



L'ESPERIENZA HPC CLOUD DI CINECA

Implementazione di un
cloud privato per workflow
di AI e di calcolo come
servizio

Autori: Debora Testi, Giuseppa Muscianisi, Mirko Cestari

Sommario

GLOSSARIO.....	3
LISTA DELLE FIGURE	3
LISTA DELLE TABELLE.....	3
INTRODUZIONE	4
HPC AND CLOUD COMPUTING – CONCETTI PRINCIPALI.....	5
HPC VS CLOUD COMPUTING.....	5
MODELLI DI SERVIZI CLOUD	6
CINECA HPC CLOUD.....	7
MODELLO DI SERVIZIO DEL CINECA HPC CLOUD	7
CASI D’USO DELL’HPC CLOUD.....	8
AUTORIZZAZIONE PROGETTI E ALLOCAZIONE RISORSE	10
STRUMENTI DI GESTIONE DELLE RISORSE CLOUD	10
INFRASTRUTTURA CINECA HPC CLOUD	12
ADA	12
UTILIZZO DELLE RISORSE CINECA HPC CLOUD	15
GAIA	18
CONCLUSIONI	21

Glossario

Acronimo	Definizione
AI	Intelligenza Artificiale
API	Interfaccia di programmazione dell'applicazione
CLI	Interfaccia a linea di comando
HPC	Calcolo ad Alte Prestazioni
HW	Hardware
IaC	Infrastructure as Code / Infrastruttura come Codice
SW	Software
VM	Macchina Virtuale

Lista delle figure

Figura 1 – Modelli di servizio del cloud computing	6
Figura 2 – Modello di servizio e responsabilità nel modello di servizio HPC Cloud di Cineca.....	8
Figura 3 – OpenStack Horizon Dashboard.....	11
Figura 4 – Evoluzione storica dell'infrastruttura Cineca HPC Cloud	12
Figura 5 – Architettura del sistema ADA cloud	13
Figura 6 – Progetti HPC cloud sul sistema ADA nel tempo	15
Figura 7 – Utenti HPC cloud sul sistema ADA nel tempo	15
Figura 8 – vCPU allocate sul sistema ADA nel tempo	16
Figura 9 – Storage allocato sul sistema ADA nel tempo	17
Figura 10 – Categorie di progetto per i progetti HPC Cloud nel 2024.....	18
Figura 11 – Infrastruttura del sistema GAIA cloud	19

Lista delle tabelle

Tabella 1 – Caratteristiche di HPC e Cloud.....	6
Tabella 2 – Servizi OpenStack disponibili su ADA.....	14

Introduzione

Negli ultimi anni l'evoluzione delle infrastrutture digitali ha portato a una convergenza sempre più marcata tra High Performance Computing (HPC) e Cloud Computing. Questa sinergia consente alle comunità scientifiche e di ricerca, così come alle industrie innovative, di affrontare i propri fabbisogni computazionali con maggiore flessibilità, scalabilità e controllo.

L'esperienza di Cineca nella progettazione e gestione di infrastrutture HPC Cloud dimostra come le tecnologie cloud, quando strettamente integrate con sistemi ad alte prestazioni, possano supportare in modo efficace workflow computazionali avanzati. Le scelte architetture e operative adottate per il Cineca HPC Cloud illustrano un modello replicabile da organizzazioni che intendono implementare un proprio cloud privato per la ricerca, per carichi di lavoro data-intensive o per l'elaborazione di dati sensibili.

Questo white paper presenta l'infrastruttura HPC Cloud realizzata al Cineca, illustrandone i concetti chiave, il modello di servizio, i casi d'uso e le soluzioni tecnologiche adottate per garantire alte prestazioni, sicurezza e adattabilità. Attraverso una panoramica dei sistemi più recenti, ADA e GAIA, il documento offre una visione complessiva delle opportunità dell'HPC Cloud per la ricerca, l'innovazione e le strategie di cloud privato in ambito industriale.

HPC and Cloud computing – concetti principali

High Performance Computing (HPC) è una tecnologia che utilizza cluster di processori potenti operanti in parallelo per elaborare enormi insiemi di dati multidimensionali e risolvere problemi complessi a velocità elevata¹. Il **cloud computing** è invece un paradigma che abilita l'accesso di rete a un insieme condiviso di risorse fisiche o virtuali, scalabili ed elastiche, con provisioning self-service e amministrazione on-demand².

HPC vs Cloud computing

Le infrastrutture HPC e Cloud presentano caratteristiche differenti (Tabella 1) e possono essere utilizzate separatamente oppure in combinazione, in base alle esigenze specifiche del workflow degli utenti.

Caratteristica	HPC	Cloud Computing
Prestazioni	Punta alle massime prestazioni possibili.	Dipende dal workload; in generale la virtualizzazione introduce un impatto contenuto.
Accesso utente	Autorizzazione gestita dallo staff HPC.	Una volta assegnate le risorse, sono gestite direttamente dall'utente.
Sistema Operativo	Scelto dallo staff HPC in base ai vincoli hardware. Gli aggiornamenti di sicurezza sono gestiti da staff HPC.	Selezionato dall'utente. Patch e aggiornamenti di sicurezza sono gestiti dall'utente.
Software Stack	Per lo più installato dallo staff HPC. Gli utenti possono installare software proprio senza privilegi di root. L'ambiente è fornito "as is".	L'utente è root sulle VM e può installare l'intero stack software necessario, modificando l'ambiente in base alle proprie esigenze.
Snapshots dell'ambiente	Non disponibile.	Gli utenti possono salvare snapshot delle VM quando necessario.
Esecuzione simulazioni	È fornito uno scheduler (es. SLURM) e occorre attendere che si liberino le risorse.	È possibile eseguire job seriali o paralleli quando e come necessario.

¹ <https://www.ibm.com/think/topics/hpc>

² "ISO/IEC 22123-1:2023(E) - Information technology - Cloud computing - Part 1: Vocabulary". International Organization for Standardization. 2023.

Simulazioni molto grandi

È fornito uno scheduler (es. SLURM) e occorre attendere che si liberino le risorse.

L'esecuzione dei workloads può essere pianificata quando e come necessario.

Tabella 1 – Caratteristiche di HPC e Cloud

Modelli di servizi cloud

Indipendentemente dalle caratteristiche hardware, le risorse cloud possono essere erogate secondo diversi modelli di servizio (Figura 1):

- **Infrastructure as a Service (IaaS)** fornisce risorse computazionali virtualizzate via rete; gli utenti possono configurarle secondo le proprie esigenze.
- **Platform as a Service (PaaS)** offre una piattaforma per sviluppare, distribuire e gestire applicazioni senza doversi occupare dell'infrastruttura sottostante.
- **Software as a Service (SaaS)** rende disponibili applicazioni software via Internet, accessibili tramite browser, senza necessità di installazione o manutenzione locale.

