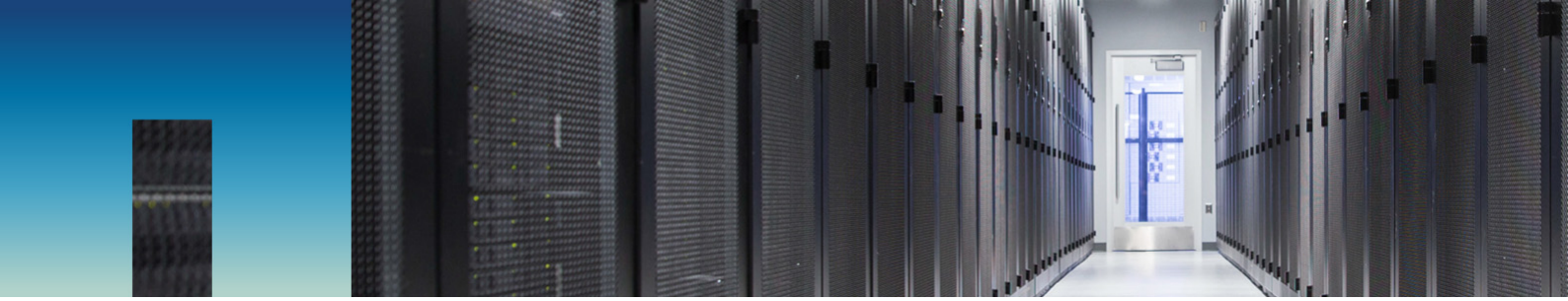


Arquitectura validada por NetApp

ONTAP AI de NetApp, con tecnología NVIDIA

Diseño de una infraestructura de IA escalable para casos de uso de aprendizaje profundo en el mundo real

David Arnette, Sundar Ranganathan, Amit Borulkar, Sung-Han Lin y Santosh Rao, NetApp Agosto de 2018 | NVA-1121

En colaboración con

NVIDIA

ÍNDICE

LISTA DE TABLAS

LISTA DE FIGURAS

1 Resumen ejecutivo

Este documento contiene información de validación para la arquitectura descrita en el documento técnico WP-7267: Infraestructura de IA escalable. El diseño de dicho documento se implementó utilizando el AFF A800 de NetApp®, un sistema FAS all-flash; servidores NVIDIA® DGX-1™ y switches Ethernet Cisco® Nexus® 3232C 100 Gb. Validamos el funcionamiento y el rendimiento de este sistema mediante herramientas estándar del sector y, a decir de los resultados de las pruebas de validación, esta arquitectura ofrece un excelente rendimiento para el entrenamiento y la inferencia. Los resultados también demuestran que queda un margen adicional de almacenamiento adecuado para el uso de varios servidores DGX-1. Además, es posible escalar de forma sencilla e independiente los recursos de computación y almacenamiento desde una configuración de medio rack hasta otra de varios racks, con un rendimiento previsible y capaz de satisfacer los requisitos de cualquier carga de trabajo de aprendizaje automático.

2 Resumen del programa

2.1 Programa Arquitectura validada por NetApp

El programa Arquitectura validada por NetApp ofrece a los clientes una arquitectura validada para soluciones de NetApp. Con Arquitectura validada por NetApp se obtiene una arquitectura de solución de NetApp que:

- Ha sido probada a conciencia.
- Tiene naturaleza prescriptiva.
- Minimiza los riesgos de implementación.
- Reduce el plazo de comercialización.

Este documento está dirigido a ingenieros de soluciones de NetApp y asociados así como a aquellos clientes con capacidad de decisión estratégica. Describe las consideraciones de diseño de arquitectura utilizadas para determinar el equipo, el cableado y la configuración específica necesaria en un entorno particular.

2.2 Solución ONTAP AI de NetApp

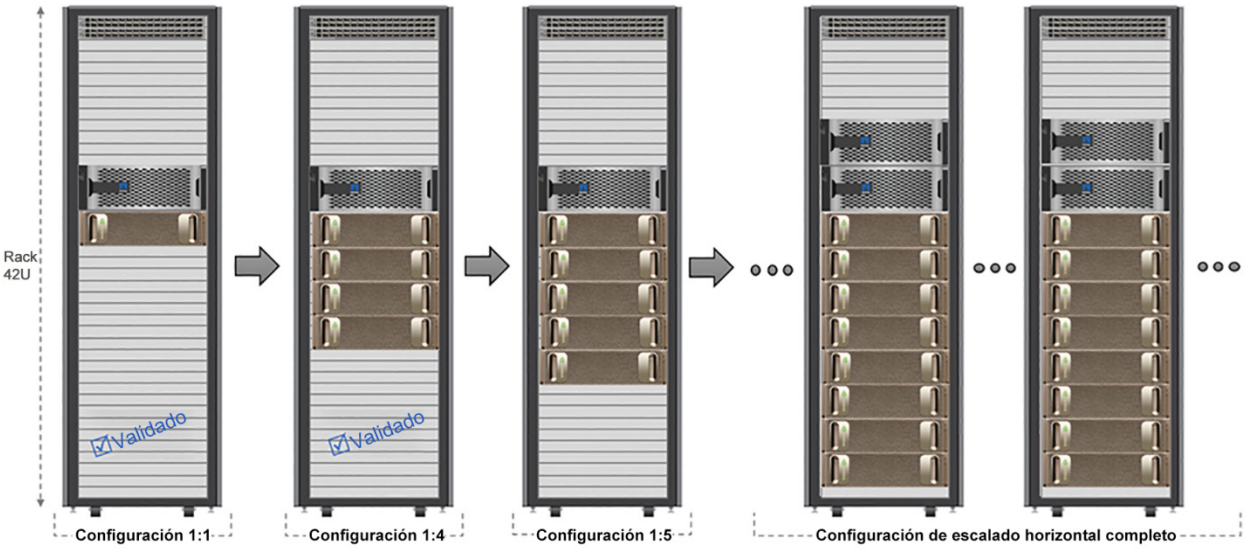
La infraestructura convergente ONTAP® AI de NetApp, con servidores NVIDIA DGX-1 y el sistema de almacenamiento de NetApp conectado al cloud, es una arquitectura desarrollada y validada por NetApp y NVIDIA. Proporciona a su organización una arquitectura prescriptiva que:

- Elimina las complejidades de diseño.
- Permite un escalado independiente de las capacidades de computación y almacenamiento.
- Permite comenzar con un sistema pequeño y escalar sin dificultades.
- Proporciona opciones de almacenamiento para distintos niveles de rendimiento y coste.

ONTAP AI de NetApp integra servidores NVIDIA DGX-1, unidades de procesamiento gráfico (GPU) NVIDIA Tesla® V100 y un sistema AFF A800 de NetApp con la mejor conexión a redes. ONTAP AI de NetApp simplifica las implementaciones de inteligencia artificial (IA) al eliminar complejidades y conjeturas en la fase de diseño. Su empresa puede comenzar con un sistema pequeño e ir creciendo sin interrupciones para gestionar de forma inteligente el movimiento de datos entre el perímetro, el núcleo y el cloud.

