Publicado en Ciudad de México, México el 15/05/2020

# [La nueva GPU de NVIDIA Ampere Data Center en plena producción](http://www.notasdeprensa.es)

## La nueva NVIDIA A100 GPU potencia hasta 20 veces la inferencia y entrenamiento de Inteligencia Artificial (IA). La primera GPU elástica y de múltiples instancias de NVIDIA unifica la capacitación, la inferencia y el análisis. Los principales fabricantes de servidores y proveedores de nube del mundo ya la adoptaron

 El día de hoy, NVIDIA Enterprise anunció que la primera GPU basada en la arquitectura de Ampere de NVIDIA®: NVIDIA A100, se está fabricando y enviando a los clientes de todo el mundo, después de recibir la mayor cantidad de pedidos previos a un lanzamiento en toda la historia de la compañía. La A100 se basa en los avances de diseño de la arquitectura NVIDIA Ampere, que ofrece el mayor salto de rendimiento de la compañía hasta la fecha dentro de sus ocho generaciones de GPUs, para unificar el entrenamiento y la inferencia de IA y aumentar el rendimiento hasta 20 veces más que los modelos precedentes. La A100 es un acelerador de carga de trabajo universal, que también está diseñado para el análisis de datos, la computación científica y los gráficos en la nube. "Las poderosas tendencias de la computación en la nube y la IA están impulsando un cambio tectónico en los diseños de los centros de datos, de modo que lo que una vez fué un mar de servidores solo de CPUs, ahora es la computación acelerada por el GPU", dijo Jensen Huang, fundador y CEO de NVIDIA. "La GPU NVIDIA A100 es un salto cuántico en el rendimiento de IA de 20 veces y un acelerador del aprendizaje automático de principio a fín, desde el análisis de datos hasta la capacitación y la inferencia. Por primera vez, las cargas de trabajo de ampliación y escalabilidad pueden acelerarse en una sola plataforma. NVIDIA A100 aumentará simultáneamente el rendimiento y reducirá el costo de los centros de datos". Las nuevas tecnologías elásticas de cómputo integradas en A100 permiten aportar la cantidad adecuada de potencia de procesamiento a cada trabajo. La capacidad de GPU de múltiples instancias permite que cada GPU A100 se divida en hasta siete instancias independientes de "mini-GPUs" para realizar tareas de inferencia más pequeñas, mientras que la tecnología de interconexión NVIDIA NVLink® interconnect technology de tercera generación, permite que múltiples GPUs A100 funcionen como una GPU gigante, para cumplir con las tareas de entrenamiento cada vez más grandes. Adopción inmediata en todo el mundoEntre las primeras empresas en aprovechar la potencia de las GPUs NVIDIA A100 está Microsoft, que aprovechará el rendimiento y la escalabilidad de A100. "Microsoft entrenó Turing Natural Language Generation, el modelo de idiomas más grande del mundo, para usar la generación actual de GPUs de NVIDIA" dijo Mikhail Parakhin, corporate vice president, Microsoft Corp. “Azure planea entrenar modelos de IA mucho más grandes utilizando miles de GPUs NVIDIA A100 a escala para impulsar lo último en lenguaje, habla, visión y multimodalidad". Otros de los primeros adoptadores incluyen laboratorios nacionales y algunas de las instituciones de educación superior e investigación más importantes del mundo, cada una de las cuales usa A100 para potenciar sus supercomputadoras de próxima generación. Cinco avances de A100La GPU NVIDIA A100 marca una revolución en el diseño técnico gracias a cinco innovaciones clave: ● Arquitectura Ampere: En el corazón de A100 se encuentra la arquitectura de GPU NVIDIA Ampere, que contiene más de 54,000 millones de transistores, lo que la hace el procesador de 7 nanómetros más grande del mundo. ● Third-generation Tensor Cores with TF32: Los núcleos Tensor de amplia adopción de NVIDIA ahora son más flexibles, rápidos