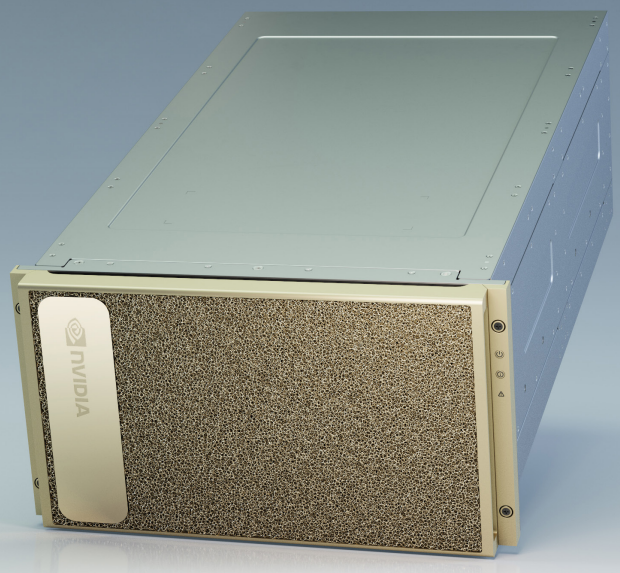
NVIDIA DGX A100
EL SISTEMA UNIVERSAL PARA LA INFRAESTRUCTURA DE IA



El desafío de escalar la IA empresarial

Toda empresa necesita transformarse con la inteligencia artificial (IA), no solo para sobrevivir, sino para prosperar en tiempos difíciles. Sin embargo, la empresa requiere una plataforma para la infraestructura de IA que mejore los enfoques tradicionales, que históricamente involucraban arquitecturas de procesamiento lentas que se aislaron mediante las cargas de trabajo de análisis, capacitación e inferencia. El antiguo enfoque generaba complejidad, aumentaba los costos, limitaba la velocidad de escala y no estaba listo para la IA moderna. Las empresas, desarrolladores, científicos de datos e investigadores necesitan una nueva plataforma que unifique todas las cargas de trabajo de IA, para simplificar la infraestructura y acelerar el retorno de la inversión.

El sistema universal para cada carga de trabajo de IA

NVIDIA DGX™ A100 es el sistema universal para cada carga de trabajo de IA, desde el análisis hasta el entrenamiento y la inferencia. DGX A100 establece un nuevo nivel para la densidad de procesamiento, ya que cuenta con 5 petaFLOPS de rendimiento de IA en un tamaño de 6U, lo que le permite reemplazar la infraestructura de procesamiento heredada con un sistema único y unificado. DGX A100 también ofrece la capacidad sin precedentes de asignar la potencia de procesamiento de forma detallada, para utilizar la capacidad de GPU de varias instancias en la GPU NVIDIA A100 con Tensor Cores, lo que permite a los administradores asignar recursos del tamaño adecuado para cargas de trabajo específicas. Esto garantiza poder llevar a cabo los trabajos más grandes y complejos, junto con los más simples y más pequeños. Al ejecutar el conjunto de software DGX con el software optimizado de NGC, la combinación de potencia de procesamiento denso y la flexibilidad completa de cargas de trabajo hacen que DGX A100 sea una opción ideal tanto para las implementaciones de un solo nodo como para los clústeres de Slurm y Kubernetes a gran escala implementados con NVIDIA DeepOps.

Acceso directo a los expertos de NVIDIA DGX

NVIDIA DGX A100 es más que un servidor, es una plataforma completa de hardware y software basada en el conocimiento adquirido del campo de pruebas de DGX más grande del mundo (NVIDIA DGX SATURNV) y el conocimiento de miles de expertos de DGX en NVIDIA. Los expertos de DGX son profesionales con fluidez en IA que ofrecen orientación prescriptiva y experiencia en diseño para ayudar a acelerar la transformación de IA. Han acumulado una gran cantidad de conocimiento y experiencia durante la última década para ayudar a maximizar el valor de tu inversión en DGX. Los expertos de DGX ayudan a garantizar que las aplicaciones críticas se pongan en funcionamiento rápidamente y se mantengan funcionando sin problemas, para acelerar enormemente la obtención de resultados.