La Figura 1 muestra la escalabilidad de la solución ONTAP AI de NetApp. El sistema AFF A800 se ha validado con cuatro servidores DGX-1 y ha demostrado tener un margen de rendimiento suficiente para admitir cinco o más servidores DGX-1 sin que ello afecte al rendimiento del almacenamiento o a la latencia. Además, la solución puede escalarse añadiendo al clúster ONTAP switches de red y parejas de controladoras de almacenamiento, hasta llegar a ocupar varios racks y ofrecer un elevadísimo rendimiento que acelere el entrenamiento y la inferencia. Este enfoque permite alterar de forma independiente las ratios de computación y almacenamiento en función del tamaño del lago de datos, los modelos de aprendizaje profundo (DL) utilizados y las métricas de rendimiento necesarias.

Figura 1) Arquitectura de la solución ONTAP AI de NetApp en escala de rack.



El número de servidores DGX-1 y de sistemas AFF por rack depende de las especificaciones de energía 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 el diseño del centro de datos.

3 Canalización de datos de aprendizaje profundo

El aprendizaje profundo es el motor que le permite detectar fraudes, mejorar la relación con los clientes, optimizar la cadena de suministros y ofrecer productos y servicios innovadores en un mercado cada vez más competitivo. El rendimiento y la precisión de los modelos de aprendizaje profundo han mejorado significativamente al aumentar el tamaño y la complejidad de la red neuronal, además de la cantidad y la calidad de los datos utilizados para educar a estos modelos.

Dados los enormes conjuntos de datos que se manejan, es esencial diseñar una infraestructura con flexibilidad para la puesta en marcha en diversos entornos. A nivel general, una puesta en marcha de aprendizaje profundo integral consta de tres fases por las que viajan los datos: el perímetro (ingesta de datos), el núcleo (clústeres de entrenamiento y lago de datos) y el cloud (archivado, organización en niveles y desarrollo y pruebas). Es un diseño muy típico de aplicaciones como el Internet de las cosas (IoT), donde los datos abarcan los tres ámbitos de la canalización.

Figura 2: Canalización de datos del perímetro al núcleo y al cloud.

La Figura 2 presenta una imagen general de los componentes en cada uno de los tres ámbitos:

- **Ingesta de datos.** La ingesta de datos suele producirse en el perímetro, por ejemplo, con la captura de los datos transmitidos por coches autónomos o dispositivos de punto de venta (POS). En función del caso práctico, es posible que deba ubicarse una infraestructura tecnológica en el punto de ingesta o cerca de él. Por ejemplo, un comercio minorista podría necesitar una huella pequeña en cada tienda que consolide los datos de varios dispositivos.
- **Preparación de los datos.** El preprocesamiento es necesario para normalizar y purgar los datos antes del entrenamiento. Se realiza en un lago de datos, posiblemente en el cloud, en forma de un nivel Amazon S3 o en sistemas de almacenamiento en las instalaciones, como un almacén de archivos o de objetos.
- **Entrenamiento.** Durante la crucial fase de entrenamiento del aprendizaje profundo, suelen copiarse datos a intervalos periódicos entre el lago de datos y el clúster de entrenamiento. Los servidores empleados en esta fase utilizan las GPU para paralelizar los cálculos, lo que crea un tremendo apetito de datos. Satisfacer las necesidades brutas de ancho de banda E/S es esencial para poder mantener un uso de GPU elevado.
- **Inferencia.** Los modelos entrenados se someten a prueba y se implementan para producción. También pueden devolverse al lago de datos para realizar nuevos ajustes en los pesos de entrada. En las aplicaciones IoT, los modelos también podrían implementarse en los dispositivos inteligentes del perímetro.
- **Archivado, organización en niveles.** Los datos fríos de iteraciones anteriores se pueden guardar de forma indefinida. Muchos equipos de IA prefieren archivar datos fríos en un almacenamiento de objetos, ya sea en un cloud público o privado.

Dependiendo de la aplicación, los modelos de aprendizaje profundo trabajan con grandes cantidades de tipos de datos (estructurados y no estructurados). Esta diferencia impone distintos requisitos al sistema de almacenamiento subyacente en cuanto al tamaño de los datos que se guardan y el número de archivos en el conjunto de datos.

Estos son algunos requisitos del almacenamiento de alto nivel:

- Capacidad de almacenamiento y recuperación de millones de archivos de forma concurrente.
- Almacenamiento y recuperación de diversos objetos de datos, por ejemplo imágenes, audio, vídeo y series históricas.
- Alto rendimiento paralelo con baja latencia para igualar la velocidad de procesamiento de las GPU.
- Gestión de datos transparente y servicios de datos que abarquen el perímetro, el núcleo y el cloud.

Combinados con la integración superior del cloud y las capacidades definidas por software de NetApp ONTAP, los sistemas AFF admiten una amplia gama de canalizaciones de datos que abarcan el perímetro, el núcleo y el cloud para el aprendizaje profundo. Este documento se centra en las soluciones para los componentes de entrenamiento e inferencia de la canalización de datos.