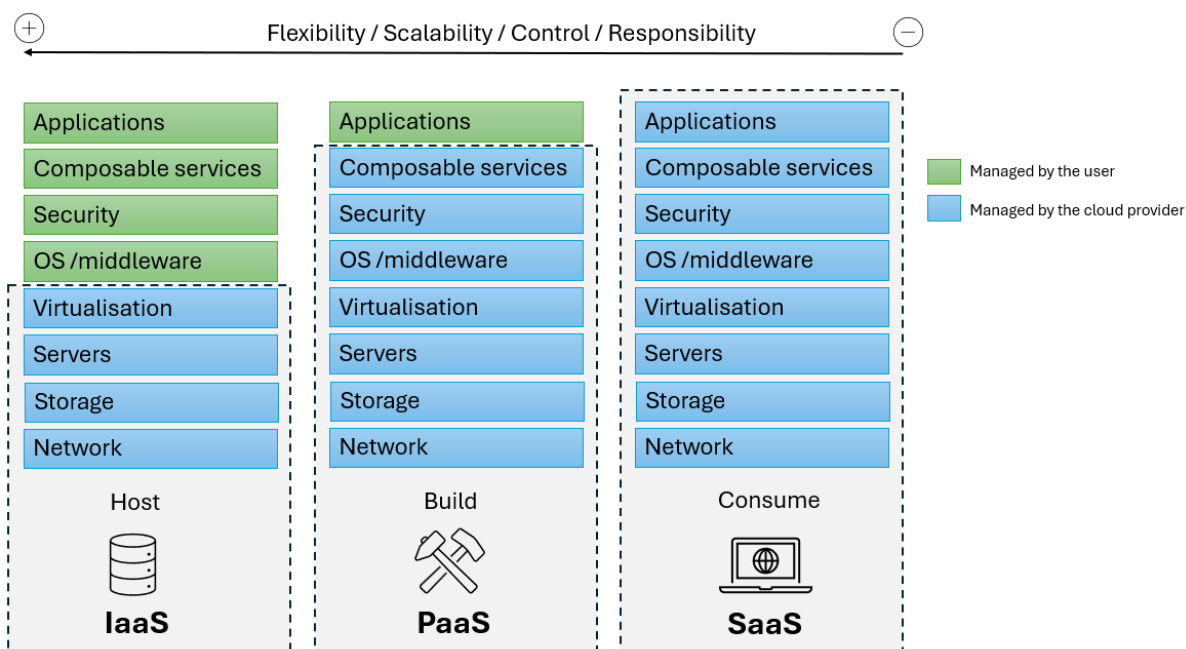


Figura 1 – Modelli di servizio del cloud computing

Cineca HPC Cloud

In Cineca, l'infrastruttura HPC Cloud complementa l'ecosistema HPC, offrendo una piattaforma strettamente integrata che combina calcolo ad alte prestazioni e calcolo flessibile.

Gli utenti possono sfruttare la potenza computazionale delle macchine HPC con il vantaggio aggiuntivo della flessibilità tipica del cloud. È possibile configurare il proprio pool di risorse in modo aderente al workflow (flessibilità), creando singole VM molto grandi o insiemi di VM più piccole in grado di comunicare tra loro.

Allo stesso tempo, le risorse possono essere scalate in funzione delle esigenze, permettendo di gestire in modo efficiente carichi di lavoro variabili (scalabilità).

Modello di servizio del Cineca HPC Cloud

L'infrastruttura HPC Cloud di Cineca è erogata secondo il modello Infrastructure as a Service (IaaS).

Nel modello IaaS, il provider cloud (Cineca) amministra gli strati hardware e di virtualizzazione dell'infrastruttura e fornisce risorse di calcolo (vCPU, storage, rete, GPU) e API di alto livello (dashboard web e strumenti CLI) che gli utenti possono utilizzare per controllare le risorse loro assegnate.

Gli utenti mantengono la piena responsabilità di ciò che configurano e installano nei propri progetti, inclusi rete, sistemi operativi e applicazioni in esecuzione sulle macchine virtuali, nonché il controllo degli accessi ai servizi e alle VM³.

³ Una chiara separazione di ruoli nell'uso del servizio cloud è definita e descritta nella sezione [CINECA Cloud model](#) della HPC Cloud User Guide.

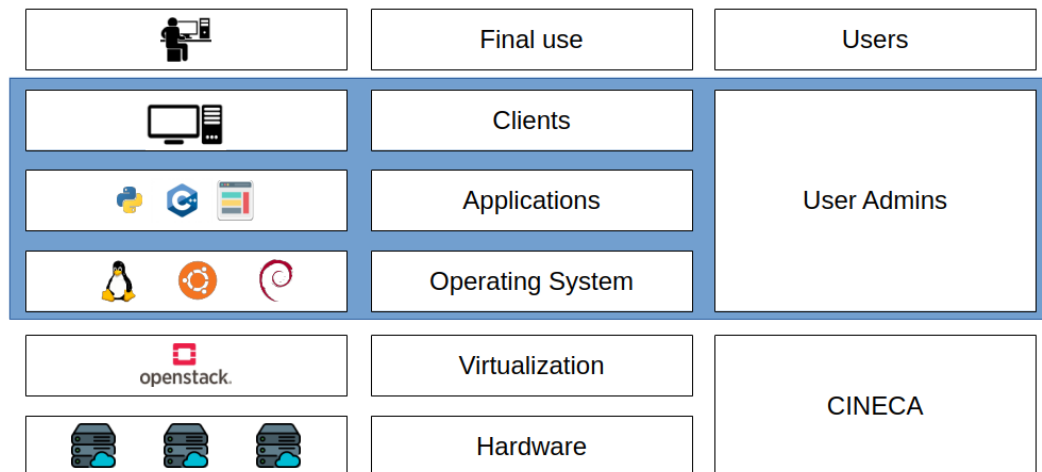


Figura 2 – Modello di servizio e responsabilità nel modello di servizio HPC Cloud di Cineca

Casi d'uso dell'HPC Cloud

Gli utenti si affidano all'infrastruttura HPC Cloud di Cineca per affrontare diverse esigenze. L'elenco seguente non è esaustivo, ma intende fornire esempi di scenari in cui HPC Cloud può risultare particolarmente utile.

