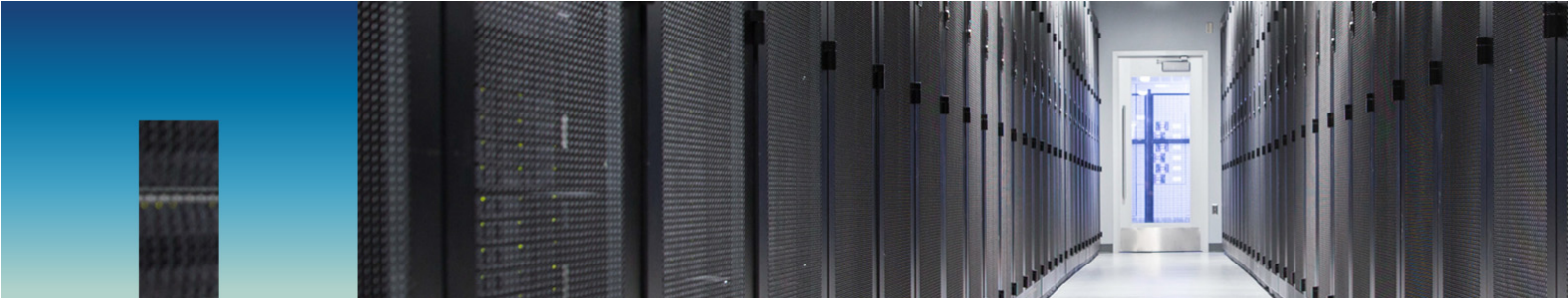


Documento técnico

Infraestructura de IA escalable

Diseño para casos prácticos de aprendizaje profundo para el mundo real

Sundar Ranganathan, NetApp
Santosh Rao, NetApp

Junio de 2018 | WP-7267

En colaboración

NVIDIA

Resumen ejecutivo

El aprendizaje profundo (DL, por sus siglas en inglés) está permitiendo avances rápidos en algunos de los mayores retos de la ciencia actual: mejores formas de curar el cáncer en medicina, detección y clasificación de partículas en física y la autonomía de nivel 5 en los vehículos sin conductor. Todos ellos tienen un elemento común: los datos. El DL se basa fundamentalmente en los datos.

Las unidades de procesamiento gráfico (GPU) permiten un nuevo nivel de conocimiento que antes no era posible. Para satisfacer las rigurosas exigencias de las GPU en una aplicación de DL, los sistemas de almacenamiento deben ser capaces de ofrecer datos constantemente a las GPU a una baja latencia y con un alto rendimiento, independientemente de si son texto, imágenes, audio o vídeo.

A medida que las organizaciones progresan desde implementaciones de DL a pequeña escala a la producción, es crucial contar con una infraestructura que pueda proporcionar un alto rendimiento y permitir un escalado fluido e independiente. El liderazgo de NVIDIA en GPU, junto con la innovación de NetApp en todos los sistemas de almacenamiento flash, forman una solución única capaz de acelerar las aplicaciones de DL.

NetApp

ÍNDICE

LISTA DE TABLAS

LISTA DE FIGURAS

1 Introducción

Cuando los líderes de tecnología se refieren al valor de los datos en sus organizaciones, los ejemplos más comunes son el aprendizaje profundo (DL) y, en términos más generales, la inteligencia artificial (IA). El DL es el motor que ya está haciendo posible la detección del fraude en el mundo financiero, el mantenimiento predictivo en la fabricación, el reconocimiento por voz en bots de soporte al cliente así como diversos niveles de autonomía en vehículos.

De cara al futuro, las aplicaciones de datos y el DL permitirán mejorar la productividad, identificar patrones esenciales y diseñar servicios, soluciones y productos disruptivos en todos los sectores imaginables. La empresa de investigación de mercados IDC prevé que la inversión en IA para software, servicios y hardware crecerá de 12000 millones de USD en 2017 a 57 600 millones de USD en 2021[1].

Los sistemas de DL aprovechan algoritmos que han mejorado de forma significativa gracias al aumento del tamaño de red neural, así como la cantidad y calidad de los datos utilizados para entrenar los modelos. En función de la aplicación específica, los modelos de DL funcionan con una gran cantidad de distintos tipos de datos ya sea texto, imágenes, audio, vídeo o datos de series temporales. Por este motivo, las aplicaciones de DL requieren una infraestructura de alto rendimiento, como la arquitectura de referencia que aquí se presenta.

La puesta en marcha de cargas de trabajo de DL en una infraestructura que haya sido creada expresamente para gestionar sus demandas únicas y crecientes y que además proporcione el tiempo de entrenamiento más rápido es vital para el éxito de la estrategia de DL de una organización.