4 Descripción general de la solución

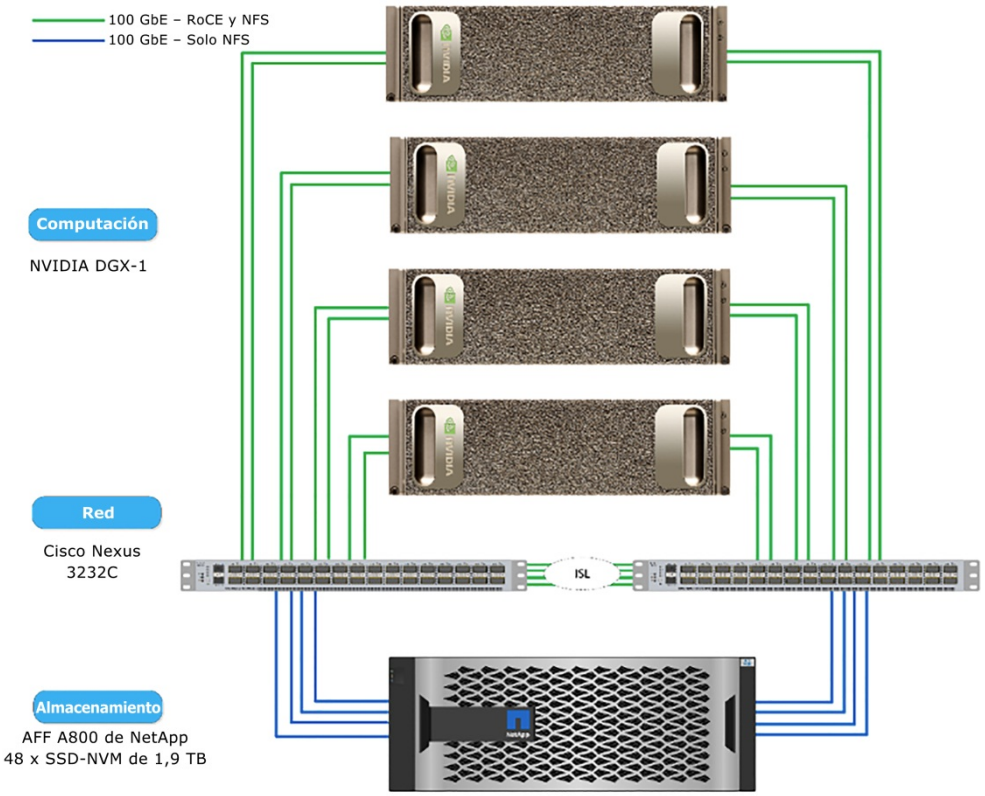
Los sistemas de aprendizaje profundo emplean algoritmos de cálculo intensivo y perfectamente adaptados a la arquitectura de las GPU de NVIDIA. Los cálculos realizados por los algoritmos de aprendizaje profundo suponen una inmensa cantidad de multiplicaciones de matrices que se ejecutan en paralelo. La 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 como el aprendizaje profundo, donde el procesamiento de datos se realiza en paralelo. Los avances en las arquitecturas de computación mediante GPU de NVIDIA (individuales o en clúster) que aprovechan el servidor DGX-1 las han convertido en las plataformas preferidas para cargas de trabajo como la computación de alto rendimiento (HPC), el aprendizaje profundo y el análisis. Para ofrecer un rendimiento maximizado en estos entornos es necesaria una infraestructura de soporte que pueda alimentar con datos las GPU de NVIDIA. Por tanto, el acceso a los conjuntos de datos debe proporcionarse con una latencia ultrabaja y un gran ancho de banda.

4.1 Tecnología de la solución

Esta solución se implementó con un sistema AFF A800 de NetApp, cuatro servidores NVIDIA DGX-1 y dos switches Ethernet Cisco Nexus 3232C 100 Gb. Cada servidor DGX-1 está conectado a los switches Nexus mediante cuatro conexiones 100 GbE que se utilizan para las comunicaciones entre GPU. Para ello se utilizan accesos remotos directos a la memoria (RDMA) sobre Ethernet convergente (RoCE). En estos enlaces también se producen las comunicaciones IP tradicionales para el acceso al almacenamiento NFS. Cada controladora de almacenamiento está conectado a los switches de red mediante cuatro enlaces 100 GbE.

Las infraestructuras HPC tradicionales utilizan RDMA en vez de InfiniBand (IB) para la conectividad entre nodos debido a su elevado ancho de banda y su baja latencia. Como la tecnología Ethernet alcanza niveles de rendimiento que antes solo eran posibles mediante IB, RoCE permite adoptar estas capacidades de forma más sencilla, ya que la tecnología Ethernet se conoce muy bien y está implementada en todos los centros de datos empresariales. La Figura 3 muestra la arquitectura básica de la solución.

Figura 3) Arquitectura validada de la solución ONTAP AI de NetApp.



4.2 Servidores DGX-1 de NVIDIA

El servidor DGX-1 es un sistema de hardware y software completo y plenamente integrado, pensado para los flujos de trabajo de aprendizaje profundo. Cada servidor DGX-1 dispone de ocho GPU Tesla V100 configuradas en una topología híbrida cubo-malla que utiliza tecnología NVIDIA NVLink™ para conseguir un ancho de banda ultraelevado y baja-latencia para la comunicación entre GPU. Esta topología es esencial para el entrenamiento de varias GPU, ya que elimina el cuello de botella asociado con las interconexiones basadas en PCIe, que no pueden ofrecer linealidad del rendimiento a medida que aumenta el número de GPU. El servidor DGX-1 también está equipado con interconexiones de red de gran ancho-de banda y baja latencia para el clustering de múltiples nodos en estructuras aptas para RDMA.

El DGX-1 dispone de la tecnología NVIDIA GPU Cloud (NGC), el registro de contenedores de NVIDIA basado en el cloud, para el software acelerado mediante GPU. NGC proporciona contenedores para los marcos de aprendizaje profundo más populares de hoy en día, como Caffe2, TensorFlow,

PyTorch, MXNet y TensorRT, que están optimizados para GPU de NVIDIA. Los contenedores integran el marco o aplicación, los controladores necesarios, bibliotecas y primitivos de comunicaciones y están optimizados para lograr el máximo rendimiento acelerado por GPU en el stack de NVIDIA. Los contenedores NGC incorporan el juego de herramientas NVIDIA CUDA, que proporciona la Biblioteca básica de subrutinas de álgebra lineal NVIDIA CUDA (cuBLAS), la Biblioteca de red neuronal profunda NVIDIA CUDA (cuDNN) y muchas cosas más. Los contenedores NGC también incluyen la Biblioteca de comunicaciones colectivas NVIDIA (NCCL) para primitivos de comunicación colectiva con varias GPU y varios nodos y que permite detectar la topología para el entrenamiento de DL. NCCL habilita la comunicación entre GPU dentro de un solo DGX-1 y entre varios servidores DGX-1.

4.3 Sistemas AFF de NetApp

AFF de NetApp es un sistema de almacenamiento de vanguardia que le permite satisfacer las necesidades de almacenamiento de su empresa con un rendimiento de primera, una flexibilidad superior, integración con el cloud y la mejor gestión de datos disponible. Los sistemas AFF han sido diseñados específicamente para flash y ayudan a acelerar, gestionar y proteger los datos esenciales para la empresa.

El sistema AFF A800 de NetApp es la primera solución NVMe integral del sector. En el caso de cargas de trabajo NAS, un solo sistema AFF A800 logra una salida de 25 GB/s para lecturas secuenciales y un millón de IOPS para pequeñas lecturas aleatorias con latencias inferiores a los 500 µs. Los sistemas AFF A800 tienen las siguientes características:

- Una enorme salida de hasta 300 GB/s y 11,4 millones de IOPS en un clúster de 24 nodos.
- Ethernet 100 Gb y conectividad con FC 32 Gb.
- Unidades de estado sólido (SSD) de 30 TB con escritura multistream (MSW).
- Alta densidad, con 2 PB en una bandeja de 2U.
- Escalado desde los 364 TB (2 nodos) hasta los 74 PB (24 nodos).
- ONTAP 9.4 de NetApp, con un completo conjunto de funciones de protección de datos y replicación que conforman la mejor gestión de datos del sector.

El segundo mejor sistema de almacenamiento en cuanto a rendimiento es el AFF A700s, con una salida de 18 GB/s para cargas de trabajo NAS y transporte 40 GbE. Los sistemas AFF A300 y AFF A220 ofrecen un rendimiento suficiente a menor precio.

