Scipion on-demand service in the cloud

IBERGRID 2018

Lisbon 11th-12th October

CSIC

instruct Integrating Biology

instruct image Processing Center

Who are we?

The Instruct Image Processing Center (I2PC)

Instruct: The European Research Infrastructure (ESFRI) for Structural Biology

Provides support and training to structural biologists in the use of image processing software

CSIC

instruct
Integrating
Biology

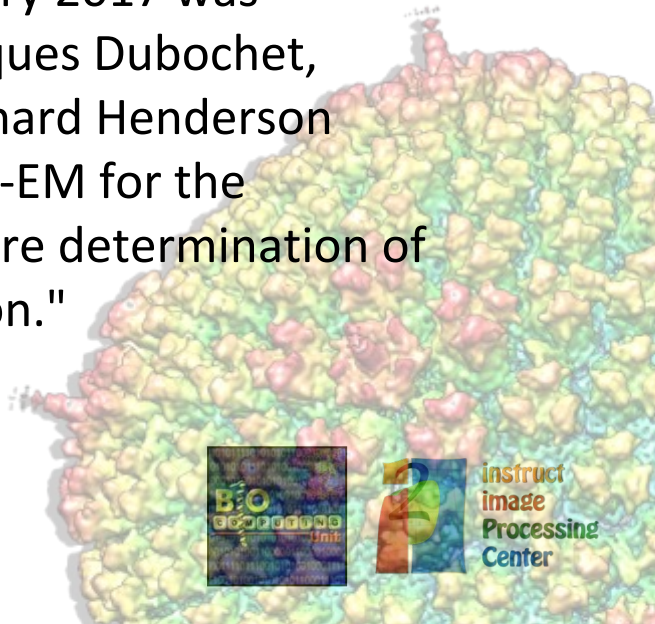
BIO COMPUTING Unit

instruct
image
Processing
Center

What is Cryo Electron Microscopy

One of the structural biology techniques at the core of the Instruct project, electron microscopy under cryogenic conditions (“cryo-EM”) is currently the fastest growing area, having been nominated “**Method of the Year** (2015)” by Nature.

Nobel Price in Chemistry 2017 was awarded jointly to Jacques Dubochet, Joachim Frank and Richard Henderson "for their work on cryo-EM for the high-resolution structure determination of biomolecules in solution."



CSIC

instruct
Integrating
Biology

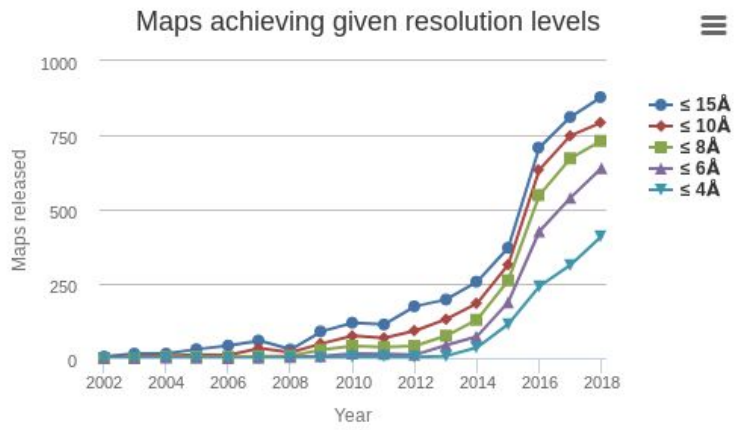
instruct
image
Processing
Center

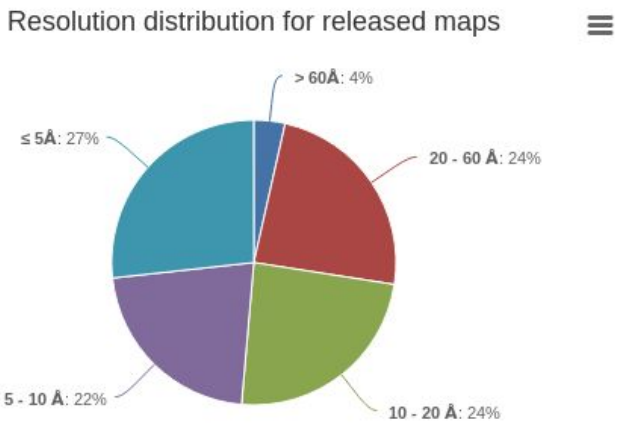
Why do we hear so much about Cryo-EM?

Because thanks to:

1) The very good performance of current microscopes
2) The very good image acquisition characteristics of Direct Electron Detector
3) The very good new software for 3D reconstruction and classification

It is possible to solve the structure of large and flexible macromolecular complexes from small amounts of not very concentrated samples.