Esta infraestructura debería satisfacer los siguientes requisitos de infraestructura de alto nivel:

- Una gran capacidad computacional para entrenar los modelos en menos tiempo.
- Almacenamiento de alto rendimiento para gestionar conjuntos de datos de gran tamaño.
- Escalar la computación y el almacenamiento de forma independiente y sin problemas.
- Gestionar los distintos tipos de tráfico de datos.
- Optimizar costes.

Es crucial contar con una infraestructura compatible con la flexibilidad del escalado horizontal y vertical. A medida que la computación y el almacenamiento se escalan, los requisitos de rendimiento pueden evolucionar y que sea necesario alterar de forma dinámica los ratios de los sistemas de computación a almacenamiento con cero tiempo de inactividad.

Este documento presenta una infraestructura escalable que incluye el servidor NVIDIA® DGX-1™ con la plataforma de unidad de procesamiento gráfico (GPU) NVIDIA® Tesla® V100[2] y el nuevo sistema de almacenamiento all-flash NetApp® A800™.

2 Computación: NVIDIA DGX-1

2.1 Computación acelerada por GPU

Los cálculos computacionales realizados por los algoritmos de DL suponen una inmensa cantidad de multiplicaciones de matrices que se ejecutan en paralelo. La arquitectura altamente paralela de las GPU modernas logra que sean sustancialmente más eficaces que las unidades centrales de procesamiento (CPU) de uso general para aplicaciones en las que el procesamiento de datos se realiza en paralelo. Los avances en las arquitecturas GPU individuales y en clúster las han convertido en la plataforma preferida para cargas de trabajo como la computación de alto rendimiento, el DL y el análisis.

[1] Fuente: IDC, Worldwide Storage for Cognitive/AI Workloads Forecast (Previsión del almacenamiento mundial para cargas de trabajo cognitivas/IA), 2018–2022
[2] Con la tecnología de la arquitectura NVIDIA Volta

2.2 Creadas para el aprendizaje profundo

Montar e integrar componentes de hardware y software estándar para DL de múltiples proveedores puede dar como resultado un aumento en la complejidad y en los tiempos de puesta en marcha, lo cual provoca que los recursos científicos de datos dediquen un esfuerzo considerable a las tareas de integración de sistemas.

Una vez que la solución se ha puesto en marcha, muchas organizaciones se encuentran con que dedican demasiados ciclos al ajuste y refinado de pila de software a medida que evolucionan sus modelos. NVIDIA, dándose cuenta de ello, ha creado la plataforma de servidor DGX-1, un sistema listo para usar de hardware y software completamente integrado para los flujos de trabajo de DL.

Cada servidor DGX-1 cuenta con ocho GPU Tesla V100 configurados en una topología híbrida de malla de cubos con NVIDIA NVLink™ que proporciona una estructura de baja latencia y un ancho de banda ultraelevado para la comunicación entre GPU básica para el entrenamiento, lo que elimina los cuellos de botella asociados a las interconexiones basadas en PCIe. El servidor DGX-1 también está equipado con interconexiones de red de ancho de banda elevado y latencia baja para clustering de múltiples nodos en estructuras aptas para RDMA.

2.3 Optimizado previamente y de clase empresarial

El servidor DGX-1 aprovecha los contenedores de software optimizados para GPU de NVIDIA GPU Cloud (NGC), incluidos los contenedores para todos los marcos de DL más populares. Los contenedores de aprendizaje profundo NGC han sido optimizados previamente en cada capa, incluidos los controladores, las bibliotecas y las primitivas de comunicación y proporcionan un rendimiento máximo para las GPU de NVIDIA. Estos contenedores integrados previamente libran a los usuarios de la agitación constante asociada a los marcos de DL de código abierto más populares en la actualidad y, de este modo, proporcionan a los equipos una pila de calidad de servicio comprobada y estable en la que se pueden basar las aplicaciones de DL de clase empresarial.

Esta solución de hardware y software completamente integrada, respaldada por la experiencia de NVIDIA, acelera las puestas en marcha de aplicaciones de DL, reduce el tiempo de entrenamiento de semanas a días u horas y aumenta la productividad de los científicos de datos, permitiéndoles invertir más tiempo en la experimentación en lugar de en la integración de sistemas y soporte.

3 Almacenamiento: NetApp AFF

A medida que las GPU son cada vez más rápidas y aumenta el tamaño y la complejidad de los conjuntos de datos, el uso de sistemas de almacenamiento de vanguardia resulta básico para eliminar los cuellos de botella y maximizar el rendimiento del sistema. Las aplicaciones de DL requieren una solución de almacenamiento que haya sido diseñada para gestionar esas cargas de trabajo paralelas masivas que necesitan un nivel elevado de concurrencia en el procesamiento de I/O para evitar bloquear las GPU cuando esperan los datos.