4.4 ONTAP 9 de NetApp

ONTAP 9 es la última generación del software de gestión del almacenamiento de NetApp y le permite modernizar su infraestructura y realizar la transición a un centro de datos preparado para el cloud. ONTAP ofrece las mejores capacidades de gestión de datos y le permite administrarlos y protegerlos con un único conjunto de herramientas independientemente de dónde residan. Los datos también pueden trasladarse libremente allí donde sea necesario, ya sea al perímetro, al núcleo o al cloud. ONTAP 9 incluye numerosas funciones que simplifican la gestión de datos, aceleran y protegen los datos esenciales y preparan su infraestructura para el futuro con arquitecturas de cloud híbrido.

Simplificar la gestión de datos

La gestión de los datos es esencial para las operaciones tecnológicas de la empresa, de modo que se utilicen recursos apropiados para las aplicaciones y conjuntos de datos. ONTAP incluye las siguientes funciones para facilitar y simplificar las operaciones y para reducir el coste total de propiedad:

- **Compactación de datos inline y deduplicación expandida.** La compactación de datos reduce el espacio perdido dentro de los bloques de almacenamiento, mientras que la deduplicación aumenta la capacidad efectiva de forma significativa.
- **Calidad de servicio (QoS) mínima, máxima y adaptativa.** Los controles granulares de QoS le permiten mantener los niveles de rendimiento para aplicaciones críticas en entornos muy compartidos.

- **FabricPool de ONTAP.** Esta función divide automáticamente los datos fríos en niveles de cara a opciones de almacenamiento en cloud privado como Amazon Web Services (AWS), Azure y la solución NetApp StorageGRID®.

Acelere y proteja sus datos

ONTAP no solo ofrece niveles de rendimiento y protección de datos superiores, sino que aumenta estas capacidades con:

- **Rendimiento y menor latencia.** ONTAP ofrece la salida más alta posible con la menor latencia posible.
- **ONTAP FlexGroup de NetApp.** Un volumen FlexGroup es un contenedor de datos de alto rendimiento que puede escalarse verticalmente de forma lineal hasta los 20 PB y los 400 000 millones de archivos y que proporciona un solo espacio de nombres que simplifica la gestión de los datos.
- **Protección de datos.** ONTAP ofrece capacidades integradas de protección de datos, con una administración común entre todas las plataformas.
- **Cifrado de volúmenes de NetApp.** ONTAP ofrece cifrado nativo en el nivel de volumen y permite la gestión de claves incorporada o externa.

Infraestructura preparada para futuros retos

ONTAP 9 le ayuda a satisfacer las exigentes y siempre cambiantes necesidades de su empresa:

- **Escalado sencillo y funcionamiento sin interrupciones.** ONTAP permite aumentar la capacidad de las controladoras existentes y escalar clústeres horizontalmente sin interrupciones. Puede empezar a utilizar tecnologías punteras como NVMe y FC 32 Gb sin necesidad de realizar costosas migraciones de datos y sin cortes.
- **Conexión de cloud.** ONTAP es el software de gestión del almacenamiento con mejor conexión de cloud e incluye opciones de almacenamiento definido por software (ONTAP Select) e instancias nativas del cloud (NetApp Cloud Volumes Service) en todos los clouds públicos.
- **Integración con aplicaciones emergentes.** ONTAP proporciona servicios de datos de clase empresarial para plataformas y aplicaciones de última generación como OpenStack, Hadoop y MongoDB, y utiliza para ello la misma infraestructura que da soporte a las aplicaciones empresariales existentes.

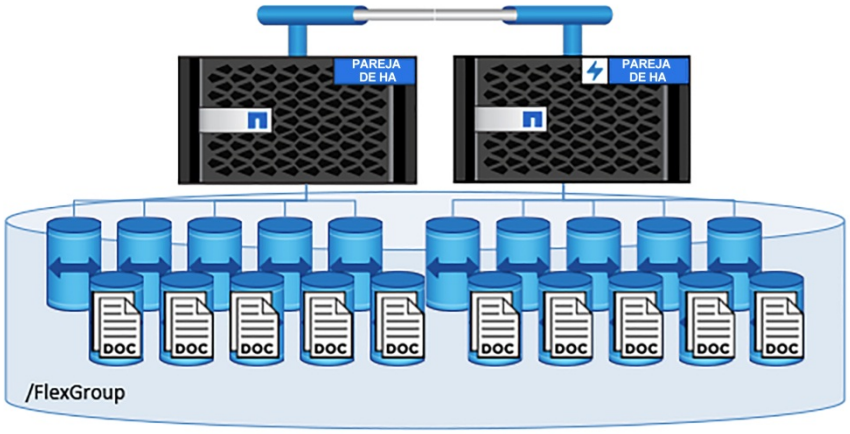
4.5 FlexGroup Volumes de NetApp

El conjunto de datos de entrenamiento suele consistir en una colección con un gran número de archivos (hasta miles de millones). Pueden ser archivos de texto, de audio, de vídeo o cualquier otra forma de datos no estructurados que deban almacenarse y procesarse para su lectura en paralelo. El sistema de almacenamiento debe guardar un gran número de pequeños archivos (potencialmente, miles de millones) y debe leerlos en paralelo, con una entrada y salida secuencial o aleatoria.

Un volumen FlexGroup (Figura 4) es un único espacio de nombres compuesto por múltiples volúmenes constituyentes y que se gestiona y actúa como un volumen NetApp FlexVol® de cara a los administradores de almacenamiento. Los archivos de un volumen de FlexGroup se asignan a volúmenes miembro individuales y no están repartidos en volúmenes o nodos. Ofrecen las siguientes capacidades:

- Los volúmenes FlexGroup tienen una enorme capacidad (varios petabytes) y una baja latencia constante para cargas de trabajo con una gran cantidad de metadatos.
- Permiten disponer de cientos de millones de archivos en un mismo espacio de nombres.
- Admiten operaciones en paralelo para cargas de trabajo NAS entre varias CPU, nodos, agregados y volúmenes FlexVol constituyentes.

Figura 4) Volúmenes FlexGroup de NetApp.



4.6 NVIDIA GPU Cloud y Trident

NVIDIA GPU Cloud (NGC) proporciona un catálogo de imágenes Docker totalmente integradas y orientadas al rendimiento para que el aprendizaje profundo aproveche al máximo las ventajas de las GPU de NVIDIA. Estas imágenes incluyen todas las dependencias necesarias, como el juego de herramientas NVIDIA CUDA y las bibliotecas DL de NVIDIA. NVIDIA ha probado, ajustado y certificado estas imágenes para su uso en servidores NVIDIA DGX-1. Además, para que pueda portar las imágenes que aprovechan las GPU, NVIDIA ha desarrollado NVIDIA Container Runtime para Docker, que permite montar en este contenedor durante el inicio los componentes del modo de usuario de los controladores de NVIDIA y las GPU.

