The HPC IBiSCo seeds for INFN experiments workflow-based cluster

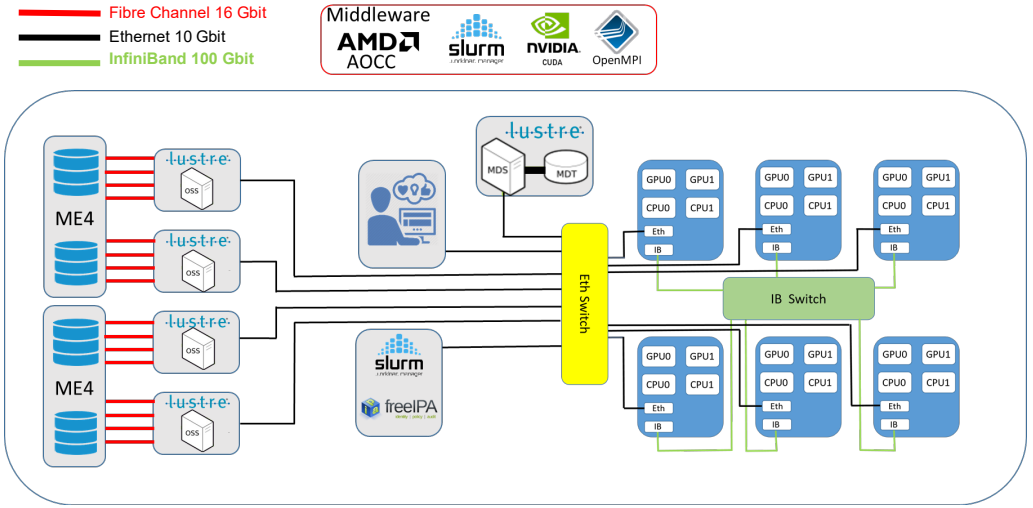
B.Spisso, G.Carlino, F.Cirotto, A.D'Onofrio, A.Doria, G.Sabella.

INFN Istituto Nazionale di Fisica Nucleare
UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale
Ministero dell'Università e della Ricerca
PON RICERCA E INNOVAZIONE 2014 - 2020
INFN NAPOLI Istituto Nazionale di Fisica Nucleare Sezione di Napoli

Introduction

In this presentation, I will highlight the valuable insights gained in high-performance computing (HPC) within the Naples INFN section, stemming from the foundational 'seeds' of the IBiSCo initiative. Following a brief overview of the cluster, I will delve into our engagement with the ATLAS user community.
This use-case exemplifies how such collaborations have been instrumental in defining the cluster's development.

Introduction to the HPC cluster of INFN Naples



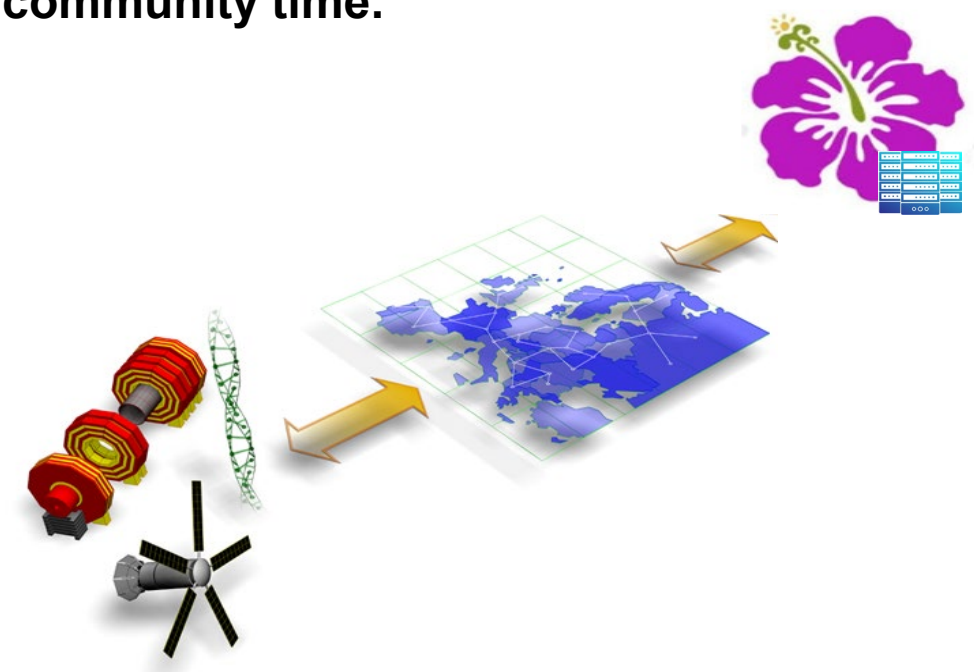
- 6x PowerEdge R7525 many-core compute nodes equipped with GPUs
- 2x Dell Powervault ME4 Storage System for short to medium term, approximately 1000 TB raw divided between all-flash and mechanical partitions.
- Infiniband 100 Gbit/s interconnection
- 2x AMD EPYC 7742 CPUs, 128 (64x2) cores @ 2.250 GHz
- 2x NVIDIA V100 16 GB PCIe 3.0 GPUs
- Main memory: 1200 GB DDR4
- 2x SATA solid-state drives of 446.63 GB 2 SATA solid-state drives of 3576.38 GB

