PIC
port d'informació
científica

JupyterHub at
Port d’Informació Científica (PIC)

Francesc Torradeflot

CS3 JupyterHub Community Technical Workshop 2024
2024-05-15

IFAE Institut de Física d’Altes Energies

BIST Barcelona Institute of Science and Technology

PIC port d'informació científica

GOBIERNO DE ESPAÑA — MINISTERIO DE CIENCIA E INNOVACIÓN

Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

cfp CIEMAT física de partículas

Outline

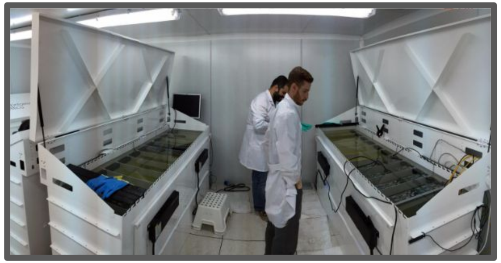
- Context: What is PIC?
- Jupyterhub at PIC
  - Early days
  - Current status
  - What’s next

What is PIC?

- PIC stands for Port d’Informació Científica
- Founded in 2003, collaboration between IFAE and CIEMAT. Located near Barcelona in the UAB campus.
- Tier-1 node of the WLCG with the mission to transfer this knowledge and technologies to other activities
- Team of 23 people (50% scientists - 50% engineers)
  - Agile teams that embed in scientific groups to
- What we do
  - R&D in methodologies and tools for advanced data analysis
  - Operate services for the preservation, analysis and sharing of data

What is PIC?

- Connectivity
  - 2x100 Gbps to Academic Network
  - 100 PB in+out per year
- Data processing services
  - Disk - dCache: 20 PB (+Ceph 3.5 PB raw)
  - Tape - Enstore: 63 PB
  - Computing - HTCondor: 12000 cores, 16 GPUs
  - Computing - Hadoop: 720 cores, 2.5 PB disk
- Facilities, ~120 kW IT
  - ~80 kW in 150 m^2 air-cooled room
  - ~40 kW in 25 m^2 liquid immersion cooling system
- Kubernetes, VMs, etc



What is PIC?

Traditionally involved in Physics experiments:

- Particle Physics
- Cosmology
- Gamma-ray Astronomy
- Gravitational Waves
- Neutrinos

Recently transferring the knowledge to other fields: bioimaging, materials sciences, health sciences, etc

WLCG
Worldwide LHC Computing Grid
ARCHIVER
ESCAPE
CMS
ATLAS
EXPERIMENT
LHCb
CERN
MAGIC
cta
cherenkov telescope array
Euclid
consortium
T2K
PAU
DARK ENERGY
SPECTROSCOPIC
INSTRUMENT
U.S. Department of Energy Office of Science
DARK ENERGY
SURVEY
LIGO
VIRGO
DUNE
DEEP UNDERGROUND
NEUTRINO EXPERIMENT
MICE

Early days

JupyterHub service was started as a testbed to:

- learn kubernetes
- provide interactive access to GPUs

It worked fine but:

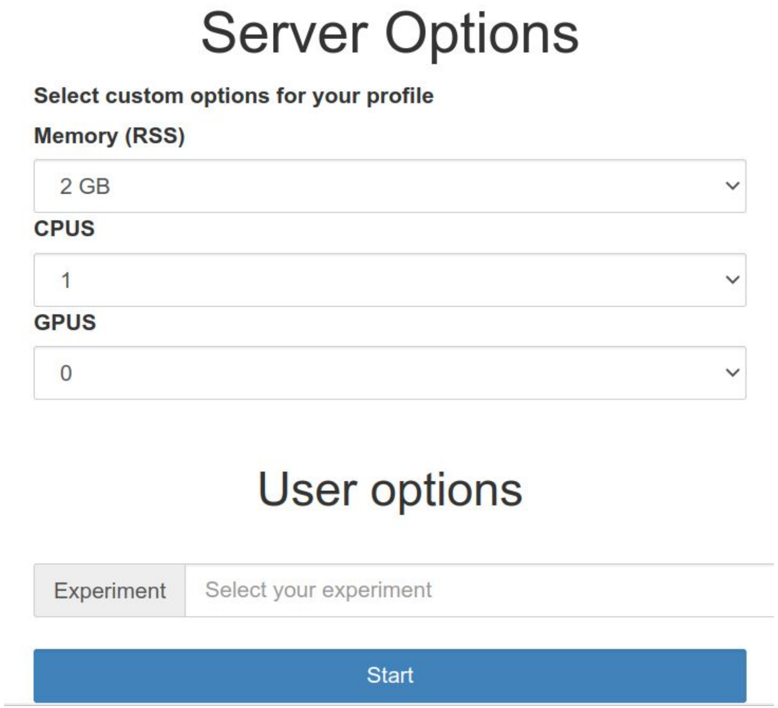
- It wasn’t integrated with our main resource manager: HTCondor
    - Independent resources -> idle resources
    - No accounting
    - No priorities

