Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

INFN

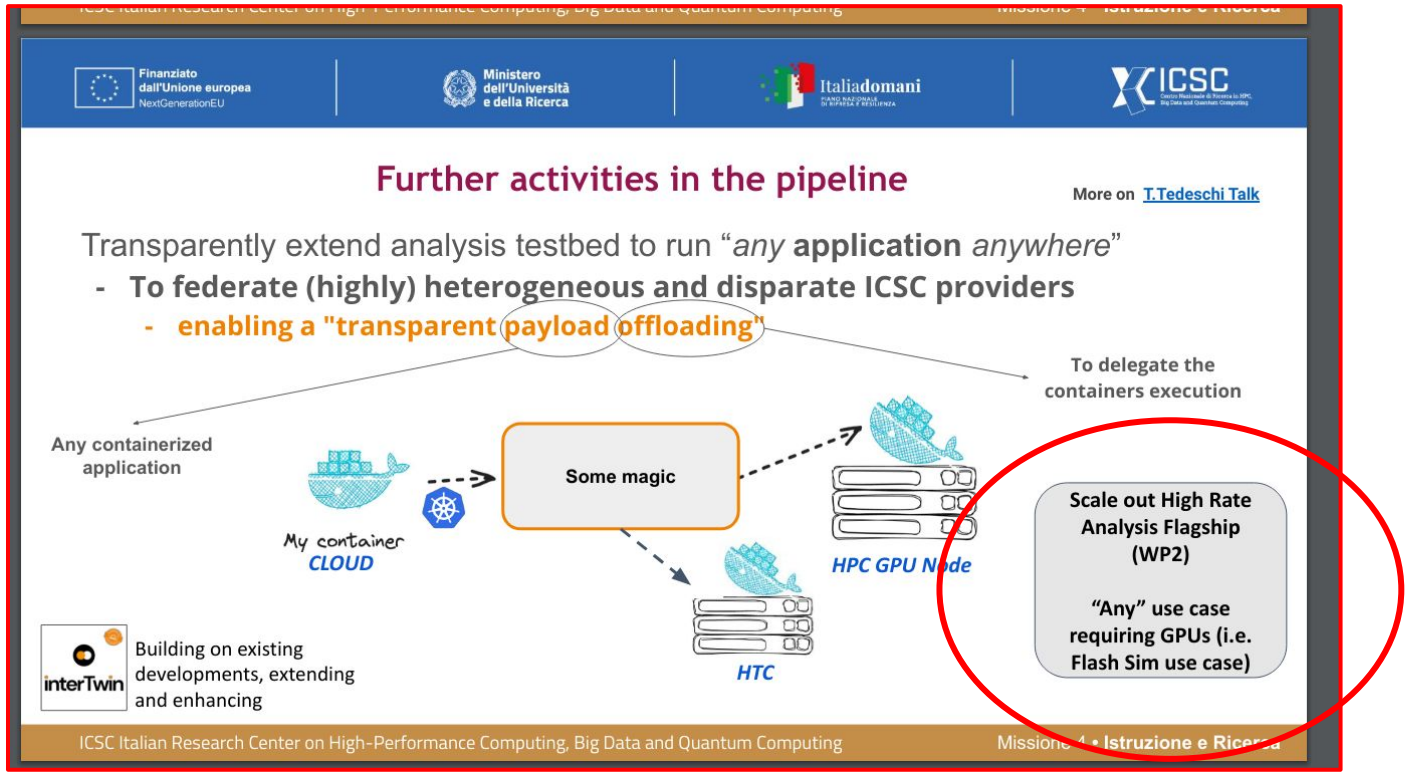
ICSC
Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