Trident, de NetApp, es un aprovisionador de almacenamiento dinámico de código abierto para Docker y Kubernetes. En combinación con NGC y con orquestadores populares como Kubernetes o Docker Swarm, Trident le permite implementar sin dificultades imágenes de contenedores DL NGC en almacenamiento de NetApp, logrando así una experiencia de nivel empresarial para la implementación de contenedores de IA. Estas implementaciones incluyen orquestación automatizada, clonado para prueba y desarrollo, pruebas mejoradas que utilizan clonado, protección y copias de cumplimiento normativo y muchos otros casos de uso de gestión de datos para las imágenes de contenedores NGC para IA y DL.

4.7 Switches de red Cisco Nexus 3232C

El switch Cisco Nexus 3232C (Figura 5) es un switch de 100 Gb/s de baja latencia, alta densidad, alto rendimiento y alta eficiencia energética, diseñado para el centro de datos. Este modelo compacto de una unidad de rack (1 RU) ofrece conmutación de capa 2 y capa 3 con velocidad de cable para todos los puertos y con una latencia de 450 ns. Es un switch que pertenece a la plataforma Cisco Nexus 3200 y ejecuta el reconocido sistema operativo Cisco NX-OS, lo que proporciona características y funciones completas y de amplia implementación. El Cisco Nexus 3232C es un switch Quad Small Form-Factor Pluggable (QSFP) y dispone de 32 puertos QSFP28. Cada puerto QSFP28 puede funcionar a 10, 25, 40, 50 y 100 Gb/s, hasta un máximo de 128 puertos a 25 Gb/s.

Figura 5) Switches Cisco Nexus compatibles con NX-OS para los estándares de Ethernet convergente mejorada y RoCE v1 y v2.

Las pruebas demuestran que esta solución consume solo la mitad de los puertos disponibles en cada switch de red. Cada switch admite hasta ocho servidores DGX-1, con puertos de acceso de almacenamiento adicionales si se precisa más potencia de GPU. Para implementaciones aún mayores, Cisco Nexus 7000 admite hasta 192 puertos de 100 GbE por switch. Como alternativa podría adoptarse una topología hoja-rama, con múltiples pares de switches Nexus 3000 conectados a un switch-rama central.

4.8 RDMA sobre Ethernet convergente

El acceso directo a la memoria (DMA) permite que subsistemas hardware como controladoras de unidades de disco, tarjetas de sonido, tarjetas gráficas y tarjetas de red accedan a la memoria del sistema para realizar operaciones de lectura y escritura sin necesidad de consumir ciclos de procesamiento de CPU. RDMA amplía esta capacidad al permitir que los adaptadores de red realicen transferencias de datos servidor a servidor entre la memoria de aplicaciones, empleando para ello una funcionalidad de copia cero sin participación alguna del sistema operativo o el controlador del dispositivo. Este enfoque reduce enormemente la sobrecarga de la CPU y la latencia al omitir el kernel en las operaciones de lectura/escritura y envío/recepción.

RoCE es la implementación más habitual de RDMA sobre Ethernet y hace uso de los nuevos estándares de Ethernet convergente mejorada (CEE). Ya está disponible como una característica estándar en muchos adaptadores de red de alto nivel, adaptadores de red convergente y switches de red. La Ethernet tradicional utiliza un mecanismo de “entrega de mejor esfuerzo” para el tráfico de red y no es apropiada para las bajas latencias y el elevado ancho de banda necesarios para la comunicación entre nodos de GPU. CEE habilita un medio de red de capa física sin pérdidas y permite la posibilidad de asignar ancho de banda a un flujo de tráfico específico en la red.

Para garantizar la entrega en orden y sin pérdidas de paquetes Ethernet, las redes CEE utilizan control del flujo prioritario (PFC) y selección de transmisión mejorada (ETS). PFC permite el envío de ventanas de pausa para cada clase de servicio (CoS) específica, por lo que puede limitar un tráfico de red específico mientras deja que otro fluya libremente. ETS permite asignar ancho de banda específico para cada CoS, lo que le proporciona un control todavía más granular del uso de la red.

La capacidad de dar prioridad a RoCE sobre el resto del tráfico permite utilizar los enlaces de 100 GbE tanto para RoCE como para el tráfico IP tradicional, por ejemplo, el tráfico de acceso al almacenamiento NFS que se utiliza en esta solución.

5 Requisitos tecnológicos

En este apartado se trata el hardware y el software empleados en la validación de esta solución. Todas las pruebas documentadas en el apartado 7, Validación de la solución, se realizaron con el hardware y el software que aquí se indican.

Nota: La configuración validada en esta arquitectura de referencia se basa en la disponibilidad de equipo de laboratorio y no en los requisitos o las limitaciones del hardware sometido a prueba.

5.1 Requisitos de hardware

La Tabla 1 enumera los componentes de hardware utilizados para validar la solución. Los componentes que usted utilice en cualquier implementación particular de esta solución pueden variar en función de sus necesidades.

Tabla 1) Requisitos de hardware.

Hardware	Cantidad
Servidores NVIDIA DGX-1 GPU	4
Sistema AFF A800 de NetApp	1 pareja de alta disponibilidad (HA), incluye 48 SSD NVMe de 1,92 TB
Switches de red Cisco Nexus 3232C	2

5.2 Requisitos de software

La Tabla 2 enumera los componentes de software necesarios para implementar la solución. Los componentes que usted utilice en cualquier implementación particular de la solución pueden variar en función de sus necesidades.

Tabla 2) Requisitos de software.

Software	Versión
ONTAP de NetApp	9,4
Firmware de switch Cisco NX-OS	7.0(3)I6(1)
Sistema operativo DGX-1 de NVIDIA	Sistema operativo Ubuntu 16.04 LTS
Plataforma contenedora Docker	18.03.1-ce [9ee9f40]
Versión del contenedor	netapp_1.7.0.2 basada en nvcr.io/nvidia/tensorflow:18.04-py2
Marco de aprendizaje automático	TensorFlow 1.7.0
Horovod	0.11.3
OpenMPI	3.1.0
Software de pruebas de rendimiento	Pruebas de TensorFlow [1b1ca8a]

