ADVANTECH
Enabling an Intelligent Planet

Comunicato stampa
Per pubblicazione immediata

Advantech lancia le soluzioni di server Edge AI per l'IA generativa

Germering (Germania), giugno 2024 - Advantech, leader globale nell'Industrial IoT, è entusiasta di annunciare un'innovativa soluzione server Edge AI per l'IA generativa, dotata della tecnologia brevettata aiDAPTIV+ di Phison. Il server AIR-520 Edge AI, alimentato da un processore AMD EPYC serie 7003, integra le unità SSD SQ ai100 AI, le schede GPU NVIDIA RTX, un SDK Edge AI e un'unità NVIDIA AI Enterprise NVIDIA AI Enterprise per fornire una soluzione pronta all'uso.

Gli strumenti di intelligenza artificiale generativa, come ad esempio i modelli linguistici di grandi dimensioni (LLM), stanno trasformando la gestione della conoscenza aziendale automatizzando l'organizzazione, il reperimento e l'analisi dei dati, aumentando così la produttività e migliorando il processo decisionale. Gli LLM personalizzati migliorano la precisione, mentre l'addestramento edge aumenta la privacy dei dati, anche se può essere più costoso. Questa soluzione supporta la messa a punto dell'LLM con 1-4 schede GPU e SSD SQ ai100 AI, consentendo alle aziende di addestrare gli LLM in modo economico e di mantenere la sicurezza dei dati sensibili a livello edge.

Quattro offerte di soluzioni per diverse applicazioni
Advantech offre quattro opzioni: AIR-520-L13B/L33B/L70B, e L70B-Plus, adatti a diverse scale e applicazioni. L13B è ideale per applicazioni in tempo reale come i chatbot e la traduzione linguistica. L33B è adatto a compiti più complessi, migliorando la produttività e l'innovazione nella creazione di contenuti. L70B eccelle nell'analisi sofisticata dei dati e nel processo decisionale per settori specializzati. Inoltre, L70B-Plus, equipaggiato con la soluzione Piattaforma software NVIDIA AI Enterprise offre SDK AI end-to-end, affidabili e ottimizzati, con supporto a lungo termine e servizi di consulenza di esperti, per garantire un'implementazione efficiente delle applicazioni aziendali.

Installazione facile e veloce Ottimizzata per l'efficienza dei costi
Tutte le soluzioni includono le unità SSD SQ ai100 AI che sfruttano la tecnologia aiDAPTIV+ di Phison. Queste unità SSD agiscono come un'estensione della vRAM della GPU, consentendo al sistema di mettere a punto gli LLM con un minimo di schede GPU. Questo approccio non solo riduce la barriera del budget, ma rende anche il server Edge AI più compatto rispetto ai tradizionali server di grandi dimensioni montati su rack. Il server AIR-520Edge AI è stato progettato per essere utilizzato in una vasta gamma di applicazioni di intelligenza artificiale. Le sue dimensioni sono paragonabili a quelle di un PC desktop e può essere montato su rack con gli accessori appropriati. Il basso profilo consente una facile implementazione di un ambiente edge AI fine-tuning, eliminando le preoccupazioni relative allo spazio e alla manutenzione.

Sviluppo GenAI in tempi rapidi con i servizi software
Oltre alle capacità di messa a punto di LLM, Advantech fornisce un SDK Edge AI con GenAI Training Studio, precaricato con modelli Llama-2 13B/33B/70B per applicazioni come chatbot e analisi dei dati. Questo semplifica e accelera l'addestramento del modello LLM specifico per il cliente e la valutazione dell'inferenza su AIR-520. Inoltre, DeviceOn di Advantech fornisce aggiornamenti software/container OTA e gestione remota, facilitando un'efficiente orchestrazione edge IA e una manutenzione a lungo termine.