Report LLM su ThinkStation PGX

Framework metodologico per sistemi conversazionali enterprise

Una collaborazione Memori, Lenovo e Araneum e i loro team

a cura di:

Fabio Lecca, Matteo Mastranza, Paolo Mazzitti, Francesco Piccolo, Mario Sebastiani, Massimo Chiriatti.

Marzo 2026

Prefazione

Nel contesto dei precedenti studi condotti sui vantaggi dell'adozione di Large Language Models open source in esecuzione locale, è emerso con chiarezza che tali modelli, a partire da una certa soglia — indicativamente dagli 8 miliardi di parametri in su — sono in grado di soddisfare i principali casi d'uso aziendali, pur presentando limitazioni intrinseche legate alla loro natura e configurazione.

Muovendoci lungo il percorso tracciato nei report precedenti, questo nuovo studio si propone di analizzare le performance degli stessi modelli, oltre ad alcuni più recenti, su un nuovo tipo di macchina caratterizzata da una nuova architettura (GB10 Grace Blackwell), creata da Nvidia con una memoria unificata da 128Gb, e dalla totale compatibilità con tutte le librerie software Nvidia, in modo da garantire alle aziende un accesso a basto costo a tutto l'insieme di strumenti AI, per poter pianificare il corretto percorso di adozione di questa tecnologia.

L'obiettivo è duplice: da un lato, valutare quantitativamente le performance dei modelli, analizzando il carico di richieste gestibili per unità di tempo in un contesto di risorse più vincolate; dall'altro, identificare lo scenario operativo più idoneo per ciascun modello analizzato, così da orientare scelte di adozione consapevoli e strategicamente fondate.

Indice

1. Introduzione

Obiettivi del test

In questo nuovo ciclo di test, l'obiettivo è quello di valutare le prestazioni della nuova macchina, destinata all'esecuzione locale di LLM, confrontandole con quelle dell'infrastruttura utilizzata nello studio passato. Alla luce dei risultati emersi precedentemente, l'analisi si concentra solo su aspetti prestazionali, legati alla gestione di carico concorrente. In particolare, l'attenzione è rivolta alla capacità del sistema di gestire richieste simultanee provenienti da più utenti, mantenendo tempi di risposta adeguati. Non sono stati quindi ripetuti i test qualitativi sulle risposte generate dai modelli, in quanto i risultati già ottenuti, hanno fornito evidenze sufficienti in merito alla qualità delle risposte prodotte dai modelli utilizzati. Inoltre, i modelli impiegati in questo studio, sono gli stessi, oppure versioni più recenti, che in linea teorica presentano prestazioni qualitative uguali o migliori rispetto ai modelli già analizzati.
Come fatto precedentemente, i test quantitativi sono stati eseguiti utilizzando due metodologie. La prima prevede l'utilizzo del benchmark integrato di vLLM, impiegato per misurare il numero di richieste completate al secondo, assumendo una lunghezza in output di 128 token. La seconda invece, prevede l'utilizzo di Apache JMeter, per simulare scenari di utilizzo reale, con l'obiettivo di determinare il numero massimo di utenti simultanei che il sistema è in grado di gestire, mantenendo prestazioni accettabili.

2. Server Lenovo

Specifiche Hardware

La workstation Lenovo ThinkStation PGX impiegata per l'esecuzione dei test è configurata con componenti hardware basati su architettura ARM e piattaforma NVIDIA Grace-Blackwell, pensati per sostenere carichi di lavoro intensivi legati all'elaborazione di LLM. Di seguito una panoramica delle principali caratteristiche tecniche.

- **Processore:** il sistema è equipaggiato con architettura **aarch64** e integra due cluster CPU basati su **Cortex-X952** e **Cortex-A725**, per un totale di 20 core fisici e 20 thread (1 thread per core). I core raggiungono rispettivamente una frequenza massima di 3.9 GHz e 2.808 GHz.
- **Memoria RAM:** il sistema dispone di 128 GB di RAM LPDDR5, con una frequenza di 8533 MT/s. L'architettura a **memoria unificata**, consente di gestire in modo efficiente le risorse tra CPU e GPU.
- **GPU:** 1x NVIDIA GB10, basata su architettura Blackwell ed è compatibile con CUDA 13.0
- **Storage:** il sistema è dotato di due dispositivi di archiviazione, un HDD esterno da 3.6 TB (utilizzato per lo storage dei modelli LLM) e un SSD NVMe ad alte prestazioni, da 931 GB

- **Sistema operativo:** la macchina esegue Ubuntu 24.04.4 LTS a 64 bit, una distribuzione Linux stabile e compatibile con ambienti HPC e AI.