Same base architecture (Slurm + Nis + Infinibad + Lustre + CUDA) as the UNINA HPC cluster
We recently moved to Lustre from standard NFS (transition is still on going)
Close but different HW and newer OS (Alma9)

Many thanks to Dr. Luisa Carraciulo for sharing her deployment experience; it served as a treasure map.

Introduction to the HPC cluster of INFN Naples

The INFN main business are the experiments and their workflow. The cluster must support their workflows in an optimized fashion to optimize **the admin and user community time.**



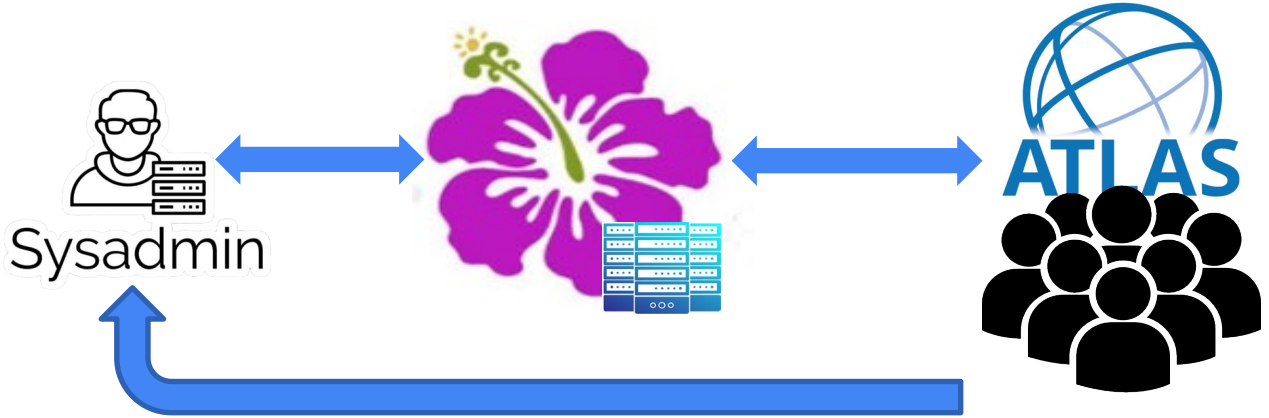
ATLAS EXPERIMENT

KM3NeT

Hyper-K

Experiments workflow-based cluster

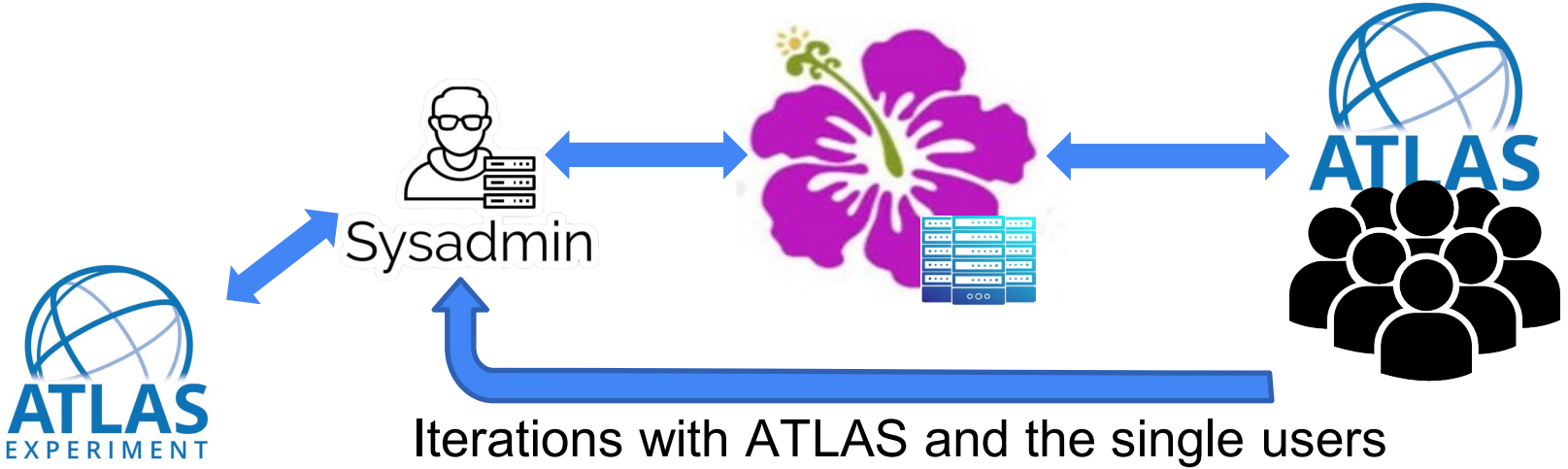
In the following, I will describe our experience and the strategy adopted for the management of workflows and software tools, taking as a case study the ATLAS experiment.



