Publicado en Ciudad de México el 04/05/2021

# [NVIDIA establece récords de inferencia para IA. Presenta las GPUs A30 y A10 para servidores empresariales](http://www.notasdeprensa.es)

## En el anuncio de los Resultados de la Inferencia en Inteligencia Artificial de MLPerf, un esquema de puntos de referencia (benchmarks) para medir qué tan rápido los sistemas pueden ejecutar modelos en una variedad de escenarios de implementación, la Plataforma de IA de NVIDIA superó todas las categorías de MLPerf, desde el Data Center hasta el Edge

 El día de hoy, NVIDIA anunció que su plataforma de inferencia de IA, recientemente ampliada con las GPU NVIDIA® A30 y A10 para servidores convencionales, ha logrado un rendimiento récord en todas las categorías en la última versión de MLPerf. MLPerf es la evaluación establecida de la industria para medir el rendimiento de la IA en diversas cargas de trabajo, que abarcan la visión de computación, las imágenes médicas, los sistemas de recomendación, el reconocimiento de voz y el procesamiento del idiomas naturales (NPL) en varios idiomas. En su debut en MLPerf, las GPUs NVIDIA A30 y A10 combinan alto rendimiento con bajo consumo de energía para brindar a las empresas opciones convencionales para una amplia gama de cargas de trabajo de inferencia de IA, entrenamiento, gráficos y computación empresariales. Se espera que Hewlett Packard Enterprise, Inspur y Lenovo integren las GPUs en sus servidores de mayor volumen a partir de este verano. NVIDIA logró estos resultados al aprovechar toda la capacidad de su plataforma de IA (que abarca una amplia variedad de GPUs y software de IA, incluidos TensorRT™ y NVIDIA Triton™ Inference Server), la cual se implementa en empresas líderes, como Microsoft, Pinterest, Postmates, USPS y WeChat. “A medida que la IA continúa transformando todas las industrias, MLPerf se está convirtiendo en una herramienta aún más importante para que las empresas tomen decisiones informadas sobre sus inversiones en infraestructura de TI”, dijo Ian Buck, Gerente General y Vicepresidente de Computación Acelerada de NVIDIA. “Ahora que los principales fabricantes de equipos originales envían resultados de MLPerf, NVIDIA y nuestros socios se están enfocando no solo en ofrecer un rendimiento líder en el mundo para la IA, sino también en democratizar la IA con una próxima ola de servidores empresariales impulsados por nuestras nuevas GPUs A30 y A10”. Resultados de MLPerfNVIDIA es la única empresa que envía los resultados de cada prueba para las categorías de Data Center y Edge, lo que ofrece los mejores resultados de rendimiento en todas las cargas de trabajo de MLPerf. Varias presentaciones también usan Triton Inference Server, que simplifica la complejidad de implementar IA en aplicaciones, dado que admite modelos de los principales frameworks, se ejecuta en GPU, así como en CPU, y se optimiza para diferentes tipos de consultas, incluidos en lotes, en tiempo real y de streaming. Las presentaciones de Triton lograron un rendimiento cercano al de las implementaciones de GPU más optimizadas, así como las implementaciones de CPU, con configuraciones comparables. NVIDIA también abrió nuevos caminos con sus presentaciones utilizando la capacidad de Múltiples instancias de GPU de la arquitectura NVIDIA Ampere al ejecutar en simultáneo las siete pruebas MLPerf Offline en una sola GPU usando siete instancias MIG. La configuración mostró un rendimiento casi idéntico en comparación con una sola instancia de MIG. Estas presentaciones demuestran el rendimiento y la versatilidad de MIG, lo que permite a los administradores de infraestructuras proporcionar cantidades adecuadas de computación de GPU para aplicaciones específicas, con el objetivo de obtener el máximo rendimiento de cada GPU del data center. Además de las presentaciones de NVIDIA, los socios de NVIDIA Alibaba, DellEMC, Fujitsu, Gigabyte, HPE, Inspur, Lenovo y Supermicro presentaron más de 360 resultados utilizando las GPUs de NVIDIA. Plataforma de IA en Expansión de NVIDIALas GPUs NVIDIA A30 y A10 son las adiciones más recientes a la plataforma de IA de NVIDIA, que incluye las GPUs con la arquitectura NVIDIA Ampere, NVIDIA Jetson™ AGX Xavier y Jetson Xavier NX, además de una batería completa de software de NVIDIA optimizado para acelerar la IA. La A30 ofrece un rendimiento versátil para servidores estándar de la industria, ya que admite una amplia gama de cargas de trabajo convencionales de computación empresarial y de inferencia de IA, como sistemas de recomendación, IA conversacional y visión de computación. La GPU NVIDIA A10 acelera la inferencia de del deep learning, la representación interactiva, el diseño asistido por computadora y los juegos en cloud, lo que permite a las empresas admitir cargas de trabajo mixtas de IA y gráficos en una infraestructura común. Usando el software de GPU virtuales de NVIDIA, se puede optimizar la gestión para mejorar la utilización y el aprovisionamiento de los desktops virtuales utilizados por diseñadores, ingenieros, artistas y científicos. “La plataforma NVIDIA Jetson, basada en el sistema en módulos NVIDIA Xavier™, proporciona un rendimiento de IA de clase de servidor en el edge, lo que permite una amplia variedad de aplicaciones en sectores como la robótica, la atención médica, el comercio minorista y las ciudades inteligentes”, dijo Marcio Gomes de Aguiar, Director de NVIDIA Enterprise para América Latina. Jetson se desarrolló con la arquitectura unificada de NVIDIA y la pila de software CUDA-X®, por lo que es la única plataforma capaz de ejecutar todas las cargas de trabajo de edge en diseños compactos mientras consume menos de 30 W de energía. DisponibilidadLas GPUs NVIDIA A100 están disponibles en los servidores de los principales fabricantes y en el cloud de los principales proveedores de servicios de cloud. Además, las GPUs A100 se ofrecen en toda la gama de sistemas NVIDIA DGX™, que incluye NVIDIA DGX Station A100, NVIDIA DGX A100 y NVIDIA DGX SuperPOD. Se espera que la A30 y la A10 estén disponibles en una amplia gama de servidores a partir de este verano, incluidos los sistemas certificados por NVIDIA™, que superan rigurosas pruebas para garantizar un alto rendimiento en diversas cargas de trabajo. Los sistema en módulos Jetson AGX Xavier y Jetson Xavier NX están disponibles mediante distribuidores de todo el mundo. NVIDIA Triton y NVIDIA TensorRT están disponibles en NGC, el catálogo de software de NVIDIA.

**Datos de contacto:**

Carlos Valencia

MKQ PR Agency

55 39 64 96 00

Nota de prensa publicada en: [https://www.mexicopress.com.mx/nvidia-establece-records-de-inferencia-para-ia](http://www.notasdeprensa.es/educalivecom-disfruta-de-cursos-presenciales-y-clases-particulares-a-traves-de-internet)

Categorías: Nacional Inteligencia Artificial y Robótica Programación Hardware Software



[**http://www.mexicopress.com.mx**](http://www.notasdeprensa.es)