Regole per l'utilizzo della piattaforma GPU/AMPERE

La piattaforma di calcolo è dotata di un server fisico in Sala Macchine (nome Ampere) con 4 GPU Nvidia A100 80GB, disponibile ai docenti, ricercatori, dottorandi e assegnisti del Dipartimento di Informatica. Per massimizzare le risorse e il numero di utenti concorrenti, si è deciso di adottare le seguenti modalità di utilizzo:

GPU RISERVATE *Risorse destinate ad un uso di calcolo intenso*

1. Sono disponibili 3 slot per macchine virtuali riservate, ognuna delle quali avrà collegata una GPU dedicata A100, 82 Gb RAM, 200Gb HD, IP pubblico (eventualmente in una rete a scelta). Il sistema operativo è Ubuntu 24.04 con drive Nvidia v.550 e sarà consegnato al richiedente l'accesso di root, il quale sarà libero di installare software, caricare dati, creare account a discrezione. Non è prevista assistenza tecnica nella gestione della macchina virtuale.
2. L'allocazione del servizio è regolato da un sistema di prenotazione in ordine di arrivo delle richieste. In base alla disponibilità in calendario e alla durata richiesta, gli utenti possono riservare o opzionare un turno settimanale.
   Il servizio di calendario è disponibile https://calampere.di.unimi.it (accesso riservato account Unimi ma tramite accreditamento, scrivendo ad admin (at) di.unimi.it). Sono state introdotte alcune regole per organizzare il servizio:

   1. la prenotazione dura 1 settimana, salvo casi eccezionali da valutarsi in accordo con la Direzione, dal lunedì alla domenica.
   2. è possibile opzionare una settimana successiva ma potrà essere revocata nel caso vi siano altre richieste con priorità maggiori (per es, laboratori con ultime richieste distanti o che non hanno mai usato il servizio)
   3. non è possibile opzionare più di una settimana alla volta.
   4. alla fine del turno, la VM viene spenta d'ufficio, senza preavviso, quindi suggeriamo sempre di concluderne l'utilizzo in anticipo. Se non vi sono prenotazioni e lo slot è libero, la VM viene lasciata accesa finché non giunge eventualmente una richiesta da altri.

3. La VM sarà conservata anche nei periodi di non utilizzo, potrà quindi essere riaccesa nel successivo turno di utilizzo. I dati presenti non saranno soggetti a backup: la conservazione dei dati è demandata al fruitore del servizio.
   Per limitare la creazione di macchine virtuali, vi preghiamo di far confluire all'interno della stessa VM un gruppo di componenti del medesimo laboratorio o di collaboratori che lavorano allo stesso progetto/gruppo di lavoro.

La configurazione **MIG** (Multi Instance GPU) di Nvidia consente di suddividere eventualmente la GPU in più frazioni (massimo 7) che possono essere sub-allocate agli altri utenti e generalmente è attiva di default.
Per usare tutta la potenza della singola GPU, senza partizionamento, disattivare il MIG mode.
Maggiori informazioni su questa configurazione sono riportate qui:

- Manuale Ufficiale https://docs.nvidia.com/datacenter/tesla/mig-user-guide
- Comandi Utili Nvidia MIG

Troubleshooting

E' possibile che alla riallocazione della VM, il MIG mode non sia più attivo perchè la scheda è cambiata: in questo caso, riattivarlo e ripartizionarlo con i comandi opportuni indicati sopra.

L'errore "*Failed to initialize NVML: Driver/library version mismatch*" è generalmente risolvibile con un update dei driver e un reboot.

GPU CONDIVISE *Risorse destinate ad un uso di calcolo semplice (esperimenti brevi e/o prototyping)*

1. E' disponibile un sito Jupyter (https://jupyter.hpc.di.unimi.it), collegato a 1 GPU Nvidia A100 ma suddivise in 7 gpu virtuali.
   Ogni sessione potrà allocarsi una o più schede di calcolo a seconda della disponibilità in quel momento: se compare l'errore "*0 nodes are available: 1 Insufficient_ nvidia.com/gpu*" significa che non vi sono ulteriori schede allocabili.
2. L'allocazione del servizio non richiede una prenotazione ed è sufficiente loggarsi tramite il proprio account di dipartimento, senza @di.unimi.it, ma accreditato da noi tramite richiesta. L'accesso tramite account di ateneo non è consentito.
3. Se è attivo un calcolo e l'utente si scollega dall'interfaccia web, il kernel continua a rimanere attivo. Se il calcolo termina, dopo un periodo di idle, il kernel viene stoppato e liberata la scheda di calcolo.
4. I dati presenti non saranno soggetti a backup: la conservazione dei dati è demandata al fruitore del servizio
5. Per garantire un'equa condivisione delle risorse, ogni utente è dotato di una quota massima di 30Gb di spazio occupato, salvo eccezioni concordate con lo staff tecnico. Nel caso in cui un account superi 1 GB di spazio occupato e non venga effettuato alcun accesso per un periodo continuativo di 12 mesi, i dati dell'utente saranno eliminati.

Lo staff tecnico può avere la necessità di effettuare interventi di manutenzione ordinaria, senza interruzione del servizio, riservandosi degli slot temporali, all'interno del calendario. Gli interventi straordinari invece, che comportano delle interruzioni del servizio, sono comunicati generalmente con una settimana di anticipo.

E' stata creata una mailing-list GPU-Ampere dove saranno inviate le comunicazioni di servizio a tutti gli utenti coinvolti.

L'uso delle risorse della piattaforma è monitorato e i dati sono raccolti per motivi analitici.

From:
https://servizi.di.unimi.it/ - **Servizi Tecnologici**
Dipartimento di Informatica

Permanent link:
https://servizi.di.unimi.it/servizi:servtec:gpu

Last update: **2025/06/27 11:30**