- Hosting di servizi generali di elaborazione e analisi dati:
 - L'infrastruttura può supportare scenari che richiedono una combinazione di prestazione e flessibilità. Gli utenti possono installare e utilizzare qualsiasi software a loro necessario nel rispetto delle "Acceptable Use Policies" del Cineca.
 - Il supporto cloud-native all'automazione (Infrastructure as Code) consente di descrivere e replicare ambienti e configurazioni.
 - Un importante requisito per i servizi IaaS è il poter disporre di backup delle istanze; questi possono automatici o attivabili su richiesta dell'utente.
- Hosting di HPC mini-cluster ad alte prestazioni che include le seguenti caratteristiche:
 - File system condiviso tra nodi di login e nodi di calcolo del cluster.
 - Provisioning di risorse GPU.
 - Stack middleware completo con librerie numeriche e di comunicazione.
 - Rete ad alte prestazioni, incluse funzionalità RDMA native.
 - Possibilità di posizionamento delle VM su host specifici per casi d'uso selezionati. (Nota, il posizionamento di VM su host specifici richiede normalmente i privilegi di amministratore).

- Hosting di servizi di gestione dati che ricevono o espongono dati via web:
 - Portali di virtual lab e piattaforme per la presentazione ed esplorazione interattiva di dataset, ad esempio mappe di previsioni meteo.
 - Servizi accessibili sia da Internet sia da intranet, in grado di ricevere e trasmettere dati.
- Hosting di workload per l'elaborazione di dati sensibili:
 - I sistemi HPC tradizionali sono risorse condivise e spesso non soddisfano tutti i requisiti necessari – crittografia, audit logging, accessi controllati, connettività di rete limitata – per ospitare legalmente o eticamente dati sensibili.
 - Le infrastrutture cloud possono invece fornire isolamento delle risorse, crittografia a riposo, controllo degli accessi e altre caratteristiche di sicurezza che le rendono adatte a questi scenari.
 - Per supportare tali casi d'uso, il team Cineca ha implementato soluzioni tecnologiche e procedurali che consentono il trattamento di dati sensibili nel rispetto del quadro normativo.
 - L'infrastruttura Cineca HPC Cloud è certificata ISO 27001 dal 2022 per i servizi HPC cloud per la ricerca in ambito life science e dal 2025 per l'erogazione di servizi IaaS per ricerca e innovazione su HPC Cloud⁴.
- Collegamento con l'infrastruttura HPC:
 - Le istanze in cloud possono agire da frontend per workload eseguiti sugli altri sistemi HPC del Cineca.
 - Tra le interfacce possibili rientrano Jupyter, Dask e Slurm.
 - Questo può includere anche l'esposizione di dati verso/da infrastrutture HPC interne al Cineca.
- Implementazione flessibile e automatizzata, tramite Kubernetes (al di sopra del livello di virtualizzazione dell'infrastruttura), di flussi di lavoro containerizzati.
 - L'infrastruttura supporta il deployment flessibile e automatizzato di workflow containerizzati attraverso Kubernetes.
 - Gli utenti possono distribuire cluster Kubernetes sia con l'orchestratore fornito dal sistema sia con modalità ad hoc personalizzate, usando strumenti come Rancher, RKE2, OKD e Juju Charms.
- Implementazione di una infrastruttura collaborativa nel progetto dell'utente (Infrastructure as Code, IaC)

⁴ L'infrastruttura Cineca HPC Cloud è certificata ISO 27001 dal 2022 per “[Servizi informatici HPC in cloud per la ricerca in ambito life science](#)” e dal 2025 per “Erogazione di servizi IaaS per ricerca e innovazione su HPC Cloud”.

- L'utente può utilizzare Terraform e Ansible per il provisioning e la gestione delle risorse cloud.

Autorizzazione progetti e allocazione risorse

Un progetto cloud è un insieme di risorse assegnate a utenti autorizzati dopo la registrazione e l'accettazione dei termini e condizioni del Cineca tramite il portale di provisioning delle risorse⁵.

In base al modello di servizio del Cineca HPC Cloud, gli utenti gestiscono in autonomia i progetti cloud (uno o più di uno) ai quali sono associati.

Ogni progetto cloud è composto da un pool di risorse virtuali (quota o budget di progetto) definito in termini di:

- Numero di vCPU: quantità totale disponibile per tutte le VM del progetto.
- GB di RAM: assegnati in funzione del numero totale di vCPU.
- GB di storage: risorsa di storage indipendente collegabile alle VM e persistente rispetto al ciclo di vita della singola VM.
- Numero di indirizzi IP pubblici: quantità totale di floating IP pubblici.

Su richiesta possono inoltre essere assegnate altre risorse, quali:

- Numero di GPU disponibili per le VM del progetto.
- Storage aggiuntivo condivisibile tra le VM del progetto.

I ricercatori di enti pubblici possono richiedere risorse HPC Cloud attraverso il programma ISCRA (Italian SuperComputing Resource Allocation)⁶.

Strumenti di gestione delle risorse cloud

Per gestire le risorse assegnate a un progetto, gli utenti possono interagire con OpenStack tramite strumenti messi a disposizione dal Cineca, come Horizon Web Dashboard e la Command Line Interface (CLI).

Horizon Dashboard è l'interfaccia principale per interagire con OpenStack e consente di monitorare l'uso delle risorse, creare nuove risorse e gestire quelle esistenti in autonomia.