Introduzione all'Offloading

G. Bianchini, D. Ciangottini, D. Spiga, T. Tedeschi

ICSC Spoke 2 WP5 meeting - 7 June 2024 - Online

Razionale: why we talk about this topic



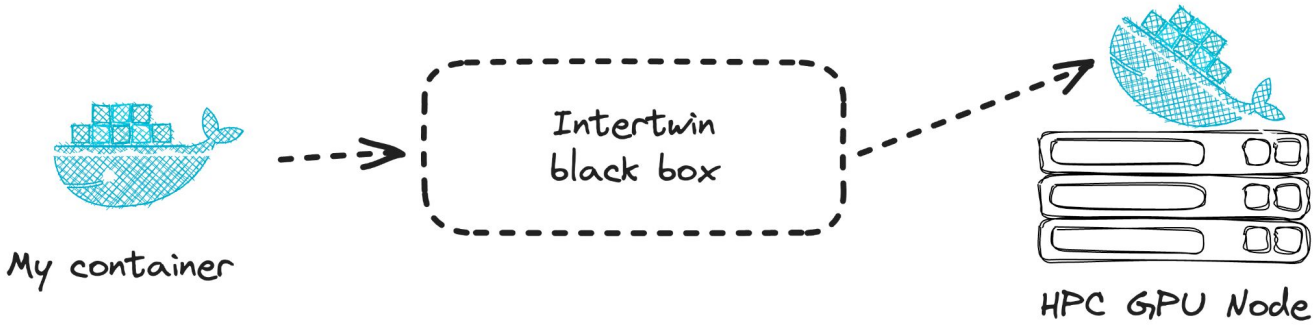
Dare seguito a quanto abbiamo nel nostro workplan

[Spoke2 Annual meeting contributo WP5]

What is offloading?

INFN

- **Delegate the execution of a container/workflow on remote resources while keeping the user interface unchanged.**
