Healthcare research: HPC and cloud.

Arnaud Ceol, arnaud.ceol@ieo.it,

WORKSHOP GARR 2019, 8-10 ottobre 2019, Rome

IEO
Istituto Europeo
di Oncologia

IEO25

IEO Hospital

Department of Experimental Oncology

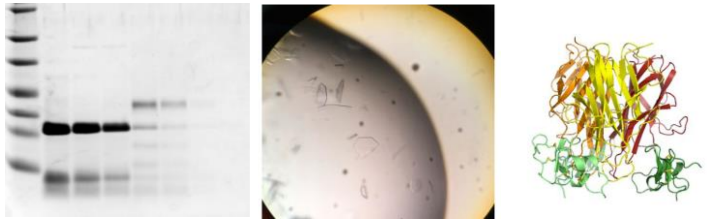
Clinical and basic research, Integration within IEO Clinical Programs, PhD program

› 20 bioinformaticians

IEO Istituto Europeo di Oncologia IEO25

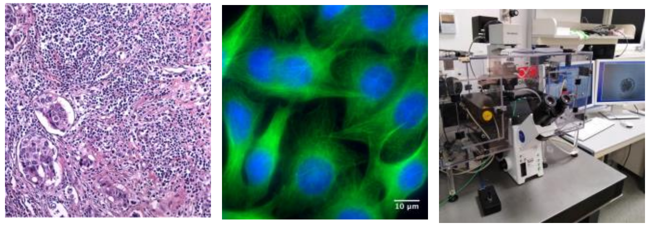
Technological units: raw & processed data

Crystallography



- Few users
- 1 TB/year
- Analyses: Memory and GPUs

imaging



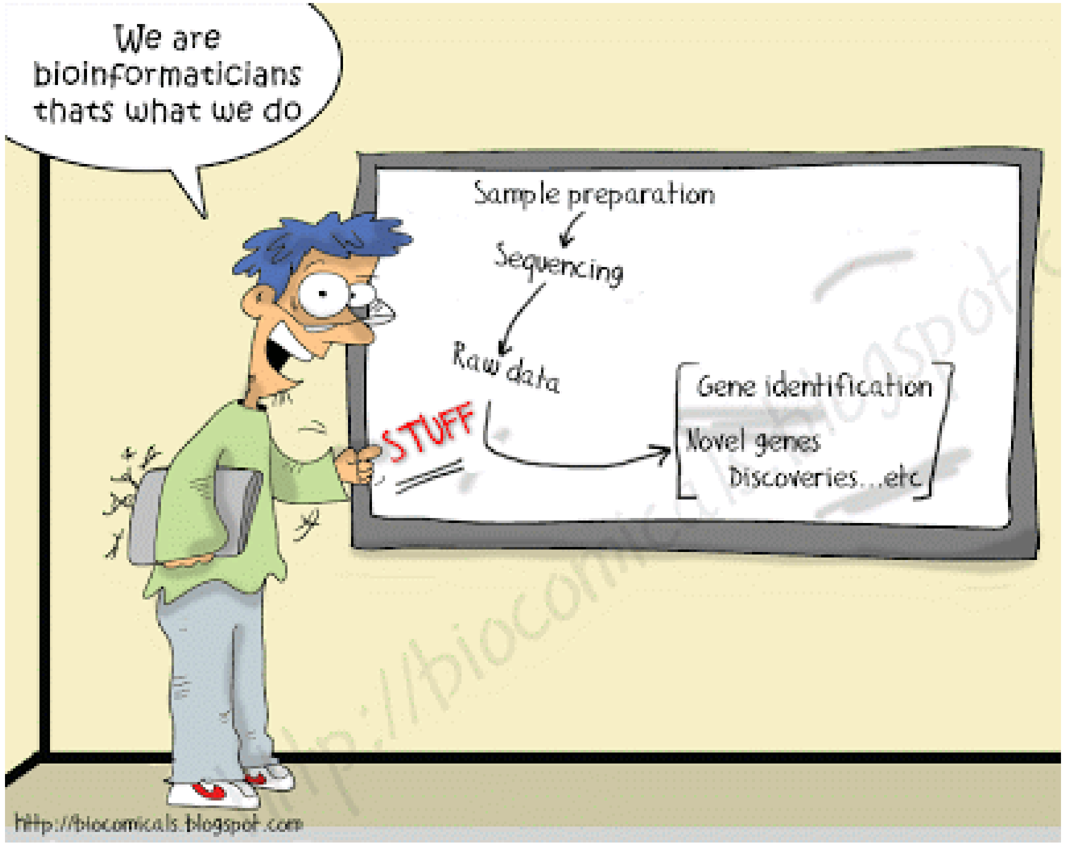
- ~50 TB/year
- ~10 users
- Big images visualization