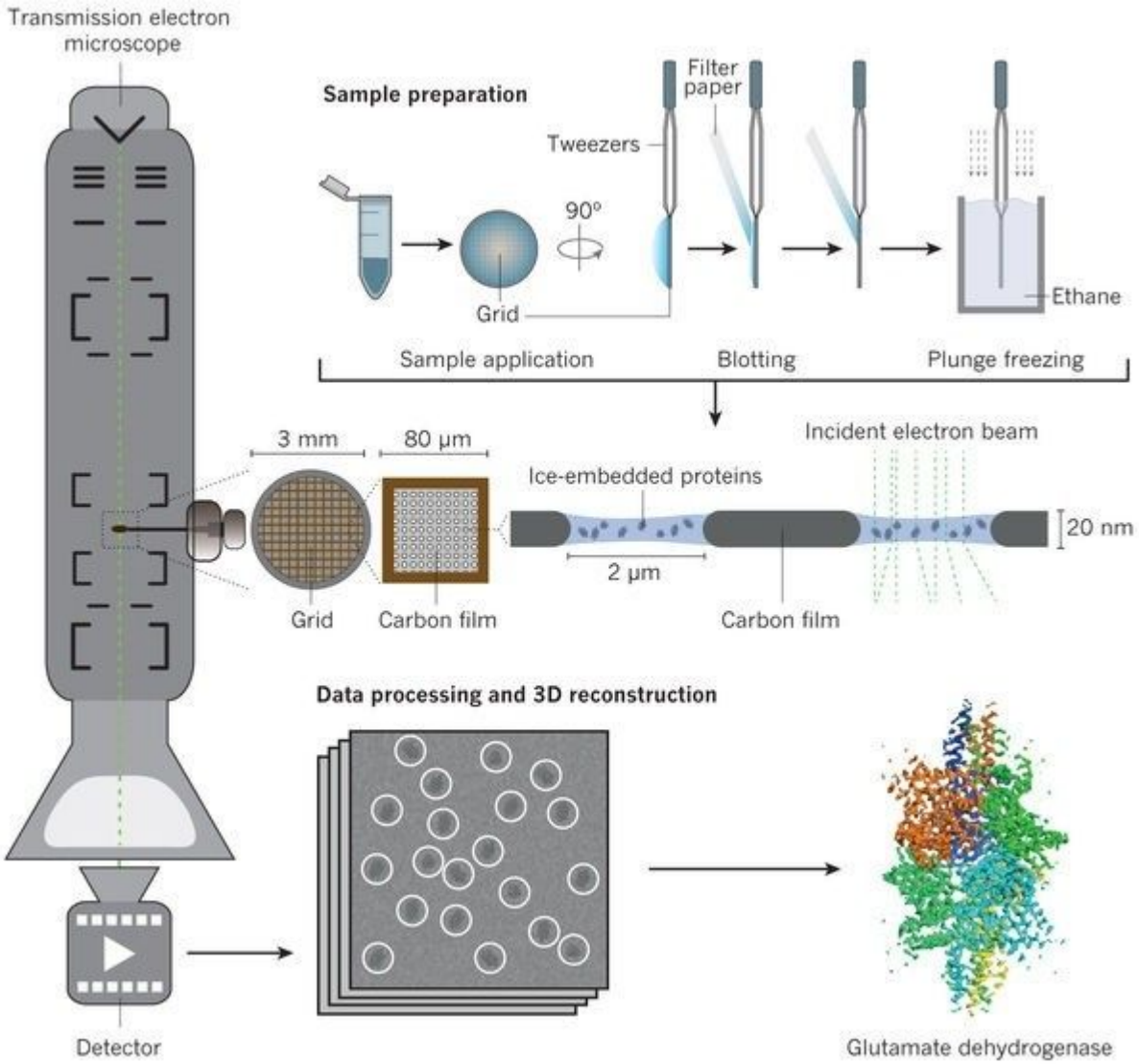
Protein Data Bank in Europe, EMBD statistics.