En muchas aplicaciones de DL, el tráfico de datos se extiende por toda la canalización de datos, desde los extremos hasta el centro y el cloud. Diseñar una arquitectura de almacenamiento requiere una metodología de gestión de datos holística, desde la ingesta y/o el análisis periférico hasta la preparación y el entrenamiento en el centro de datos central para el archivado en el cloud. Comprender los requisitos de rendimiento, las características de los distintos conjuntos de datos y los servicios de datos necesarios resulta crucial.

Una solución de almacenamiento ideal para los flujos de trabajo de DL debe destacar en los siguientes requisitos de alto nivel.

3.1 Alto rendimiento

Los cuellos de botella de una infraestructura de DL normalmente se producen durante la fase de entrenamiento, cuando se requiere un ancho de banda de I/O elevado con un paralelismo de I/O masivo para garantizar un uso de la GPU elevado y constante. Esto se traduce en la capacidad de la arquitectura de almacenamiento de proporcionar un rendimiento de alto nivel a la vez que se mantiene un perfil de baja latencia que, a su vez, se traduce en soporte para las estructuras de red de alta velocidad.

Un único sistema NetApp A800 da cabida a un rendimiento de 25 GB/s para lecturas secuenciales y un millón de IOPS en pequeñas lecturas aleatorias en latencias inferiores a 500µs[3]. Además, lo que distingue al A800 es su compatibilidad con las redes de 100 GbE[4], lo que permite acelerar el movimiento de datos y también fomenta el equilibrio en el sistema de entrenamiento general, ya que el DGX-1 es compatible con RDMA de 100 GbE para la interconexión de clústeres. El sistema A700s de NetApp es compatible con múltiples vínculos 40 GbE para proporcionar un rendimiento máximo de 18 GB/s.

3.2 Escalabilidad

Los conjuntos de datos de gran tamaño son importantes para aumentar la precisión de los modelos. Las puestas en marcha de DL que empiezan a menor escala (unos pocos terabytes de almacenamiento) pueden requerir próximamente un escalado horizontal a diversos petabytes. Además, las necesidades de rendimiento pueden variar en función del modelo de entrenamiento utilizado y la aplicación final, por lo que podrían requerir un escalado de la computación y/o el almacenamiento independiente. Diseñar una arquitectura de sistemas robusta en un entorno a escala de rack permite un escalado independiente.

Los sistemas A800 y A700 de NetApp pueden escalarse de forma independiente, sin problemas ni interrupciones desde 2 nodos (364,8 TB) a un clúster de 24 nodos (74,8 PB con el A800, 39,7 PB con el A700). El uso de volúmenes de ONTAP® FlexGroup™ permite una gestión de datos en un volumen lógico de espacio de nombres único de 20 PB, que admite más de 400 000 millones de archivos. Para las capacidades de clúster superiores a los 20 PB, se pueden crear múltiples FlexGroups que abarquen la capacidad necesaria.

3.3 Integración de plataforma robusta

A medida que las organizaciones aceleran su tasa de recogida de datos, es más evidente la necesidad la automatización de dichos datos. El uso de contenedores es una de las formas de lograrlo: permite una puesta en marcha más rápida separando las dependencias de las aplicaciones de la capa dispositivo-unidad y del sistema operativo. Una gestión de datos sencilla y eficiente es básica para reducir el tiempo de entrenamiento.

Trident™ es un orquestador de almacenamiento dinámico de NetApp para imágenes de contenedor completamente integrado con Docker™ y Kubernetes™. En combinación con NVIDIA GPU Cloud (NGC) y orquestadores populares como Kubernetes o Docker Swarm, Trident permite a los clientes poner en marcha sin problemas imágenes de contenedor IA/DL NGC en almacenamiento de NetApp, proporcionando una experiencia de nivel empresarial para las puestas en marcha de contenedores de IA. Esto incluye la orquestación automática, el clonado para pruebas y desarrollo, las pruebas de renovación de NGC con clonado, la copias de protección y cumplimiento de normativas y muchos otros casos prácticos de gestión de datos para las imágenes de contenedor de IA de NGC.

El tráfico de datos en DL puede consistir en millones de archivos (imágenes, vídeo/audio, archivos de texto). Los sistemas de archivos de red (NFS) son la opción ideal para la prestación de un alto rendimiento en una gran variedad de cargas de trabajo, ya que pueden gestionar bien tanto I/O aleatorias como secuenciales. Cuando se utiliza con los volúmenes de ONTAP FlexGroup, NetApp AFF puede proporcionar un alto rendimiento para cargas de trabajo de archivos pequeños que abarquen diversos sistemas de almacenamiento.