- Example:
  - "I have my own ML training container, and I want to run it on a node with 4 A100s"



How can we implement offloading?

INFN

We want a NATIVE integration with the Kubernetes primitives, acting underneath as a virtual node.

N.B. We aim to use Kubernetes as the workhorse for the "offloading", NOT as the user interface though

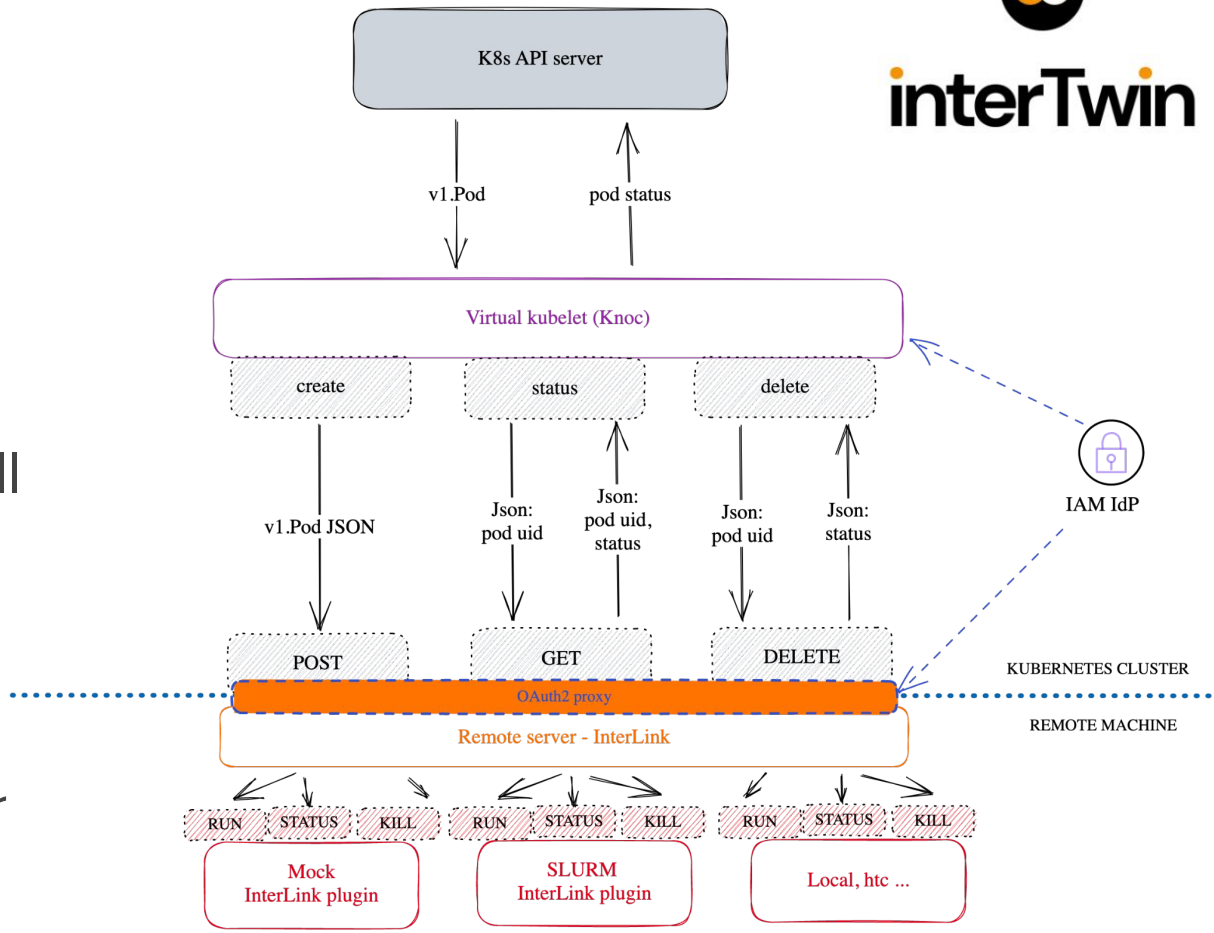
Offloading should be transparent for the users