CSIC

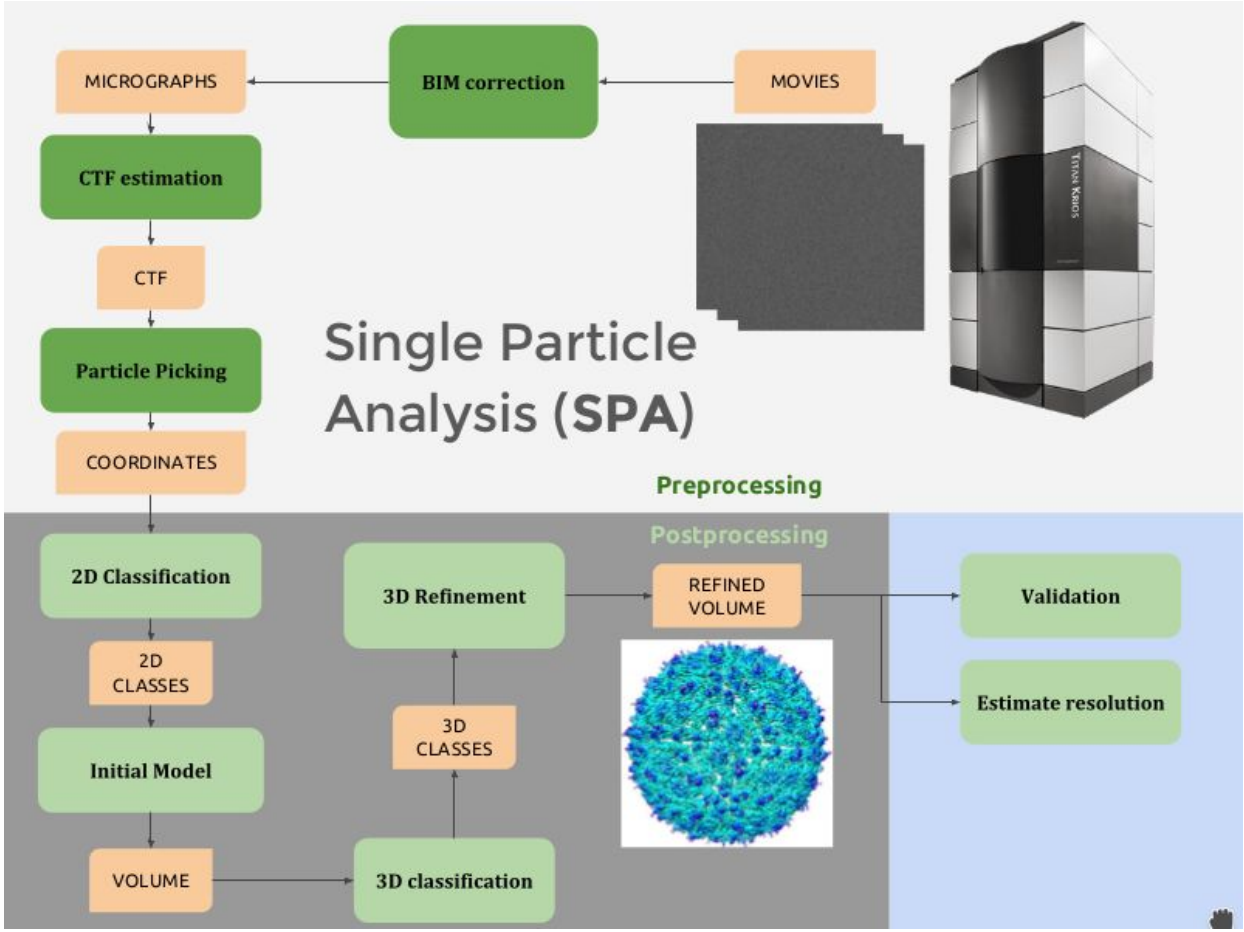
instruct Integrating Biology

instruct image Processing Center

How does Cryo-EM work?

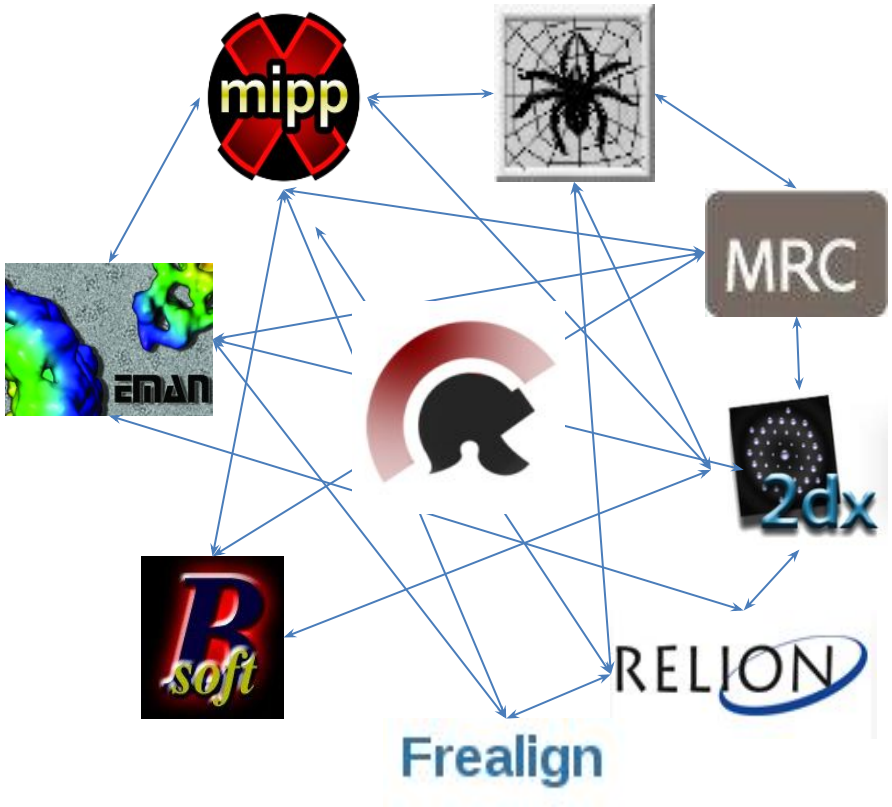


Typical Cryo-EM Workflow
Single Particle Analysis (SPA)



