¿Cómo optimiza IBM las GPU para mejorar la eficiencia y el rendimiento de las cargas de trabajo de inferencia de LLM de Gen AI?

"Permitir que Turbonomic amplíe y reduzca verticalmente nuestros servidores de inferencia LLM me ha permitido dedicar menos tiempo a monitorear el rendimiento.

Tom Morris
, jefe de infraestructura y operaciones de la plataforma de IA
Investigación, IBM

Leer la historia completa

El equipo de IBM® Big AI Models (BAM), que respalda el entorno principal de investigación y desarrollo para que los equipos de ingeniería prueben y perfeccionen sus proyectos de IA gen, vio una oportunidad de mejora. A medida que más proyectos pasaban por la etapa de prueba, el equipo reconoció la importancia de utilizar de manera óptima cada instancia para evitar el desperdicio de recursos.

Para optimizar sus recursos de GPU y gestionar sus instancias de LLM Kubernetes, el equipo de IBM BAM implementó IBM Turbonomic®, un software avanzado de gestión de recursos de aplicaciones.

Como software de IBM diseñado explícitamente para optimizar la nube híbrida, incluidas las aplicaciones en contenedores, las máquinas virtuales y las nubes públicas, IBM Turbonomic proporcionó una integración perfecta dentro de la infraestructura existente.

Dotación dinámica de recursos de GPU en Kubernetes con IBM Turbonomic

- Aumento de 5,3 veces en los recursos de GPU inactivos, lo que los hace disponibles para la nueva demanda
- 2 veces el rendimiento de la GPU sin degradar el rendimiento de la latencia
- 960+ acciones de recursos automatizadas tomadas en las cargas de trabajo de inferencia de IA

Componentes de la solución:

IBM Turbonomic®