6 Arquitectura de la solución

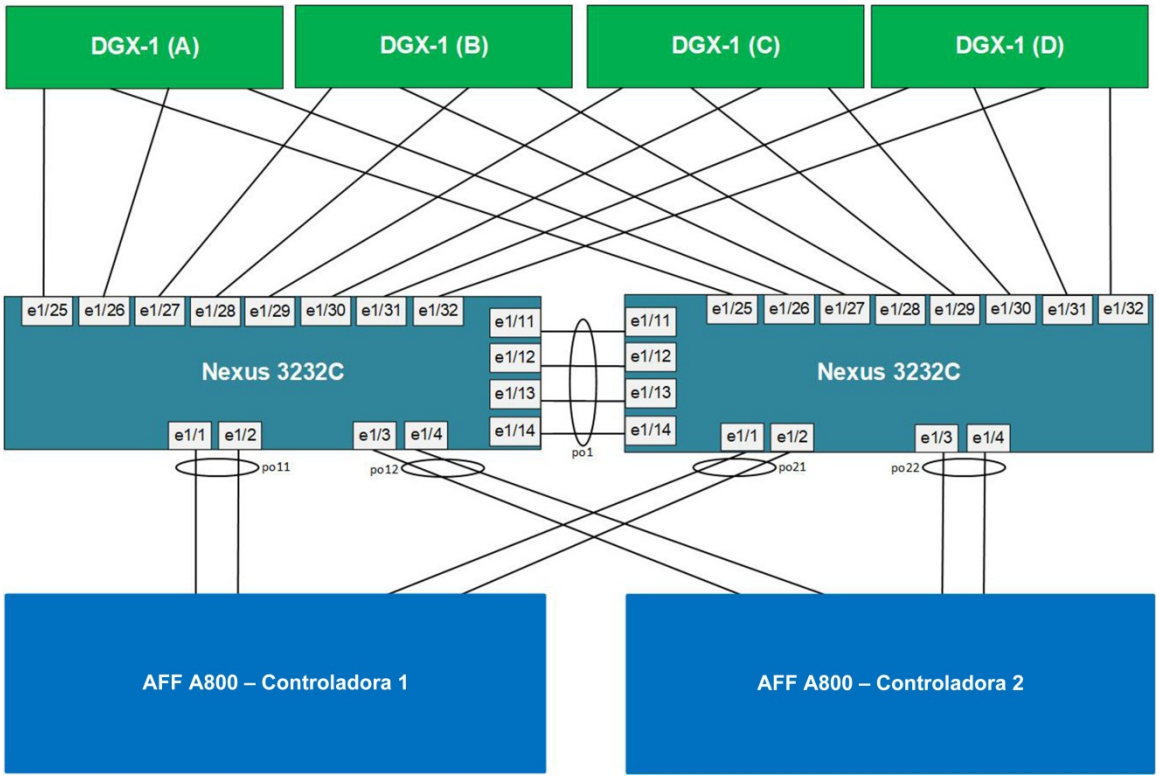
Se ha validado que esta arquitectura cumple los requisitos para la ejecución de cargas de trabajo de aprendizaje profundo. Esta validación permite a los expertos en datos implementar marcos y aplicaciones de aprendizaje profundo en una infraestructura prevalidada, lo que ayuda a eliminar riesgos y permite a la empresa centrarse en la obtención de información valiosa a partir de sus datos. Además, esta arquitectura también puede ofrecer un rendimiento del almacenamiento excepcional para otras cargas de trabajo HPC sin necesidad de realizar ninguna modificación o de ajustar la infraestructura.

6.1 Topología de la red y configuración del switch

Esta solución emplea RoCE en vez de IB para lograr la conectividad de elevado ancho de banda y baja latencia necesaria para la comunicación entre servidores DGX-1. Los switches Cisco Nexus son compatibles con RoCE al implementar la tecnología PFC, que permite a los usuarios dar prioridad al tráfico RoCE sobre el tráfico IP tradicional en un vínculo compartido, así como utilizar los enlaces de 100 GbE al mismo tiempo para RoCE y para IP.

Esta arquitectura utiliza un par de switches Cisco Nexus 3232C Ethernet 100 Gb para la principal red interclúster y de acceso al almacenamiento. Se conectan entre ellos mediante cuatro puertos de red de 100 Gb configurados como un canal de puerto estándar. Este canal de puerto de enlace entre switches (ISL) permite que el tráfico fluya entre los switches cuando se produce un fallo en el enlace con el host o con el sistema de almacenamiento. Cada host se conecta a los switches Nexus mediante un par de vínculos activo-pasivo y, para proporcionar redundancia en la capa de enlace, cada controladora de almacenamiento se conecta a cada switch Nexus mediante un canal de puerto LACP de dos puertos. La Figura 6 muestra la configuración de switches y puertos de la red.

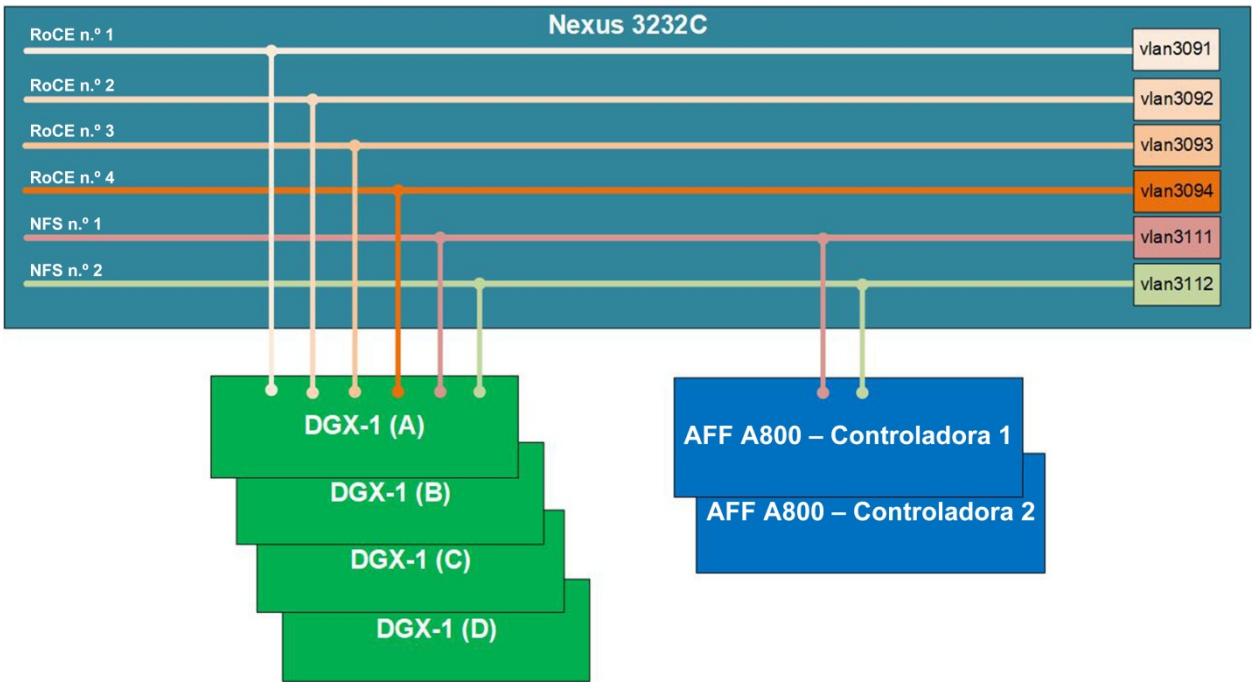
Figura 6) Configuración de switches y puertos de la red.