Plethora of EM software packages: Our answer "Scipion" Workflow Integrator

Bringing software integration to EM in workflows

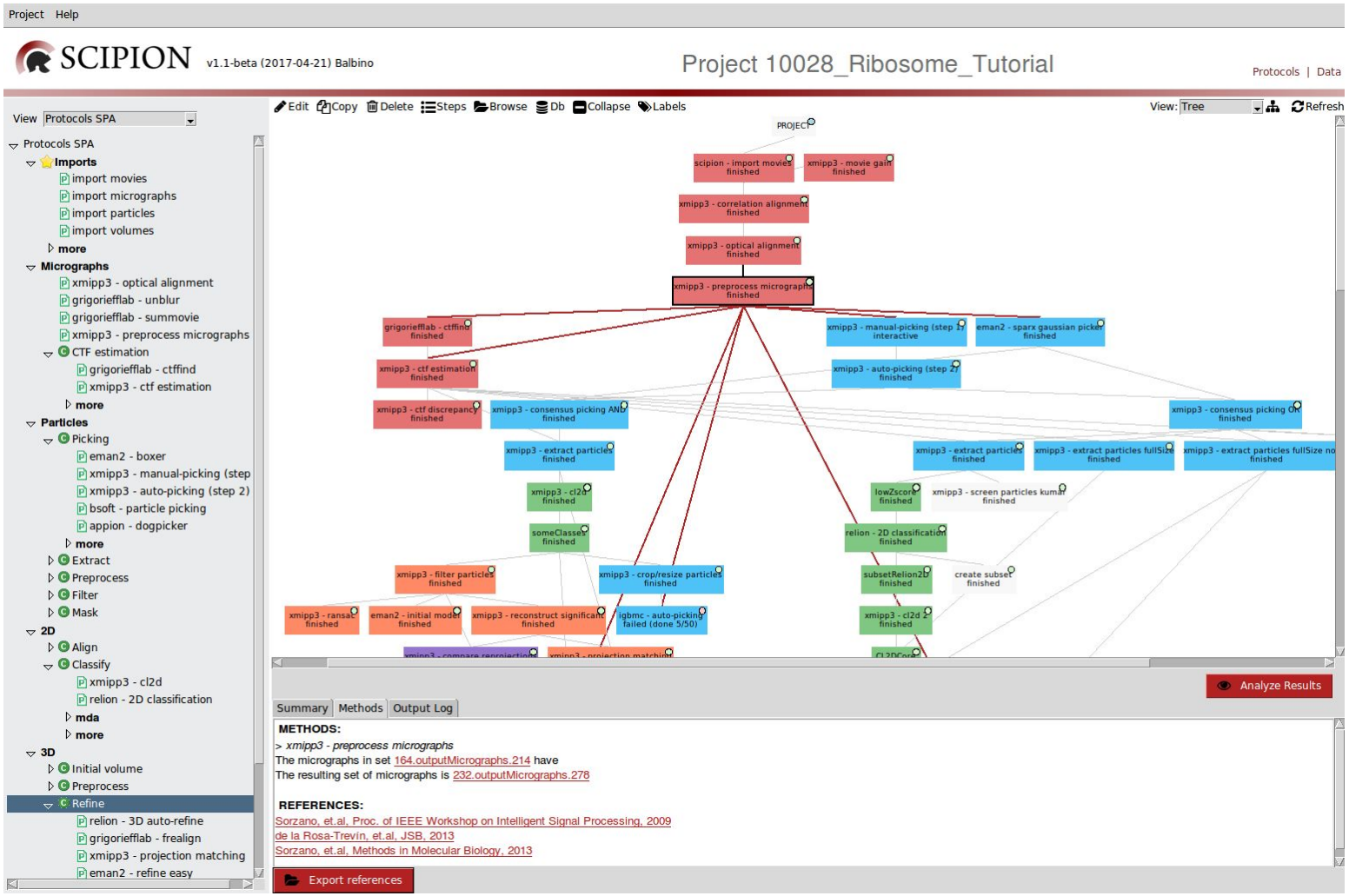


CSIC

instruct Integrating Biology

instruct image Processing Center

Scipion Framework



CSIC

instruct Integrating Biology

BIO COMPUTING Unit

instruct image Processing Center

Scipion Framework

Scipion encapsulates:

- Parallelization: By each EM program or by Scipion (OpenMPI)
- Environment setup, libraries
- Batch system submission: Scipion template
- Use of GPUs
  - Different requirements by EM package
  - Scipion homogeneous variables setup
- Logging
- Workflow automation

Scipion distributions

- Binaries or source code + EM packages autoinstall
- ScipionCloud
  - Public AMI on AWS EC2 and Virtual Appliance on EGI AppDB
    - Ubuntu 16.04 LTS
    - Scipion release 1.2 (source git)
    - Most important EM packages compiled with CUDA (GPU support)
    - Nvidia driver + cuda toolkit (7.5 & 8.0)
    - TurboVNC + VirtualGL + noVNC (remote desktop)
    - Starcluster (only AWS)
  - Puppet + Cloudify (Westlife project)

CSIC instruct Integrating Biology BIO COMPUTING Unit instruct image Processing Center

Scipion for Cryo-EM facilities
Run workflows automatically in streaming

Acquisition

Microscope

ScipionBox (cluster, workstation,..)

- Import Movies
- Movie Alignment
- CTF Estimation
- Particle Picking
- Extraction
- 2D classification (dev)

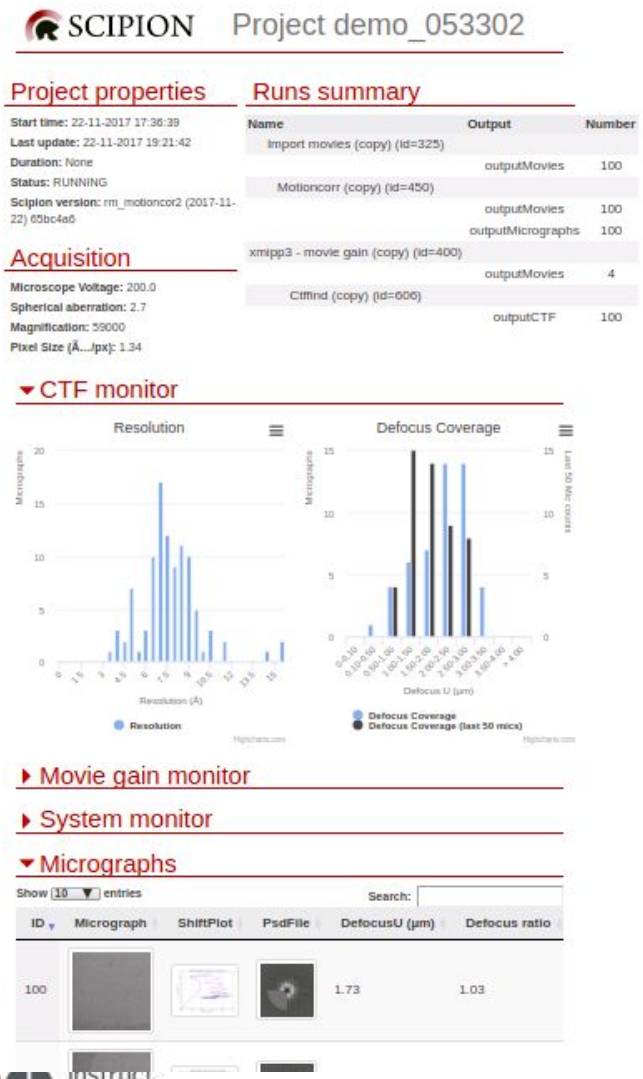
CSIC

instruct Integrating Biology

BIO COMPUTING Unit

instruct image Processing Center

Scipion for Cryo-EM facilities



- Report what has been done in the facility
- Track system status, memory, gpu, cpu, network
- Raise alarms when thresholds reached.
- HTML report and alarms
- Customized workflows
- Programmatically accessible (python API)