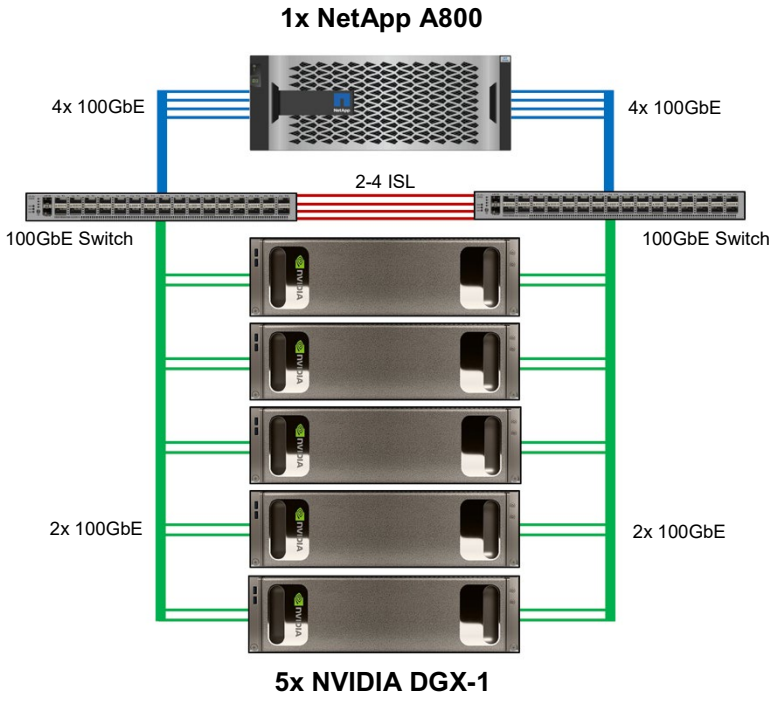
[3] https://blog.netapp.com/the-future-is-here-ai-ready-cloud-connected-all-flash-storage-with-nvme/
[4] https://www.netapp.com/es/products/storage-systems/all-flash-array/aff-a-series.aspx#technical-specifications

4 Arquitectura de referencia

Diseñar una infraestructura que pueda proporcionar un rendimiento de I/O constante con una baja latencia para explotar el paralelismo de I/O de las GPU y además pueda escalar la computación y los sistemas de almacenamiento sin interrupciones es crucial para lograr un tiempo de entrenamiento más rápido. Estos requisitos se traducen en la necesidad de que el sistema de almacenamiento sea compatible con una estructura de red de baja latencia, alta velocidad y con un ancho de banda elevado para poder maximizar el rendimiento y mantener múltiples servidores de DGX-1 con la entrada de datos constante que se necesita para la ejecución de cada entrenamiento de DL.

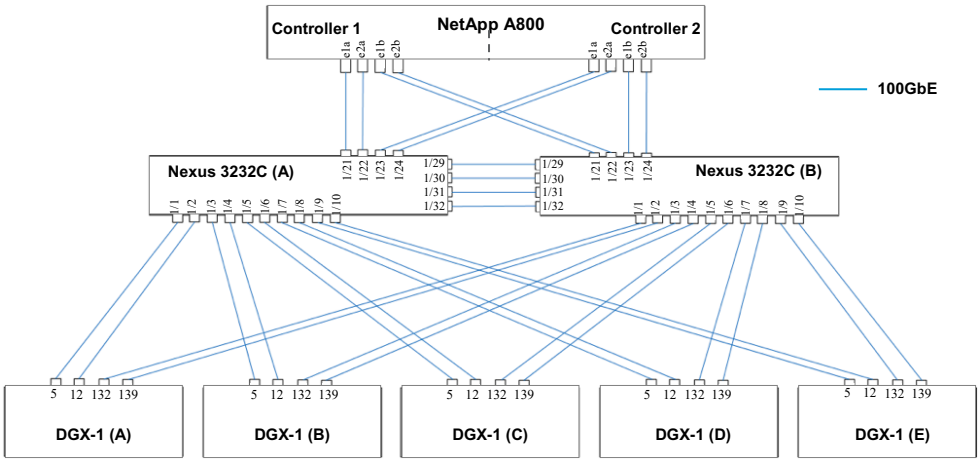
La Figura 1 muestra una arquitectura de NetApp en una configuración 1:5 con cinco servidores DGX-1 alimentados con un par de alta disponibilidad (HA) A800 mediante dos switches. Cada servidor DGX-1 se conecta con cada uno de los dos switches mediante dos enlaces de 100 GbE. El A800 se conecta con cada switch mediante cuatro enlaces de 100 GbE. Los switches pueden tener de dos a cuatro enlaces de 100 Gb internos diseñados para los casos de conmutación al nodo de respaldo. El diseño de alta disponibilidad es activo-activo, por lo que es posible mantener un rendimiento máximo en todas las conexiones de red en caso de ausencia de fallo.