⁵ <http://userdb.hpc.cineca.it>

⁶ <https://www.hpc.cineca.it/hpc-access/access-cineca-resources/iscra-projects/>

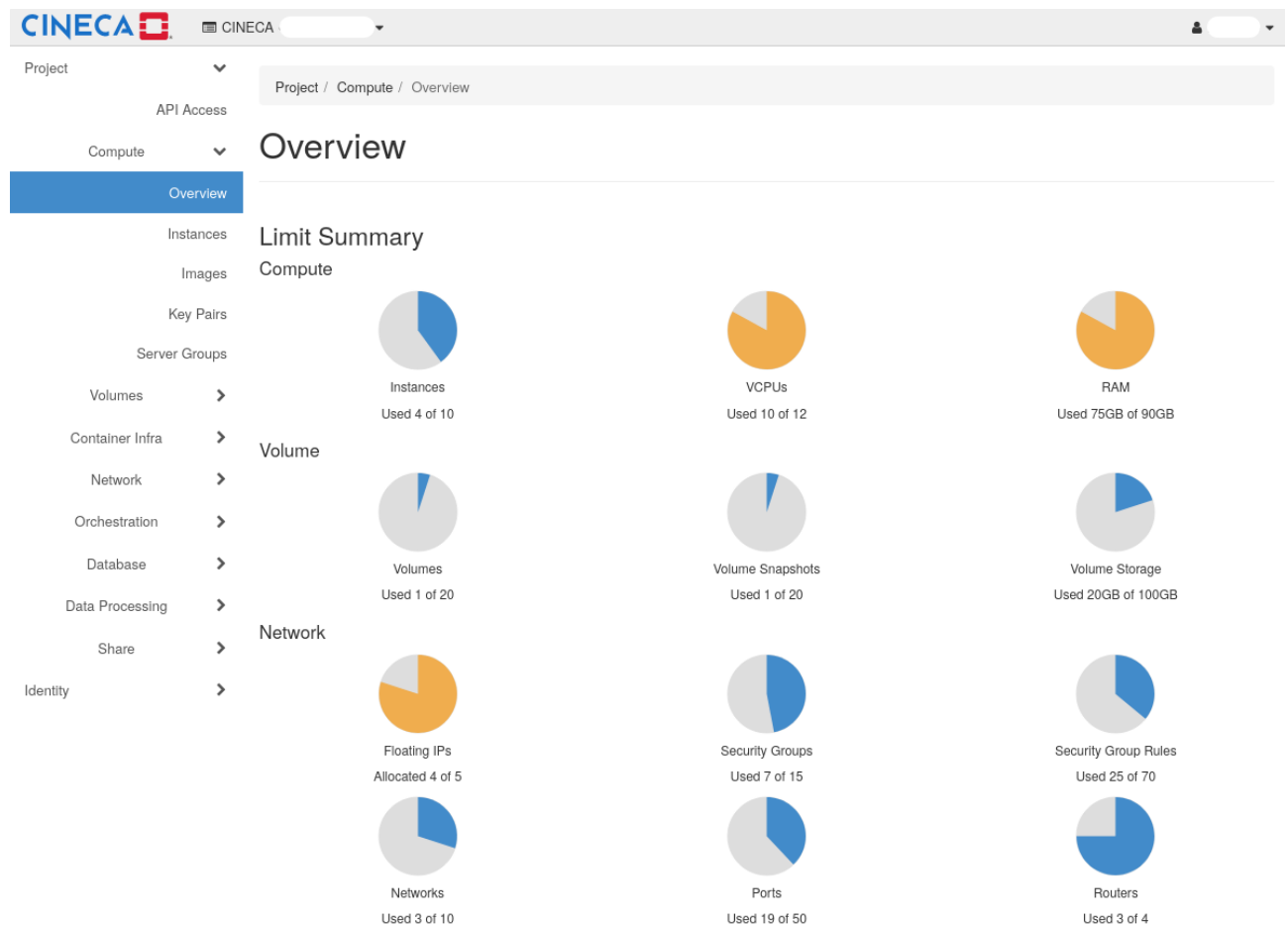


Figura 3 – OpenStack Horizon Dashboard

La **Command Line Interface** (CLI) di OpenStack consente di eseguire da terminale le stesse operazioni disponibili via dashboard – ad esempio creare istanze, volumi e reti – ed è uno strumento particolarmente utile per l'automazione e la gestione efficiente delle operazioni ricorrenti.

Infrastruttura Cineca HPC Cloud

L'esperienza di Cineca in ambito cloud computing HPC è iniziata nel 2015 (Figura 4), quando alcune comunità scientifiche hanno espresso l'esigenza di disporre di un ambiente più flessibile e isolato per configurare ed eseguire i propri workflow.

Il tradizionale cluster HPC multiutente non era il contesto ideale per questi scenari, e così è stata realizzata la prima infrastruttura HPC Cloud. Dopo questa prima esperienza positiva, sono stati introdotti altri sistemi, con incremento della potenza degli hypervisor e con storage CEPH dedicato.

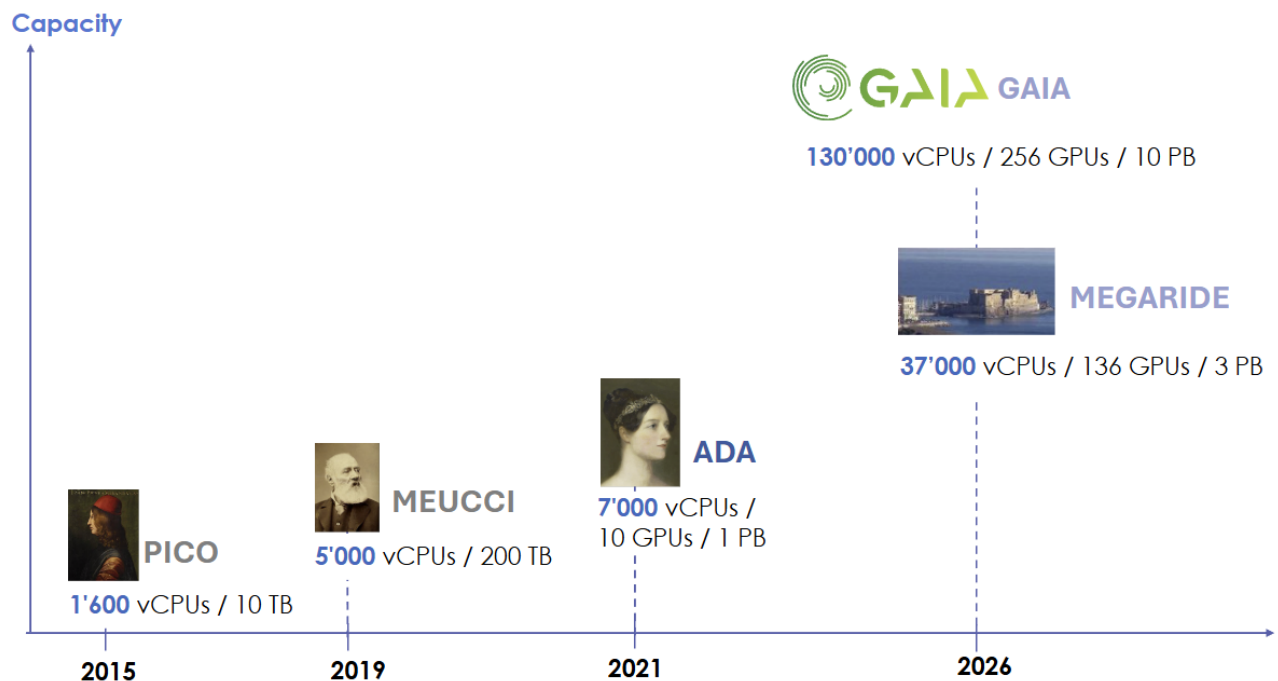


Figura 4 – Evoluzione storica dell'infrastruttura Cineca HPC Cloud

ADA

Nel 2021 è stato installato il sistema di supercalcolo Tier-1 GALILEO100 che fa parte dell'infrastruttura Fenix finanziata dal progetto ICEI Human Brain⁷. L'hardware è stato selezionato per massimizzare le prestazioni e per supportare workflow complessi che combinano ambienti utente flessibili e simulazioni HPC tradizionali. Si tratta di un

⁷ <https://fenix-ri.eu/>

sistema complesso che può essere utilizzato sia per simulazioni di calcolo HPC standard, sia per attività di cloud computing e di calcolo interattivo.