CSIC

Instruct Integrating Biology

BIO COMPUTING Unit

instruct image Processing Center

Scipion on-demand service
for Instruct users of Cryo-EM facilities

CSIC

instruct
Integrating
Biology

BO
COMPUTING
Unit

instruct
image
Processing
Center

Main goal

Our goal is that all the complex processing steps after Cryo-EM Facilities preprocessing could be performed with ScipionCloud for users with no IT knowledge at all.

The computational requirements are within limits, but the complexity of dealing with multiple software is beyond most users capacities. We want to “democratize” the processing of Cryo-EM data.

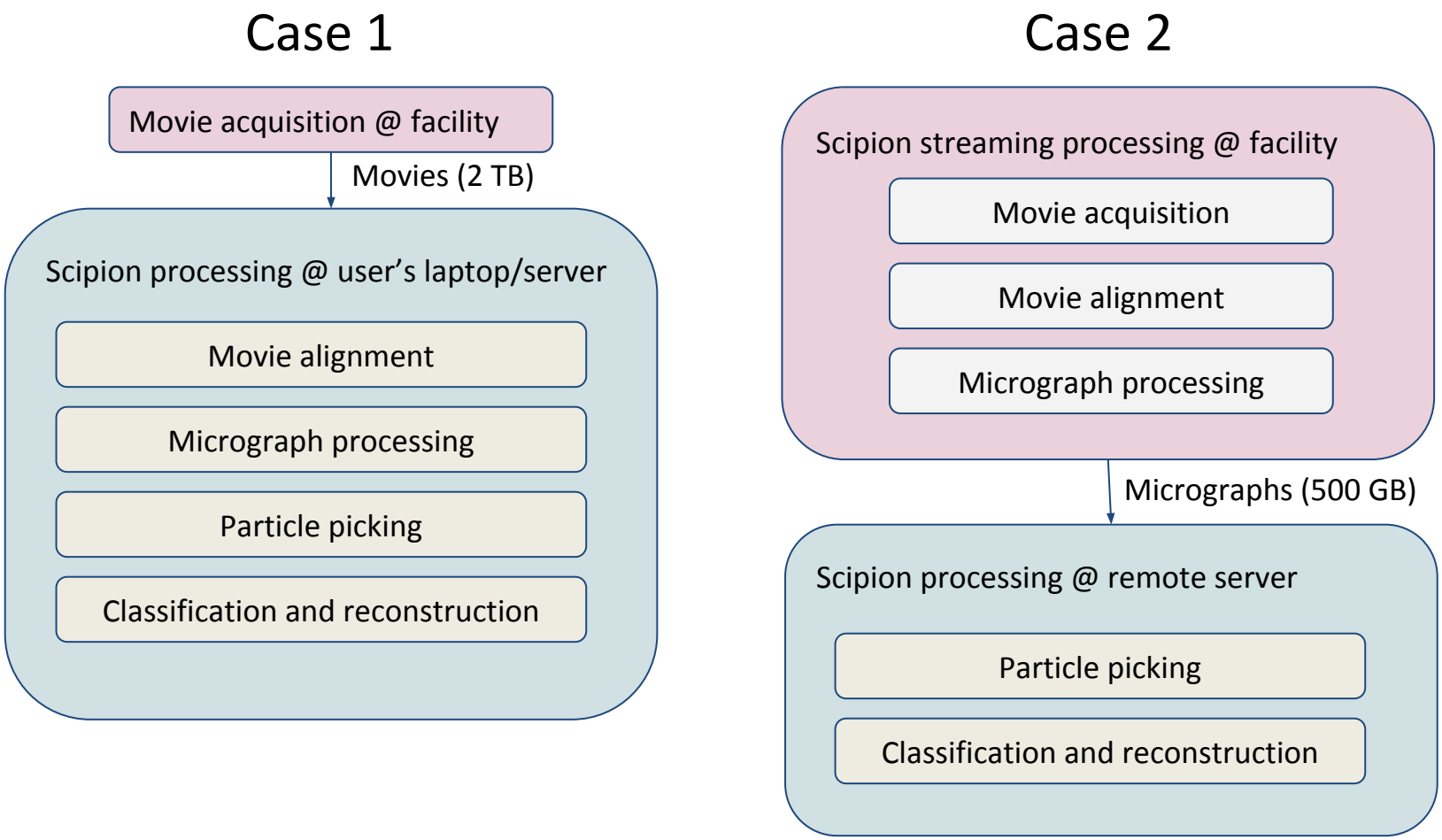
CSIC

instruct Integrating Biology

BO COMPUTING Unit

instruct image Processing Center

Scipion processing use cases

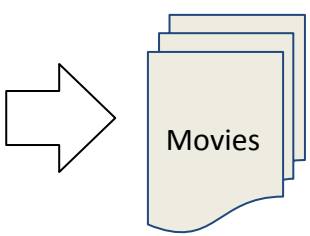


CSIC

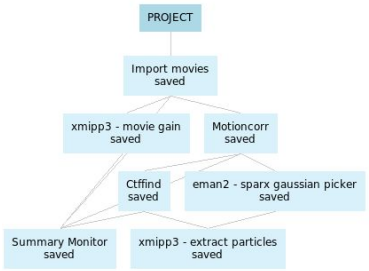
instruct Integrating Biology

instruct image Processing Center

First step @ Instruct Cryo-EM facilities

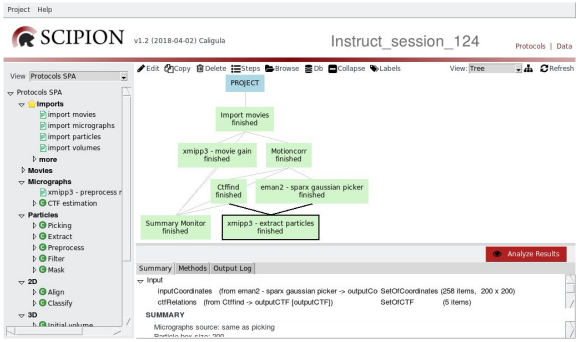


+



.json

(includes Scipion and EM packagesversions)



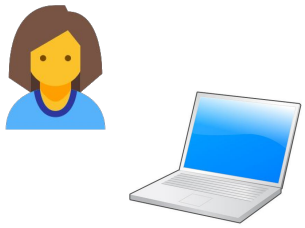
Instruct Project: Acquiring data on streaming mode with Scipion.
User leaves the facility with raw data (movies) and Scipion project with some steps executed.

Second step @ I2PC or user's location

1st approach:
Done manually by I2PC staff

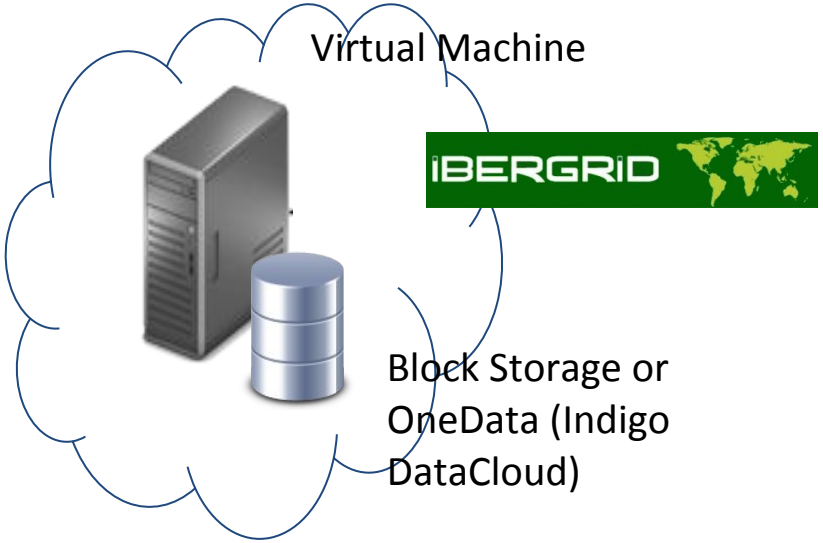
2nd approach: Done automatically by user through a web portal (already under development)

Authentication (TBD):
- VO enmr.eu ?
- Robot certificate on portal
- Instruct users login



1. Create Virtual Machine using cloudify+puppet customized with json info.
Specify a **time limit for the project**.

Cloudify+puppet and portal solution already implemented but other options might be considered.



CSIC
instruct Integrating Biology
BIO COMPUTING Unit
instruct image Processing Center

Scipion in the Cloud

User Friendly CryoEM Data Analysis from Anywhere