Figura 1) Arquitectura de referencia en una configuración 1:5.



La Figura 2 muestra la conectividad de la arquitectura a nivel de puertos. Esta arquitectura consiste en dos switches Cisco Nexus 3232C 100 GbE, así como un switch de gestión de 1 GbE (que no se muestra). Existen cinco servidores DGX-1 conectados a la red con cuatro enlaces de 100 GbE cada uno, con dos enlaces desde cada DGX-1 conectados a cada switch Nexus. El almacenamiento lo proporciona un par de alta disponibilidad de A800 de NetApp y cada una de las dos controladoras de almacenamiento está conectada a cada switch de Nexus mediante dos enlaces de 100 GbE.

Figura 2) Diagrama de red de una conectividad a nivel de puerto para una configuración 1:5.



Con A700, la arquitectura mostrada en la Figura 1 cambia a una configuración 1:4 (1 A700 : 4 DGX-1), con cuatro enlaces de 40 GbE desde los A700 a cada switch del lado del almacenamiento y dos enlaces de 100 GbE desde cada DGX-1 a cada uno de los dos switches. Además de 100 GbE, el sistema A800 también admite 40 GbE. Ambas arquitecturas pueden escalarse de forma vertical y horizontal sin tiempos de inactividad a medida que crece el tamaño de los conjuntos de datos.

5 Escalado del rack: empiece con poco y escale a lo grande

El escalado horizontal significa que a medida que el entorno de almacenamiento crece se añaden nodos de computación y/o capacidad de almacenamiento adicional sin problemas al pool de recursos que reside en una infraestructura de almacenamiento compartida. Las conexiones de host y de cliente, así como los almacenes de datos, pueden moverse sin problema a cualquier lugar del pool de recursos. Por tanto, las cargas de trabajo existentes pueden equilibrarse fácilmente entre los recursos disponibles y las cargas de trabajo nuevas pueden ponerse en marcha fácilmente. Las actualizaciones tecnológicas (añadir o sustituir bandejas de unidades y/o controladoras de almacenamiento) se realizan mientras el entorno sigue en línea y continúa ofreciendo datos.

NetApp ha combinado la potencia computacional de los servidores DGX-1 con la arquitectura de alto rendimiento de los sistemas A800 y A700 para ofrecer una solución atractiva que permita a las organizaciones poner en marcha flujos de trabajo de DL en pocas horas y escalar horizontalmente sin problemas según sea necesario.

Las organizaciones que se inician en el DL pueden empezar con una configuración 1:1 y escalar a medida que crecen los datos a una configuración 1:5 y más allá en el modo de escalado horizontal. La Tabla 1 destaca el escalado de capacidad y rendimiento que puede lograrse con diversas configuraciones con DGX-1 y A800 con ONTAP 9.4.

Tabla 1) Métricas de capacidad y rendimiento en casos de escalado horizontal con A800.

Número de **A800** Sistemas de almacenamiento	Núm. de servidores DGX-1	Rendimiento	Típica Capacidad bruta[5]	Capacidad bruta con expansión[5]
1 pareja de alta disponibilidad	5	25 GB/s	364,8 TB	6,2 PB

La información de la Tabla 1 se basa en las métricas de rendimiento de A800 y ONTAP 9.4. Cada A800 proporciona un rendimiento de 25 GB/s y es capaz de gestionar el tráfico de sistemas 5 DGX-1 a la vez que proporciona la opción de escalar a un almacenamiento de 6,2 PB.

Tabla 2) Métricas de capacidad y rendimiento en casos de escalado horizontal con A700.