Iterations with the single users

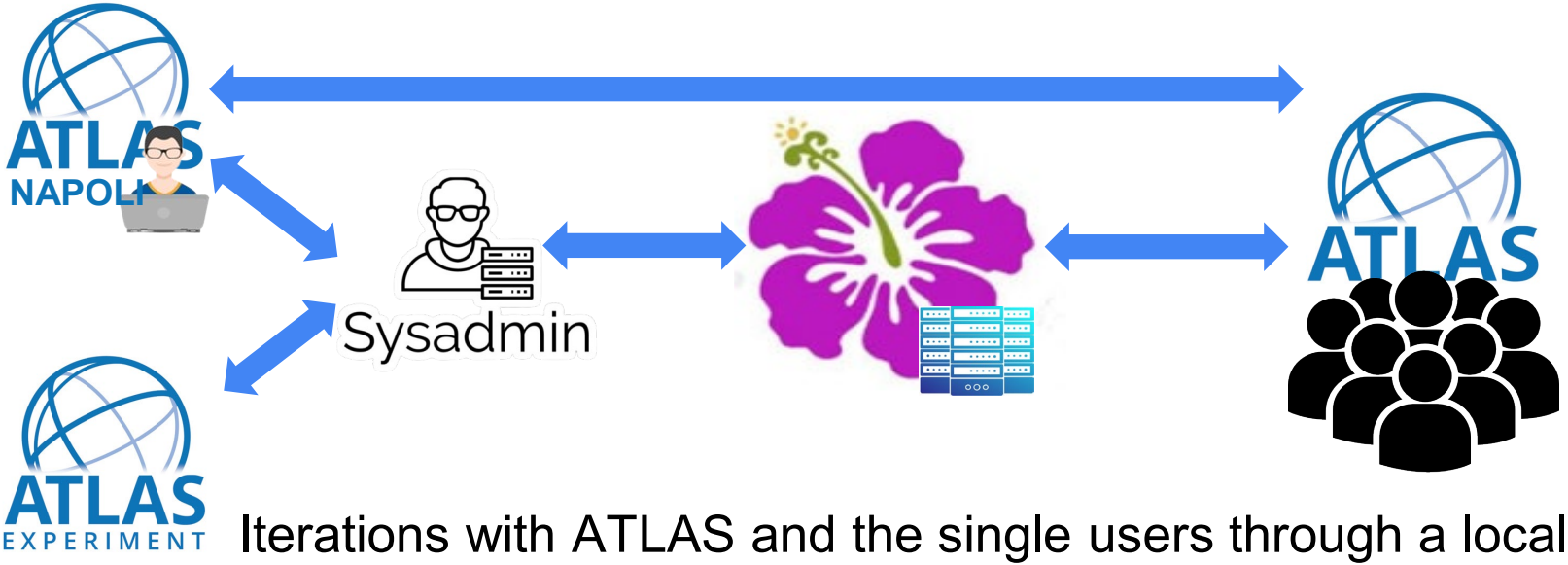
Experiments workflow-based cluster

In the following, I will describe our experience and the strategy adopted for the management of workflows and software tools, taking as a case study the ATLAS experiment.



Experiments workflow-based cluster

In the following, I will describe our experience and the strategy adopted for the management of workflows and software tools, taking as a case study the ATLAS experiment.



Iterations with ATLAS and the single users through a local group responsible

Implementation of tools for workflow support

From the previous model, in addition to the basic tools for the ATLAS workflows (such as **CvmFS, Miniconda, Apptainer**, etc…), there have been several requests emerging that were specific to the local group.
Among the most requested there were:

- **Jupyter HUB on the UI**
- **Jupyter Notebook on the nodes**
- **Dask**
- **A bridge to the HTC part of ATLAS**

Implementation of tools for workflow support