A possible solution: InterLink

InterLink aims to provide an abstraction for the execution of a Kubernetes pod on any remote resource capable of managing a container execution lifecycle.

The project consists of two main components:

- **A Kubernetes Virtual Node**: based on the VirtualKubelet technology. Translating request for a kubernetes pod execution into a remote call to the interLink API server.
- **The interLink API server**: a modular and pluggable REST server where you can create your own container manager plugin (called sidecar), or use the existing ones: remote docker execution on a remote host, singularity Container on a remote SLURM or **HTCondor batch system**, etc...

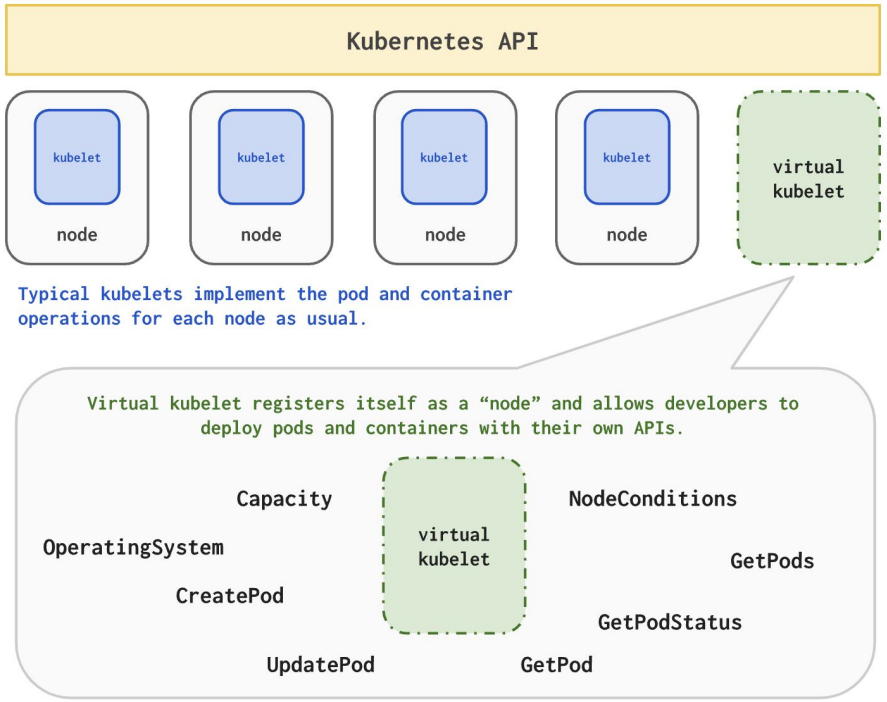


https://github.com/interTwin-eu/interLink

Components: VK

https://virtual-kubelet.io/

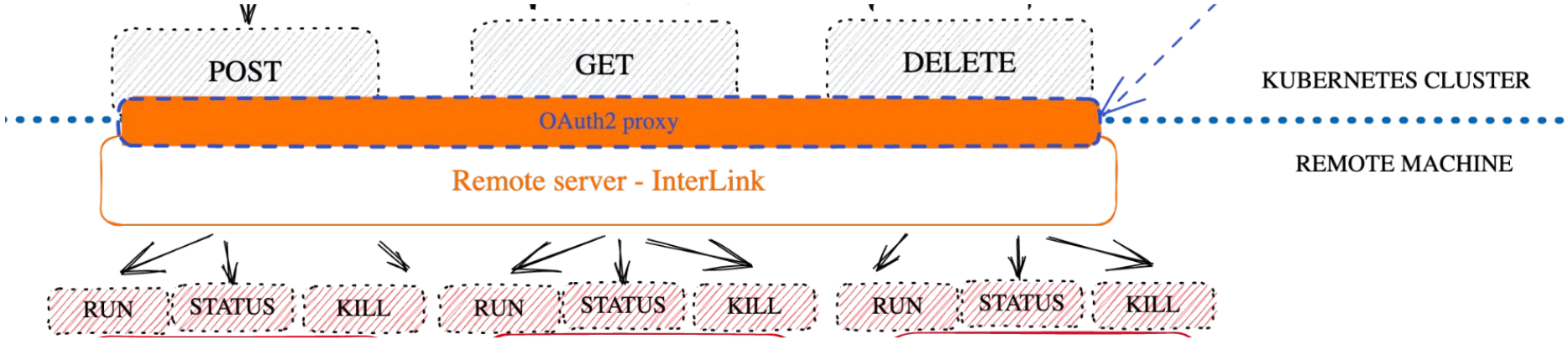
- **Virtual kubelet (VK)**:
  - "Open-source Kubernetes kubelet implementation that masquerades as a kubelet. This allows Kubernetes nodes to be backed by Virtual Kubelet providers"
- Can be imagined as a translation layer:
  - "I take your pod and run your container wherever I want"
- Registers virtual node and pulls work to run
- The pod lifecycle is managed via interlink rest calls
- Oauth2 via service token kept "refreshed"



Components: Interlink + Oauth2 proxy

INFN

- Oauth2 proxy: authN with IAM and authZ configurable on aud and groups
- "Digests" and manipulates calls from VK to the sidecar
- Self contained binary, distributable on all OS without dependencies



Components: Sidecar/Plugin

INFN

- Agent that must expose a REST with defined specs, but which can be implemented in the language and with the methods you prefer:
  - creation of the pod: run local docker or submit a job on htc, slurm etc
  - collect the execution states
  - collect and forward logs upon request
  - kill
- Existing plugins: local Docker (Go), SIurm (Go), HTCondor (python), ARC (python), Kubernetes (python), Kueue (python)

