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

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

Evolving High Rate Analysis infrastructure with seamless offloading on different types of providers

Tommaso Tedeschi on behalf of WP2.5

Spoke2 Annual Meeting - 18-20 December 2023 - CINECA

The context

INFN

- Pushed by the enormous increase in HEP experiments computing resources requests from 2029/30 onwards (HL-LHC above all), a new analysis paradigm is arising:
  - **High-rate declarative interactive or quasi-interactive data analysis approach**
    - enabled by the usage of slimmed (flat) data formats
    - based cutting-edge analysis tools (ROOT's RDataFrame, Coffea, …) which scale up thanks to industry-standard data science backends (Dask)
- **The development of infrastructural solutions to implement such a new model is done inside WP2 and WP5:**
  - the infrastructure is developed using a use case-driven approach and tested with real-world analyses:
    - see Adelina's talk
  - **and synergically with Spoke0**:
    - adopting/proposing infrastructural solutions

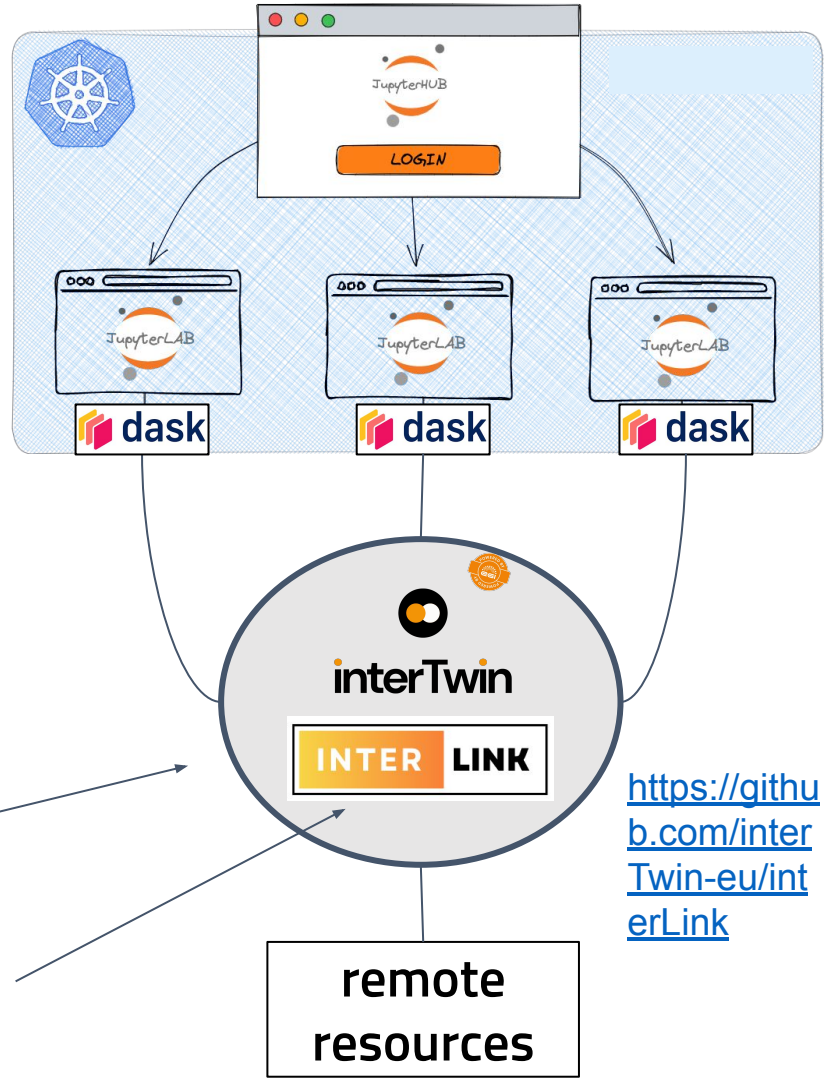
Motivations and goals

Current general-purpose infrastructure (see Gianluca's talk):