genomics

- > 20 TB/year
- cpus and memory for data preparation and analyses

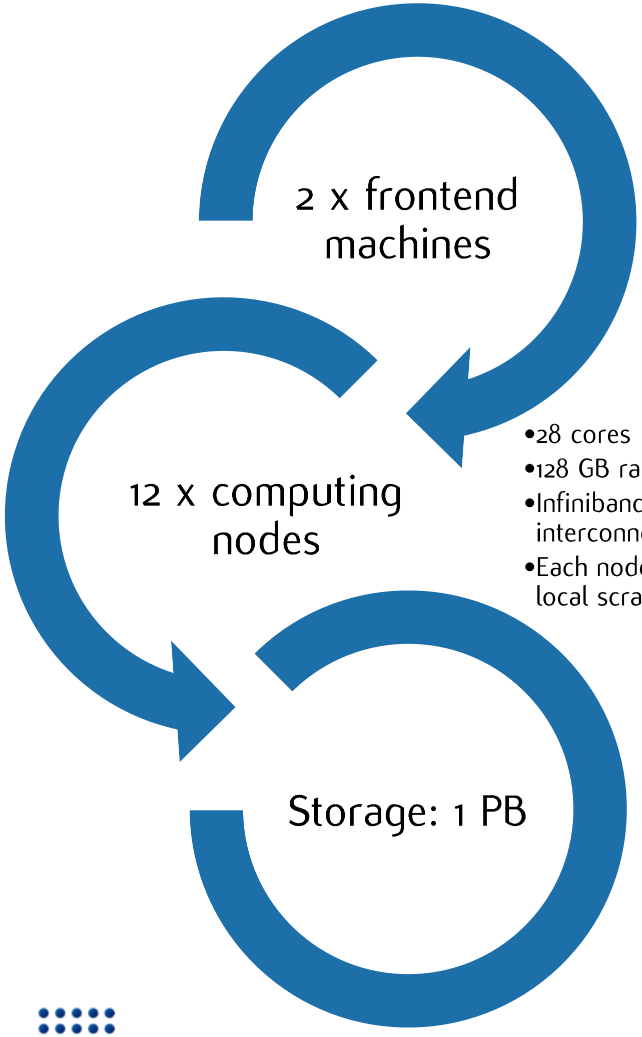
Proteomics, Radiomics, …

- > 100 TB/year ?
- CPUs, GPUs



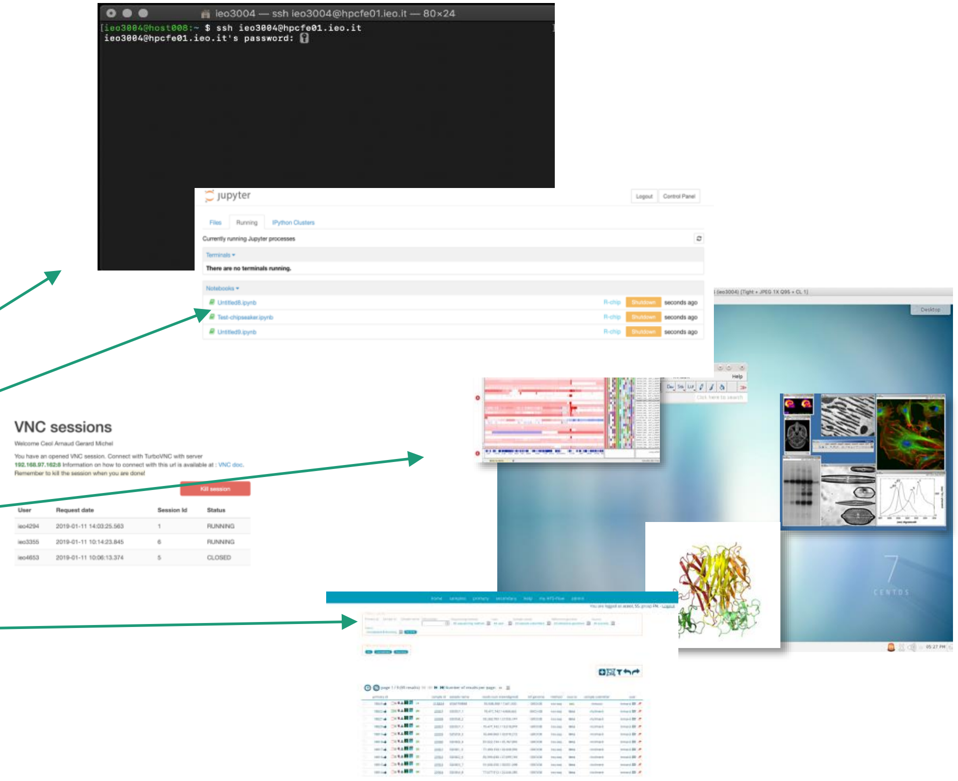
Dry-lab Researchers (bioinformaticians): analyze the data

The bioinfo (HPC) platform



- 24 cores
- 128 GB ram
- Nvidia Quadro M4000

- 28 cores
- 128 GB ram
- Infiniband low latency interconnection
- Each node with 1.8TB of local scratch



```
ieo3004 — ssh ieo3004@hpcfe01.ieo.it — 80×24
[ieo3004@host008:~ $ ssh ieo3004@hpcfe01.ieo.it
 ieo3004@hpcfe01.ieo.it's password:
```

Currently running Jupyter processes

Terminals ▾

There are no terminals running.

Notebooks ▾

Untitled8.ipynb	R-chip	Shutdown	seconds ago
Test-chipseaker.ipynb	R-chip	Shutdown	seconds ago
Untitled9.ipynb	R-chip	Shutdown	seconds ago

VNC sessions

Welcome Ceol Arnaud Gerard Michel

You have an opened VNC session. Connect with TurboVNC with server 192.168.97.162:8 Information on how to connect with this url is available at : VNC doc.
Remember to kill the session when you are done!

User	Request date	Session Id	Status
ieo4294	2019-01-11 14:03:25.563	1	RUNNING
ieo3355	2019-01-11 10:14:23.845	6	RUNNING
ieo4653	2019-01-11 10:06:13.374	5	CLOSED