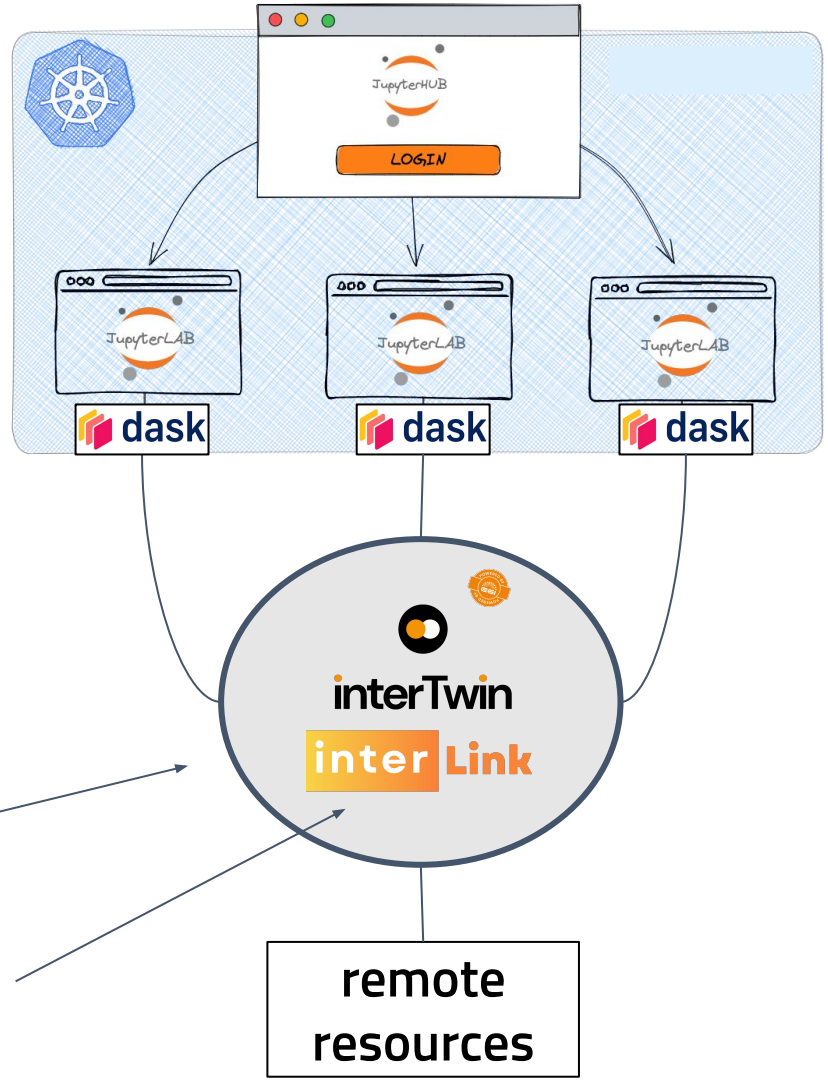
Offloading high rate analysis platform

Current general-purpose infrastructure:

- analyzers can scale up computations **within the cluster that possibly scale within the provider**
- Potentially, **a huge amount of users with diverse use cases may join**

Plan to enable the platform to dynamically exploit all kinds of resources (HTC, HPC, Cloud) transparently for the user

- looking for a synergy with active developments in this context, **to delegate container execution on remote resources** while keeping the very same user interface
  - Possible solution: **InterLink**, which provides execution of a Kubernetes pod on almost any remote resource



