FONDAZIONE BRUNO KESSLER
MAGAZINE

L’algoritmo del risparmio

26 Settembre 2025

Il Centro per la Cybersecurity di FBK ha concluso il primo servizio del progetto InnovAction, testando un modello di GPU-as-a-Service per ottimizzare i costi di questa preziosa risorsa. L’attività è stata finanziata dal progetto alla PMI Clastix. Funziona e fa una bella differenza.

Nell’ambito dell’iniziativa InnovAction, che mette al centro la trasformazione delle imprese italiane attraverso l’adozione di tecnologie avanzate e sostenibili, FBK – con il Centro Digital Industry e il Centro per la Cybersecurity – offre servizi, che rientrano nella categoria “Test Before Invest”, descritti nell’apposito catalogo. Questa modalità permette alle imprese di sperimentare tecnologie innovative in un ambiente controllato, prima di procedere alla loro implementazione definitiva, riducendo così i rischi associati agli investimenti.

Grazie a InnovAction, le aziende possono accedere all’offerta di servizi del progetto, usufruendo di Aiuti di Stato che permettono la copertura fino ad un massimo del 100% dei costi. Il progetto InnovAction ha infatti ricevuto il *Seal of Excellence* nell’ambito di una call europea del programma Digital Europe per la selezione degli European Digital Innovation Hubs (**EDIHs**), con l’obiettivo di sostenere competitività e crescita attraverso l’innovazione. Attivo dal 2023 al 2026, il progetto è cofinanziato dal Ministero delle Imprese e del Made in Italy (MIMIT), dall’Unione Europea tramite il programma Next Generation EU e dai fondi del Piano Nazionale di Ripresa e Resilienza (**PNRR**).

Un esempio concreto di sviluppo e applicazione di un servizio realizzato da FBK e finanziato tramite il progetto InnovAction riguarda la collaborazione del Centro per la Cybersecurity con una PMI, Clastix, per ottimizzare l’allocazione delle risorse computazionali nel contesto del Cloud Computing, con particolare attenzione alle GPU.

Le GPU (unità di elaborazione grafica) sono essenziali per la fase di training delle tecniche di Intelligenza Artificiale che altrimenti richiederebbero tempi di elaborazione proibitivi con le tradizionali CPU. Il modello “GPU as a Service” (**GPUaaS**) è un servizio cloud che offre accesso on-demand a potenti unità di elaborazione grafica (GPU) tramite internet, consentendo alle aziende di utilizzare la potenza di calcolo parallelo delle GPU per carichi di lavoro intensivi come l’intelligenza artificiale, il machine learning, e il rendering 3D, senza dover acquistare e gestire l’hardware costoso e complesso. Questo modello permette di scalare rapidamente l’infrastruttura in base alle esigenze e di pagare solo per il tempo e le risorse effettivamente utilizzate, rendendo più accessibili le tecnologie che offrono elevate capacità di calcolo.

L’IA energivora e la necessità di aguzzare l’ingegno

Con l’avvento dell’Intelligenza Artificiale, che si sta affermando come la più grande rivoluzione tecnologica degli ultimi anni, la necessità di un utilizzo sempre piu’ intensivo di GPU per addestrare modelli e servire utenti comporta un aumento significativo dei costi per le aziende e un elevato consumo energetico.

I consumi da uso di AI, in particolare per la fase di training e creazione di modelli, rischiano di essere una fonte di problemi importanti se le tecniche di AI verranno adottate sempre di più. Basti pensare che i data center vengono posizionati vicino a laghi, per il raffreddamento, e/o vicino a centrali elettriche, per l’energia. Ridurre il consumo energetico dovuto all’uso di GPU (che sono energivore) può contribuire, insieme ad altre soluzioni, a tenere insieme i vantaggi dell’applicazione di tecniche di AI e un ridotto impatto sull’ambiente.

Problem solving

L’attività del progetto si è concentrata sulla progettazione di un **algoritmo di ottimizzazione** per migliorare l’efficienza nell’utilizzo delle GPU al fine di diminuire i costi di utilizzo, mantenendo contemporaneamente inalterata la qualità di servizio richiesta.