3. Modelli usati per i test

Nello studio sono stati analizzati diversi LLM open source per identificarne l'efficacia nel nostro caso di studio

LLM usati

Modello	Azienda	Context window (Tokens)	Memoria occupata (approx)
Qwen3-32B	Qwen	128k*	30 GB
Qwen3-30B-A3B	Qwen	128k*	28 GB
Qwen3-14B	Qwen	128k*	27.5 GB
Qwen3-8B	Qwen	128k*	7.64 GB
Qwen3-4B	Qwen	128k*	3.8 GB
Qwen3-1.7B	Qwen	32k	1.6GB
Qwen3-0.6B	Qwen	32k	523MB
granite-3.1-8b-instruct	ibm-granite	32k*	7.61GB
granite-3.3-8b-instruct	ibm-granite	128k	7.61GB
granite-3.3-2b-instruct	ibm-granite	128k	2.36GB
gemma-3-27b-it	google	128k	17GB
Llama-3.1-8B-Instruct	meta-llama	128k	4.5GB
Llama-3.3-70B-Instruct-NVFP4*	nvidia	128k	43GB
Llama-3.2-3B	meta-llama	128k	2GB

*modello di meta-llama, ma quantizzato da nvidia

Modello	Azienda	Context window (Tokens)	Memoria occupata (approx)
gpt-oss-20b	OpenAI	128k	16 GB
gpt-oss-120b	OpenAI	128k	60 GB
Qwen3-Next-80B-A3B-Instruct	Qwen	256k	80 GB

NVIDIA-Nemotron-3-Nano-30B-A3B-BF16	Nvidia	1M	60 GB
phi-4 (14 B)	Microsoft	16k	12 GB

Proprietà dei modelli

Questi modelli hanno delle proprietà che li caratterizzano:

- **Dimensione e memoria**: modelli più grandi offrono prestazioni migliori in termini di accuratezza ma richiedono molta più memoria. I modelli più piccoli sono meno esigenti ma adatti a compiti più semplici.
- **Finestra di contesto**: maggiore è la finestra di contesto, più le richieste possono essere articolate. Tuttavia, l'elaborazione di input molto lunghi può ridurre la qualità delle risposte.

Con le risorse hardware a disposizione, non è stato possibile testare modelli di dimensioni superiori, così da garantire un livello adeguato di concorrenza nell'ambiente di esecuzione previsto.

4. Benchmark vLLM: token/sec

Questo capitolo fornisce un'analisi quantitativa delle performance dei modelli testati, utilizzando un dataset di circa **53k conversazioni ShareGPT** derivate da un dataset iniziale di ~100 k interazioni presente su Hugging Face:
https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered/blob/main/ShareGPT_V3_unfiltered_cleaned_split.json

vLLM

vLLM è una libreria open-source progettata per l'inferenza e il servizio di modelli di linguaggio di grandi dimensioni (LLM) in modo efficiente e ad alte prestazioni. Con il comando *serve* si possono istanziare i LLM installati per interrogarli e valutarne le prestazioni.

Metodologia di benchmark

I test sono stati condotti utilizzando input standardizzati e un limite massimo di token per la risposta (**Output-len**) impostato a 128 token. Per ogni modello, sono stati raccolti i seguenti indicatori di performance:

- **throughput (requests/s)**: il numero di richieste completate al secondo, che rappresenta un indicatore diretto della capacità del modello di gestire carichi multiutente.
- **total tokens/s**: il totale dei token processati al secondo, inclusi input e output.

- **output tokens/s**: il numero di token generati dal modello al secondo, un indicatore dell'efficienza del modello nella produzione di risposte.

Per avviare i test è stato predisposto un container Docker con l'immagine di NVIDIA vLLM *nvcr.io/nvidia/vllm:25.11-py3,* per garantire un ambiente di esecuzione riproducibile e ottimizzato per la GPU a disposizione. Il container ha accesso diretto alla GPU ed è stato configurato con i parametri di sistema necessari per gestire carichi di lavoro ad alta intensità di memoria. Il benchmark è stato eseguito tramite il comando *vllm bench throughput*, che sottopone il modello all'intero dataset di conversazioni, misurando le metriche di throughput descritte in precedenza.