Early days

- Integration with other PIC services wasn’t straightforward
  - Access to massive storage with POSIX permissions
  - Importing user HOME
  - Alerts / Monitoring
- Maintenance was hard
  - Custom images and Helm charts apart from the python/conda environment
  - Newbies to k8s
- Looking into the future
  - Dask clusters
  - Connectivity with Hadoop cluster

Current status: Overview

- Launch a jupyter notebook server on PIC's HTC cluster using jupyterhub and batchspawner
- User-defined resources
  - CPUs
  - Memory
  - GPUs
- Choose experiment for accounting and POSIX permissions
- Managed with puppet & gitlab CI/CD
- High priority jobs to minimize waiting time



Server Options

Select custom options for your profile

Memory (RSS)

2 GB

CPUS

1

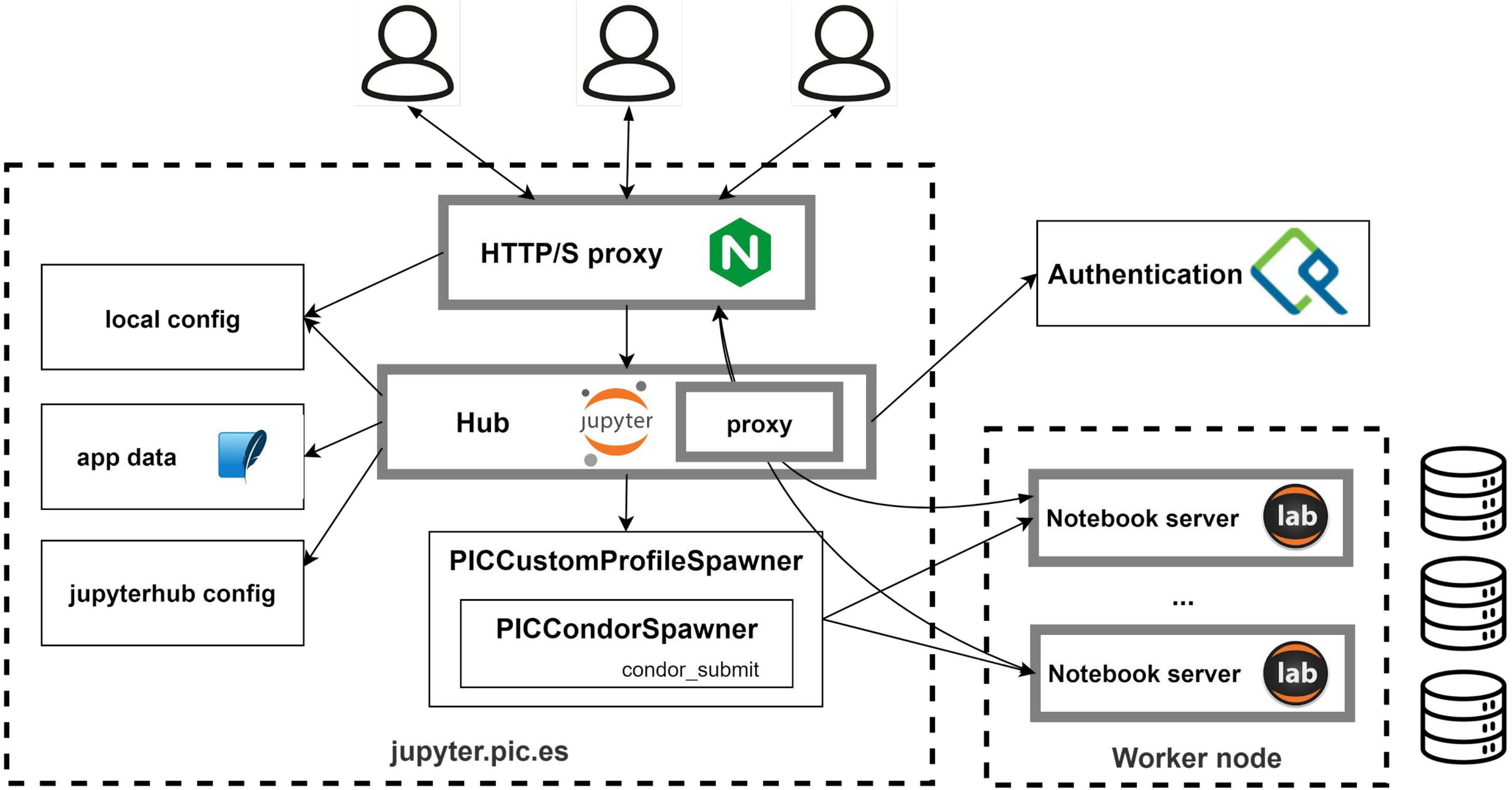
GPUS

0

User options

Experiment: Select your experiment

Current status: Overview



Current status: environment

We provide a python environment with the most common scientific libraries

Python 3.11 Numpy 1.24 Matplotlib 3.7 pandas 2.0 scipy 1.10

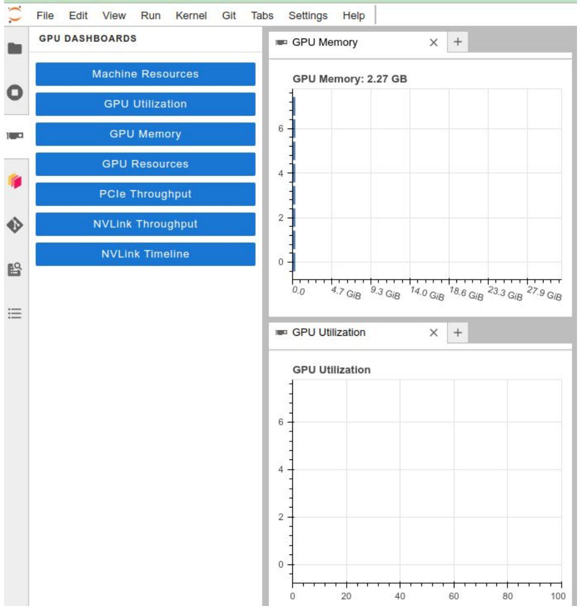
And some additions

astropy scikit-learn scikit-image Dask pillow

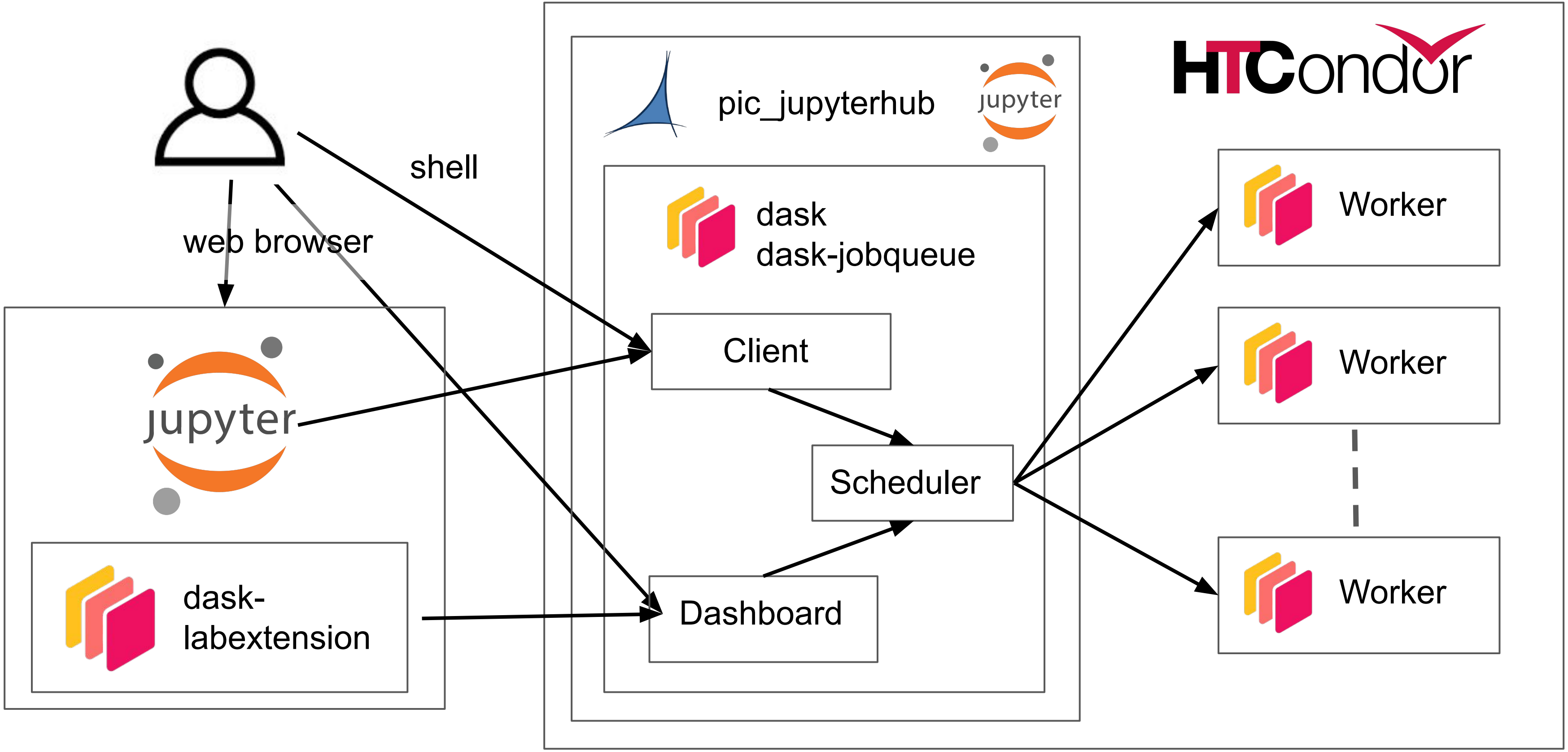
seaborn bokeh plotly statsmodels jupyter stack