ESPECIFICACIONES DEL SISTEMA

GPU	**8 GPU NVIDIA A100 con núcleos Tensor**
Memoria de GPU	**320 GB en total**
Rendimiento	**5 petaFLOPS de AI** **10 petaOPS de INT8**
NVIDIA NVSwitches	**6**
Uso de energía del sistema	**6,5 kW máx.**
CPU	**AMD Rome 7742 doble, 128 núcleos en total, 2,25 GHz (básico), 3,4 GHz (potencia máxima)**
Memoria del sistema	**1 TB**
Red	**8 Mellanox ConnectX-6 VPI de un solo puerto** **HDR InfiniBand de 200 Gb/s** **1 VPI Mellanox ConnectX-6 de doble puerto** **Ethernet de 10/25/50/100/200 Gb/s**
Almacenamiento	**Sistema operativo: 2 unidades de M.2 NVME de 1,92 TB** **Almacenamiento interno: Unidades NVME U.2 de 15 TB (4x3,84 TB) U.2**
Software	**Sistema operativo Ubuntu Linux**
Peso del sistema	**271 lbs (123 kg)**
Peso del sistema empaquetado	**315 lbs (143 kg)**
Medidas del sistema	**Altura: 10,4 pulgadas (264 mm)** **Ancho: 19 pulgadas (482,3 mm) MÁX.** **Largo: 35,3 pulgadas (897,1 mm) MÁX.**
Rango de temperatura en funcionamiento	**5 a 30 °C (41 a 86 °F)**

Acelerar el proceso de presentación de solución

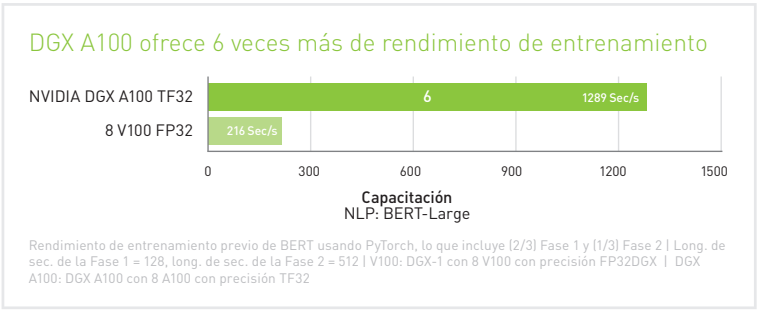
NVIDIA DGX A100 incluye 8 GPU NVIDIA A100 con Tensor Cores, lo que proporciona a los usuarios aceleración incomparable. Además, se optimizó completamente para el software NVIDIA CUDA-X™ y el conjunto de soluciones integral de NVIDIA para el data center. Las GPU NVIDIA A100 incluyen una nueva precisión, TF32, que funciona igual que FP32, pero proporciona 20 más de FLOPS para IA en comparación con la generación anterior. Lo mejor de todo, no se requieren cambios de código para obtener esta aceleración. Y cuando se usa la precisión mixta automática de NVIDIA, A100 ofrece el doble de aumento en el rendimiento con solo una línea adicional de código que usa la precisión FP16. La GPU A100 también tiene un ancho de banda de memoria de 1,6 terabytes por segundo (TB/s) líder en su clase, que representa un aumento de más del 70 % con respecto a la última generación. Además, la GPU A100 tiene mucho más memoria en el chip, incluida una memoria caché de nivel 2 de 40 MB que es casi 7 veces más grande que la generación anterior, lo que maximiza el rendimiento de procesamiento. DGX A100 también presenta la próxima generación de NVIDIA NVLink™, que duplica el ancho de banda directo de GPU a GPU a 600 gigabytes por segundo (GB/s), casi 10 veces más alto que PCIe Gen 4, y un nuevo NVIDIA NVSwitch que es 2 veces más rápido que la última generación. Esta potencia sin precedentes acelera el lanzamiento de soluciones, lo que permite a los usuarios abordar desafíos que antes no eran posibles o prácticos.