- 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

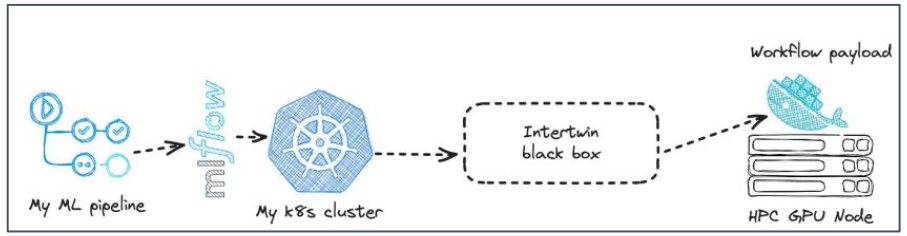


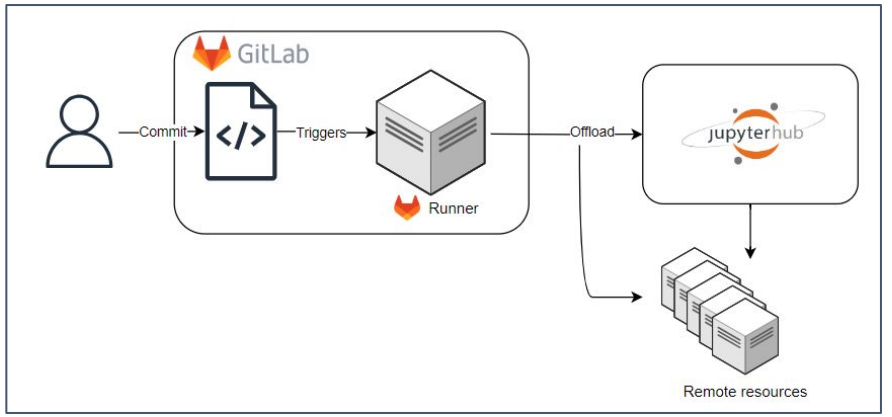
https://github.com/interTwin-eu/interLink

What use cases we want to enable

INFN

- **Unlock full power of cutting-edge analysis tools**
  - Speed-up of factor O(10-100) for HEP analysis workflows
- **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
- Enable **CI/CD as a trigger for analysis execution**
  - see Matteo's talk
- …





Roadmap

INFN

- This activity will follow WP2 scientific activities (and more that will come):
  - easing KPI achievements
    - see Francesco and Tommaso's talk
- The idea is to start from what has been done in WP5 and extend it:
  - interact with WP1 and WP6
- Getting real with prototypes and testbeds:
  - using ICSC resources to come
- Document everything with portability and reproducibility in mind:
  - everything in a single place
    - https://github.com/ICSC-Spoke2-repo/HighRateAnalysis-WP5

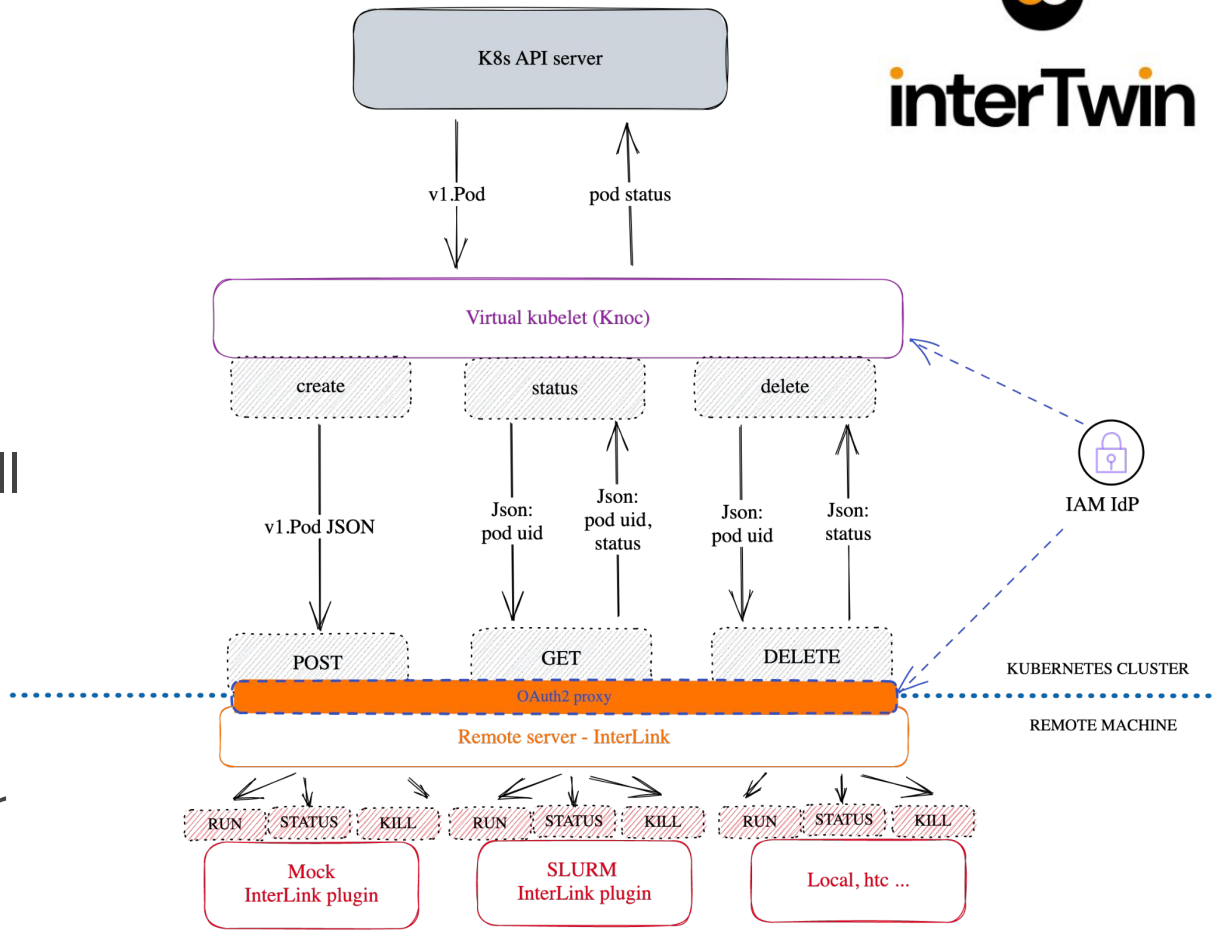
Backup

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...

