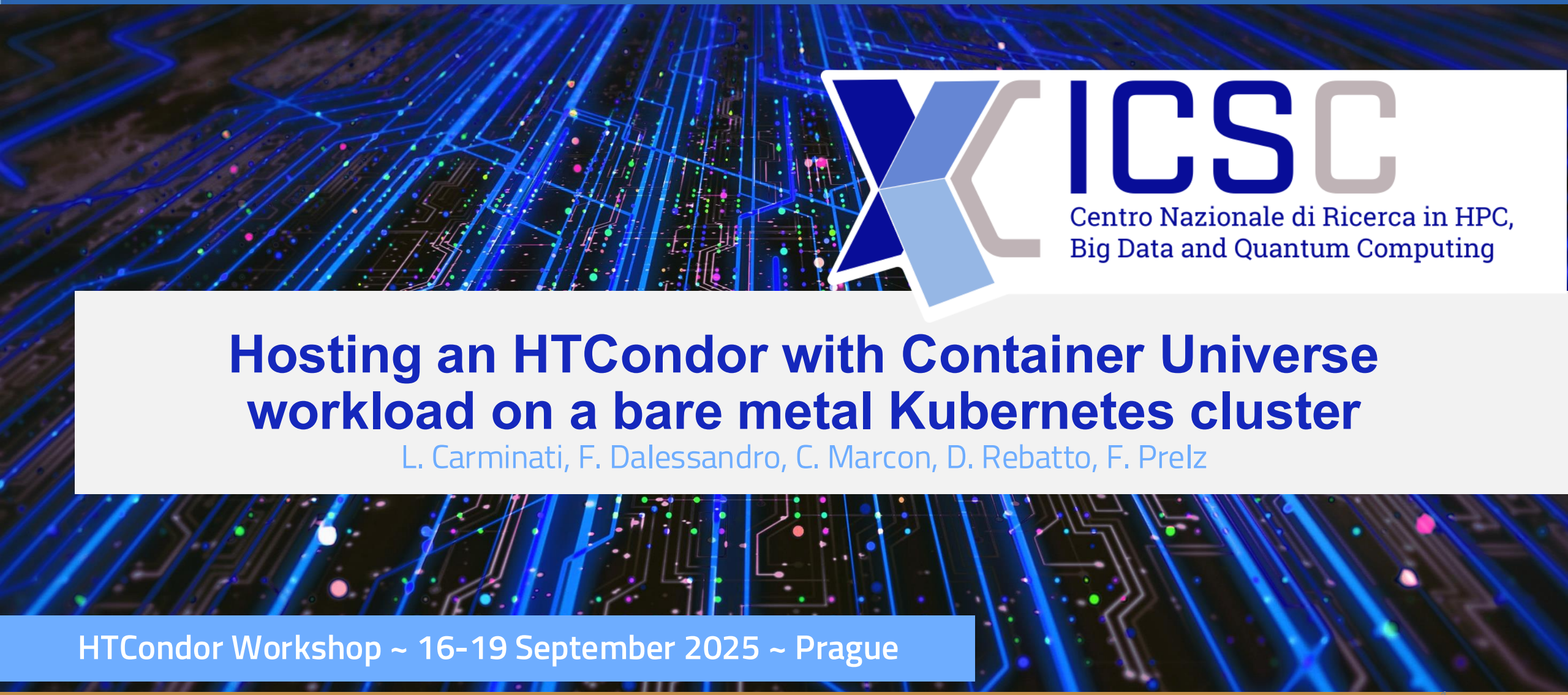
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing



Hosting an HTCondor with Container Universe workload on a bare metal Kubernetes cluster

L. Carminati, F. Dalessandro, C. Marcon, D. Rebatto, F. Prelz

HTCondor Workshop ~ 16-19 September 2025 ~ Prague

Introduction

- With its **Datacloud infrastructure**, INFN can rely on a scalable and nation-wide network of federated cloud sites;
- On this foundation, the PNRR-funded (National Recovery and Resilience Plan) **ICSC project** ("Italian Center for SuperComputing") has been established to **provide a federated cloud computing infrastructure** and **conduct R&D activities**;
- The **Milano computing center** has increased its capacity as **part of the ICSC efforts:**
  - 6 DELL chassis C6400 each equipped with 4 PowerEdge C6525 servers already available;
  - 4 Lenovo ThinkSystem SR665 V3;
  - Additional resources are planned: 6.25 PB storage capacity (CEPH + traditional).
- This contribution describes how to deploy an **HTCondor cluster on Kubernetes to run jobs in a Container Universe**, using Docker and Apptainer runtimes.