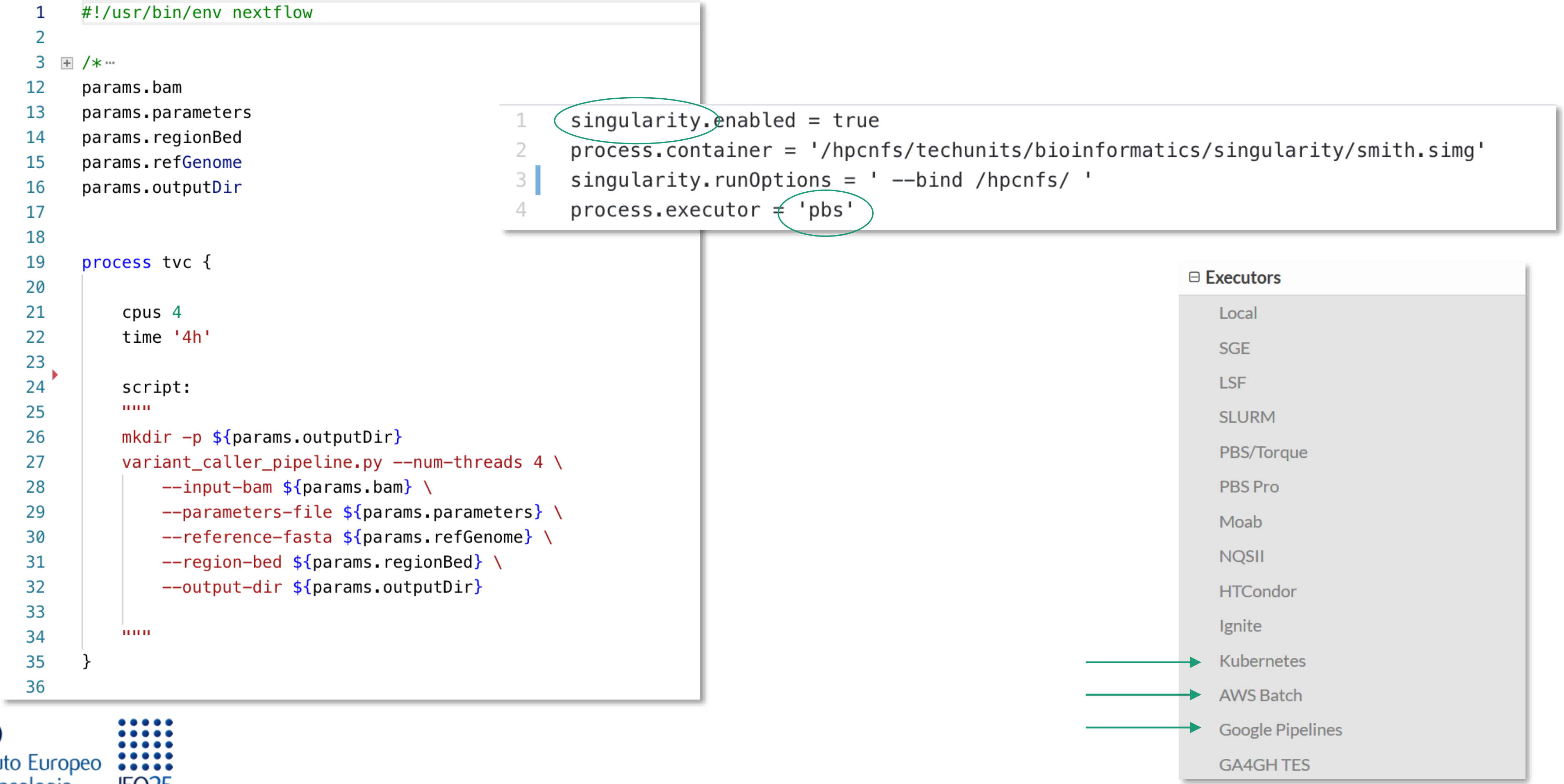
We encourage the usage of:
- **Containers**: 100% reproducible results
- pipeline managers (Nextflow, Snakemake)

IEO Istituto Europeo di Oncologia IEO25

nextflow

nextflow

https://www.nextflow.io/



```
#!/usr/bin/env nextflow

/*⋯
params.bam
params.parameters
params.regionBed
params.refGenome
params.outputDir


process tvc {

    cpus 4
    time '4h'

    script:
    """
    mkdir -p ${params.outputDir}
    variant_caller_pipeline.py --num-threads 4 \
        --input-bam ${params.bam} \
        --parameters-file ${params.parameters} \
        --reference-fasta ${params.refGenome} \
        --region-bed ${params.regionBed} \
        --output-dir ${params.outputDir}

    """
}

