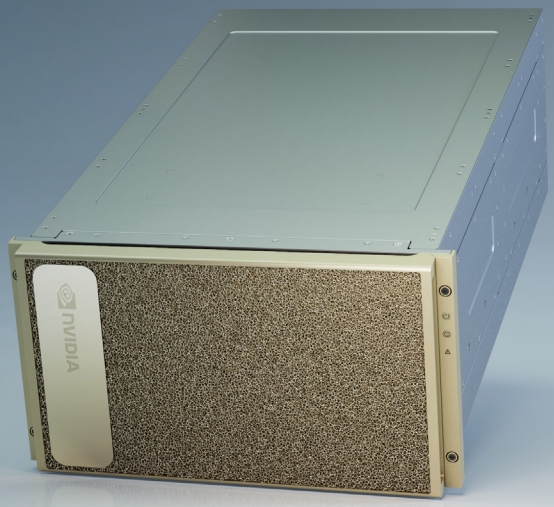
HOJA TÉCNICA

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 computación lentas y aisladas por las cargas de trabajo de análisis, entrenamiento 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 AI, 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 formato de 6U, lo que le permite reemplazar la infraestructura de computación 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 Múltiples Instancias (MIG) en la GPU NVIDIA A100 Tensor Core. Esto les permite a los administradores asignar recursos del tamaño adecuado para cargas de trabajo específicas.

DGX A100 está disponible con hasta 640 gigabytes (GB) de memoria GPU total, lo que aumenta el rendimiento en trabajos de entrenamiento a gran escala hasta 3 veces y duplica el tamaño de las instancias MIG. Por lo tanto, puede abordar los trabajos más grandes y complejos, junto con los más simples y más pequeños. Al ejecutar la pila de software DGX con el software optimizado de NVIDIA NGC™, la combinación de potencia de computación densa 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 Bright Cluster Manager.

Nivel de Asistencia y Conocimiento Incomparables

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 utilizan la IA con fluidez y han acumulado una gran cantidad de conocimiento y experiencia durante la última década para ayudar a maximizar el valor de una 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

	NVIDIA DGX A100 640GB	
GPU	**8 GPU NVIDIA A100 80GB Tensor Core**	
Memoria de GPU	**640 GB en total**	
Rendimiento	**5 petaFLOPS de IA** **10 petaOPS, INT8**	
NVIDIA NVSwitches	**6**	
Uso de energía del sistema	**6.5 kW máx.**	
CPU	**Dual AMD Rome 7742, 128 núcleos en total, 2.25 GHz (básico), 3.4 GHz (potencia máxima)**	
Memoria del Sistema	**2 TB**	
Redes	**8 NVIDIA ConnectX- 7 de un solo puerto** **InfiniBand de 200 Gb/s** **2 VPI NVIDIA ConnectX- 7 de doble puerto** **Ethernet de 10/25/50/ 100/200 Gb/s**	**8 NVIDIA ConnectX-6 VPI de un solo puerto** **InfiniBand de 200 Gb/s** **2 VPI NVIDIA ConnectX- 6 de doble puerto** **Ethernet de 10/25/50/100/ 200 Gb/s**
Almacenamiento	**SO: 2 unidades de M.2 NVME de 1.92 TB** **Almacenamiento Interno: Unidades NVME de 30 TB (8 de 3.84 TB)**	
Software	**Ubuntu Linux OS** **También es compatible con: Red Hat Enterprise Linux CentOS**	
Peso del Sistema	**271.5 lbs (123.16 kg) máx.**	
Peso del Sistema Empaquetado	**359.7 lbs (163.16 kg) máx.**	
Medidas del Sistema	**Altura: 10.4 in (264.0 mm)** **Ancho: 19.0 in (482.3 mm) máx.** **Longitud: 35.3 in (897.1 mm) máx.**	
Rango de Temperatura en Funcionamiento	**5 a 30 °C (41 a 86 °F)**	

Acelerar el Proceso de Lanzamiento de Soluciones

NVIDIA DGX A100 incluye 8 GPU NVIDIA A100 Tensor Core, 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 ofrecen precisión Tensor Float 32 (TF32), el formato de precisión predeterminado para los frameworks de IA TensorFlow y PyTorch. Esto es igual que FP32, pero proporciona 20 veces más operaciones flotantes por segundo (FLOPS) para la IA en comparación con la generación anterior. Lo mejor de todo es que no se requieren cambios de código para lograr esta aceleración.

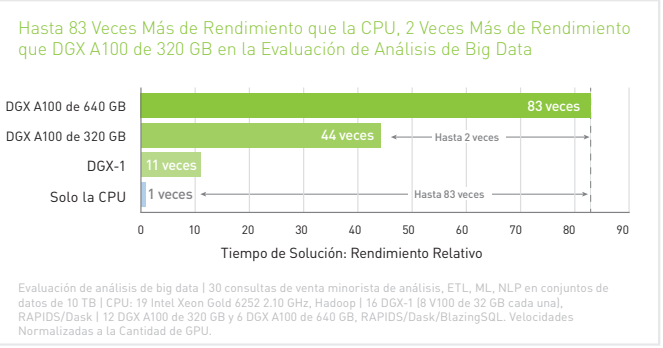
La GPU A100 de 80 GB aumenta el ancho de banda de la memoria de la GPU un 30 por ciento en comparación con la GPU A100 de 40 GB, lo que la convierte en la primera del mundo con 2 terabytes por segundo (TB/s). También tiene significativamente más memoria en el chip que la GPU NVIDIA de la generación anterior, incluido un caché de nivel 2 de 40 megabytes (MB) que es casi 7 veces más grande, lo que maximiza el rendimiento de la computación. DGX A100 también incluye 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 la 4.ª generación de PCIe, 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 AI más seguro del mundo para empresas

NVIDIA DGX A100 ofrece una postura de seguridad 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 placa de GPU y las unidades autocifradas, 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 Incomparable del Data Center con las Redes de NVIDIA

Con la arquitectura de E/S más rápida de cualquier sistema DGX, NVIDIA DGX A100 es el componente fundamental para grandes grupos de clústeres de IA como NVIDIA DGX SuperPOD™, el plan empresarial para la infraestructura de AI escalable. DGX A100 cuenta con 8 adaptadores NVIDIA ConnectX®-7 de puerto único para la agrupación en clúster y hasta dos adaptador esConnectX- 7 VPI de doble puerto (InfiniBand o Ethernet) para el almacenamiento y la red, todos con capacidad de 200 Gb/s. Con la conectividad ConnectX-7 con los switches NVIDIA Quantum-2 InfiniBand, DGX SuperPOD se puede



construir con menos switches y cables, lo que ahorra CAPEX y OPEX en la infraestructura del data center. 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 IA 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, está disponible una cartera de soluciones de infraestructura que incorpora lo mejor de la arquitectura de referencia NVIDIA DGX POD™. Estas soluciones se entregan como ofertas totalmente integradas y listas para implementar a través de nuestra red de socios de NVIDIA (NPN). Por lo tanto, hacen que las implementaciones de IA en el centro de datos sean más simples y rápidas.

Más información

Para obtener más información sobre NVIDIA DGX A100, visita **www.nvidia.com/dgxa100**