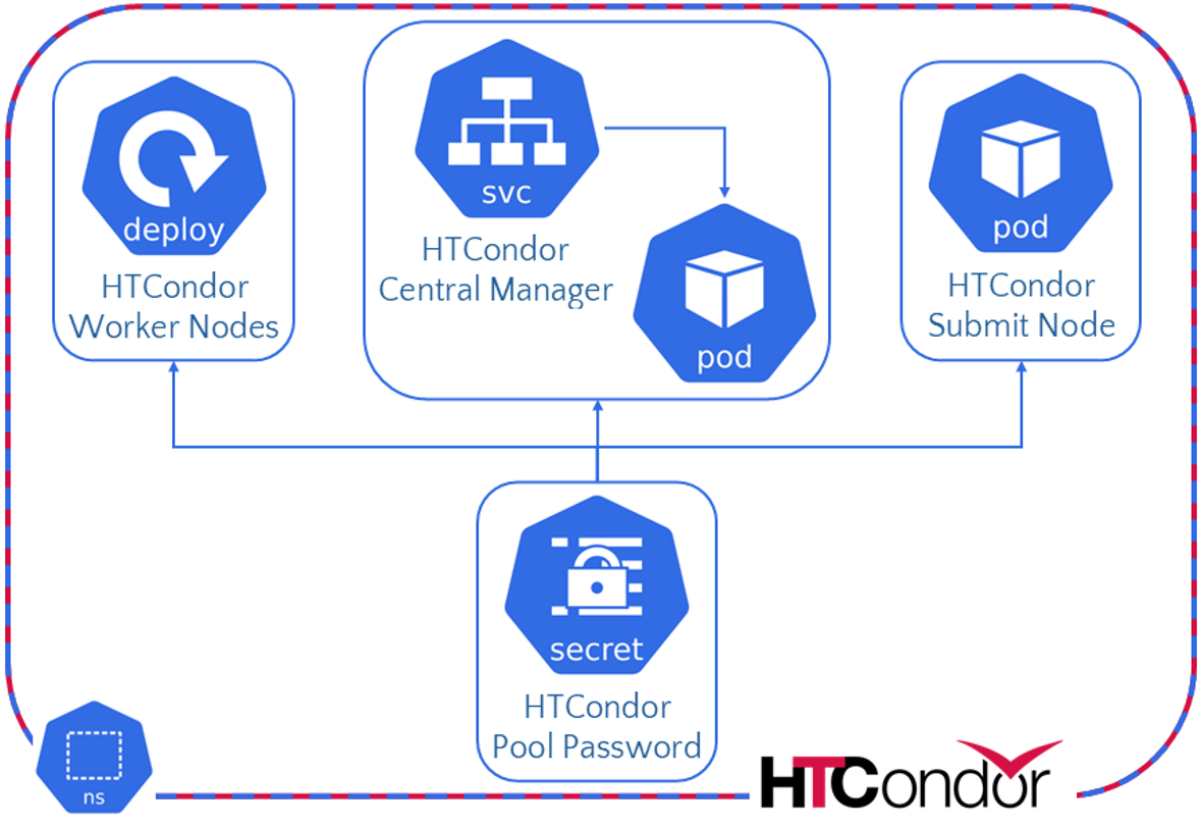
Kubernetes cluster on bare metal

INFN Milano opted for **a bare metal k8s cluster:**

- Rancher **RKE2** [1] **coupled to Foreman** [2] + **Puppet** [3] to install and maintain k8s on the nodes;
- **Dual stack IPv4/IPv6** network is configured;
- **3 control planes** managed by a load balancer and **24 worker nodes**;
- **40 TB of NFS** shares on bare metal available to the Kubernetes pods;
- Additional **CEPH storage** will be delivered;
- **Grid tools** will also provide access to distributed storage resources.