```

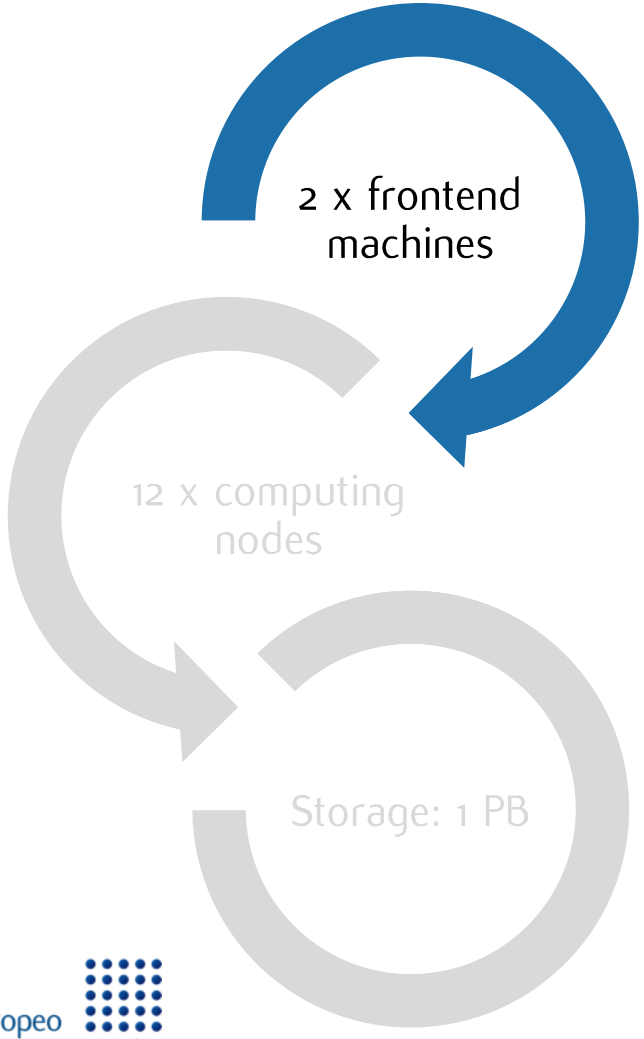
```
singularity.enabled = true
process.container = '/hpcnfs/techunits/bioinformatics/singularity/smith.simg'
singularity.runOptions = ' --bind /hpcnfs/ '
process.executor = 'pbs'
```

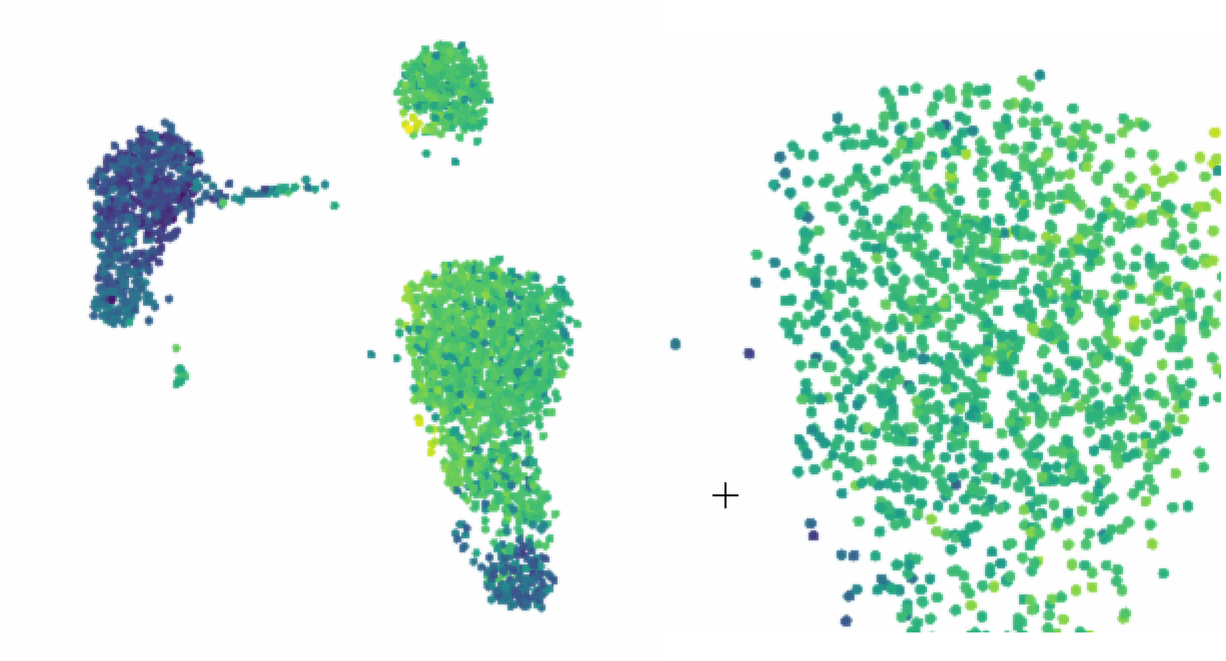
⊟ Executors

- Local
- SGE
- LSF
- SLURM
- PBS/Torque
- PBS Pro
- Moab
- NQSII
- HTCondor
- Ignite
- Kubernetes
- AWS Batch
- Google Pipelines
- GA4GH TES