Comando utilizzato per far partire i test:

```
docker run -it --rm   --gpus all   --ipc=host   --ulimit memlock=-1
--ulimit stack=67108864  -e HF_TOKEN="$TOKEN"
-v ~/test:/tmp/dataset   -v
~/.cache/huggingface:/root/.cache/huggingface
nvcr.io/nvidia/vllm:25.11-py3   vllm bench throughput
--dataset /tmp/dataset/ShareGPT_V3_unfiltered_cleaned_split.json --
output-len 128     --model $MODEL_NAME
```

Modello	Requests/s	Total tokens/s	Output tokens/s
ibm-granite/granite-3.1-8b-instruct	7,21	2.750,44	923,03
ibm-granite/granite-3.3-2b-instruct	18,42	7.027,33	2.358,33
ibm-granite/granite-3.3-8b-instruct	7,11	2.710,95	909,78
Qwen/Qwen3-0.6B	26,94	9.625,56	3.447,96
Qwen/Qwen3-1.7B	19,96	7.133,98	2.555,45
Qwen/Qwen3-4B	11,47	4.097,13	1.467,63
Qwen/Qwen3-8B	8,05	2.877,32	1.030,68
Qwen/Qwen3-14B	4,76	1.699,90	608,92
Qwen/Qwen3-32B	2,09	746,05	267,24
Qwen/Qwen3-30B-A3B	4,73	1.691,26	605,82
google/gemma-3-27b-it	2,37	882,06	303,97
meta-llama/Llama-3.2-3B	16,42	5.783,81	2.101,23
nvidia/Llama-3.3-70B-Instruct-NVFP4	1,56	549,66	199,73
meta-llama/Meta-Llama-3.1-8B-Instruct	8,60	3.029,60	1.100,64

Comparazione con la precedente infrastruttura

Per capire la potenza di questa nuova macchina, sono stati comparati i benchmark ottenuti con il server precedente. A differenza della scorsa analisi, i test sono stati eseguiti utilizzando gli stessi LLM, per garantire una comparazione omogenea.

	Macchina	Request/s	Performance Ratio
ibm-granite/granite-3.1-8b-instruct	H200	85,46	0,08
	GB10	7,21	
ibm-granite/granite-3.3-2b-instruct	H200	115,08	0,16
	GB10	18,42	
ibm-granite/granite-3.3-8b-instruct	H200	82,42	0,09
	GB10	7,11	
Qwen/Qwen3-0.6B	H200	150,43	0,18
	GB10	26,94	
Qwen/Qwen3-1.7B	H200	134,34	0,15
	GB10	19,96	
Qwen/Qwen3-4B	H200	108,78	0,11
	GB10	11,47	
Qwen/Qwen3-8B	H200	90,91	0,09
	GB10	8,05	
Qwen/Qwen3-14B	H200	65,5	0,07
	GB10	4,76	
Qwen/Qwen3-32B	H200	35,24	0,06
	GB10	2,09	
Qwen/Qwen3-30B-A3B	H200	75,35	0,06
	GB10	4,73	
google/gemma-3-27b-it	H200	28,24	0,08
	GB10	2,37	

meta-llama/Llama-3.2-3B	H200	129,53	0,13
	GB10	16,42	
meta-llama/Llama-3.3-70B-Instruct	H200	18,77	0,08
	GB10	1,56	
meta-llama/Llama-3.1-8B-Instruct	H200	88,73	0,10
	GB10	8,60	

Test modelli più recenti

Come già specificato in precedenza, i test sono stati condotti anche su dei modelli più recenti e di seguito ci sono i risultati ottenuti.

Modello	Requests/s	Total tokens/s	Output tokens/s	Output-len
openai/gpt-oss-20b	9,81	3479,56	1255,34	128
openai/gpt-oss-120b	3,76	1333,69	481,16	128
nvidia/NVIDIA-Nemotron-3-Nano-30B-A3B-BF16	3,47	1262,32	443,85	128
microsoft/phi-4	5,02	1778,32	642,82	128

Risultati

Dai confronti emerge che le prestazioni di questa nuova macchina, risultano significativamente inferiori rispetto all'architettura basata su H200, risultato atteso in quanto si tratta della macchina meno costosa disponibile sul mercato.

Per la maggior parte dei modelli, il throughput raggiunge il 6% - 10% di quella della H200, evidenziando un divario prestazionale nell'ordine di circa 10-15x. Solo per i modelli più piccoli, come Qwen 0.6B e 1.7B, il gap si riduce a circa 5x, ma rimane comunque una differenza rilevante rispetto alla architettura H200 come riferimento.