El sistema de IA más seguro del mundo para empresas

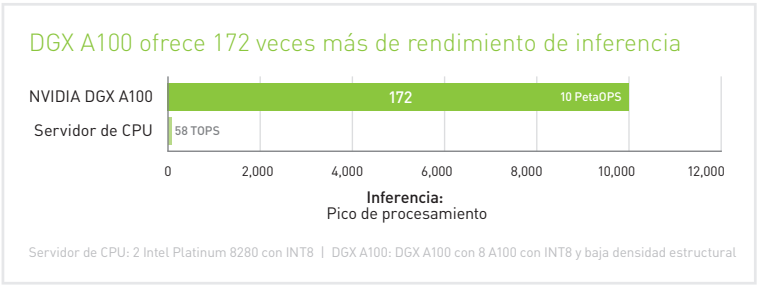
NVIDIA DGX A100 ofrece la postura de seguridad más sólida para tu empresa de IA, con un enfoque de múltiples capas que asegura todos los componentes principales de hardware y software. DGX A100 se extiende a todo el controlador de administración de la placa base (BMC), la placa de CPU, la GPU, las unidades autocifradas y el arranque seguro, por lo que tiene seguridad incorporada, lo que permite a los usuarios centrarse en los resultados en lugar de la evaluación y mitigación de amenazas.

Escalabilidad en el data center incomparable con Mellanox

Con la arquitectura de I/O más rápida de cualquier sistema DGX, NVIDIA DGX A100 es el componente fundamental para grandes grupos de inteligencia artificial como NVIDIA DGX SuperPOD™, el plan empresarial para la infraestructura de IA escalable. DGX A100 cuenta con 8 adaptadores InfiniBand Mellanox ConnectX-6 VPI HDR de un solo puerto para la agrupación en clúster y 1 adaptador Ethernet ConnectX-6



Rendimiento de entrenamiento previo de BERT usando PyTorch, lo que incluye (2/3) Fase 1 y (1/3) Fase 2 | Long. de sec. de la Fase 1 = 128, long. de sec. de la Fase 2 = 512 | V100: DGX-1 con 8 V100 con precisión FP32DGX | DGX A100: DGX A100 con 8 A100 con precisión TF32



Servidor de CPU: 2 Intel Platinum 8280 con INT8 | DGX A100: DGX A100 con 8 A100 con INT8 y baja densidad estructural

3000 servidores de CPU en comparación con 4 DGX A100 | Conjunto de datos de análisis común publicados: 128.000 millones Edges, 2,6 TB Graph

VPI de doble puerto para el almacenamiento y la red, todos con capacidad de 200 Gb/s. La combinación de computación acelerada por GPU masiva con las optimizaciones de hardware y software de redes de última generación significa que DGX A100 puede escalar a cientos o miles de nodos para cumplir con los mayores desafíos, como la inteligencia artificial conversacional y la clasificación de imágenes a gran escala.

Soluciones de infraestructura probadas creadas con líderes confiables del data center

En combinación con los principales proveedores de tecnología de almacenamiento y redes, ofrecemos una cartera de soluciones de infraestructura que incorporan lo mejor de la arquitectura de referencia NVIDIA DGX POD™. Estas soluciones hacen que las implementaciones de IA en el centro de datos sean más simples y rápidas para TI, ya que se entregan como ofertas totalmente integradas y listas para implementar a través de nuestra red de socios de NVIDIA.

Para obtener más información sobre la VDI acelerada por NVIDIA, visita **www.nvidia.com/DGXA100**

© 2020 NVIDIA Corporation. Todos los derechos reservados. NVIDIA, el logotipo de NVIDIA, NVIDIA DGX A100, NVLink, DGX SuperPOD, DGX POD y CUDA son marcas comerciales y/o marcas comerciales registradas de NVIDIA Corporation. Todos los nombres de productos y compañías son marcas comerciales o marcas comerciales registradas de los respectivos propietarios a los que están asociados. Las características, los precios, la disponibilidad y las especificaciones están sujetas a cambios sin previo aviso. MAY20

NVIDIA