HPC: limitations, and how the cloud can help

Limitation(s) 1: no admin, no Docker, no web-serving, no outside access

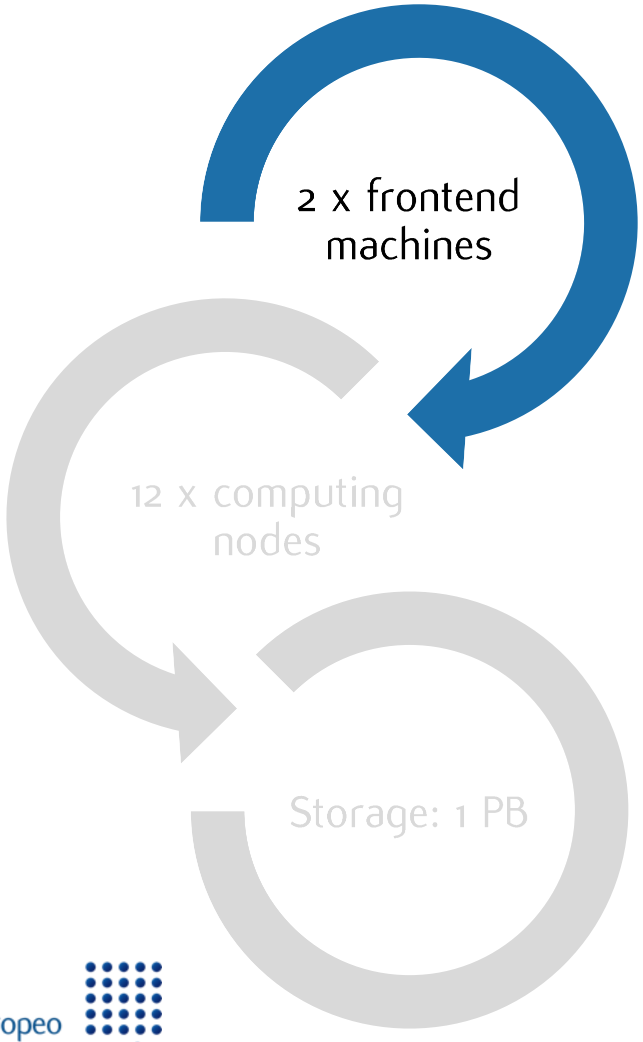




Example: testing docker-based web application for the visualization and analyses of single cell sequencing

cell×gene

Limitation 1: no admin and web serving



Cloud solution →

Virtual Machine 1:
Managed by the user:
Docker + floating IP

openstack

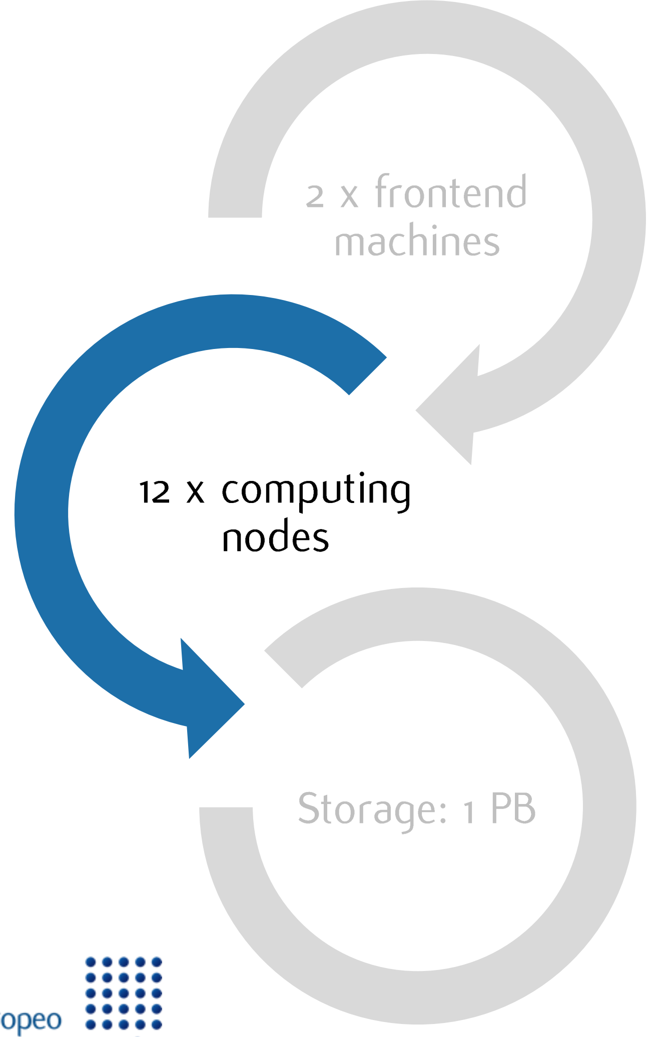
Virtual Machine 2:
Provide to all users of the institute a "HPC-like" image.