Questi risultati vanno quindi analizzati nel contesto di una macchina entry-level che ha la capacità di eseguire gli stessi modelli delle macchine più costose e quindi permette di effettuare dei test locali abbassando notevolmente il costo di ingresso nel mondo AI e permettendo di pianificare un corretto percorso di adozione di queste tecnologie.

5. Test di prestazioni: utenti simultanei

L'obiettivo principale del test è identificare il giusto compromesso tra tre fattori fondamentali:

- **Qualità delle risposte**: ovvero la capacità del modello di fornire output coerenti, rilevanti e precisi rispetto alla richiesta
- **Tempo di risposta**: parametro cruciale in applicazioni interattive o real-time, che misura la reattività percepita dal punto di vista dell'utente
- **Numero di richieste gestibili nell'unità di tempo**: indicatore diretto della scalabilità del sistema, fondamentale per mantenere performance accettabili anche in scenari ad alto traffico.

Questi tre elementi sono correlati tra loro: migliorare la qualità richiede modelli più complessi e pesanti, che però aumentano i tempi di risposta e riducono il throughput complessivo. Al contrario, l'utilizzo di modelli più leggeri migliora scalabilità e latenza, ma può compromettere la qualità delle risposte. L'equilibrio tra questi fattori è quindi essenziale per garantire un'esperienza d'uso ottimale all'interno di un contesto come quello analizzato.

Per arrivare a quantificare un numero di utenti simultanei il più possibile corrispondente con un caso d'uso reale, sono state effettuate le seguenti considerazioni:

- la "Qualità delle risposte" è stata assicurata selezionando un modello di riferimento per i test, nel caso della Thinkstation PGX è stato selezionato Qwen/Qwen3-8B, che è un modello entry-level che ha dato risultati accettabili con AIsuru
- è stata definita una soglia di tempo di risposta accettabile (100 secondi al massimo per ottenere una risposta complessa), in modo da garantire la qualità percepita dall'utente
- è stato quindi eseguito un insieme di test aumentando via via il numero di utenti simultanei, in modo da arrivare al numero massimo di richieste simultanee che rispettano il parametro del tempo di attesa (cosiddetto "punto di rottura"),
- è stato utilizzato un prompt tipico di AIsuru costituito da 50.000 caratteri prevalentemente in lingua italiana (circa 12.500 token), utilizzando tale dimensione come valore medio, tenendo conto che alcune richieste possono essere più semplici ed altre più complesse e richiedendo più iterazioni, come ad esempio quelle che prevedono il *tool calling.*

Test Apache JMeter

In quest'ottica, il test di performance è stato progettato con carichi progressivi, fino a raggiungere un punto di effettivo stress dell'infrastruttura, identificato nel carico di circa 10 utenti simultanei.

Modello	GPU Mem	Num thread	Tempo totale	Throughput	Media	Min	Max
Qwen/Qwen3-8B	8,17Gb	1	00:01:14	0.01 req/s	73965	73965	73965
Qwen/Qwen3-8B	8,17Gb	2	00:01:32	0.01 req/s	82592	73965	91219

Qwen/Qwen3-8B	8,17Gb	3	00:01:33	0.01 req/s	87224	77628	92228
Qwen/Qwen3-8B	8,17Gb	4	00:01:43	0.01 req/s	96015	83584	102886

Modello	GPU Mem	Num thread	Tempo totale	Throughput	Media	Min	Max
openai/gpt-oss-20b	13.72Gb	1	00:00:22	0.02 req/s	21565	21565	21565
openai/gpt-oss-20b	13.72Gb	2	00:00:23	0.1 req/s	25022	21350	28694
openai/gpt-oss-20b	13.72Gb	3	00:00:43	0.1 req/s	32457	25375	32756
openai/gpt-oss-20b	13.72Gb	4	00:00:44	0.1 req/s	32287	25590	43657
openai/gpt-oss-20b	13.72Gb	8	00:00:50	0.2 req/s	36201	30014	49548
openai/gpt-oss-20b	13.72Gb	16	00:00:51	0.3 req/s	43879	37452	50433

Modello	GPU Mem	Num thread	Tempo totale	Throughput	Media	Min	Max
openai/gpt-oss-120b	65.97Gb	1	00:00:29	0.02 req/s	28896	28896	28896
openai/gpt-oss-120b	65.97Gb	4	00:00:44	0.1 req/s	42191	39732	43499
openai/gpt-oss-120b	65.97Gb	8	00:01:01	0.1 req/s	54968	51273	60132
openai/gpt-oss-120b	65.97Gb	16	00:01:18	0.2 req/s	72542	62026	77040

In questo caso con i modelli oss-gpt-20b e oss-gpt-120b si osserva che la macchina risulta gestire situazioni di parallelismo in maniera molto efficiente.