Flagship AI - AI_INFN Platform use case

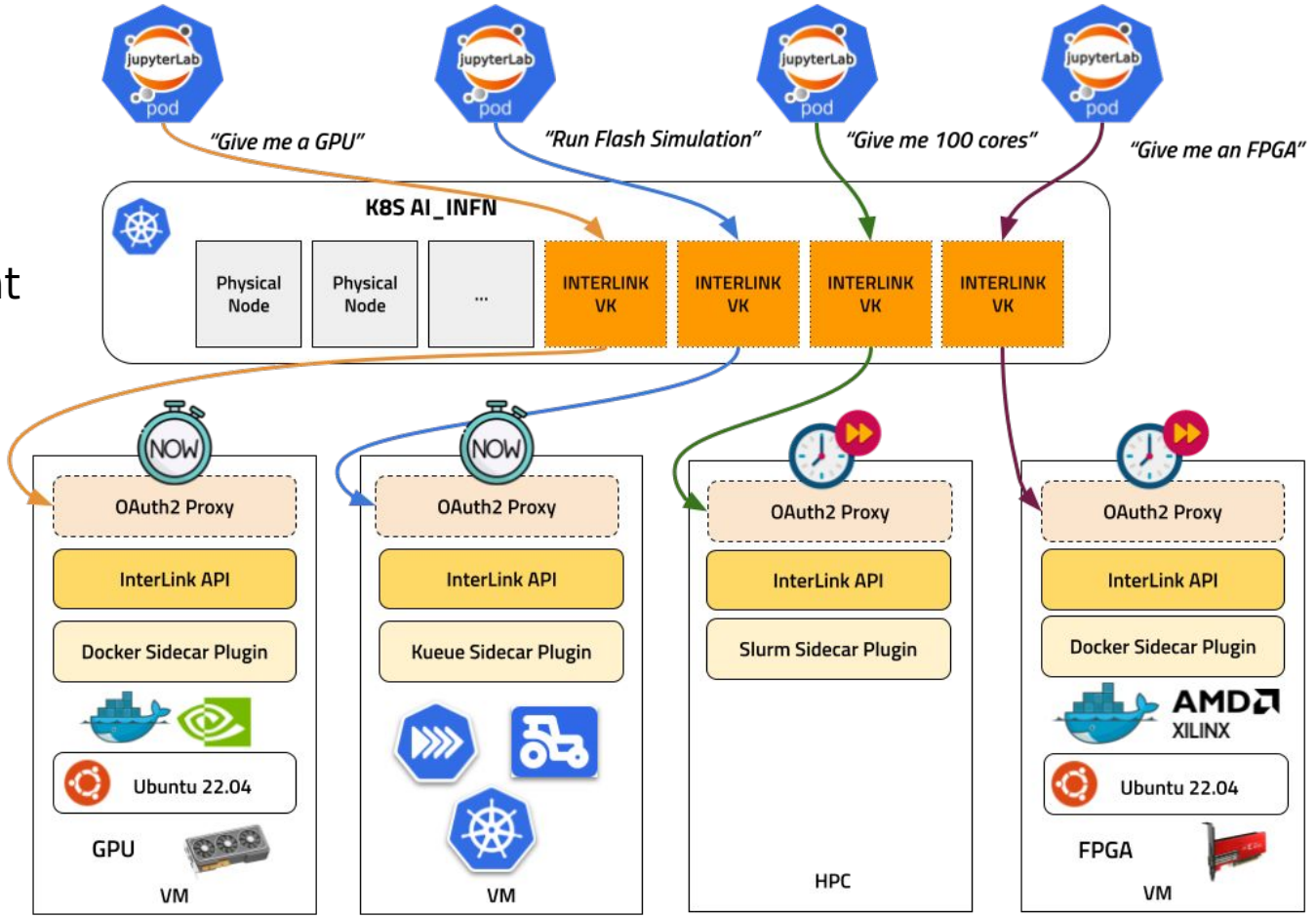
INFN

- The **AI_INFN platform provides a complex cloud-native use case** to test InterLink with:
  - Interactive access via offloading
  - Heterogeneous computing (CPU, GPU, FPGA...)

- We are proceeding in parallel with the development of two plugins (**docker and kueue**) to demonstrate decoupling from the backend.

- **The docker plugin** has already been validated for GPU provisioning and **can similarly support FPGA provisioning**.

- In concrete actions, the AI_INFN platform could already leverage this offloading system for spawning JupyterLab instances with some limitations (NFS).



InterLink: development context and ICSC related activities

INFN

The technical solution (interLink) has been initially prototyped by INFN in the context of the interTwin EU Funded project and is now enhanced within the ICSC development/research programme.
In particular

- **It is part of the Spoke0** infrastructural toolkits. As such it is under consolidation, testing and improvement
- **It is part of the Spoke2 - WP5 work plan**
  - in this respect there is a ongoing integration effort to extend the High rate analysis platform over HTC/HPC computing resources
- **Also part of the Spoke3 integration plan**
  - idea is to benefit of the interLink capabilities to offer highly dynamic access to specialized HW (i.e. over Leonardo)
  - integrating offloading with data retrieval from the data-lake prototype

Many fruitful sinergies should lead toward a generic technical solution, versatile and extendible based on specific needs.