La componente HPC Cloud, denominata ADA, è isolata rispetto alla parte scalabile del sistema; tuttavia, per alcuni casi d'uso selezionati, è possibile predisporre interazioni con la parte HPC. Inoltre, il file system Lustre ("Capacity Tier" in Figura 5) può essere condiviso con le macchine virtuali ospitate come livello aggiuntivo di hot storage.

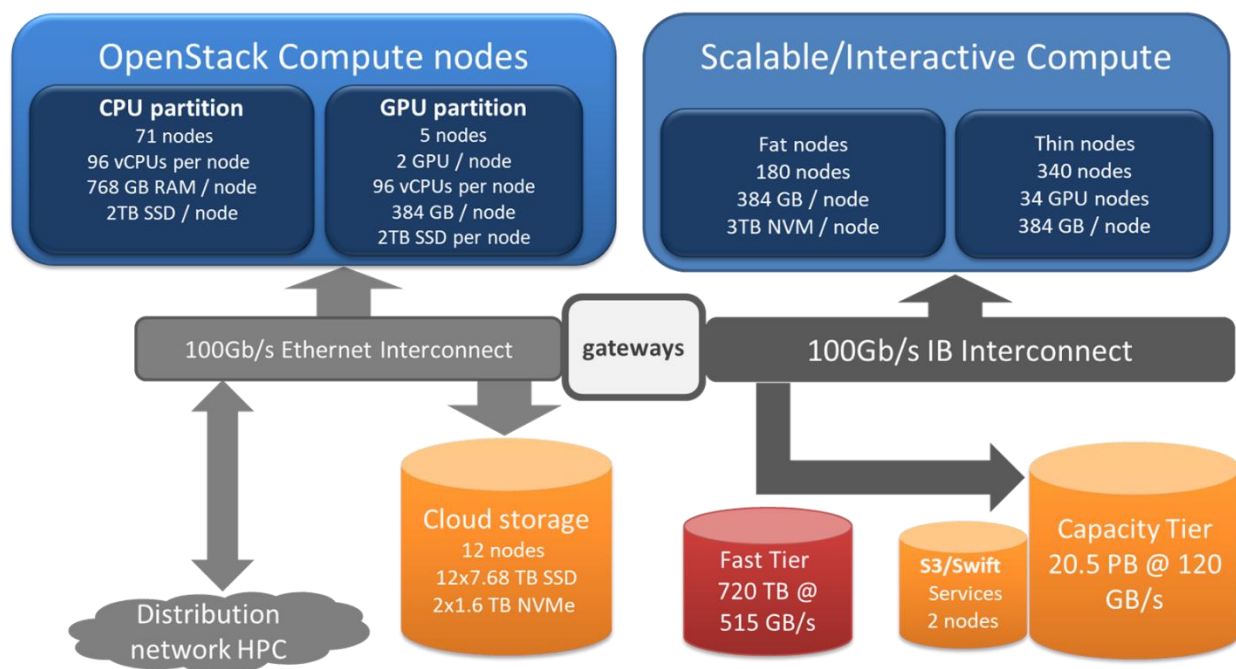


Figura 5 – Architettura del sistema ADA cloud

L'infrastruttura ADA Cloud Computing è basata su OpenStack versione Zed⁸ ed è composta da:

- 3 nodi fisici con ruolo di Cloud Controller, che ospitano i servizi core di OpenStack.
- 76 nodi fisici con ruolo di Compute Node, che forniscono risorse di calcolo, memoria, storage e rete per le macchine virtuali.

Tutti i nodi sono interconnessi tramite rete Ethernet a 100 Gbps basata su tecnologia Mellanox.

⁸ <https://docs.openstack.org/zed/>

Lo storage, connesso direttamente e dedicato integralmente a OpenStack, è costituito da 12 server con 1 PB di dischi raw complessivi. Lo storage è full-flash NVMe/SSD ed è gestito da un cluster CEPH configurato per fornire servizi di object storage, block storage e file system storage.

Per sostenere casi d'uso ad alte prestazioni non è consentita oversubscription delle risorse.

Oltre ai servizi fondamentali di OpenStack⁹, su ADA sono disponibili ulteriori servizi (Tabella 2):

Servizio OpenStack	Ruolo
Nova	Virtualizzazione CPU e creazione/gestione delle VM
Cinder	Gestione del block storage e dei volumi
Neutron	Gestione della rete interna ai progetti e verso reti esterne
Glance	Gestione delle immagini delle VM e degli snapshot
Horizon	Dashboard di OpenStack che fornisce una interfaccia ad OpenStack
Keystone	Gestione delle identità, fornitura di servizi di autenticazione e autorizzazione per tutti i servizi OpenStack.
Barbican	Servizio di gestione delle chiavi a supporto della crittografia per casi d'uso che riguardano dati sensibili.
Manila	Servizio di file system condiviso tra le VM
Octavia	Bilanciamento del carico come servizio (LBaaS)
Trove	Gestione del database come servizio (DBaaS)

Tabella 2 – Servizi OpenStack disponibili su ADA

Le architetture software e hardware sono implementate in piena ridondanza per evitare Single Point of Failure. Questa impostazione ha consentito un'elevata disponibilità del sistema, con un uptime medio del 99,98% per i progetti cloud nel corso del 2025.

⁹ <https://www.openstack.org/software/project-navigator/openstack-components#openstack-services>

ADA è particolarmente adatto a progetti con dati in tempo reale, a workload che trattano dati sensibili e a scenari che richiedono di pubblicare sul web grandi quantità di dati, ad esempio provenienti da simulazioni HPC.

Utilizzo delle risorse Cineca HPC Cloud

La domanda di risorse HPC Cloud è cresciuta in modo costante negli anni (Figura 6 and Figura 7).

