Migration of the AI_INFN hardware assets to **tier1 zone**

L. Anderlini, S. Dal Pra, D. Michelotto

AI_INFN
Artificial Intelligence technologies
for INFN research

March, 4th 2024

Quick recap on the motivation

The migration to the technopole of the CNAF hardware assets is upcoming.

During the migration the "*Software Development and Distributed System*" (sdds) cloud zone will be merged in the "Tier1" zone.

Merging the two zones represents a **logical reorganization** of the resources that must be accomplished **before** moving physically to the new site.
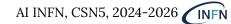
AI_INFN resources are currently in sdds zone and must be *moved* to tier1.

Live migration of the VM is possible **if:**

- they **do not have accelerators** (including GPUs),
- **hardware resources are sufficient** to temporarily double the request.

Plan: **Monday, March 4th**

- Warn users:
  - via ml-infn-csn5 AND ai-infn-csn5 mailing lists
  - adding a large red warning in the login page of the .212 deployment
  - **All deployments from the PaaS must be cleared by the users**
- Stop new deployments in the AI_INFN tenancy in M1, M2 and M3
- Live migrate all CPU-only deployments from M3 → M2

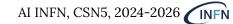# Plan: **Tuesday, March 5th**

Reserved for:

- managing exception in the live migrations M3 → M2
- leaving users the opportunity for a last-minute backup

Plan: **Wednesday, March 6th**

- **All Virtual Machines on M1 and M3 are deleted (including their filesystem, with all the contained data),** this includes:
  - All VMs with T4 GPUs
  - All VMs with RTX5000 GPUs
  - Most VMs with A100 GPUs (excluding the production VM .212)

- M1 and M3 servers are switched off
- One nVME drive is moved from M1 → M3
- M1 and M3 are brought back online in tier1 zone
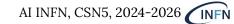- Verify flavours

Plan: **Thursday, March 7ᵗʰ**

- New deployments are re-enabled on M1 and M3
- The new production platform (hub.ai.cloud.infn.it) is deployed
- Group structure is imported from .212 production setup
- Users are encouraged to **start migrating from .212 → hub.ai.cloud.infn.it**
- M2 remains online to phase-out the .212 production deployment, together with the A30 deployment and FPGA VMs
- One or two A100 GPUs will be unusable for insufficient CPU (as we are using CPU to handle live migrations from M2)

Plan: **Monday, March 11th**

- Report on the migration at the AI_INFN meeting
- If no objection is raised during the meeting, **the .212 deployment will be removed from network** (remove FIP assignment)

Plan: **Tuesday, March 12ᵗʰ**

- **All VMs on M2 (.212, A30 and FPGAs) are deleted (including their filesystem, with all the contained data)**
- M2 is switched-off and migrated to tier1 zone
- **S*ome* CPU-only VMs is live-migrated from M3 → M2 to free CPU for the A100 deployments**
- **Remaining A100 GPUs on M3 are joined to the hub service**

The migration is over.