6. Modelli NIM di Nvidia

La licenza AI Enterprise di Nvidia consente di accedere al repository NGC (NVIDIA Container Registry), che contiene immagini Docker certificate che espongono modelli e motori di inferenza ottimizzati per le architetture NVIDIA con servizi web compatibili con OpenAI mediante i NIM (Nvidia Inference Microservices), usufruendo di un supporto costante che prevede anche aggiornamenti tempestivi e verifiche di sicurezza delle immagini.
Nel caso della Thinkstation PGX, in particolare risulta di grande interesse la possibilità di accedere a modelli ottimizzati quantizzati con la tecnica NVFP4, che è un formato floating point supportato nativamente dall'architettura GB10 Blackwell. Tra i modelli disponibili sono presenti:

- NVIDIA-Nemotron-Nano-9B-v2-DGX-Spark
- Qwen3-32B NIM for DGX Spark

- Llama-3.1-8b-Instruct-DGX-Spark

7. Dimensionamento di RAM e CPU

Il dimensionamento della RAM e della CPU di un server è sempre un tema molto importante nel caso di applicazioni classiche per determinare il numero massimo di utenti che possono utilizzare il sistema. Per questo motivo, durante tutti i test è stato monitorato l'utilizzo di RAM e CPU per capire se l'inferenza porta con sé anche un consumo di risorse significativo, tale da richiedere una correzione nei parametri di dimensionamento. Durante i test non si è osservato un consumo aggiuntivo significativo della memoria unificata rispetto a quella utilizzata dal modello.
E' tuttavia importante osservare che la memoria unificata della Thinkstation PGX è molto preziosa quindi occorre monitorare costantemente il memory footprint delle applicazioni per evitare di sottrarre memoria alla GPU, eventualmente si consiglia di spostare la parte applicativa su un'altra macchina meno costosa e utilizzare al meglio la GPU istanziando anche più modelli fino ad occupare tutta la memoria disponibile per l'inferenza.

8. Conclusioni

Lo studio corrente è stato completato a pochi mesi di distanza dal precedente, e nel frattempo abbiamo avuto a disposizione un'infrastruttura notevolmente più performante rispetto al costo dell'hardware, nonché una generazione successiva per tutti i modelli open source che erano stati testati a gennaio 2025. Questi due fattori lasciano ben sperare sul fatto che le soluzioni LLM on premise saranno via via sempre più utilizzabili in ambiti lavorativi dove la privacy e la disponibilità dei dati aziendali sia un fattore determinante per la scelta di tale soluzione.

La nuova infrastruttura, pur mostrando delle performance ridotte rispetto alla precedente (costituita dalle schede H200), in realtà democratizza l'accesso ai modelli LLM locali in quanto il costo di accesso è notevolmente minore, inoltre la RAM unificata da 128Gb consente l'utilizzo di modelli con maggior numero di parametri. Sono stati infatti identificati dei modelli open source che si integrano perfettamente con la piattaforma AISuru, forniscono una qualità delle risposte più che soddisfacente (in particolare Nemotron-3-nano-30B-A3B-nvfp4, gpt-oss-20b o gpt-oss-120b e Qwen3.5-35B-A3B) e consentono anche un alto grado di parallelismo delle richieste, in modo da consentire a un piccolo gruppo di utenti aziendali simultanei (1-5) di interagire correttamente con gli agenti AIsuru.

La Thinkstation PGX si è dimostrata molto versatile in quanto è stato possibile testare tanti progetti di diversa natura oltre all'inferenza di LLM, proprio in virtù della sua compatibilità con tutte le librerie NVIDIA, per i seguenti compiti, tutti eseguiti con elaborazione esclusivamente locale:

- traduzione di testi mono e multilingue, anche in modalità batch
- sintesi vocale di testo, anche in più lingue
- generazione di modelli 3d da immagini
- generazione di modelli 3d da prompt
- generazione di immagini da prompt
- clonazione della voce umana, sintesi vocale con la voce clonata

- generazione di filmati con lip-sync per la produzione di contenuti audiovisivi educativi
- fine tuning di modelli