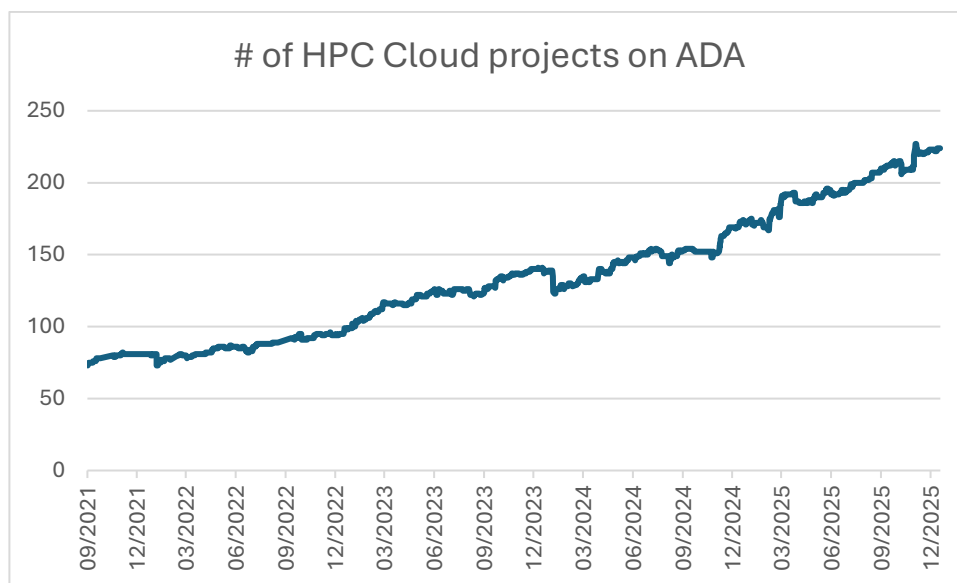


Figura 6 – Progetti HPC cloud sul sistema ADA nel tempo

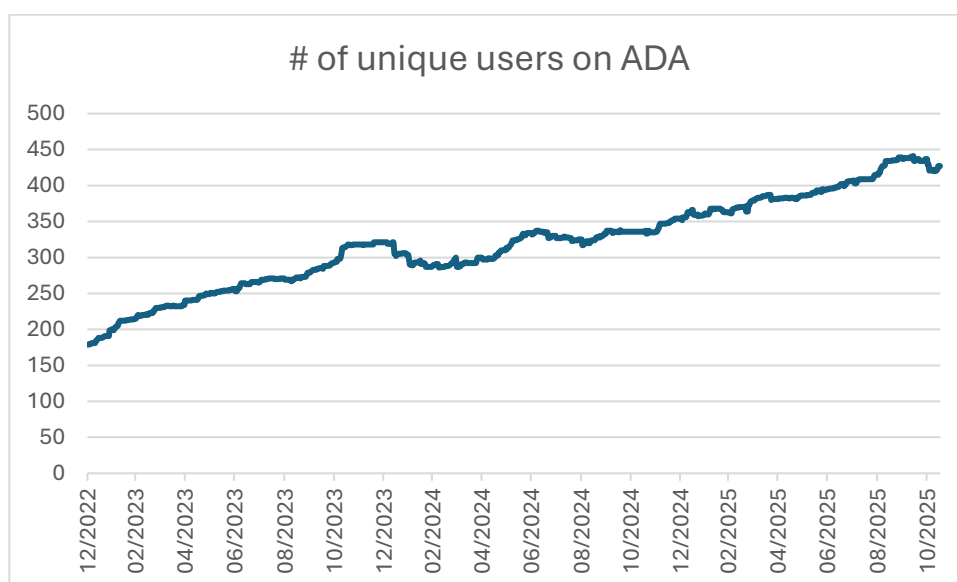


Figura 7 – Utenti HPC cloud sul sistema ADA nel tempo

L'aumento del numero di progetti ha comportato un incremento significativo del numero di vCPU e dello spazio di archiviazione allocati per i progetti su ADA (Figura 8 and Figura 9).