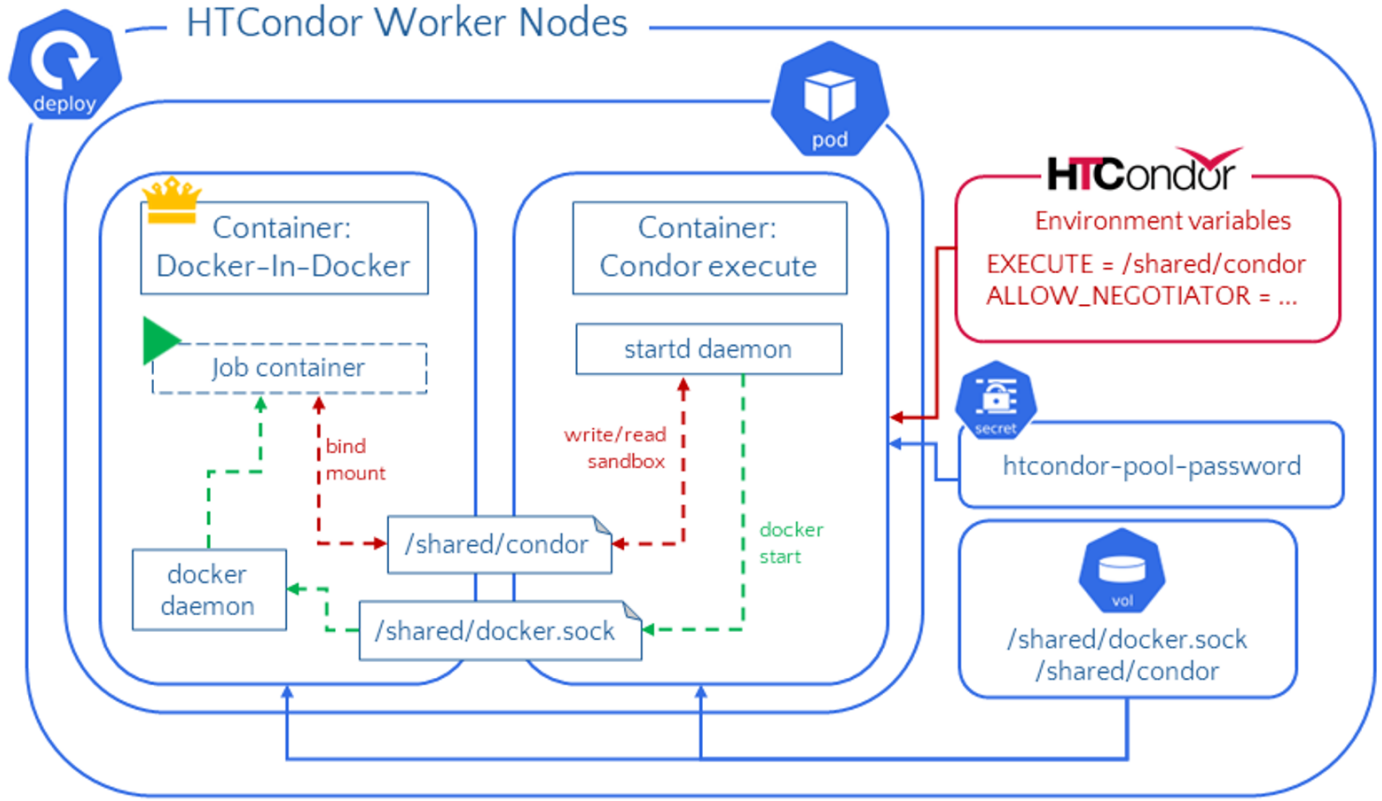
HTCondor on Kubernetes

- An **HTCondor cluster** can be deployed entirely **on k8s** by:
  - creating a **Service** to expose the **Central Manager;**
  - defining a **Deployment** acting as a scalable template for **Worker Nodes**;
  - (optional) starting a **Pod** with the **Submit Node** role.
- Alternatively, the **Deployment** can be attached to an **existing HTCondor Cluster;**
- The analysis software is **packaged in a Docker image** and does not require external dependencies or changes to the users' workflows.