Current Status: GPUs

- 16 GPUs available at PIC
  - gpu01: **8 x RTX 2080 Ti**, available via jupyter and HTCondor with preemption
  - gpu05: **8 x V100**, available via HTCondor with preemption, and a subset of 4 available via jupyter
- GPU dashboards in jupyterlab show the GPU usage
- No GPU libraries in the base environment
- Number of GPUs is going to scale up by the end of 2024



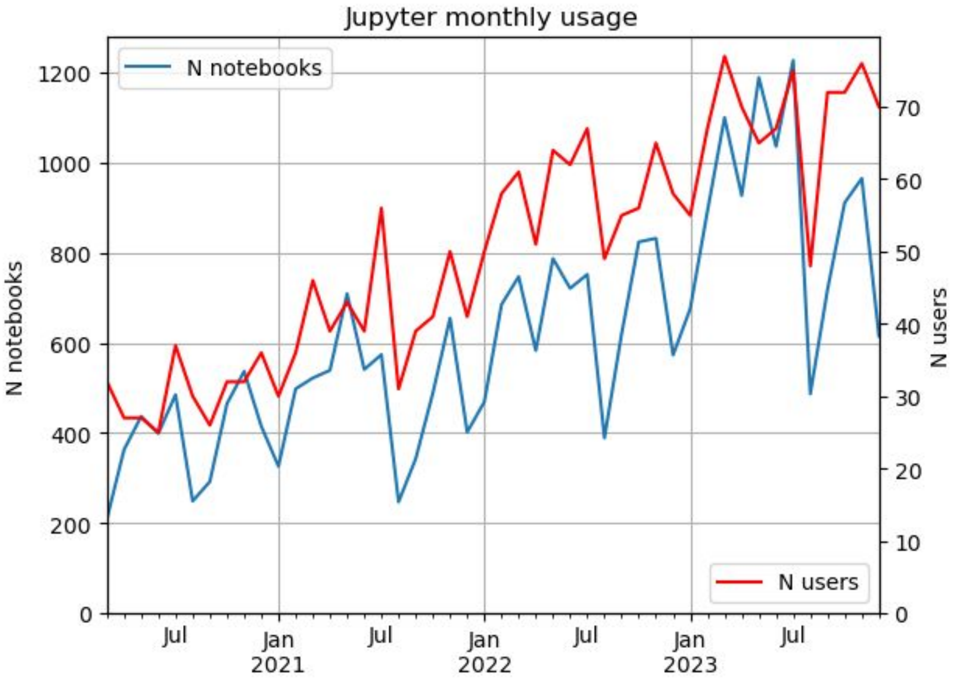
Current Status: Dask



PIC
port d'informació
científica

Current Status: usage

- Usage steadily increasing
  - From ~10 to >30 notebooks/day
  - From 25 to 60 active users
- Insignificant resource consumption
  - ~0.01% of total walltime norm
  - Low efficiency ~15%



Data baked by J.Casals

What’s next?

The JupyterHub service is very stable.

Well integrated into PIC’s main services → **low maintenance**

Flexible for users to use their own software → **few feature requests**

But there’s still some roadmap ahead

What’s next?

- Environment update
  - jupyterlab 4
  - Rucio-jupyterlab extension
- Integration with PIC’s Hadoop Cluster
- Improve Desktop interface
- Provide notebooks to non-typical users
  - 100s of students changing every semester
  - eduGAIN integration
  - This is why we are here
  - Go back to kubernetes (?!)

Thank you!

IFAE Institut de Física d'Altes Energies
BIST Barcelona Institute of Science and Technology
PIC port d'informació científica
GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas
cfp CIEMAT física de partículas