y fáciles de usar. Sus capacidades expandidas incluyen nuevos TF32 for AI,, que permiten un rendimiento de IA de hasta 20 veces mayor, con precisión FP32, sin cambiar el código. Además, los núcleos Tensor ahora permiten FP64, para lograr hasta 2.5 veces más de procesamiento de datos que la generación anterior, para las aplicaciones de HPC. ● GPU de múltiples instancias: MIG, una nueva función técnica, que permite que una sola GPU A100 se particione en hasta siete GPUs por separado, para poder proporcionar diferentes grados de procesamiento para trabajos de diferentes tamaños, lo que permite una utilización óptima y el máximo retorno de la inversión. ● Third-generation NVIDIA NVLink ™: Duplica la conectividad de alta velocidad entre las GPUs para proporcionar escalabilidad de rendimiento eficiente en un servidor. ● Baja densidad estructural: Esta nueva técnica de eficacia aprovecha la naturaleza ineherentemente escasa de los cálculos de IA para duplicar el rendimiento. Juntas, estas nuevas características hacen que la NVIDIA A100 sea ideal para cargas de trabajo diversas y exigentes, que incluyen entrenamiento e inferencia de IA, así como simulación científica, IA conversacional, sistemas de recomendación, genómica, análisis de datos de alto rendimiento, modelos sísmicos y pronósticos financieros. El modelo NVIDIA A100 está disponible en nuevos sistemas, y pronto llegará a la nubeEl NVIDIA DGX™ A100 system, que también se presentó hoy, cuenta con ocho GPU NVIDIA A100 interconectadas con NVIDIA NVLink. Ya está disponible mediante NVIDIA y socios aprobados. Alibaba Cloud, AWS, Baidu Cloud, Google Cloud, Oracle and Tencent Cloud están planeando ofrecer los servicios basados en la A100. Además, se lanzará una amplia variedad de servidores con A100 mediante los principales fabricantes de sistemas del mundo, como: , including Atos, Dell Technologies, Fujitsu, GIGABYTE, H3C, HPE, Inspur, Lenovo, Quanta/QCT and Supermicro. Para ayudar a acelerar el desarrollo de servidores de sus socios, NVIDIA ha creado HGX A100, un componente fundamental de servidores en la forma de bloques de construcción, en la forma de zócalos integrados en múltiples configuraciones de GPUs. La HGX A100 de cuatro GPUs ofrece una interconexión completa entre las GPUs con NVLink, mientras que la configuración de ocho GPUs ofrece un ancho de banda completo de GPU a GPU, a través de NVIDIA NVSwitch™. HGX A100, con la nueva tecnología MIG y su arquitectura de GPU de múltiples instancias, se puede configurar como 56 GPUs pequeñas, cada una más rápida que NVIDIA T4, hasta un servidor gigante de ocho GPUs con 10 petaflops de rendimiento de inteligencia artificial. Acerca de NVIDIALa invención de la GPU por parte de NVIDIA en 1999 generó el crecimiento del mercado de juegos para PC, redefinió los gráficos por computación modernos y revolucionó la computación en paralelo. Recientemente, el deep learning de GPU impulsó la inteligencia artificial moderna, la próxima era de la computación, ya que la GPU ocupa el papel del cerebro de las computadoras, los robots y los automóviles de conducción autónoma que pueden percibir y comprender al mundo. Encontrar más información en http://nvidianews.nvidia.com/. Contacto para medios:Carlos Valencia.MKQ.55 39 64 96 00.cvalencia@mkq.mx

**Datos de contacto:**

Carlos Valencia Escárcega

MKQ Agencia de Relaciones Públicas

52 (55) 39 64 96 00

Nota de prensa publicada en: [https://www.mexicopress.com.mx/la-nueva-gpu-de-nvidia-ampere-data-center-en](http://www.notasdeprensa.es/educalivecom-disfruta-de-cursos-presenciales-y-clases-particulares-a-traves-de-internet)

Categorías: Finanzas Inteligencia Artificial y Robótica Programación Hardware Software Ciudad de México Otras Industrias



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