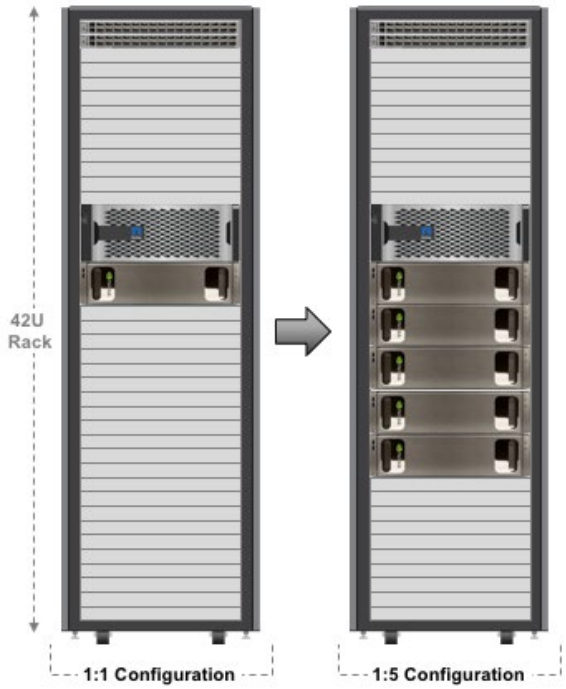
Número de **A700s** Sistemas de almacenamiento	Núm. de servidores DGX-1	Rendimiento	Típica Capacidad bruta[5]	Capacidad bruta con expansión[5]
1 pareja de alta disponibilidad	4	18 GB/s	367,2 TB	3,3 PB

La información de la Tabla 2 se basa en las métricas de rendimiento de A700 y ONTAP 9.4. El sistema A700 admite un rendimiento de 18 GB/s y es ideal para una configuración 1:4 inicial.

Las organizaciones que necesitan empezar con un espacio de almacenamiento y coste menores pueden utilizar los sistemas de almacenamiento A300 o A200 de NetApp que también admiten escalabilidad fluida.

Según la información de escalado de la Tabla 1, la Figura 3 ilustra cómo se puede escalar una configuración 1:1 a una puesta en marcha de configuración 1:5 en un centro de datos. Este método proporciona la flexibilidad de alterar los ratios de computación a almacenamiento en función del tamaño del lago de datos, los modelos de DL utilizados y las métricas de rendimiento necesarias.

Figura 3) Escalado a nivel de rack de una configuración 1:1 a 1:5 con A800.



El número de servidores de DGX-1 y de sistemas de almacenamiento de AFF por rack depende de las especificaciones de potencia y refrigeración del rack que se esté utilizando. La ubicación final de los sistemas depende del análisis dinámico fluido computacional, la gestión del flujo de aire y del diseño del centro de datos.

[5] https://www.netapp.com/es/media/ds-3582.pdf

6 Pruebas de rendimiento

Las pruebas de rendimiento de TensorFlow se realizaron en una configuración 1:1 (un servidor DGX-1 y un sistema de almacenamiento A700) con un conjunto de datos ImageNet (143 GB) almacenado en un volumen FlexGroup en el sistema A700. El sistema de archivos elegido para estas pruebas fue NFSv3.

Configuración del entorno:

- Sistema operativo: Ubuntu 16.04 LTS
- Docker: 18.03.1-ce [9ee9f40]
- Dockerfile: nvcr.io/nvidia/tensorflow:18.04-py2
- Marco: Tensorflow 1.7.0
- Pruebas de rendimiento: Tensorflow Benchmarks [26a8b0a]

Como parte de nuestras pruebas iniciales, ejecutamos pruebas de rendimiento basadas en datos sintéticos para estudiar el rendimiento de las GPU sin una canalización TensorFlow potencial ni cuellos de botella relacionados con el almacenamiento. El entrenamiento se realizó con datos sintéticos tanto en los núcleos CUDA como en los núcleos Tensor para todos los modelos en la prueba de rendimiento TensorFlow.

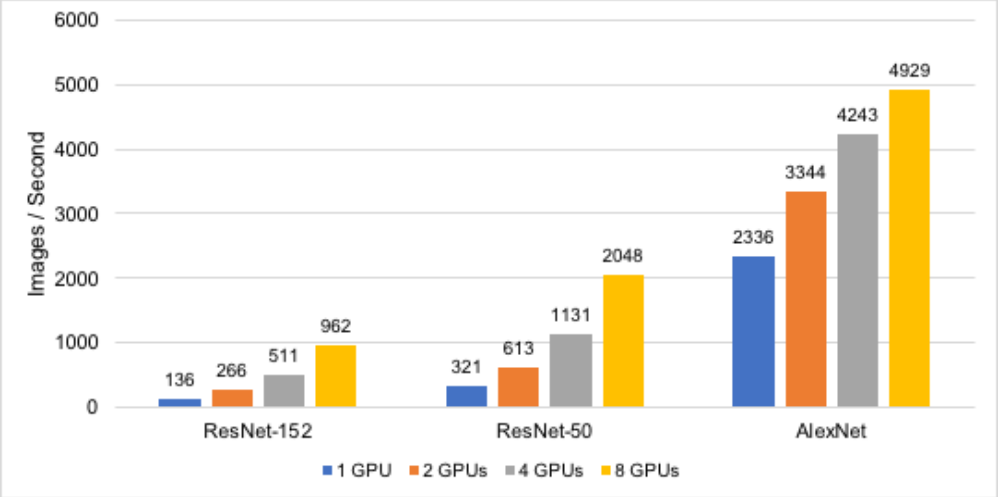
En el siguiente paso, se utilizaron datos reales con distorsión para todas las pruebas.