From the previous model, in addition to the basic tools for the ATLAS workflows (such as **CvmFS, Miniconda, Apptainer**, etc…), there have been several requests emerging that were specific to the local group.
Among the most requested there were:

- **Jupyter HUB on the UI**
- **Jupyter Notebook on the nodes**
- **Dask**
- **A bridge to the HTC part of ATLAS**

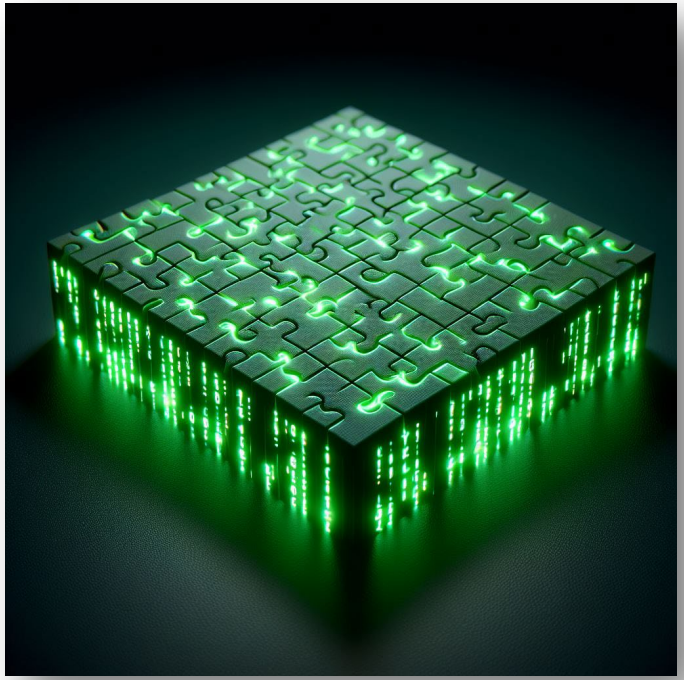
Implementation of tools for workflow support

From the previous model, in addition to the basic tools for the ATLAS workflows (such as **CvmFS, Miniconda, Apptainer**, etc…), there have been several requests emerging that were specific to the local group.
Among the most requested there were:

- **Jupyter HUB on the UI**
- **Jupyter Notebook on the nodes**
- **Dask**
- **A bridge to the HTC part of ATLAS**

Implementation of tools for workflow support

So, I started to collect pieces of the puzzle:



Can run scripts prior and after a job is launched (prolog and epilog mechanism)

Can redirect an HTTP/s service from one of the node to the UI.