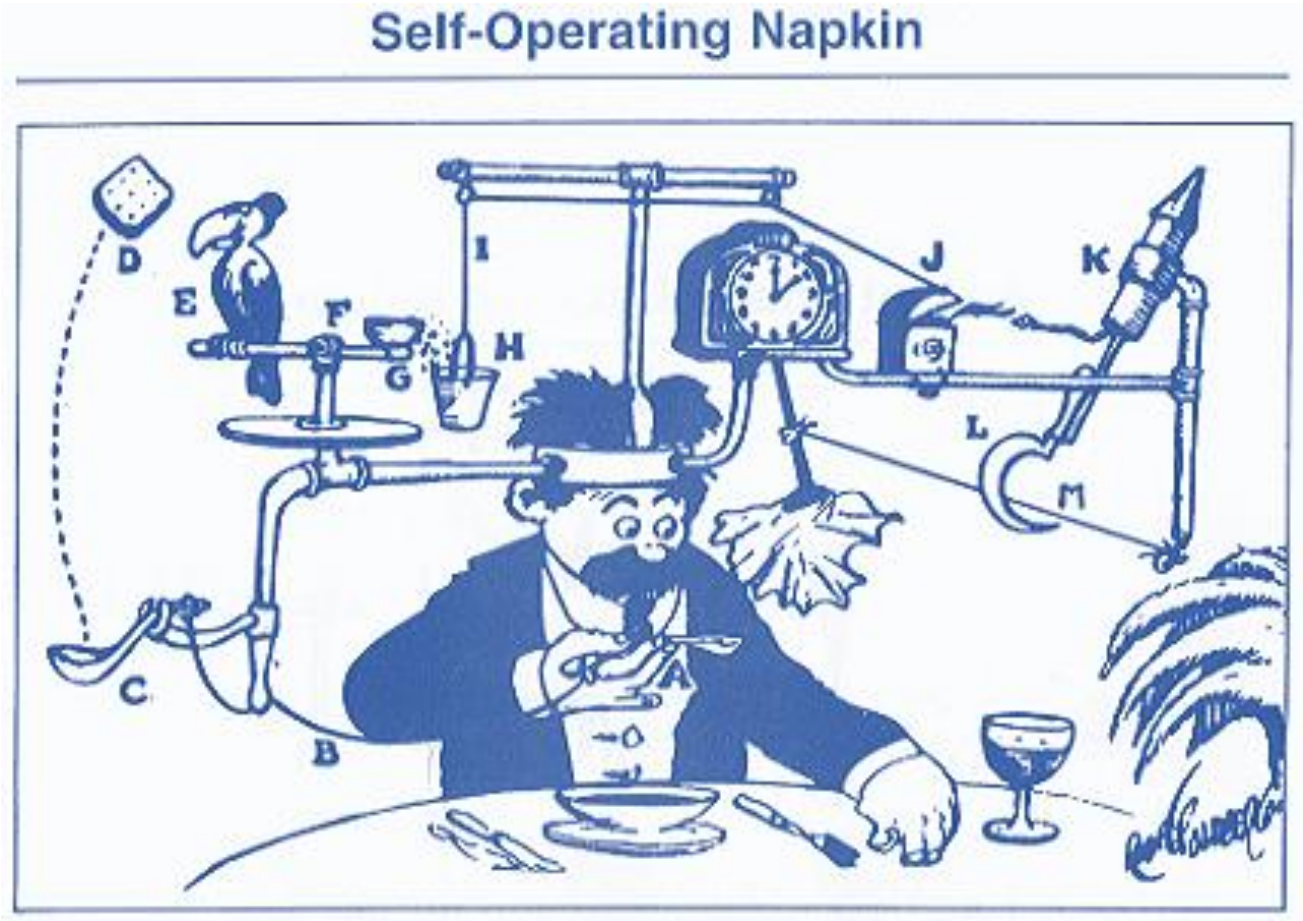
Container Universe with Docker runtime

- **HTCondor worker nodes → k8s deployments**;
- HTCondor jobs are configured to run in the **Container Universe**, in which the job container is spawned by the **HTC Execute node**;
- The HTC **Docker** runtime is provisioned via the **Docker-In-Docker** (DIND) container [4];
- The DIND container requires **privileged mode**;
- The Docker daemon is made available to the execute container over a **shared volume** (k8s emptyDir type);
- The same volume is used to share the HTC **job input and output files** between the **startd** and the **DIND containers.**



