage
Technical Use Cases (TUCs)

Translating scientific needs to steering of ELIXIR Compute Platform

- Services from ELIXIR nodes
- Specific ELIXIR requirements

23 scientific technology translations
Can I use MMG pipeline on my data?

**TUCs ID 1, 4, 5, 7, 8**

---

**Example scientific use of Compute platform using ELIXIR AAI compatible services**

<table>
<thead>
<tr>
<th>Task 4.3.1. User identities &amp; Service access (AAI)</th>
<th>Task 4.3.2. Cloud &amp; Compute</th>
<th>Task 4.3.3. Data storage &amp; transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relying (ELIXIR) Service</td>
<td>Run scientific software environment</td>
<td>Move data between endpoints</td>
</tr>
<tr>
<td>Get (Life Scientist) ID</td>
<td>(Cloud) file storage (computable)</td>
<td></td>
</tr>
<tr>
<td>ELIXIR Compute Services</td>
<td>(Virtual) servers</td>
<td>Data resources</td>
</tr>
<tr>
<td>ELIXIR ID proxy</td>
<td>(Cloud) file storage</td>
<td>Data transfer technologies</td>
</tr>
<tr>
<td>Service access rights</td>
<td></td>
<td></td>
</tr>
<tr>
<td>External Services</td>
<td>Cloud &amp; Compute server providers</td>
<td></td>
</tr>
<tr>
<td>Identity federation</td>
<td>Cloud &amp; Compute federation</td>
<td></td>
</tr>
<tr>
<td>Identity providers</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. ELIXIR COMPUTE DEMONSTRATOR WITH MARINE METAGENOMICS
First user story

Running the Spark based ELIXIR Norway (UIT) marine metagenomics pipeline at ELIXIR Finland (CSC) cloud and using the ELIXIR AAI to identify the user.
Use Marine metagenomics service request, part 1

Biochemist with raw sequence data authenticates to Marine metagenomics portal (https://metapipe.uit.no/) offered by ELIXIR using ELIXIR ID in order to analyse dataset with the workflow of bioinformatics tools.
Please sign in

META-pipe uses the Elixir AAI. You may use it to sign in using your university credentials or your Google account.

Log in with Elixir AAI
ELIXIR COMPUTE perspective

- ELIXIR ID is provided by ELIXIR Authentication and Authorisation services (AAI).
  - ELIXIR AAI uses user's home institution to confirm electronically who they are.
  - Social media identities like Google can also be connected to ELIXIR ID, but these identities alone may not be considered sufficiently reliable (real user identity not properly confirmed) and some service providers may refuse to give access to their services for users who do not provide more trustworthy electronic identity.
  - ELIXIR Proxy IdP is used to combine the individual federated identities users may have/use
ELIXIR AAI (draft)

External authentication (e-infrastructures)

Relying services:
- EGA
- wiki
- Cloud
- Intranet
- Data archive

ELIXIR AAI:
- Credential translation
- ELIXIR Proxy IdP
- ELIXIR Directory

External authentication (e-infrastructures):
- eduGAIN IdPs
- Common IdPs

ELIXIR AAI services:
- Dataset authorisation management
- Group/role management
- Bona fide management
- Attribute self-management

Dataset authorisation management
Group/role management
Bona fide management
Attribute self-management
Once the user has successfully logged in, they

- Upload their dataset to be analysed
- Execute analysis on their dataset by launching the pre-configured data analysis

These computations enrich understanding on the raw marine metagenome data derived from the samples.
Start Meta-pipe

Select dataset

- Upload new
- Choose existing

Select parameters

- Executor
- Cutoff: 500

Start Meta-pipe
Start Meta-pipe

Select dataset

Choose a different dataset

MVIS.fas
https://storage.metapipe.uit.no/435569bc57/inputs/MVIS.fas-ivjvz5t-h503qneuja3webmg3nmi