openstack

JAAS

The cloud allows to create testing environment & publish web servers/services

Limitation 2: cores, GPU, RAM

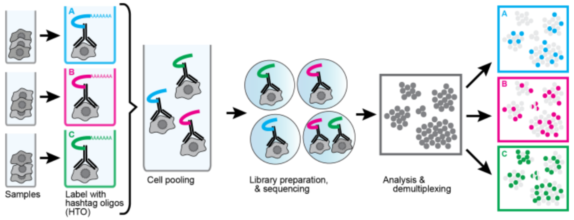


2 x frontend machines

12 x computing nodes

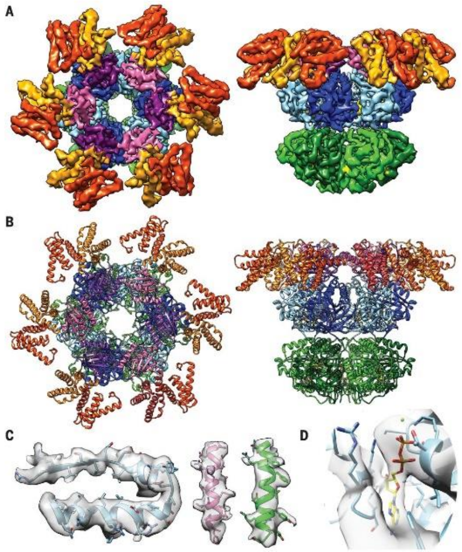
Storage: 1 PB

Example 1: CRISPR screening ortoghonal interference of 1M cells with multiplexing-enabling. **5TB** raw data to be analized **100000 cpu/hour** and for the integration and analysis we need a **512GB ram** machine


A
B
C
Samples
Label with hashtag oligos (HTO)
Cell pooling
Library preparation, & sequencing
Analysis & demultiplexing

Example 2: DeepVariant is an analysis pipeline that uses a deep neural network to call genetic variants from next-generation DNA sequencing data: needs VM with **64 CPUs 240 GB RAM**

Example 3: Cryo-EM analyze large, complex and flexible structures. 1 project = **3 weeks to 6 months** with **2 GPUs**