Container Universe with Docker runtime

- **HTCondor worker nodes → k8s deployments**;
- HTCondor jobs are configured to run in the **Container Universe**, in which the job container is spawned by the **HTC Execute node**;
- The HTC **Docker** runtime is provisioned via the **Docker-In-Docker** (DIND) container [4];
- The DIND container requires **privileged mode**;
- The Docker daemon is made available to the execute container over a **shared volume** (k8s emptyDir type);
- The same volume is used to share the HTC **job input and output files** between the **startd** and the **DIND containers.**



Container Universe with Apptainer

- **HTCondor worker nodes → k8s deployments**;
- HTCondor jobs are configured to run in the **Container Universe**, in which the job container is spawned by the **HTC Execute node**;
- **Privileged mode is required** by the HTCondor execute container;
- Unlike Docker, Apptainer does not require a running daemon:
  - Simpler k8s setup;
  - **No need for a** DIND equivalent **sidecar container**;
  - Apptainer installed on the execute container at startup.

A representative use case

- As a representative use case, an ATLAS workflow aimed at generating Higgs boson production events via gluon-gluon (ggH) fusion with NNLO accuracy has been tested with **four configurations**:
  - Apptainer jobs on baremetal;
  - Apptainer jobs inside Kubernetes;
  - Docker jobs on baremetal;
  - Docker jobs inside Kubernetes.
- The added virtualization layer **has not hindered the job's performance**:
  - Apptainer on Baremetal --> 15956 s+/- 0.8%;
  - Apptainer on k8s --> 15991 s +/- 0.4%;
  - Docker on Baremetal --> 15857 +/- 0.3%;
  - Docker on k8s --> 15897 s +/- 0.2%.

Conclusions & Outlook

- Analysis **workflow completed successfully** with HTCondor Container Universe on k8s;
- **Proof-of-concept can be scaled up** to more users and/or more complex workflows;
- **Up and down-scaling policies** of the HTC worker deployment will be defined based on performance studies and ad-hoc metrics;
- The added flexibility of running in a kubernetes environment, despite the additional virtualization layer, has not hindered the job's performance.

Bibliography

[1] https://baltig.infn.it/infn-cloud/wp1/rke2

[2] https://theforeman.org/

[3] https://www.puppet.com/

[4] https://indico.cern.ch/event/936993/contributions/4022096/attachments/2109009/3547280/k8s.pdf

Backup

HTCondor with Docker Universe

- An **HTCondor cluster** has been installed on k8s cluster;
- As a case study, an ATLAS workflow aimed at generating Higgs boson production events via gluon-gluon (ggH) fusion with NNLO accuracy has been considered;
- The analysis software is **packaged in a Docker image** and does not require external dependencies or changes to the users' workflows.

HTC role	K8s implementation		Docker image
Central Manager	1 service		---
	1 pod		htcondor/cm:24.4.0-el9
Submit	1 pod		htcondor/submit:24.4.0-el9
Execute	1 deployment	1 pod 2 containers	htcondor/execute:24.4.0-el9
			docker:dind

ICSC @ INFN Milano: resources

The **Milano computing center** has increased its capacity as **part of the ICSC efforts (budget 2 Meuro, ~2000 cores)**;

- 6 DELL chassis C6400 each equipped with 4 **PowerEdge C6525 servers**;
- 4 Lenovo ThinkSystem SR665 V3;

DELL PowerEdge C6525	
CPU	2x CPU AMD EPYC 7413 (24 cores)
RAM	16x 32 GB RAM DDR4 3200 MT/s
SATA storage	2x 890 GB SSD (4 RAID1 pairs)
SAS storage	2x 3.5 TB SSD (4 RAID1 pairs)
Networking	1x Broadcom Adv. Dual 10Gb Ethernet

Lenovo ThinkSystem SR665 V3	
CPU	2x CPU AMD EPYC 9654 (96 cores)
RAM	1.6 TB DDR5
NVMe storage	2x 960 GB SSD
	3x 7.68 TB SSD

- Additional resources are planned:
  - 6.25 PB storage capacity (CEPH + traditional).