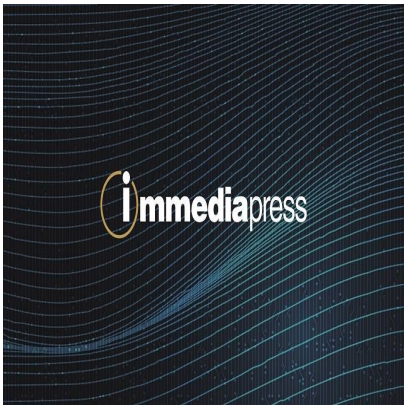


Supermicro lancia KV Cache Solution con IBM e NVIDIA dimostrando significativi vantaggi sulle prestazioni di inferenza

Descrizione

default watermark

COMUNICATO STAMPA â?? CONTENUTO PROMOZIONALE

SAN JOSE, California, 5 nov 2026 /PRNewswire/ â?? Supermicro, Inc. (NASDAQ: SMCI), un fornitore di soluzioni IT, IA, Enterprise, Storage e 5G/Edge Total con Data Center Building Block SolutionsÂ® (DCBBS), ha annunciato oggi una nuova soluzione di archiviazione di cache KV con IBM e NVIDIA. La soluzione congiunta integra server di archiviazione all-flash Supermicro Petascale utilizzando processori AMD EPYC 9005 Series (â??Torinoâ?•), software IBM Storage Scale, insieme al software di framework di inferenza NVIDIA Dynamo e alla rete Ethernet Spectrum-X.

Insieme, queste tecnologie possono ridurre notevolmente il tempo necessario ai modelli di intelligenza artificiale per iniziare a generare una risposta, un vantaggio fondamentale, poichÃ© le imprese adottano flussi di lavoro di IA agentici e a lungo termine. Estendere la capacitÃ di cache KV oltre la memoria GPU locale puÃ² migliorare lâ??utilizzo della GPU e la velocitÃ effettiva di inferenza, consentendo al contempo unâ??elevata disponibilitÃ del contenuto della cache KV e una scalabilitÃ non distruttiva della capacitÃ di cache KV. Il risultato Ã¨ unâ??architettura di inferenza piÃ¹ scalabile e resiliente che offre esperienze di IA piÃ¹ veloci e reattive.

â??Mentre i carichi di lavoro dellâ??IA si espandono dalla formazione allâ??inferenza, le imprese e i neocloud che implementano carichi di lavoro di inferenza ad alto rendimento, caratteristici dellâ??IA agentica e di quella generativa a piÃ¹ turni a lungo termine, devono rimuovere i colli di bottiglia per migliorare lâ??efficienza della GPUâ?•, ha affermato Charles Liang, presidente e CEO di Supermicro. â??La nostra soluzione con IBM e NVIDIA dimostra chiaramente questi vantaggi in termini di prestazioniâ?•.

Per un video-riassunto di questa pubblicazione, clicca qui.

I carichi di lavoro di inferenza sono costituiti da due fasi: precompilazione e decodifica. Durante la precompilazione, il testo in ingresso viene convertito in token ed elaborato dal modello linguistico di grandi dimensioni (Large Language Model, LLM) per generare gli stati chiave-valore (KV) intermedi necessari per lâ??inferenza. Conservando questi stati KV precedentemente calcolati in una cache KV persistente, il sistema puÃ² riutilizzarli invece di ricalcolare lo stesso contesto per le richieste successive. I flussi di lavoro agentici, gli assistenti di codifica, la generazione aumentata dal recupero delle informazioni (Retrieval-Augmented Generation, RAG) a lungo contesto, lâ??IA multimodale e le applicazioni a piÃ¹ turni possono generare uno stato KV sostanziale man mano che elaborano grandi basi di codice, documenti, conversazioni, immagini e altri contesti, superando rapidamente la memoria GPU e CPU e determinando la necessitÃ di scaricare la cache KV su una rete di archiviazione condivisa ad alta capacitÃ .

La soluzione di cache KV di Supermicro, IBM e NVIDIA fornisce questo livello di archiviazione persistente, contribuendo a migliorare lâ??efficienza di inferenza, la velocitÃ effettiva e la scalabilitÃ riducendo al contempo il calcolo ridondante.