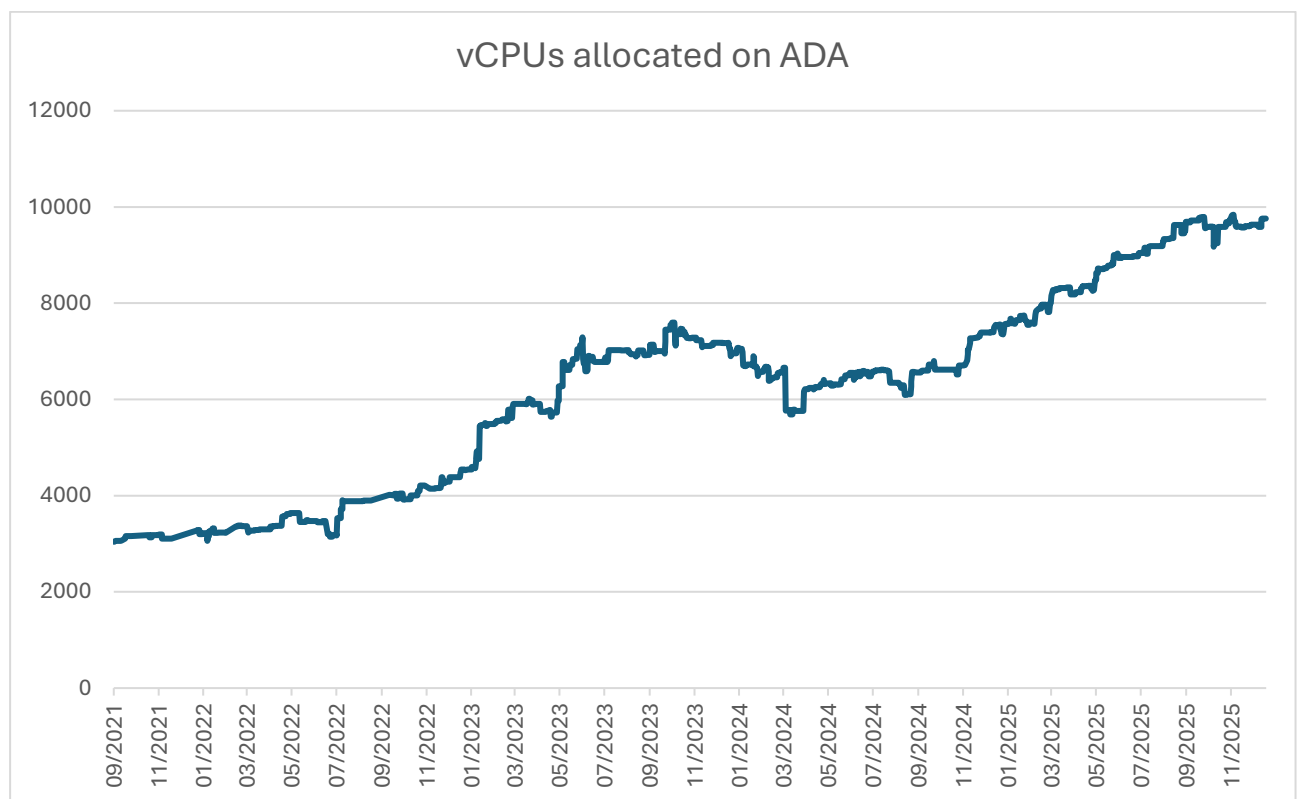


Figura 8 – vCPU allocate sul sistema ADA nel tempo

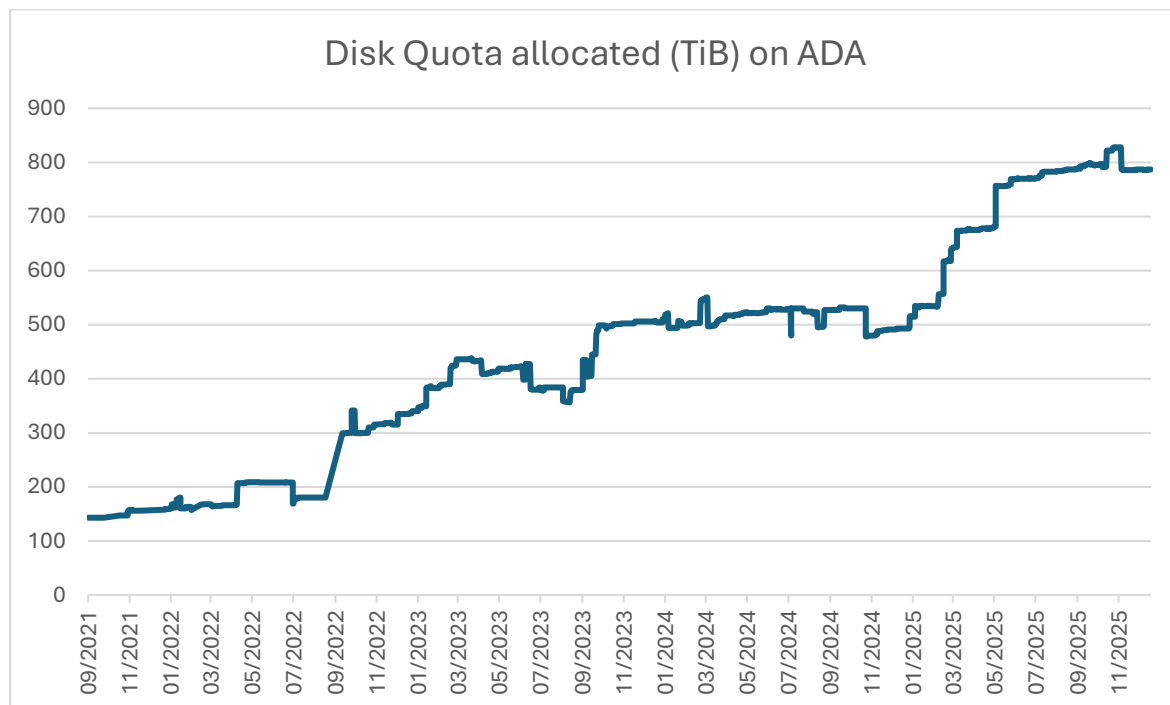


Figura 9 – Storage allocato sul sistema ADA nel tempo

La maggior parte delle risorse è stata destinata a: i) attività di ricerca finanziate da progetti UE (28%), ii) ricercatori italiani tramite il programma ISCRA¹⁰ (26%) e iii) lavori nell'ambito di contratti o accordi con istituzioni (sia pubbliche che private) (Figura10**Error! Reference source not found.**).

¹⁰ <https://www.hpc.cineca.it/hpc-access/access-cineca-resources/iscra-projects/>

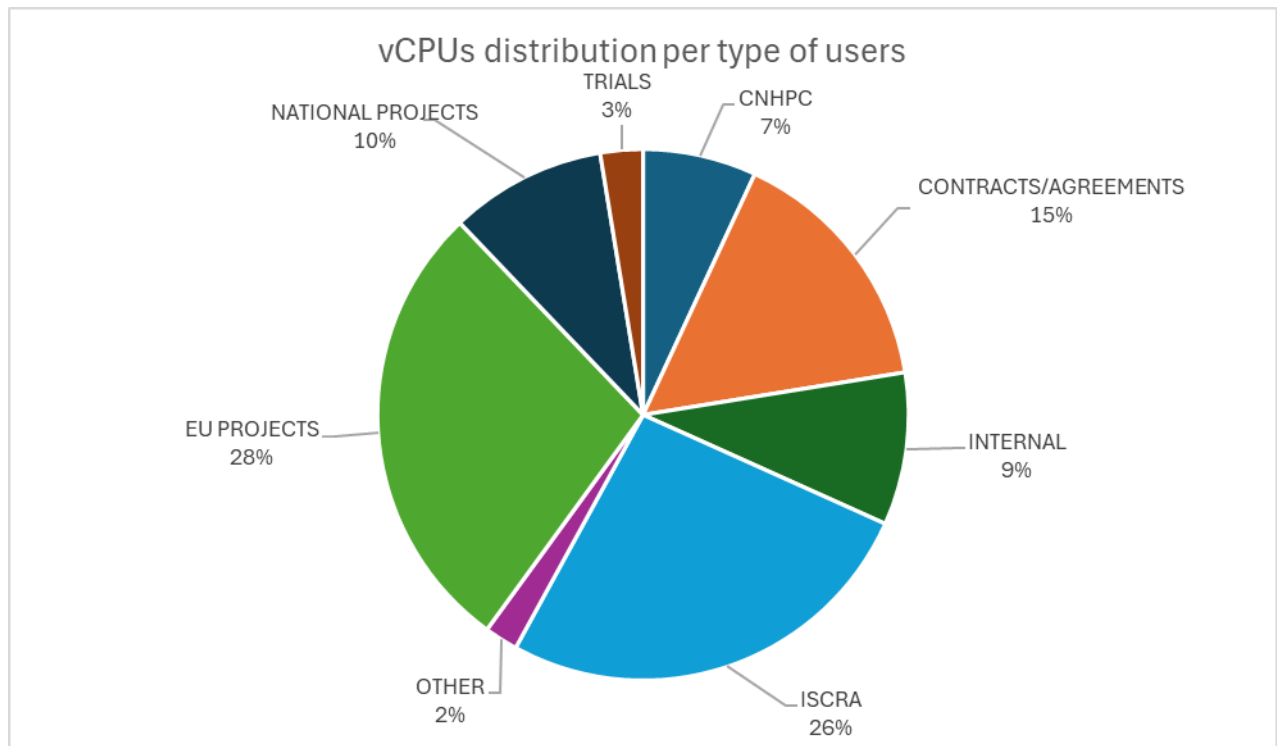


Figura10 10 – Categorie di progetto per i progetti HPC Cloud nel 2024

GAIA

Considerata l'elevata domanda di risorse HPC Cloud nel 2024 e, più in generale, il costante aumento della domanda negli anni, è stata pianificata la realizzazione di una nuova infrastruttura cloud. GAIA è la nuova infrastruttura cloud prevista per il 2026 come risultato del progetto Terabit¹¹.