Los puntos siguientes detallan los aspectos más destacados de las pruebas:

- El rendimiento del entrenamiento del modelo se mide como el número de imágenes procesadas por segundo.
- Para demostrar las tasas de entrenamiento que se pueden lograr, se seleccionaron tres modelos populares que representan distintos grados de complejidad computacional y precisión predictiva: ResNet-152, ResNet-50 y AlexNet.
- Para reflejar escenarios de entrenamiento de modelos realistas, se permitió la distorsión (pasos de procesamiento previo de imágenes) para forzar el sistema desde el punto de vista del procesamiento de la GPU y del almacenamiento.
- Las métricas de rendimiento se midieron con un número diverso de GPU activadas en el servidor DGX-1.
- La utilización de la GPU se mantuvo cerca del 100 % durante todo el periodo de entrenamiento, lo cual indica que el sistema A700 es capaz de ofrecer datos con rapidez suficiente a las GPU y, a la vez, mantener una tasa de entrenamiento elevada.
- Se eligió el modelo AlexNet por su consumo de I/O más intensivo para forzar la canalización e ilustrar los casos de uso extremos. Es posible que este modelo no sea el más preciso y se considera restrictivo en los escenarios de escalado.
- El tamaño de lote utilizado: 64 para ResNet-152 y ResNet-50 y 512 para AlexNet.

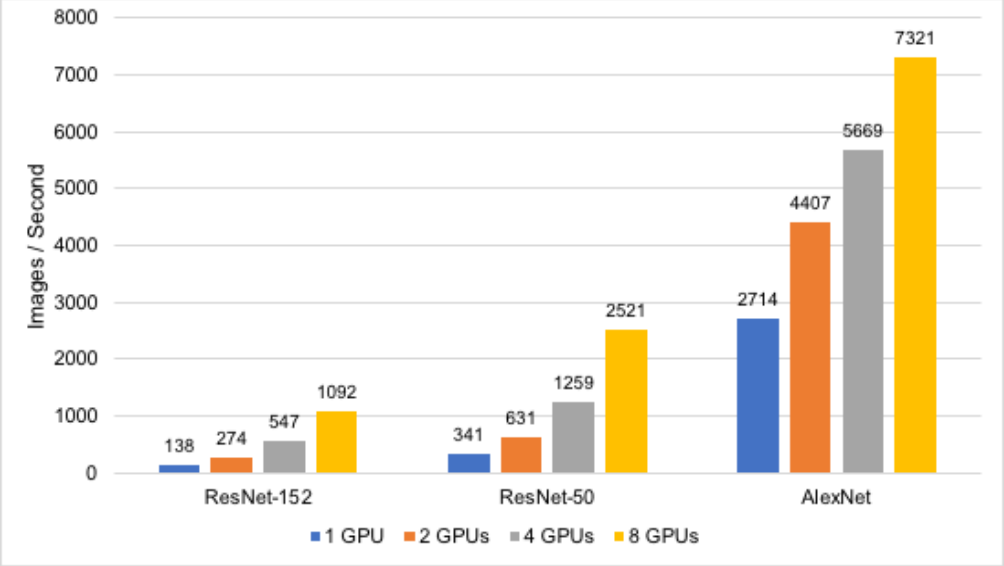
Las Figuras 4 y 5 ilustran el rendimiento de entrenamiento medido con cada uno de los tres modelos de DL con uno, dos, cuatro y ocho GPU.

Figura 4) Tasas de entrenamiento con datos reales y distorsiones de imagen activadas.



* Datos redondeados al valor decimal más cercano

Figura 5) Tasas de entrenamiento con datos reales y distorsiones e imagen desactivadas.



* Datos redondeados al valor decimal más cercano

La Figura 6 ilustra la utilización de la GPU lograda durante el entrenamiento del modelo ResNet-50 con 8 GPU. La curva verde muestra la suma de las utilizaciones de las 8 GPU y la curva naranja indica el rendimiento de lectura del A700. Para mantener una utilización elevada de la GPU y una tasa de entrenamiento de ~2500 imágenes por segundo, el rendimiento de lectura del A700 alcanza los ~300 MB/s. Se necesitaron ~500 segundos para cargar el conjunto de datos de 143 GB en la memoria del DGX-1. Antes y después de la marca de 500 segundos, la utilización de la GPU y las tasas de entrenamiento eran idénticas. Esto indica que no se ha producido ningún cuello de botella de I/O del almacenamiento u otros en la canalización de la transferencia de GPU a esta tasa de entrenamiento.