Can be accessed via token authentication and can be configured to use an arbitrary port.

A single piece remains to be found…

Implementation of tools for workflow support

A way for the user to turn on and off the entire mechanism

After some digging, I found a possible way out:

```
srun - Run parallel jobs
```

.
.
.

```
--comment=<string>
       An arbitrary comment. This option applies to job allocations.
```

Easy enough for the users and most of all, **the variable SLURM_JOB_COMMENT is visible across all the cluster**

Implementation of tools for workflow support

Putting all together and a few moments later…

```
[bspisso@ibisco-ui ~]$ srun -w ibisco-gpu03 --partition=gpus  --cpus-per-task=1
--gpus-per-node=1  --comment="JUPYTER" --pty bash -l
```

```
--comment="JUPYTER"
```

Implementation of tools for workflow support

Putting all together and a few moments later…

```
[bspisso@ibisco-ui ~]$ srun -w ibisco-gpu03 --partition=gpus  --cpus-per-task=1
--gpus-per-node=1  --comment="JUPYTER" --pty bash -l
```

```
--comment="JUPYTER"
```

```
Jupyter notebook started on port 18608
On you favorite broswer go to  http://ibisco-ui.na.infn.it:
18608?token=846c4240db0a55c23daff8f89d92fc92a338292b257b7f16
```

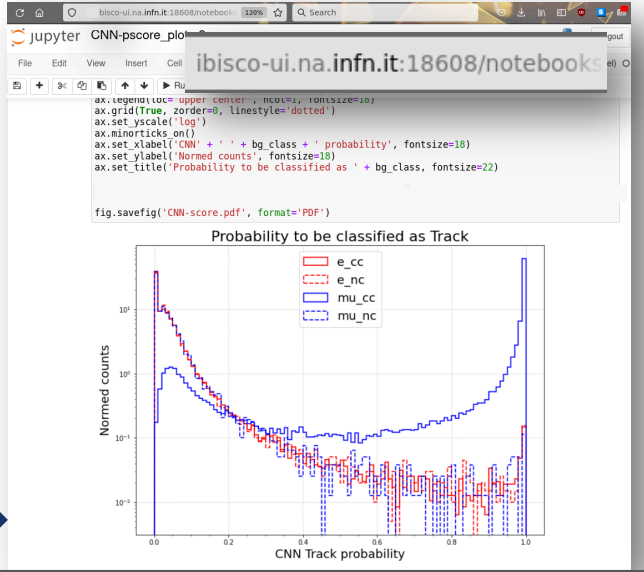
Implementation of tools for workflow support

Putting all together and a few moments later…

```
[bspisso@ibisco-ui ~]$ srun -w ibisco-gpu03 --partition=gpus  --cpus-per-task=1
--gpus-per-node=1  --comment="JUPYTER" --pty bash -l
```

```
--comment="JUPYTER"
```

```
Jupyter notebook started on port 18608
On you favorite broswer go to  http://ibisco-ui.na.infn.it:
18608?token=846c4240db0a55c23daff8f89d92fc92a338292b257b7f16
```



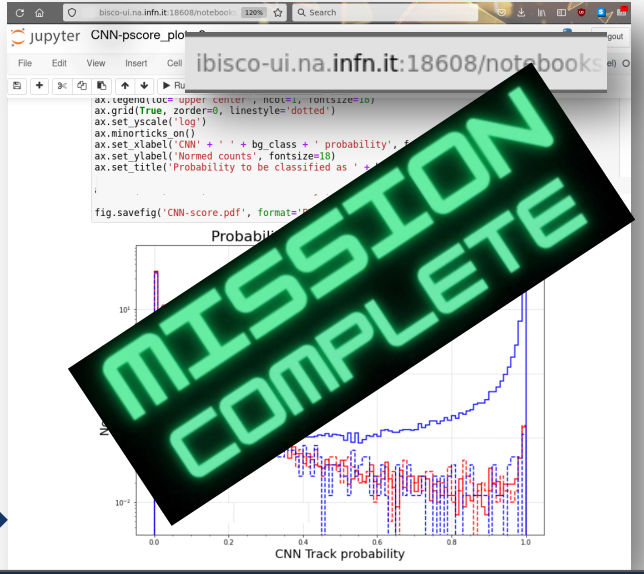
Implementation of tools for workflow support

