Publicado en Ciudad de México el 23/09/2021

NVIDIA amplía su liderazgo en Rendimiento de Inferencias de IA, con resultados de MLPerf

NVIDIA Amplía su Liderazgo en Rendimiento de Inferencias de IA, con Resultados de Debut en Servidores Basados en Arm. Los últimos puntos de referencia de MLPerf muestran que NVIDIA ha extendido sus altas marcas de agua en rendimiento y eficiencia energética para la inferencia de inteligencia artificial para Arm, así como para computadoras x86.

Es la tercera vez consecutiva que NVIDIA establece récords en rendimiento y eficiencia energética en las pruebas de inferencia de MLPerf, un grupo de evaluación comparativa de la industria formado en mayo de 2018.

Y es la primera vez que las pruebas de la categoría del data center se ejecutan en un sistema basado en Arm, lo que brinda a los usuarios más opciones para implementar la IA, la tecnología más transformadora de nuestro tiempo.

Cuento de la Cinta
Los sistemas impulsados por la plataforma de IA de NVIDIA superaron las siete pruebas de rendimiento de inferencia en la última ronda con sistemas de NVIDIA y nueve de sus socios del ecosistema, incluidos Alibaba, Dell Technologies, Fujitsu, GIGABYTE, Hewlett Packard Enterprise, Inspur, Lenovo, Nettrix y Supermicro.

La inferencia es lo que sucede cuando una computadora ejecuta un software de inteligencia artificial para reconocer un objeto o hacer una predicción. Es un proceso que utiliza un modelo de deep learning para filtrar datos y encontrar resultados que ningún ser humano podría capturar.

Los puntos de referencia de inferencia de MLPerf se basan en las cargas de trabajo y los escenarios de IA más populares de la actualidad, que abarcan la visión por computadora, las imágenes médicas, el procesamiento del lenguaje natural, los sistemas de recomendación, el aprendizaje por refuerzo y más. Por lo tanto, independientemente de las aplicaciones de inteligencia artificial que implementen, los usuarios pueden establecer sus propios registros con NVIDIA.

¿Por qué es importante el Rendimiento?
Los modelos y conjuntos de datos de IA continúan creciendo a medida que los casos de uso de IA se expanden desde el data center hasta el edge y más allá. Es por eso que los usuarios necesitan un rendimiento que sea confiable y flexible de implementar.

MLPerf brinda a los usuarios la confianza para tomar decisiones de compra informadas. Cuenta con el respaldo de docenas de líderes de la industria, incluidos: Alibaba, Arm, Baidu, Google, Intel y NVIDIA,

por lo que las pruebas son transparentes y objetivas.

Flexionando Arm para la IA Empresarial
La arquitectura Arm está comenzando a abrirse paso en los data centers de todo el mundo, en parte gracias a su eficiencia energética, incrementos de desempeño y expandiendo el ecosistema de software.

De hecho, el servidor basado en Arm que se usó, fue incluso más rápido que un sistema x86 similar en una de las pruebas.

NVIDIA tiene una larga tradición de compatibilidad con todas las arquitecturas de CPU, por lo que están orgullosos de haber ayudado a Arm a demostrar su destreza en inteligencia artificial en un punto de referencia de la industria revisado por sus pares.

“Arm, como miembro fundador de MLCommons, está comprometido con el proceso de creación de estándares y puntos de referencia para abordar mejor los desafíos e inspirar la innovación en la industria de la computación acelerada”, dijo David Lecomber, Director Senior de HPC y Herramientas en Arm.

“Los últimos resultados de inferencia demuestran la preparación de los sistemas basados en Arm impulsados por CPUs basadas en Arm y GPUs de NVIDIA para abordar una amplia gama de cargas de trabajo de IA en el centro de datos”, agregó.

Los Socios Muestran sus Poderes de IA
La tecnología de IA de NVIDIA está respaldada por un ecosistema grande y en continuo crecimiento.

Siete fabricantes de equipos originales presentaron un total de 22 plataformas aceleradas por GPUs en los últimos puntos de referencia. La mayoría de estos modelos de servidor están certificados por NVIDIA y están validados para ejecutar una amplia gama de cargas de trabajo aceleradas. Y muchos de ellos son compatibles con NVIDIA AI Enterprise, software lanzado oficialmente el mes pasado.

Sus socios que participaron en esta ronda fueron Dell Technologies, Fujitsu, Hewlett Packard Enterprise, Inspur, Lenovo, Nettrix y Supermicro, así como el proveedor de servicios en el cloud Alibaba.

El Poder del Software
Un ingrediente clave del éxito de la IA de NVIDIA en todos los casos de éxito es la batería de software completa de NVIDIA.

Para la inferencia, eso incluye modelos de IA previamente entrenados para una amplia variedad de

casos de éxito. El Kit de Herramientas NVIDIA TAO personaliza esos modelos para aplicaciones específicas mediante el aprendizaje por transferencia.

El software NVIDIA TensorRT optimiza los modelos de IA para que hagan un mejor uso de memoria y se ejecuten más rápido. Se habitualmente para las pruebas de MLPerf y está disponible tanto para sistemas x86 como para los basados ??en Arm.

También se empleó el software para el Servidor de Inferencia NVIDIA Triton y la capacidad de GPU de Instancias Múltiples (MIG) en estos puntos de referencia. Ofrecen a todos los desarrolladores el tipo de rendimiento que normalmente requiere de codificadores expertos.

Gracias a las mejoras continuas en esta bateríaa de software, NVIDIA logró ganancias de hasta un 20 por ciento en rendimiento y un 15 por ciento en eficiencia energética con respecto a los anteriores puntos de referencia de inferencia de MLPerf hace solo cuatro meses.

Todo el software que se usó en las últimas pruebas está disponible en el repositorio MLPerf, por lo que cualquiera puede reproducir los resultados de referencia. Continuamente se agrega este código a los frameworks y contenedores de deep learning disponibles en NGC, el centro de software para aplicaciones de GPUs.

Es parte de una oferta de IA de batería completa, compatible con todas las arquitecturas de un procesador, probada en los últimos puntos de referencia de la industria y disponible para abordar trabajos reales de IA en la actualidad.

“Es muy gratificante ver cómo la División Enterprise de NVIDIA ha emergido como líder en soluciones de inteligencia artificial en todo el mundo. NVIDIA seguirá innovando en otras plataformas que aportan a diferentes sectores”, enfatizó Marcio Aguiar, Director de la División Enterprise de NVIDIA para Latinoamérica.

Para obtener más información sobre la plataforma de inferencia de NVIDIA, consultar su descripción general de la tecnología de inferencia de NVIDIA.

Por Dave Salvator

Datos de contacto:
Carlos Valencia
MKQ PR Agency
55 39 64 96 00