Figura 6) Utilización de la GPU y rendimiento de lectura del A700 con ResNet-50 a una velocidad de ~2500 imágenes/segundo.

7 Conclusión

La IA requiere una potencia computacional masiva y una arquitectura de infraestructura que pueda seguirle el ritmo. Las disciplinas como el DL exigen un rendimiento extremo para poder alimentar estos algoritmos que requieren una gran cantidad de datos. A medida que la IA se vaya convirtiendo en una funcionalidad básica empresarial, casi todas las organizaciones dependerán de la información generada por los grandes conjuntos de datos que generan. Para cumplir con sus objetivos, necesitarán una infraestructura que pueda ponerse en marcha rápidamente, proporcionar el rendimiento paralelo necesario y escalarse y gestionarse con facilidad.

El servidor DGX-1 ha permitido averiguar información de los lagos de datos masivos acelerando la velocidad de entrenamiento de los modelos. Al combinar un hardware de GPU de vanguardia con contenedores optimizados para GPU de NGC se consigue poner en marcha las aplicaciones de DL de forma rápida y eficaz.

Como líder en el sector de NFS, NetApp cuenta con un paquete de productos con la familia AFF, ONTAP, FlexGroup, Trident y funciones de eficacia del almacenamiento líderes para satisfacer los requisitos de rendimiento paralelos masivos de las aplicaciones de DL junto con la experiencia práctica en la puesta en marcha de soluciones de IA.

NetApp se ha asociado con NVIDIA para presentar una arquitectura de escala de rack que permite a las organizaciones empezar con poco y aumentar la infraestructura sin problemas a medida que aumenta el número de proyectos y el tamaño de los conjuntos de datos. Esta arquitectura ha sido diseñada para liberar a las organizaciones de tener que tratar con un aumento en la complejidad de la infraestructura y ayudarles a centrarse en el desarrollo de mejores aplicaciones de DL. Adoptar estas soluciones de IA capacita a las empresas para satisfacer incluso los requisitos de rendimiento más exigentes, abriendo una nueva era de aplicaciones inteligentes.

8 Apéndice: lista de componentes

La Tabla 3 enumera los componentes que se han utilizado para los diseños de arquitectura descritos en este informe.

Tabla 3) Lista de componentes.

Componente	Cantidad	Descripción
Servidor base	1	Placa base Dual Intel Xeon CPU con x2 9,6 GT/s QPI, 8 canales con 2 DPC DDR4, Chipset Intel X99, AST2400 BMC
	1	Placa base GPU compatible con 8 módulos SXM2 (malla de cubos híbrida) y 4 ranuras PCIE x16 para NCI InfiniBand
Puertos de conexión	1	Puerto 10/100BASE-T IPMI
	1	Puerto de serie RS232
	2	Puertos USB 3.0
CPU	2	Intel Xeon E5-2698 v4, 20 núcleos, 2,2 GHz, 135 W
GPU	8	Tesla V100: 1 petaFLOPS, precisión mixta 32 GB de memoria por GPU 40960 núcleos NVIDIA CUDA® 5120 núcleos NVIDIA Tensor
Memoria del sistema	16	32 GB DDR4 LRDIMM (512 GB total)
Controlador SAS Raid	1	LSI SAS 3108 RAID Mezzanine de 8 puertos
Almacenamiento (RAID 0) (Data)	4	1,92 TB, 6 Gb/s, SATA 3.0 SSD
Almacenamiento (Sistema operativo)	1	480 GB, 6 Gb/s, SATA 3.0 SSD
10 GbE NIC	1	Adaptador de red Mezzanine, puerto dual, 10GBASE-T
Ethernet / InfiniBand EDR NIC	4	Mellanox ConnectX-4 VPI MCX455A-ECAT de puerto único, x16 PCIe
A800/A700/A300 de NetApp	1	Cabina all-flash, 1 pareja de alta disponibilidad
Cisco Nexus 3232C	2	Switches Ethernet de 100 Gb
Cisco Nexus 3048-TP	1	Switch de gestión Ethernet de 1 Gb

Consulte el apartado de la Herramienta de Matriz de Interoperabilidad (IMT) en el sitio web de soporte de NetApp para confirmar que las versiones exactas del producto y las funciones descritas en este documento son compatibles con su entorno concreto. La herramienta IMT de NetApp define los componentes y las versiones del producto que pueden utilizarse para crear configuraciones que sean compatibles con NetApp. Los resultados específicos dependen de la instalación que realice cada cliente de acuerdo con las especificaciones publicadas.