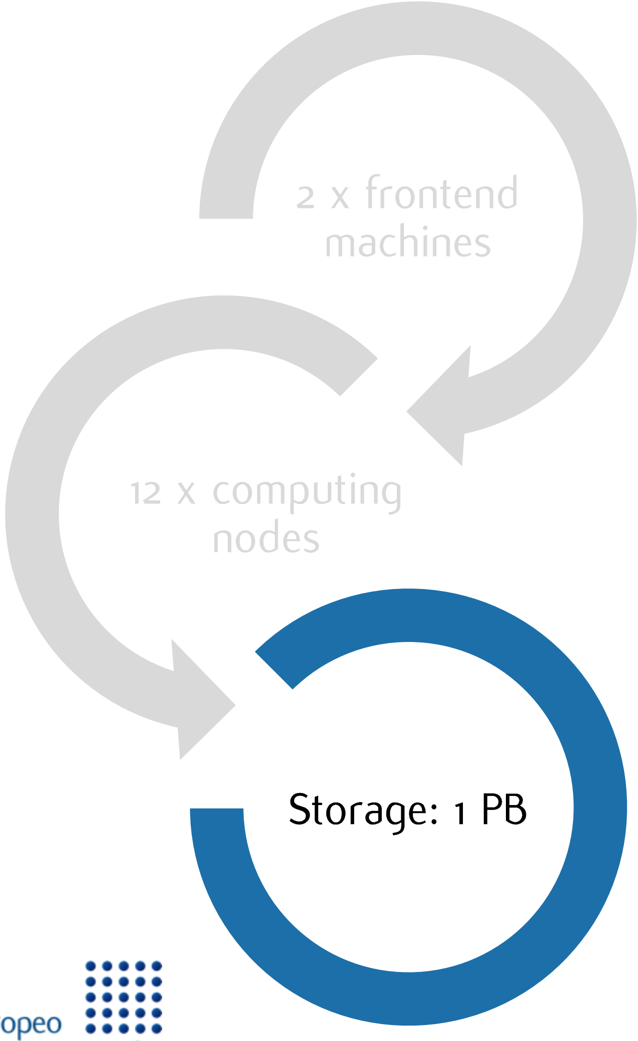
Cloud solution →

openstack™

IEO Istituto Europeo di Oncologia

IEO25

Limitation 3: storage of archives



Storage: Isilon (1PB)

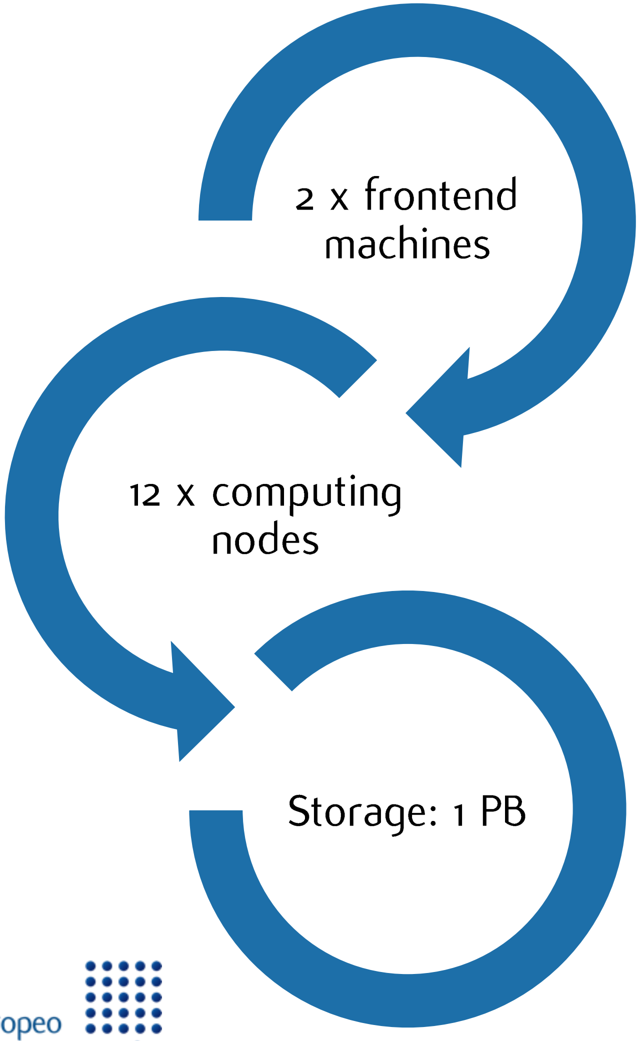
Growth rate: ~ 100 TB/year (only raw data), but will increase.

Cold data: > 300 TB
- 90 % will not be used anymore
- cannot be deleted (publication, patents, etc.)

Cloud solution →

Cloud storage:
-> admin-based management
-> user-based management

Limitation 1: exchange and collaboration



HPC and storage accessible only from the IEO network (intranet).

Cloud solution →

RCLONE

ceph

Summary and further considerations

Cloud for: freedom, BIG stuff, long term

Virtual data centers
for collaborative projects

Clinical/biomedical data → Solutions & guidance are welcome

IEO25

1994-2019
25 anni di ricerca e innovazione
per la lotta al cancro,
25 anni di Istituto Europeo di Oncologia

Arnaud Ceol, arnaud.ceol@ieo.it