A. Křenek, R. Peša, J. Ketner, V. Holer, D. Kouřil, D. Antoš, L. Del Caño Novales, P. Conesca Mingo

West-Life
Structures for Life

Scipion

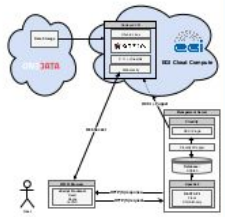
Scipion is an image processing framework, originally designed as a desktop PC application, for obtaining **3D models of macromolecular complexes using Electron Microscopy**. It integrates several software packages allowing scientists to execute workflows **combining different software tools**, while taking care of formats and conversions. The prevalence of Scipion stems from a thorough integration of multiple software packages with the software. Moreover, all steps are tracked and they can be reproduced later on.

Why in Cloud?

Hardware requirements for real-world EM data analysis scenario may reach **beyond usual desktop PC**. Many **software dependencies** also jeopardise its portability, especially in the world of fragmented desktop operating systems. Last but not least, CryoEM **datasets are large** typically, which complicates their efficient sharing on desktop PCs. We address all these issues by providing an **integrated environment** which deploys the application in the cloud and exposes a **web-only interface** to the user.

Architecture

Scipion deployment, including all the depencencies and VM configuration, is described in a **TOSCA blueprint** and a set of **Puppet recipes**. Those are processed with **Cloudify** to set up the cloud node. In this way we avoid dealing with large pre-canned VM images which are difficult to maintain and to copy over network.



The VM setup with Cloudify and its access with VNC client is still too complicated for the typical user, therefore we wrap it with a thin, application specific software layer, which manages the deployments and their lifecycle, handles errors etc.

The user is exposed to a one-stop web interface (developed using **React/Redux** framework) where the he/she manages the deployed machines and gets **instant access to the remote VM desktop** in the web browser.