Se han aprovisionado varias LAN virtuales (VLAN) para dar servicio tanto al tráfico de RoCE como al del almacenamiento NFS. Hay cuatro VLAN dedicadas al tráfico RoCE y dos dedicadas al tráfico del almacenamiento NFS. Se utilizan cuatro VLAN discretas y otros tantos intervalos IP para proporcionar enrutamiento simétrico para cada conexión RoCE. El stack de software NVIDIA gestiona la agregación de ancho de banda y la tolerancia a fallos en estas conexiones. Para el acceso al almacenamiento, esta solución utiliza NFSv3, que no admite el acceso multivía, por lo que se utilizan dos VLAN para permitir la existencia de varios montajes NFS dedicados. Este enfoque no proporciona ninguna tolerancia a fallos adicional, pero sí permite el uso de varios enlaces con el fin de aumentar el ancho de banda disponible. PFC se configura en los switches de modo que las cuatro VLAN para RoCE se asignan a la clase prioritaria, mientras que las VLAN para NFS se asignan a la clase de mejor esfuerzo predeterminada. Todas las VLAN se configuran para marcos jumbo, con un tamaño máximo de unidad de transmisión (MTU) de 9000.

Los puertos de los switches para los servidores DGX-1 se configuran como puertos troncales y se permiten todas las VLAN para RoCE y NFS. Los puertos-canales configurados para las controladoras del sistema de almacenamiento también son troncales, pero solo se permiten las VLAN para NFS. La Figura 7 muestra la conectividad VLAN para el servidor DGX-1 y los puertos del sistema de almacenamiento.

Figura 7) Conectividad VLAN para DGX-1 y los puertos del sistema de almacenamiento.



Con el fin de proporcionar un servicio prioritario al tráfico RoCE, el adaptador de red del host asigna un valor CoS de 4 al tráfico de las VLAN para RoCE. El switch se configura con una política QoS que proporciona servicio sin pérdidas para el tráfico con este valor de CoS. Al tráfico NFS se le asigna el valor CoS predeterminado de 0, lo que se corresponde con la política QoS predeterminada del switch y proporciona un servicio de mejor esfuerzo.

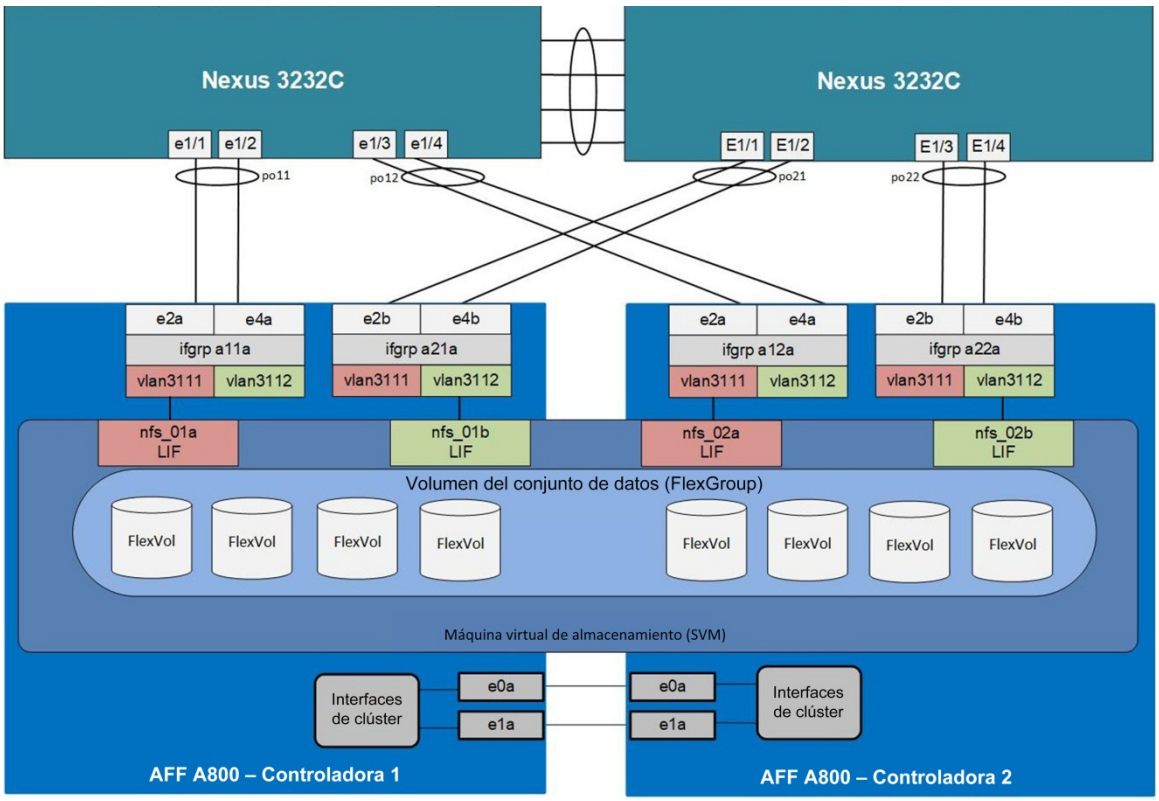
A continuación, se habilita PFC en todos los puertos DGX-1, lo que les permite enviar ventanas de pausa para clases de servicio específicas con el fin de eliminar las congestiones en el switch. Al utilizar ETS para asignar el 95 % del ancho de banda al tráfico RoCE en caso de congestión, esta configuración permite una asignación de recursos dinámica entre el tráfico RoCE y NFS, al tiempo que se da prioridad a la comunicación entre nodos. También se puede modificar de forma dinámica la asignación de ancho de banda si se la quiere optimizar para cargas de trabajo que requieran un mayor rendimiento de almacenamiento y menor comunicación entre nodos.

6.2 Configuración del sistema de almacenamiento

Con el fin de satisfacer los requisitos de red de almacenamiento de cualquier posible carga de trabajo en esta arquitectura, cada controladora de almacenamiento está aprovisionada con cuatro puertos de 100 GbE además de los puertos incorporados necesarios para la interconexión del clúster de almacenamiento. La Figura 8 muestra la configuración del sistema de almacenamiento. Cada controladora está configurada con un grupo de interfaces LACP de dos puertos (ifgrp en la Figura 8) para cada switch. Estos grupos de interfaces proporcionan a cada switch hasta 200 Gb/s de conectividad resiliente para el acceso a datos. Para el acceso al almacenamiento NFS se aprovisionan dos VLAN que nacen en los switches y se conectan con cada uno de estos grupos de interfaces. Esta configuración permite el acceso concurrente a los datos mediante varias interfaces desde cada uno de los hosts, lo que mejora el ancho de banda potencial del que disponen.

Todos los accesos a datos desde el sistema de almacenamiento se realizan mediante el acceso NFS desde una máquina virtual de almacenamiento (SVM) dedicada a esta carga de trabajo. La SVM está configurada con un total de cuatro interfaces lógicas (LIF), dos para cada VLAN de almacenamiento. Cada grupo de interfaces hospeda un LIF, lo que resulta en un LIF por VLAN en cada controladora con un grupo de interfaces dedicado para cada VLAN. Sin embargo, ambas VLAN están conectadas a ambos grupos de interfaz en cada controladora. Esta configuración permite que cada LIF pueda conmutar por error a otro grupo de interfaces en la misma controladora, de modo que ambas controladoras permanezcan activas en caso de un fallo de red.