What's next

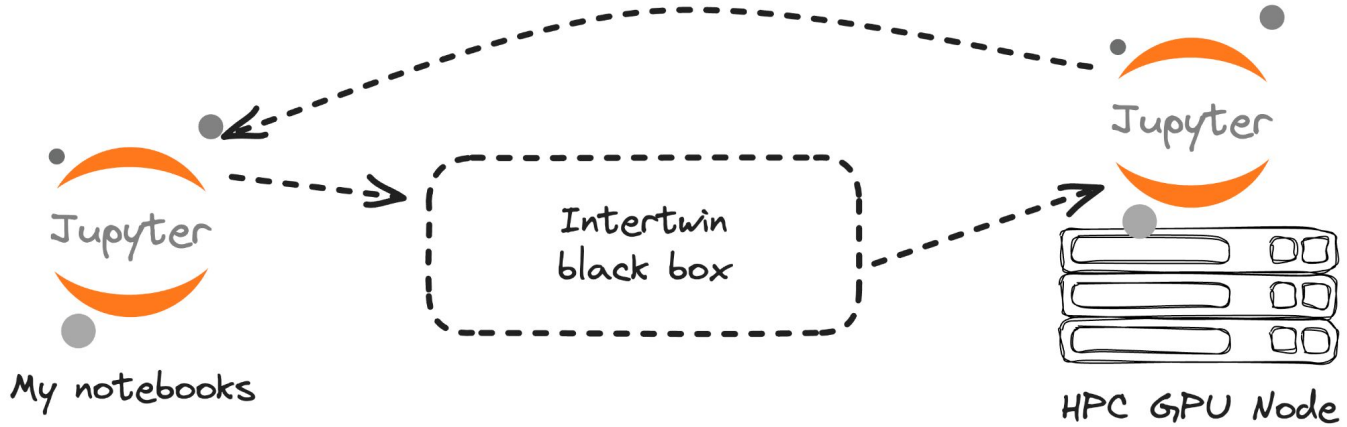
- WHEN RAC resources become available:
  - Prepare the setup
  - Tutorial/hands-on on how to access them
- BEFORE RAC resources become available:
  - Discuss and converge on the integration with current high-rate platform
  - Come up with a prototype
    - Test on resources:
      - HTC for high-rate platform scaling out
      - HPC for AI/ML to allow access to many GPUs

BACK UP

Use cases

INFN

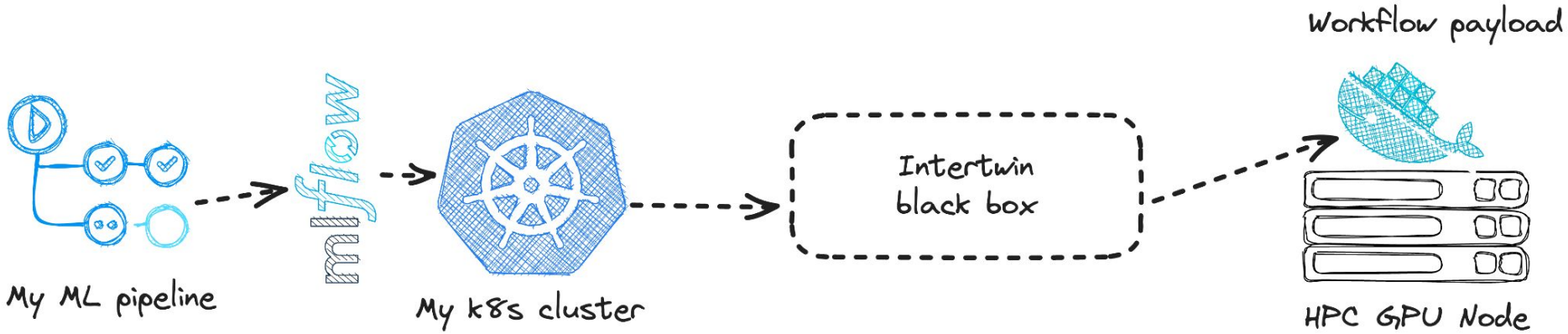
- Interactive session on HPC nodes:
  - I have my image to work interactively and I want through a JupyterHUB to select a node with GPU on HPC to run it