Putting all together and a few moments later…

```
[bspisso@ibisco-ui ~]$ srun -w ibisco-gpu03 --partition=gpus  --cpus-per-task=1
--gpus-per-node=1  --comment="JUPYTER" --pty bash -l
```

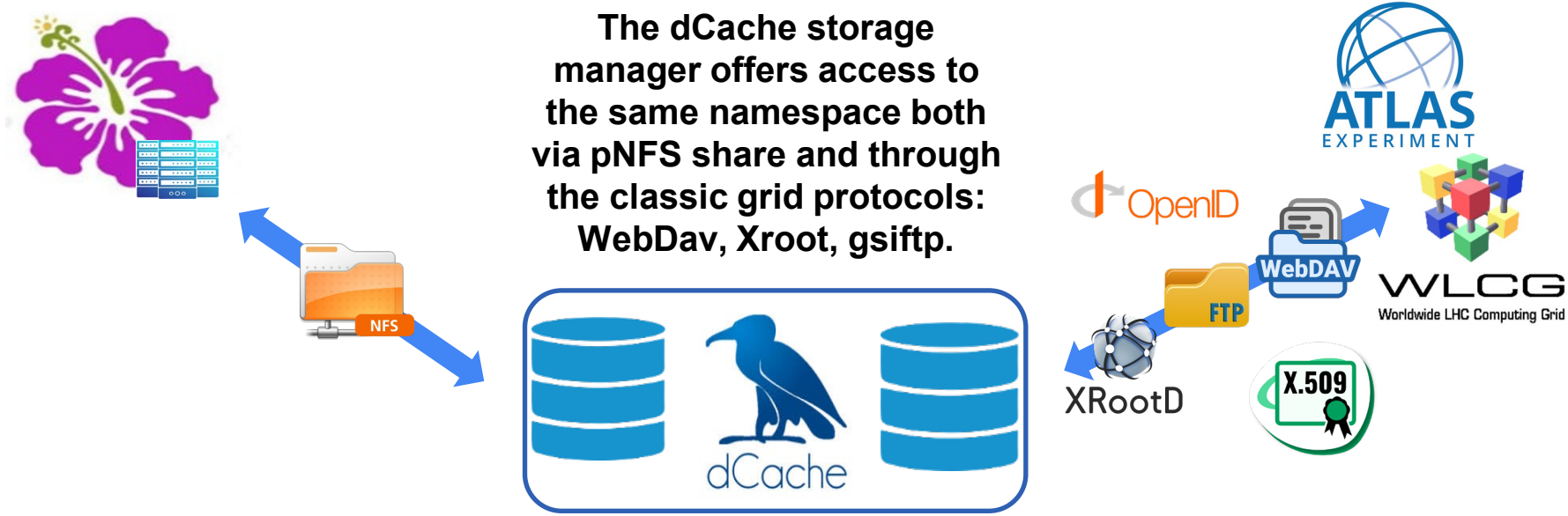
```
--comment="JUPYTER"
```

```
Jupyter notebook started on port 18608
On you favorite broswer go to  http://ibisco-ui.na.infn.it:
18608?token=846c4240db0a55c23daff8f89d92fc92a338292b257b7f16
```



An experimental HTC-HPC workflow

Furthermore, in addition to the native tools of the ATLAS HTC workflows (such as Rucio, GFAL, etc.), I was seeking a tighter integration between the two realms. **I started with the data using dCache!**