Select parameters

Executor

Cutoff
500

Start Meta-pipe
Start Meta-pipe

Select dataset

Choose a different dataset

MVIS.fas

https://storage.metapipe.uit.no/43559bc57finputs/MVIS.fas-fvijyvz5t-h503qneuja3webmglnmi

Select parameters

Executor

cPouta

Cutoff

500

Start Meta-pipe
META-pipe
ELIXIR COMPUTE perspective

- Marine metagenomics portal divides the data analysis into independent compute tasks that are assigned to worker nodes.
  - Job is divided into multiple parts. There is one (Spark) job per marine metagenomics dataset to be analyzed.
  - The execution environment at the ELIXIR cloud contacts the portal coordination server to request jobs tagged for that execution environment.
ELIXIR COMPUTE perspective

Data needed for compute is transferred to target computable storage space seen by the worker nodes.

In the demonstrator worker nodes are executed on ELIXIR Finland node cloud service.

In future instances the worker nodes will be run in other sites, too.
Cluster consists of a master node and several workers. META-pipe files are stored in a shared file system. META-pipe tasks are submitted to scalable Spark cluster and run as Spark jobs.

The bastion host at ELIXIR Finland acts as a gateway between the Spark cluster and the task manager in ELIXIR Norway.
Marine metagenomics service, part 3

Once computations finish the data is returned to the portal, and the user can download the enriched results for further analysis or visualization using separate tools e.g. Krona and Artemis
ELIXIR COMPUTE perspective

Data is collected together for download by user.
Second user story

Running the ELIXIR EMBL EBI marine metagenomics pipeline at the EBI Cloud setup.
The initial need

- Provide one-click deployments to cloud infrastructures
- Use industry-standard, widely adopted, Open Source tools
  (easy to use is also a good plus, of course!)
- Basically, don’t reinvent the wheel once again
The EBI Cloud Portal

- Provides a RESTful API serving multiple clients:
  - Main *EBI Cloud Portal* web client + project-branded portals
  - Programmatic clients

- **Single Sign On** authentication through *Elixir AAI*
EBI Cloud Portal and Elixir AAI at work
The EBI Cloud Portal - Defining applications

- Wraps around:
  - Terraform *(Infrastructure Provisioning)*
  - Ansible *(Configuration)*

- A *common* language for applications definition

- Uses git as a way to *collaborate* & *version infrastructure*
The EBI Cloud Portal

[Image of the EBI Cloud Portal interface]
Cloud credentials

EBI Embassy
- Provider: OSTACK
- Fields:
  - OS_USERNAME
  - OS_TENANT_NAME
  - OS_AUTH_URL
  - OS_PASSWORD

My Amazon
- Provider: AWS
- Fields:
  - AWS_ACCESS_KEY_ID
  - AWS_SECRET_ACCESS_KEY
  - AWS_DEFAULT_REGION

http://my.homepage.me
Technology and Science Integration
Persistence: cloud storage
Compute: Defined by applications

**GridFTP Server**

A simple GridFTP infrastructure to transfer data

- Version 0.1
- Contact dario@ebi.ac.uk

**Available for**

- AWS
- OSTACK
- GCP

**OpenLava cluster**

An OpenLava cluster for AWS and OpenStack

- Version 0.1
- Contact dario@ebi.ac.uk

**Available for**

- AWS
- OSTACK
- GCP
The second wave of needs/requirements

- Provision *pipelines*, not only infrastructure

- Switch from IaaS to PaaS (*kind of*)

- If possible, make pipeline sharing *fast, easy and controllable*

- Re-use the EBI Cloud Portal as much as possible
  
  *We really don’t like re-inventing the wheel! :-)*
The idea

• Use **images** to share pipelines!

1. Deploy an OpenLava cluster
2. Attach the pipeline as a volume via a pre-baked image
3. Fire off the processing via Ansible as part of the deployment

• **Plug-and-play** pipelines!
The deployment card - General info

Metagenomics Pipeline

A cloud deployment of the EBI Metagenomics Pipeline

Version
0.1

Contact
dario@ebi.ac.uk

Available for
OSTACK
GCP
The deployment card - Inputs

Source
https://github.com/EMBL-EBI-TSY/mg...