GPU support

Display of reconstructed 3D electron density map requires **hardware accelerated rendering**. A GPU has to be attached to the virtual machine in the cloud, typically via PCI-passthrough. The VM runs a headless X11 server with hardware accelerated OpenGL 3D rendering using the GPU. Scipion is run in VNC server. OpenGL 3D rendering calls are intercepted by a preloaded **VirtualGL** library, they are **redirected to the accelerated X11 server**, rendered to off-screen windows, and the resulting pixmaps are copied back. The same GPU is used for **accelerated computation** too.

OneData integration

Scipion project is a set of files in a single folder with typical size of **10–10,000 GB**. We integrate the cloud setup with OneData storage, shielding the user from the need to copy these datasets around manually.

OneData provides a web interface for basic data manipulation. In the VM, a FUSE client for Linux is used to mount the same workspace. Due to performance and stability reasons, the work at the VM is done on a local data copy, which is **periodically synchronized back** to OneData.

User authentication

The management server is interfaced with the **West-life authentication infrastructure** (AAI), following recommendations of AARC project. However, these mechanisms contradict the concepts of RESTful operation of the application based on the React framework. We follow a tradeoff approach, generating a JSON Web Token in an authenticated area (where the user provides his/her password), and using the token to access the remaining parts of the interface. The mechanism is completely **transparent to the user**.

Access to the OneData web interface uses West-Life AAI too. However, in order to keep control on credential delegation, the user is required to generate a specific access token specific to the Scipion project, and to paste it into a dedicated field in the deployment form. The token is passed to the VM and used there to act on behalf of the user.

Acknowledgement

This work was done in the West-life project, part of e-Infrastructures for virtual research environments (VRE), project Number 675858

Computing resources were provided by CERIT Scientific Cloud centre within project no. LM2015085.

HORIZON 2020 e-infrastructure

Contact: ljocha@ics.muni.cz

Further information at https://bit.ly/2y132Gu

CSIC

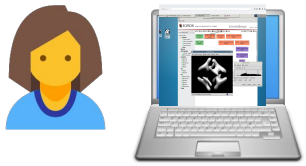
instruct Integrating Biology

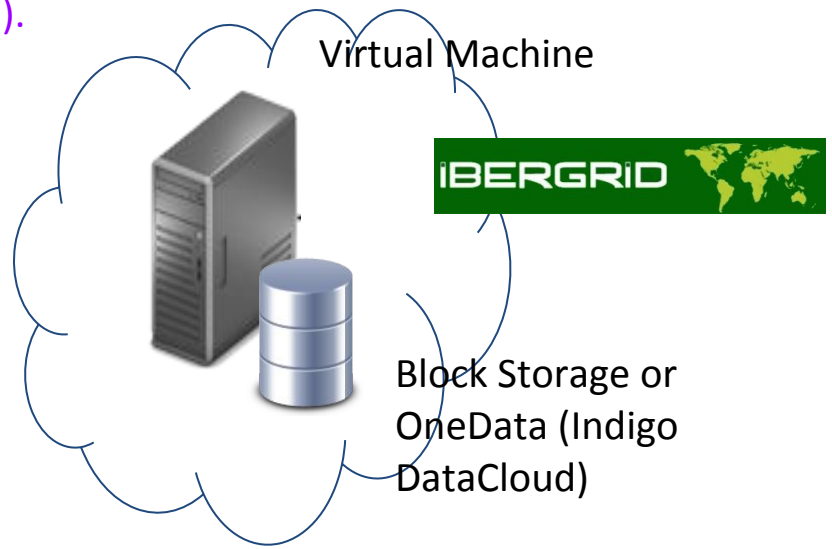
BIO COMPUTING Unit

instruct image Processing Center

Third step @ user's location

2. Copy Scipion project (raw data in principle not needed).





3. Connect to remote desktop through a browser and continues processing.

4. When processing finishes or project expires user should download results (and project).
If a publication is produced user should acknowledge Instruct and IberGRID provider.

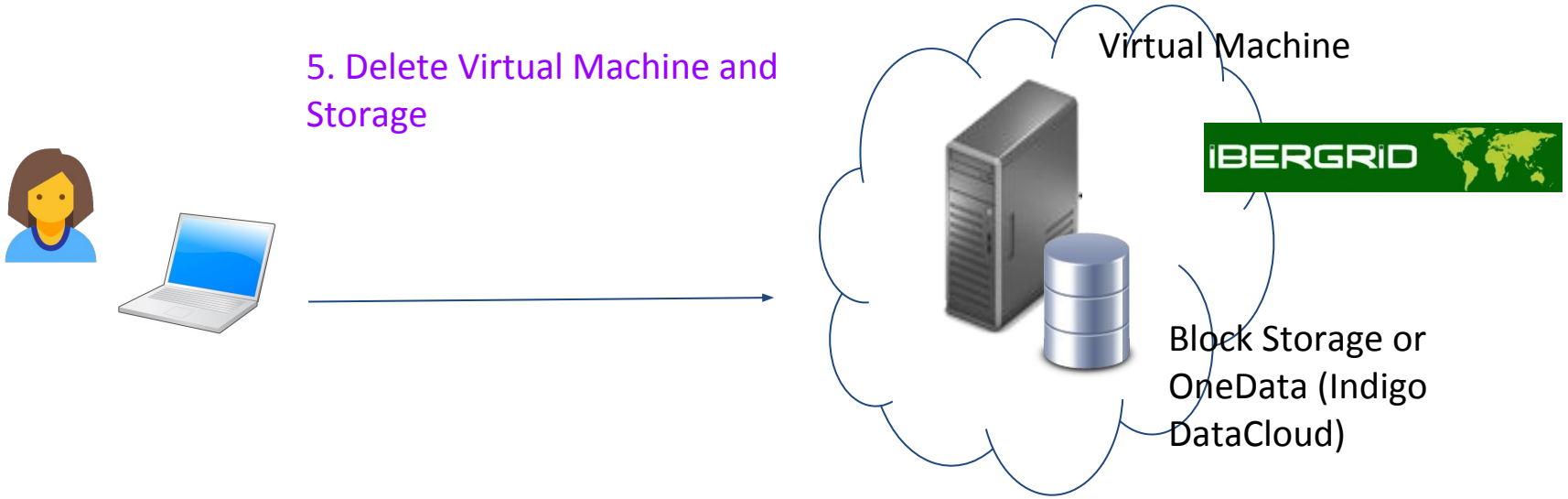
CSIC

instruct
Integrating
Biology

BIO COMPUTING Unit

instruct
image
Processing
Center

Final step @ I2PC (or automatic)



