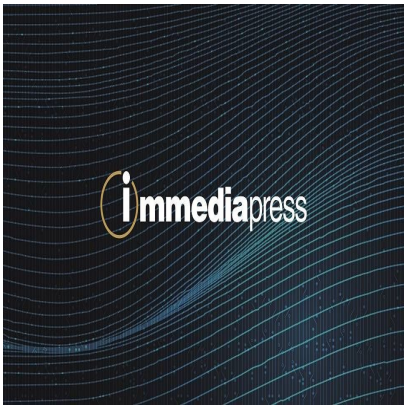

Huawei presenta OceanStor M900 Context Memory Storage per accelerare lâ??inferenza IA nei data center hyperscale

Descrizione

COMUNICATO STAMPA â?? CONTENUTO PROMOZIONALE

SHANGHAI, 18 settembre 2026 /PRNewswire/ â?? In occasione dello HUAWEI CONNECT 2026, David Wang, Vicepresidente del Consiglio di Amministrazione e Presidente a rotazione di Huawei, ha presentato ufficialmente OceanStor M900 Context Memory Storage durante il suo keynote. Progettato per lâ??inferenza IA nei data center hyperscale, il prodotto fornisce ai SuperPoD uno spazio di memoria completamente condiviso che offre capacitÃ su scala PB e prestazioni di livello TB/s. CiÃ² segna un cambiamento nellâ??infrastruttura IA da un modello basato sul calcolo a unâ??elevata collaborazione tra calcolo, rete e archiviazione. Questo contribuirÃ a liberare la potenza di calcolo dei SuperPoD.

Il 2026 ha visto lâ??accelerazione della transizione dellâ??IA dalle innovazioni tecnologiche allâ??implementazione su larga scala. Le applicazioni di IA si sono evolute dai chatbot agli agenti in grado di svolgere autonomamente compiti complessi. Questi agenti sono ampiamente adottati nei settori strategici, segnando lâ??inizio dellâ??era dellâ??IA agentica.

Man mano che i modelli di grandi dimensioni raggiungono i 10 mila miliardi di parametri, i SuperPoD stanno diventando la scelta ottimale per le infrastrutture di IA. I modelli di grandi dimensioni tradizionali supportano giÃ finestre di contesto superiori a un milione di token, lâ??inferenza multi-turn e le attivitÃ complesse sono diventate la norma e i dati della cache KV generati durante lâ??inferenza continuano a crescere. Queste tendenze hanno spinto la memoria on-chip e la DRAM oltre i loro limiti in termini di capacitÃ e redditivitÃ . Ã? ormai obiettivo comune nel settore costruire un sistema di archiviazione a piÃ¹ livelli che coordini la memoria on-chip, la DRAM e gli SSD per creare uno spazio di memoria completamente condiviso con capacitÃ estesa.

Huawei ha presentato OceanStor M900 Context Memory Storage per superare i rallentamenti nella capacitÃ di memoria nel contesto ultra-lungo e nellâ??inferenza multi-turn. OceanStor M900 utilizza la rete UnifiedBus per creare una cache KV multilivello globale su scala PB con connessioni one-hop. CiÃ² libera completamente il potenziale di potenza di calcolo dei SuperPoD e accelera lâ??inferenza IA

nei data center hyperscale. OceanStor M900 Context Memory Storage ha tre funzionalitÃ chiave:

Superare i limiti di capacitÃ per potenziare lâ??IA su larga scala con memoria estesa

Alimentata dalla rete di interconnessione UnifiedBus ad alta velocitÃ , la cache KV realizza il pooling globale e la condivisione con archiviazione a piÃ¹ livelli. La cache KV dei SuperPoD viene estesa dalla memoria on-chip e dalla DRAM agli SSD, consentendo a un singolo cluster di fornire 64 PB di capacitÃ . La capacitÃ di cache KV disponibile per NPU viene aggiornata da gigabyte a terabyte, consentendo di memorizzare, condividere e riutilizzare piÃ¹ contesto. CiÃ² aumenta significativamente il rapporto di hit della cache KV.

Potenziare le prestazioni di inferenza per liberare completamente la potenza di calcolo

OceanStor M900 Ã¨ la prima architettura del settore ad integrare la CPU, lâ??unitÃ di controllo di rete e lâ??unitÃ di controllo NAND. Fornisce semantica KV nativa per consentire una connessione one-hop dallâ??NPU del SuperPoD agli SSD. CiÃ² elimina la necessitÃ di conversione del protocollo e di inoltro della CPU, riducendo drasticamente la latenza di accesso da millisecondi a 60 microsecondi, una riduzione del 90%. Un singolo cluster fornisce 40 TB/s di larghezza di banda di accesso aggregata, 1,5 volte superiore rispetto alle soluzioni peer. In tipici scenari di programmazione IA, questa architettura raddoppia il throughput dei token del cluster di inferenza e dimezza il tempo necessario al primo token, convertendo la potenza di calcolo in produttivitÃ .

Riduzione dei costi dei token per consentire unâ??adozione su larga scala ed economicamente efficiente dellâ??IA

OceanStor M900 utilizza la prima tecnologia di archiviazione adattativa KV-aware del settore, che prevede i cicli di vita della cache KV in base al valore dei dati e distribuisce in modo intelligente i dati tra i supporti di archiviazione. Questa tecnologia consente fino a 24 scritture su disco al giorno (DWPD), estendendo la resistenza degli SSD di 16 volte e garantendo la stabilitÃ per tre anni. Riducendo i costi di sostituzione dei media e di manutenzione e gestione, riduce i costi a lungo termine delle infrastrutture di inferenza IA su larga scala e consente unâ??adozione piÃ¹ rapida dellâ??IA.

Man mano che lâ??IA si espande nei principali sistemi di produzione in tutti i tipi di industrie, lâ??infrastruttura dellâ??IA si sta evolvendo da un modello basato sul calcolo verso una collaborazione piÃ¹ stretta tra calcolo, rete e archiviazione. Lâ??archiviazione della memoria di contesto sarÃ essenziale per migliorare continuamente la capacitÃ e lâ??efficienza di accesso delle cache KV di inferenza hyperscale.

View original content:https://www.prnewswire.com/news-releases/huawei-presenta-oceanstor-m900-context-memory-storage-per-accelerare-linferenza-ia-nei-data-center-hyperscale-302883305.html

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
Settembre 18, 2026
Autore
redazione