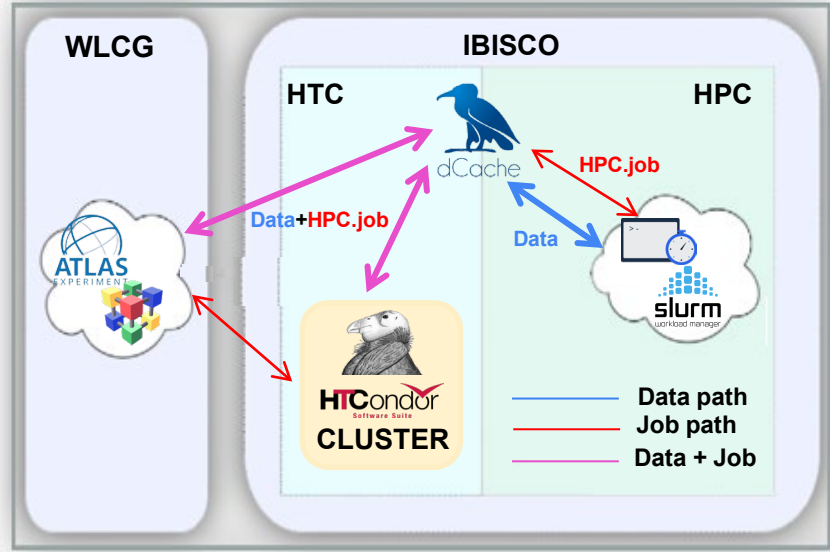
The dCache storage manager offers access to the same namespace both via pNFS share and through the classic grid protocols: WebDav, Xroot, gsiftp.

A GRID (HTC) job can natively utilize a storage space as its source or transfer its output to one that is directly accessible by Slurm

An experimental HTC-HPC workflow

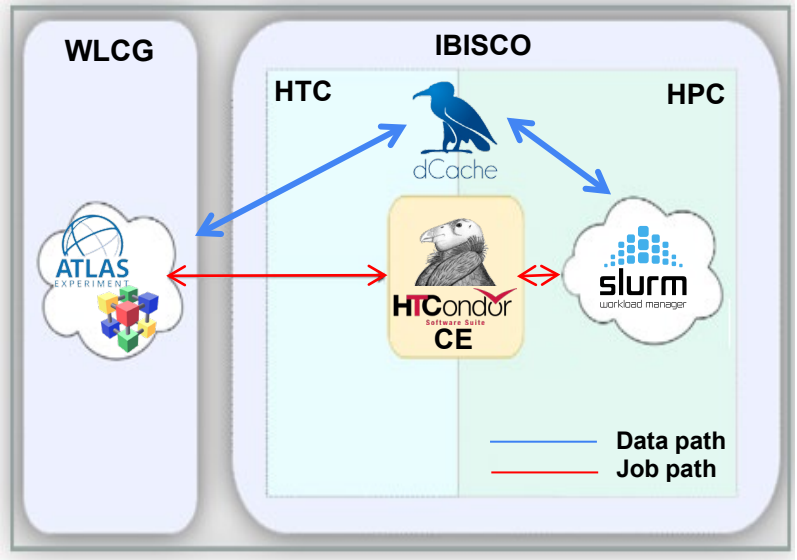
Regarding job integration, we are exploring two branches:

HPC backfill through dCache



A predetermined dCache location will be periodically monitored by the HPC cluster through a cron daemon, looking for a Slurm job.

HPC backfill with HTCondor



Through a conversion layer, it is possible to submit a Slurm job via an HTCondor CE using a JDL file.