Volumes

This application requires no attached volumes

Inputs

This application requires to enter the following values
ENA_ID

nodes

Outputs

This application will generate the following values
- MASTER_IP

OSTACK  GCP
Providing the inputs

Source
https://github.com/EMBL-EBI-TS/mg...

Volumes
This application requires no attached volumes

Inputs
This application requires to enter the following values
RNA175504
nodes
10

Outputs
This application will generate the following values
- MASTER_UP

OSTACK GCP
Deploying...

Metagenomics Pipeline

- **Reference**: TSI14279223038423

- **Status**: STARTING

  Waiting for provider data...

- **Volumes**: There are no attached volumes for this deployment

- **Inputs**:
  - **ENA_ID**: R1NA175504
  - **nodes**: 10

- **Outputs**: There are no generated outputs for this deployment
Ansible at work

PLAY [MG Pipeline]   

TASK [Force Ansible to reconnect]   
 changed: [104.155.99.246]   

TASK [Find the OpenLava bin folder]   
 ok: [104.155.99.246]   

TASK [Save the path into a variable]   
 ok: [104.155.99.246]   

TASK [Ensure result directory is in place]   
 changed: [104.155.99.246]   

TASK [Launch pipeline]   
 changed: [104.155.99.246]   

el psy kosh
OpenLava, up & running

```bash
[centos@tsi1479223038423-master ~]$ bhosts
HOST_NAME     STATUS   JL/U  MAX  NJOBS  RUN  SSUSP  USUSP  RSV
tsi1479223038423-m closed  -  16  1  0  0  0  0  0
tsi1479223038423-n ok  -  16  0  0  0  0  0  0
```

(continued on the right)
Processing the jobs

```
[centos@tsi1479223038423-master ~]$ bjobs -a
JOBID  USER STAT    QUEUE    FROM_HOST  EXEC_HOST   JOB_NAME   SUBMIT_TIME
102    centos RUN  normal   tsi14792230 tsi14792230 *FASTQ.dat Nov 15 15:23
101    centos DONE normal   tsi14792230 tsi14792230 *se -l 100 Nov 15 15:22
103    centos DONE normal   tsi14792230 tsi14792230 CRDGLF Nov 15 15:24
[centos@tsi1479223038423-master ~]$ 
```
Cleaning up…
What we learned (I)

• Deploy customised compute on-demand is feasible, but:
  
  • **Pre-bake** whenever possible
  
  • Put variables everywhere - *flexibility is everything*
  
  • Keep a sharp *separation* between infra & workloads
What we learned (II)

• Using images to **share** pipeline works!

  • **Cuts** the deployment time

  • Image (thus pipelines) **versioning**

  • No **scaling** issue

  • Users free to pick the volume **performance**
3. NEXT STEPS
Up next -- Help for Diagnostics: Human Access-controlled Data

- The next priority use case is supporting distributed data collection and release to authorized individual users from the European Genome-phenome Archive (EGA) archive, and to partner downstream secure data analysis on EGA-external clouds with support of ELIXIR Compute platform.
Conclusions

- An automated and scalable cluster setups from ELIXIR nodes are technologically suitable for loads in bioinformatics software and data pipelines for marine metagenomics
  - (Open Stack) cloud cluster suitable for MMG currently requires four commands and less than 15 minutes to deploy in ELIXIR Finland, and similar level of automation is available at the EBI cloud
- Encapsulation of scientific software pipelines in containers and images increase their portability and provide extensibility to ELIXIR’s distributed infrastructure
- Challenge remaining is streamlining cloud access authorization to users and service providers (identified with ELIXIR AAI) at the Europe-level to support biological information and research
ELIXIR COMPUTE PLATFORM

https://www.elixir-europe.org/platforms/compute