In dettaglio, l’algoritmo permette di ottimizzare l’utilizzo di GPU da parte di applicativi di Intelligenza Artificiale (Large Language Models) in modo che il costo complessivo a carico di chi affitta queste GPU (dai Cloud Providers) sia più basso.

Si tratta di una sorta di problema di **impacchettamento** (bin packing in inglese) dove bisogna trovare la strategia migliore per impacchettare il maggior numero di oggetti (gli applicativi di IA) nel minor numero di contenitori (le GPU) o in GPU che costano meno. Una ulteriore complessità del problema, rispetto a un semplice bin packing, riguarda il vincolo di mantenere elevata la qualità del servizio offerto, senza cioè incorrere in prestazioni degradate.

Al progetto, che si è concluso nel mese di luglio 2025, hanno lavorato **Silvio Cretti e Marco Zambianco** dell’Unità Distributed AI for dependable cyberSecuritY (DAISY) del Centro per la Cybersecurity.

Colmare le distanze

“La maggiore difficoltà – commenta **Cretti** *– è stata capirsi e riuscire a parlare lo stesso linguaggio, perché spesso l’approccio pragmatico dell’azienda si scontra con quello più teorico del centro di ricerca. Il gergo è diverso e spesso le stesse parole hanno significati diversi in contesti differenti oppure sono utilizzate solo in un contesto e non in un altro (per esempio i concetti di simulazione, ottimizzazione, orchestrazione). Inoltre, solitamente l’azienda ha un problema pratico da risolvere (voglio diminuire i costi in un contesto molto specifico) mentre il centro di ricerca ha spesso una soluzione generica che non sempre si adatta a quello specifico contesto. Adattarsi non è facile: si potrebbe dire che* ***coesistono un “problema alla ricerca di una soluzione” e “una soluzione alla ricerca di un problema”****. Nel progetto siamo riusciti a trovare un punto di incontro, non perfetto per nessuno ma abbastanza buono per entrambi.”*

Vantaggi reciproci

"Mentre il centro di ricerca può dimostrare – aggiunge Cretti – *che un proprio risultato "funziona" non solo sulla carta o in un ambiente di simulazione, ma in un contesto simile a quello operativo, l'azienda ha ricevuto una Proof of Concept che un problema (i costi di utilizzo delle GPU) può essere mitigato grazie a un algoritmo di ottimizzazione:* ***con l'algoritmo sviluppato da FBK si riesce a ottenere un risparmio "teorico" di circa il 20% sui costi di esercizio pur mantenendo delle prestazioni adeguate****."*

"Gli algoritmi di gestione energetica per le GPU come quelli sviluppati nel progetto – commenta **Silvio Ranise**, Direttore del Centro FBK per la Cybersecurity – *possono contribuire alla messa in sicurezza della fase di training delle moderne tecniche di AI. Funzionano come un "controllore del traffico" che impedisce a singoli processi di monopolizzare le risorse, bloccando gli attacchi di esaurimento (denial of service) delle risorse che rallentano o fermano le computazioni. Un altro vantaggio di algoritmi così concepiti consiste nel rendere il consumo di energia della GPU più uniforme, nascondendo fluttuazioni che gli attaccanti possono usare per inferire dati di training e il funzionamento del modello, contrastando così gli attacchi basati sull'analisi del consumo energetico (power side-channel)."*

Cover foto: Silvio Cretti e Marco Zambianco del Centro Cybersecurity di FBK

LINK

https://magazine.fbk.eu/it/news/lalgoritmo-del-risparmio/

TAG

- #3d
- #ai
- #algoritmo
- #clastix
- #cloud computing
- #cybersicurezza
- #daisy
- #edih
- #GPU
- #GPUaaS
- #industriadigitale
- #innovaction
- #intelligenzaartificiale

- #llm
- #Machine learning
- #PMI
- #pnrr
- #proofofconcept
- #testbeforeinvest

AUTORI

- Giancarlo Sciascia