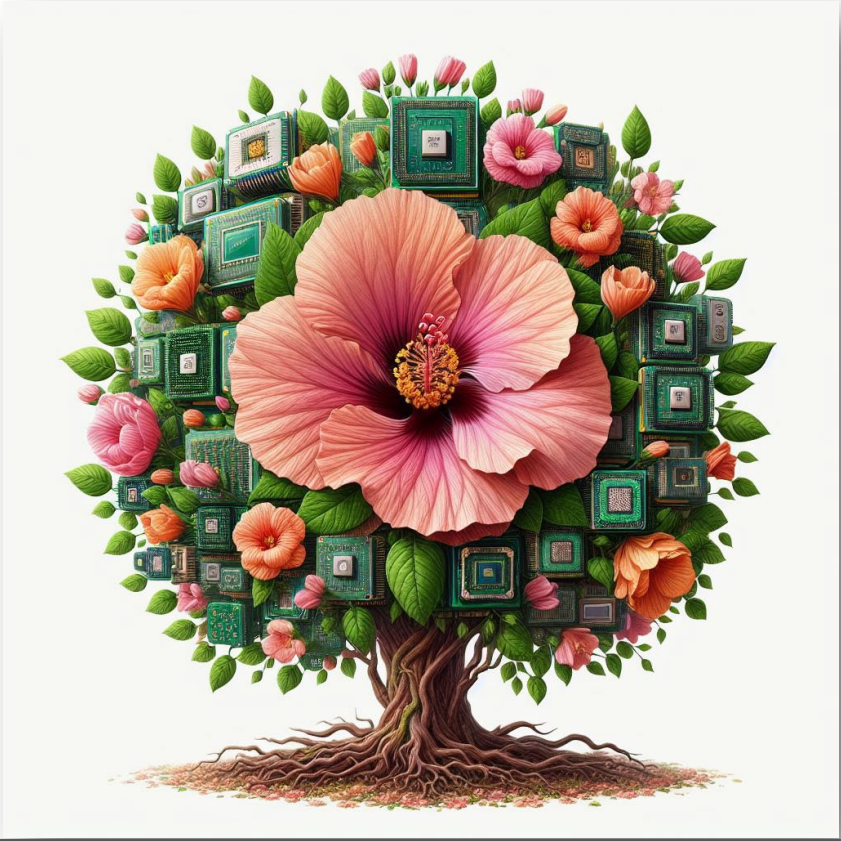
Conclusion and growth

The seeds of IBiSCo have flourished, nurturing a tree rich with knowledge and expertise in high-performance computing, which continues to thrive and broaden its reach.

We raised from the scratch a working HPC cluster witch we are now tuning on the communities needs without renounce to experiment new solutions.

We are still rooting, so there is much to do:

- Complete the authentication via national SSO
- Complete the transition from the previous NFS to the Lustre file system
- Scale up the storage bandwidth from 10 to 100 Gbit/s
- And go on with the experimentation



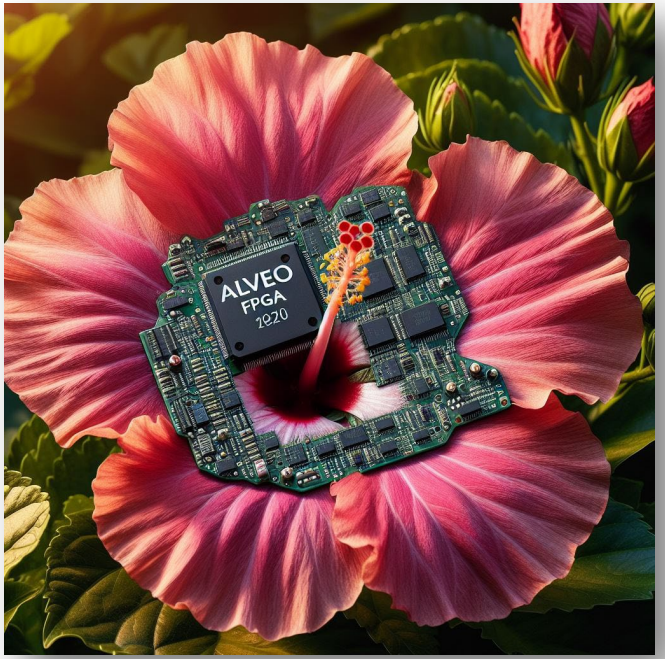
Conclusion and growth

Waiting for the next bloom.

Hardware expected at the end of 2024

Thanks to the PNRR funding, the HPC division of the Naples section will secure additional resources:

- New **InfiniBand** CPU nodes
- An **4x Nvidia H100 GPU** equipped machine
- 2x **FPGA** nodes each equipped 4 **Xilinx U55C Alveo** boards



Thank you for your attention