I principali vantaggi per i clienti della soluzione di cache KV includono:

default watermark

â??Le imprese hanno bisogno di un modo fondamentalmente migliore per ridimensionare lâ??inferenza dellâ??IA. Insieme, IBM, Supermicro e NVIDIA stanno offrendo una soluzione di cache KV ad alte prestazioni e resiliente che aiuta i clienti a massimizzare i loro investimenti GPU, scalare indipendentemente dal calcolo e migliorare significativamente lâ??efficienza delle infrastrutture â??, ha dichiarato Sam Werner, GM, IBM Storage. â??Questa collaborazione rappresenta un passo importante per rendere lâ??inferenza dellâ??IA aziendale piÃ¹ scalabile, affidabile e convenienteâ?•.

â??Lâ??IA agentica e applicazioni a lungo contesto dipendono dal riutilizzo del contesto in modo efficiente in molti passaggi e NVIDIA Dynamo aiuta a orchestrare lâ??inferenza distribuita in modo che le imprese possano ottenere una velocitÃ effettiva piÃ¹ fruibile da ogni GPUâ?•, ha affermato Shruti Koparkar, NVIDIA AI Inference Product Lead. â??Accoppiando NVIDIA Dynamo e Spectrum-X Ethernet con server di archiviazione Supermicro Petascale e IBM Storage Scale, questa soluzione di cache KV crea un livello di archiviazione condivisa ad alte prestazioni in grado di ridurre il calcolo ridondante, migliorare il time-to-first-token e aiutare i clienti a scalare i servizi di inferenza dellâ??IA resilientiâ?•.

Per ulteriori informazioni sulle soluzioni Supermicro e IBM clicca qui e leggi lâ??IBM Redbook, Context without Limits: Una piattaforma KV Cache ad alte prestazioni per lâ??inferenza dellâ??IA su larga scala, clicca qui.

A proposito di Super Micro Computer, Inc.Supermicro (NASDAQ: SMCI) Ã¨ leader mondiale nelle soluzioni IT complete, ottimizzate per le applicazioni. Fondata e operativa a San Jose, in California, Supermicro si impegna a introdurre per prima sul mercato innovazioni destinate alle infrastrutture IT aziendali, cloud, IA e 5G Telco/Edge. Lâ??azienda offre soluzioni IT complete che comprendono server, sistemi per lâ??IA, archiviazione, IoT, sistemi switch, software e servizi di assistenza. Lâ??esperienza di Supermicro nella progettazione di schede madri, alimentatori e chassis favorisce ulteriormente lo sviluppo e la produzione di innovazioni di nuova generazione, dal cloud allâ??edge, per i clienti di tutto il mondo. I nostri prodotti sono progettati e realizzati internamente (negli Stati Uniti, a Taiwan e nei Paesi Bassi), sfruttando le operazioni globali per ottenere economie di scala ed efficienza, e sono ottimizzati per migliorare il costo totale di proprietÃ (TCO) e ridurre lâ??impatto ambientale

(Green Computing). Il pluripremiato portfolio di Server Building Block SolutionsÂ® permette ai clienti di ottimizzare le proprie soluzioni in base al carico di lavoro e allâ??applicazione specifica, scegliendo tra unâ??ampia gamma di sistemi basati sui nostri blocchi modulari flessibili e riutilizzabili, che supportano una gamma completa di fattori di forma, processori, memorie, GPU, archiviazione, networking, alimentazione e soluzioni di raffreddamento (aria condizionata, raffreddamento ad aria libera o a liquido).

Supermicro, Server Building Block Solutions e We Keep IT Green sono marchi commerciali e/o marchi registrati di Super Micro Computer, Inc.

Tutti gli altri marchi, nomi e loghi sono di proprietÃ dei rispettivi proprietari.

Risultati dettagliati sono disponibili in IBM Redbook:
https://www.redbooks.ibm.com/docs/MD260021/MD260021.html
1

default watermark

Copyright 2026 PR Newswire. All Rights Reserved.

COMUNICATO STAMPA â?? CONTENUTO PROMOZIONALE: Immediapress Ã¨ un servizio di diffusione di comunicati stampa in testo originale redatto direttamente dallâ??ente che lo emette. Lâ??Adnkronos e Immediapress non sono responsabili per i contenuti dei comunicati trasmessi

â??

immediapress/pr-newswire

Categoria

1. Comunicati

Tag

1. ImmediaPress

Data di creazione
Novembre 5, 2026
Autore
redazione