Graid Technology lanza soluciones de IA agéntica para eliminar cuellos de botella de caché KV

Comunicados | 22/04/2026 | 00:00

Desde inferencia en el edge hasta NVIDIA STX, infraestructura de caché KV diseñada específicamente para un rendimiento consistente a escala

Desde inferencia en el edge hasta NVIDIA STX, infraestructura de caché KV diseñada específicamente para un rendimiento consistente a escala

Graid Technology, pionera en almacenamiento NVMe acelerado por GPU, anunció hoy su Agentic AI Storage Portfolio: una familia de soluciones de caché KV diseñadas específicamente para eliminar el cuello de botella de almacenamiento que frena la IA de producción "always-on" (siempre activa). El porfolio abarca tres niveles de despliegue: KV Cache Server, KV Cache Rack y KV Cache Platform, todos basados en la tecnología SupremeRAID™. KV Cache Platform, el nivel más alto del porfolio, está alineado específicamente con la arquitectura de referencia STX de NVIDIA, con ejecución nativa en DPU BlueField-4 en la hoja de ruta para la segunda mitad de 2026.

A medida que la IA agéntica pasa de la experimentación a la producción, los supuestos de infraestructura que sustentaban la inferencia de una sola pasada han dejado de ser válidos. Los modelos que ejecutan tareas continuas de múltiples pasos y mantienen contexto durante horas de operación generan demandas de caché KV que sobrepasan la memoria HBM de las GPU. El resultado: picos de latencia de hasta 18x, utilización de GPU tan baja como el 50 % y fallos a nivel de modelo, incluyendo alucinaciones y degradación del razonamiento, que son difíciles de detectar y costosos de recuperar.

SupremeRAID™ aborda esto directamente, agregando hasta 32 unidades NVMe en un único pool virtual de 280 GB/s, evitando la CPU mediante GPU Direct Storage y ofreciendo lecturas de caché KV en 1,3 ms — 77 veces más rápido que NVMe estándar. Los tres niveles del porfolio llevan esta capacidad a cada escala de despliegue:

KV Cache Server – aceleración NVMe de nodo único para servidores de inferencia individuales y despliegues de IA en el edge. Disponible ahora.

KV Cache Rack – soluciones a escala de rack validadas por socios y co-diseñadas con

fabricantes líderes de servidores para clústeres empresariales multi-GPU. Disponible ahora.

KV Cache Platform – diseñada específicamente para la arquitectura de referencia STX de NVIDIA, con ejecución nativa en DPU BlueField-4 y expansión de almacenamiento a escala de rack en la hoja de ruta.

"Hace un año, en GTC 2025, Jensen Huang predijo que el almacenamiento se convertiría por primera vez en acelerado por GPU. Este año, NVIDIA convirtió ese concepto en una arquitectura con STX y CMX", dijo Leander Yu, CEO de Graid Technology. "Nuestro porfolio de caché KV está construido precisamente para este momento, ofreciendo el rendimiento de almacenamiento que exige la IA agéntica, con una economía propia de la capa de almacenamiento".

Para empresas y equipos de infraestructura que evalúan despliegues de IA agéntica, la arquitectura completa de despliegue, las especificaciones técnicas y los detalles de compatibilidad con NVIDIA STX están disponibles en el documento de solución: Cartera de productos de almacenamiento con IA de Graid Technology Agentic: soluciones de caché KV diseñadas específicamente para la inferencia a gran escala.

Sobre Graid Technology

Graid Technology está construyendo la base de almacenamiento para el futuro de la IA, la empresa y la computación de alto rendimiento. Como creadora de SupremeRAID™, el primer y único RAID basado en GPU del mundo, y administrador global de Intel® Virtual RAID on CPU (Intel® VROC), Graid Technology ofrece soluciones RAID flexibles que maximizan el rendimiento NVMe al tiempo que garantizan una protección de datos resiliente y escalable para la infraestructura de datos moderna. Con sede en Silicon Valley y operaciones globales e I+D en Taiwán, Graid Technology está impulsando la innovación en RAID para la próxima generación de cargas de trabajo intensivas en datos. Para más información, visitar graidtech.com.

Para obtener más información sobre las soluciones de inteligencia artificial de Graid Technology, visitar graidtech.com/ai.

Autor: Redacció

https://gironanoticies.com/comunicado/294024-graid-technology-lanza-soluciones-de-ia-agentica-para-eliminar-cuellos-de-botella-de-cache-kv.htm