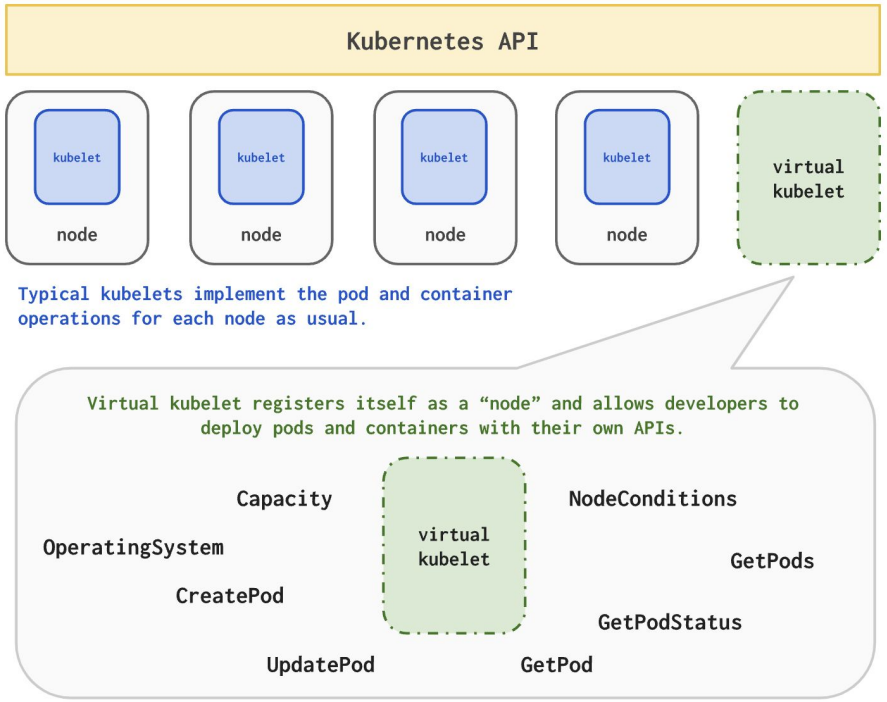


https://github.com/interTwin-eu/interLink

Components: VK

https://virtual-kubelet.io/

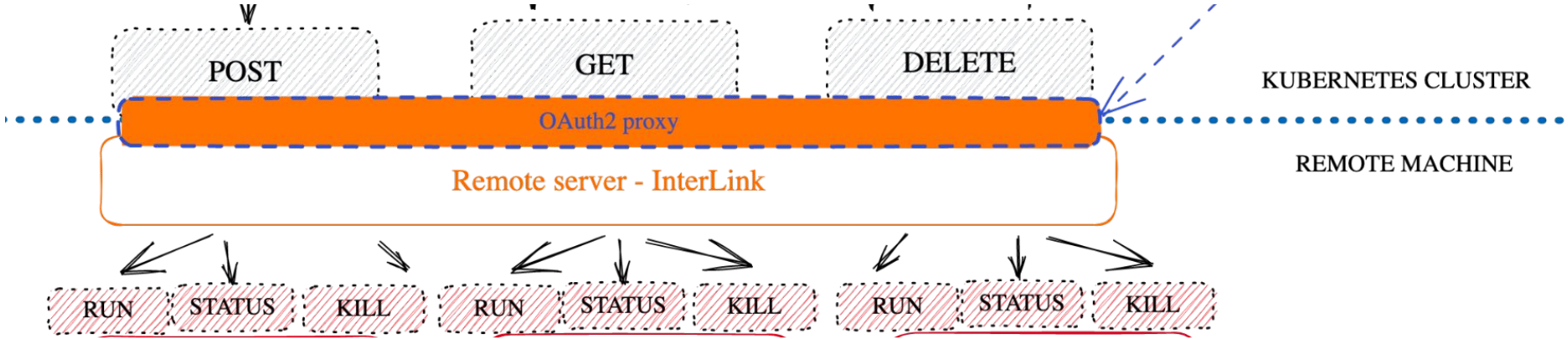
- **Virtual kubelet (VK)**:
  - "Open-source Kubernetes kubelet implementation that masquerades as a kubelet. This allows Kubernetes nodes to be backed by Virtual Kubelet providers"
- Can be imagined as a translation layer:
  - "I take your pod and run your container wherever I want"
- Registers virtual 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

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
- At the moment sidecar with local docker and slurm are implemented in go, HTcondor in python

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