Lo scopo dell'infrastruttura GAIA Cloud è quello di fornire:

- un numero maggiore di vCPU a disposizione degli utenti, oltre 100.000;
- una partizione GPU adatta sia ai carichi di lavoro HPC che a quelli ottimizzati per IA/Grafica;
- uno storage CEPH dedicato, in parte ottimizzato per IOPS e in parte per capacità, adatto a bilanciare il numero di vCPU disponibili;
- un'infrastruttura di storage data lake multiprotocollo (file, blocchi, oggetti) per la conservazione dei dati a lungo termine e l'elaborazione interattiva.

¹¹ <https://www.terabit-project.it/it/>

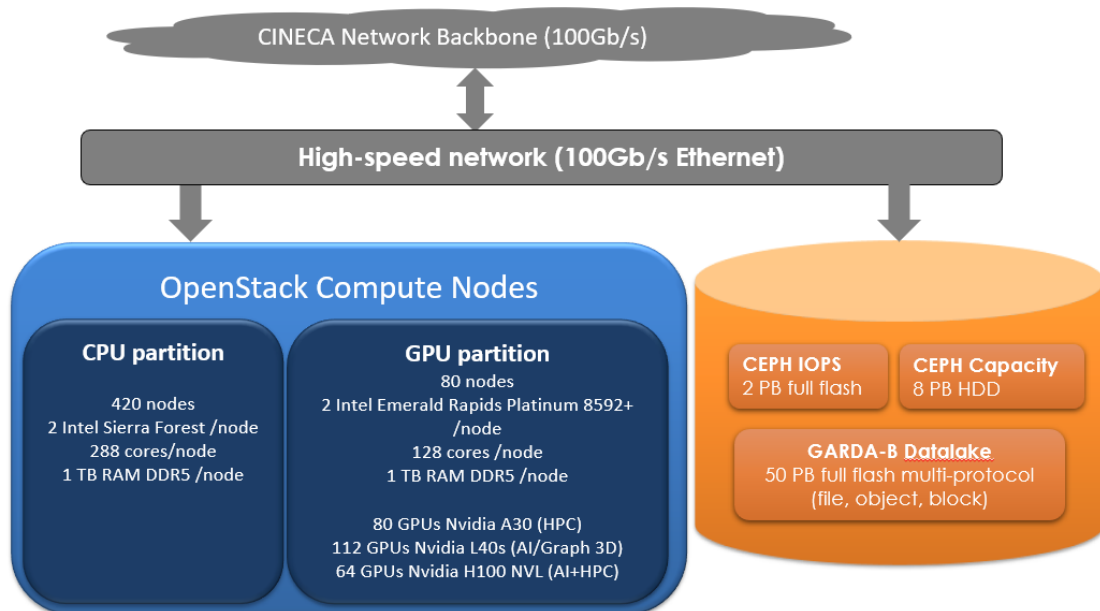


Figura 11 11 – Infrastruttura del sistema GAIA cloud

L'infrastruttura di cloud computing GAIA si basa su OpenStack, come i precedenti sistemi cloud, e includerà un'ampia gamma di servizi OpenStack.

La partizione CPU è dotata di nodi server standard a doppio socket (2 CPU) basati su Intel Xeon 6780E, per un totale di 288 core fisici per nodo.

La partizione GPU è dotata di CPU Intel Xeon 8592+, per un totale di 128 core. Inoltre, questa partizione offre GPU NVIDIA di fascia alta sia per carichi di lavoro di intelligenza artificiale e rendering 3D (L40) sia per casi d'uso scientifici che richiedono aritmetica FP64 a doppia precisione (A30 e H100).

Lo storage è composto da due livelli principali:

- CEPH
 - Partizione CEPH ad alte prestazioni, basata su soluzione full flash
 - CEPH ad alta capacità, basata su unità HDD
- Data lake GARDA-B
 - Soluzione full flash multiprotocollo

Mentre il livello CEPH servirà principalmente l'infrastruttura virtuale (cloud OpenStack) fornendo storage a blocchi per le macchine virtuali e storage per i servizi interni, il data

lake sarà trasversale al cloud e alle altre infrastrutture di supercalcolo grazie ai protocolli NFS e S3. Il nome "GARDA" è un riferimento al lago più grande d'Italia e la "B" sta per Bologna, poiché un'infrastruttura equivalente verrà installata nel data center CINECA di Napoli per fornire un'infrastruttura di storage ridondante.

GAIA diventerà la più grande infrastruttura cloud per la ricerca in Italia e si posizionerà tra le migliori infrastrutture cloud accademiche in Europa.

Conclusioni

L'infrastruttura HPC Cloud di Cineca rappresenta un passo avanti strategico nel panorama computazionale italiano, offrendo una combinazione unica di potenza, flessibilità e sicurezza.

L'infrastruttura HPC Cloud sviluppata presso Cineca rappresenta un ambiente maturo, scalabile e sicuro, in grado di supportare una vasta gamma di flussi di lavoro computazionali scientifici e industriali.

Le piattaforme ADA e GAIA dimostrano il costante impegno di Cineca nel soddisfare le esigenze in continua evoluzione della comunità scientifica, fornendo ambienti isolati per l'elaborazione di dati sensibili, supporto per carichi di lavoro di intelligenza artificiale e funzionalità per il calcolo interattivo. In un contesto di domanda di risorse cloud in costante crescita, il modello HPC Cloud si pone come pilastro fondamentale per la ricerca europea, promuovendo l'accesso equo, la sostenibilità e l'innovazione.

Oltre a supportare la comunità scientifica, questa esperienza fornisce un esempio concreto di come le organizzazioni possano implementare un modello di cloud privato su misura per le proprie esigenze operative e normative. I principi architetturali adottati, come i livelli di virtualizzazione, l'isolamento delle risorse, gli ambienti flessibili gestiti dagli utenti e le funzionalità avanzate di storage e networking, possono costituire un modello di riferimento per le industrie che cercano piattaforme di calcolo robuste, sicure e personalizzabili.