Server requirements for processing

Typical Cryo-EM project:

- Disk (BS)
  - Data (movies) ~ 1-2 TBs. No need to copy to BS, only if user would need to reprocess them. For some superresolution movies movie sets can be much bigger (up to 12 TBs).
  - Scipion project ~ 500 GB - 1 TB. Of course this depends on number of workflow steps.
- # CPUs from 12-32.
- RAM 20 - 60 GB (depends on software package used).
- GPU powered. Nvidia devices with compute capability higher than 3.5 and RAM at least 8 GB. It is not mandatory but new algorithms require it.
- Processing time depends (2 weeks - 2 months with and without 1 GPU).

Tests on IFCA OpenStack Cloud

July 2018

Process a movies dataset (case 1)

Initial aim: Process EMPIAR dataset 10061 (beta-galactosidade to 2.2 A map resolution) which contains 1539 superresolution movies gain corrected (12.4 TB).

Final aim: Process a subset of the movies (97, around 800 MB) to obtain a low resolution map using the most demanded algorithms (involving GPU usage).

Contacted several IberGRID sites but only IFCA could provide GPU cloud machines.

CSIC

instruct Integrating Biology

BIO COMPUTING Unit

instruct image Processing Center

GPU flavours at IFCA

GPU Flavors

ID	Name	CPU Family and other HW	vCPUs	RAM (MB)	Disk (GB)	Eph (GB)	Hard Extra	Public
2defa7e9-782b-4d37-b272-885f53556966	g1.xlarge	Intel Xeon X5550 2,67Ghz 4 GPU Nvidia GT200GL	1	5000	60	150	1 GPU	FALSE
33b9d3a5-aa68-4c37-8da7-c50583b8f684	g1.2xlarge		2	10000	60	150	2 GPU	FALSE
52be20f1-27d3-4dc5-8cc7-bf18d518230d	g1.4xlarge		4	20000	60	300	4 GPU	FALSE
9afca810-3beb-47be-a393-cb034f0fb648	g2.large	Intel Xeon E5-2670 2,6Ghz GPU Nvidia Titan X	8	60000	60	150	1 GPU	FALSE
47dcdee9-84da-4e43-baa7-bb67b3a53ed4	g3.large	Intel Xeon E5-2620 2Ghz GPU Nvidia 1080ti	12	22000	60	150	1 GPU	FALSE
1acdb78b-8ba9-440c-94a6-a0a952b58a1b	g4.large	Intel Xeon E5-2603 1,7Ghz GPU Nvidia 1080ti	1	2500	30	100	1 GPU	FALSE
cf1aeea5-33b8-4bf3-b843-93117e59f5f8	g4.xlarge		2	3750	60	200	2 GPU	FALSE
b13bf655-a4d4-48f2-a75a-232a36ad813a	g4.2xlarge		4	7500	60	200	4 GPU	FALSE
0b5aedfc-dcbf-4d23-adef-75933d654713	g4.4xlarge		8	15000	60	200	8 GPU	FALSE
dfe62bb7-ae93-4288-8bef-050d8e8fbdf5	g4.6xlarge		12	30000	60	200	10 GPU	FALSE

CSIC instruct Integrating Biology

BIO COMPUTING Unit instruct image Processing Center

Tests on g4.4xlarge

VM with 8 CPUs, 15 GB RAM and 4 GPUs.

External 1 TB BS.

ScipionCloud-GPU image on EGI AppDB with some upgrades.

Impossible, **not enough RAM** for such superresolution movies.

Tests on g3.xlarge

VM with 12 CPUs, 22 GB RAM and 1 GPUs.

Same external 1 TB BS.

ScipionCloud-GPU image on EGI AppDB with some upgrades.

Same input data.

Whole workflow run obtaining a low resolution map (23 A). It might be improved but probably not to reach enough resolution to get map that could be published.

Conclusions

- At least flavour g2.xlarge (8 CPUs, 60 GB RAM and 1 GPU) is necessary to process cryoEM superresolution data (not yet tested).
- Other possibility will be to use a non GPU flavour although it would take longer time (2 months instead of 2 weeks). Users do not like this in general cause new algorithms run on GPU.
- On the proposed service movies are not considered so in principle huge storage requirements are not needed.
- Better use of cloud resources could be achieved by using different VM flavours in different steps: Picking + Extracting on a non GPU flavour and classification and refinement on a powerful GPU flavour.

CSIC

instruct Integrating Biology

BIO COMPUTING Unit

instruct image Processing Center

Acknowledgments

- Ales Krenek & Radim Pesa (Mazarykova University)
- Enol Fernandez & Giuseppe La Rocca (EGI)
- Alvaro López y Pablo Orviz (IFCA)

References

- INSTRUCT
- Instruct Image Processing Center (I2PC)
- Scipion project
- INSTRUCT case on IBERGRID
- Scipion cloudify project (WestLife)
- WestLife project