Use cases

INFN

- Offloading managed automatically by frameworks capable of speaking K8s:
  - I have my pipeline (mlflow, snakemake, kubeflow, argo workflows etc…) I want to declaratively indicate to the fwk that that step must be executed on a special and remote node



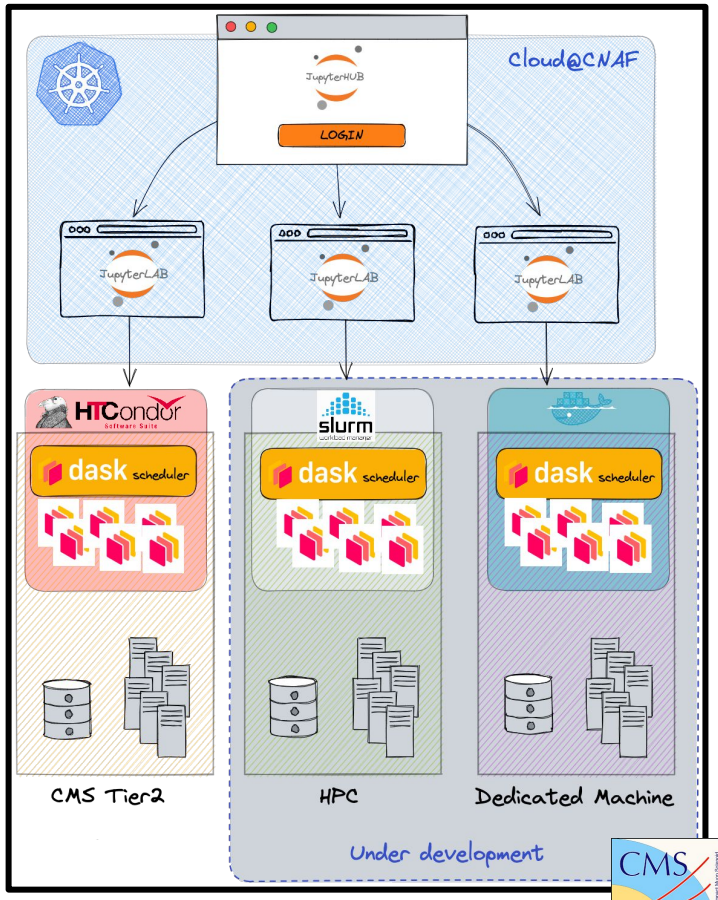
An inspiring use-case: INFN AF analysis offload

INFN

- **INFN Analysis Facility offload on Italian Tier2 sites**:
  - Deployment of Dask clusters on remote resources via RemoteHTCondor (Dask-jobqueue plugin)
  - "Pilot" wn jobs on Italian Tier2 production HTCondor queues via Interlink + HTCondor sidecar
  - Dedicated slot on all sites to contribute for a "seed" of resources available for AF user DASK cluster bootstraping
  - Scaling of the static quota based on active users
  - Additional workers will follow normal batch submission
  - Making this dynamically adapting based on the user "pressure"

What users see

What the offloading hides to the user



Flagship AI ...

INFN

- **Easy GPU access**:
  - seamless access to HPC centers
  - ML training triggered via workflow automation, e.g. ML pipelining tools (Kubeflow, MLflow, ...)
  - many GPUs at once == more/faster hyperparameter optimization