Figura 8) Configuración del sistema de almacenamiento.

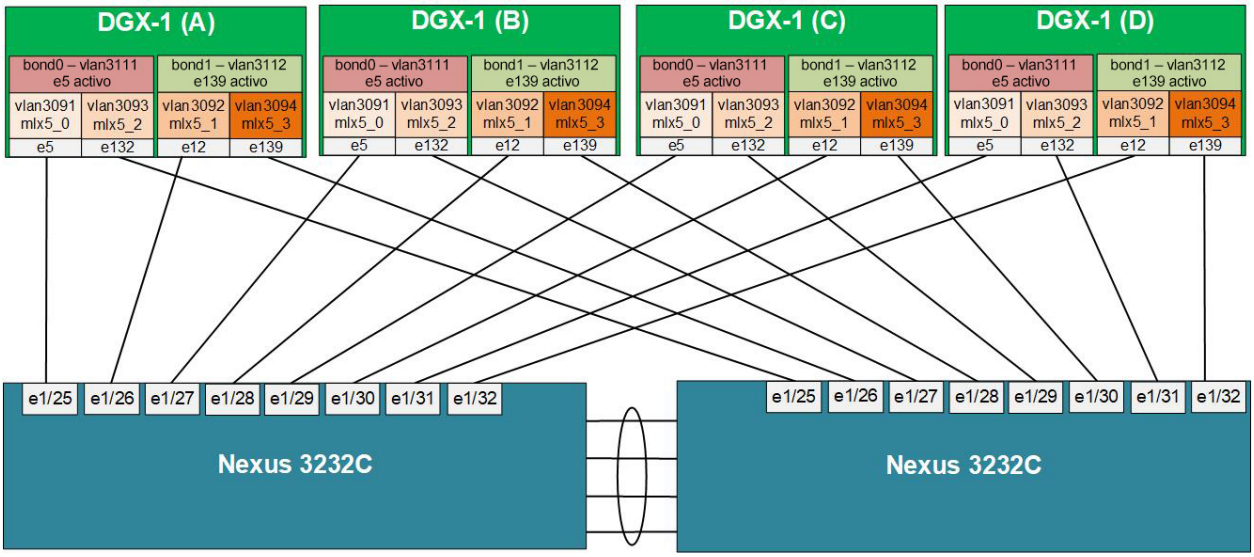


Para el aprovisionamiento de almacenamiento lógico, la solución utiliza un volumen FlexGroup. Dicho volumen proporciona un único grupo de almacenamiento que se distribuye entre los nodos del clúster de almacenamiento. Cada controladora aloja un agregado de 46 particiones de disco; ambas controladoras comparten todos los discos. Cuando el FlexGroup se implementa en la SVM de datos, en cada agregado se aprovisionan varios volúmenes FlexVol que luego se combinan en el FlexGroup. Con este enfoque, el sistema de almacenamiento puede proporcionar un único grupo de almacenamiento que puede escalarse verticalmente hasta la máxima capacidad de la cabina, lo que proporciona un rendimiento excepcional al aprovecharse todos los SSD de la cabina de forma concurrente. Los clientes NFS pueden acceder al FlexGroup como un solo punto de montaje mediante cualquiera de los LIF aprovisionados para la SVM. Se puede aumentar la capacidad y el ancho de banda de acceso a clientes con solo añadir más nodos al clúster de almacenamiento.

6.3 Configuración de host

Para la conectividad de red, cada DGX-1 está aprovisionado con cuatro tarjetas de interfaz de red de un puerto Mellanox ConnectX4. Estas tarjetas operan a velocidad Ethernet de hasta 100 Gb y son compatibles con RoCE, por lo que suponen una alternativa más barata que IB para las aplicaciones de interconexión de clústeres. Cada puerto 100 Gb se configura como un puerto troncal en el switch apropiado, con cuatro VLAN RoCE y dos VLAN NFS en cada uno. La Figura 9 muestra la configuración de puertos de red y VLAN en los hosts DGX-1.

Figura 9) Configuración de puertos de red y VLAN en los hosts DGX-1.



Para la conectividad RoCE, cada puerto físico hospeda una interfaz VLAN y una dirección IP en una de las cuatro VLAN para RoCE. Los controladores Mellanox se configuran para aplicar un valor CoS de red de 4 a cada VLAN para RoCE. PFC se configura en los switches para garantizar a la clase RoCE un servicio prioritario y sin pérdidas. RoCE no admite la agregación de varios enlaces en una única conexión lógica, pero el software de comunicación NVIDIA NCCL puede utilizar varios enlaces para agregar ancho de banda y ofrecer tolerancia a fallos.

Para el acceso al almacenamiento NFS, se utiliza un enlace a cada switch para crear dos vínculos activo-pasivo. Cada vínculo hospeda una interfaz VLAN y una dirección IP en una de las dos VLAN de NFS y cada puerto activo del vínculo está conectado a un switch diferente. Esta configuración proporciona hasta 100 Gb de ancho de banda en cada VLAN de NFS y proporciona redundancia en caso de cualquier fallo de switch o de enlace con el host. Con el fin de proporcionar un rendimiento óptimo para las conexiones RoCE, todo el tráfico NFS se asigna a la clase QoS de mejor esfuerzo predeterminada. Todas las interfaces físicas y las interfaces de vínculo están configuradas con un valor MTU de 9000.

Para aumentar el rendimiento del acceso a los datos, se realizan varios montajes NFSv3 entre el servidor DGX-1 y el sistema de almacenamiento. Cada servidor DGX-1 está configurado con dos VLAN de NFS, con una interfaz IP en cada VLAN. El volumen FlexGroup en el sistema AFF A800 se monta en cada una de estas VLAN de cada DGX-1, lo que proporciona conexiones totalmente independientes entre el servidor y el sistema de almacenamiento. Aunque un solo montaje NFS es capaz de proporcionar el rendimiento necesario para esta carga de trabajo, se definen varios puntos de montaje para que otras cargas que requieran un uso más intensivo del almacenamiento puedan disponer de ancho de banda adicional.

7 Validación de la solución

Este apartado describe las pruebas que realizamos para validar el funcionamiento y el rendimiento de esta solución. Desarrollamos todas las pruebas que se describen aquí con el equipo específico y el software enumerados en el apartado 5, Requisitos tecnológicos.

7.1 Plan de la prueba de validación

La solución se validó mediante pruebas estándar con distintas configuraciones de computación, con el fin de demostrar la escalabilidad de la arquitectura. El conjunto de datos ImageNet se hospedó en el sistema AFF A800 utilizando un único volumen FlexGroup al que accedían mediante NFSv3 hasta cuatro servidores DGX-1, tal y como recomienda NVIDIA para el acceso a almacenamiento externo. Se utilizó TensorFlow como marco de aprendizaje automático para todos los modelos probados. Para cada caso de prueba se capturaron las métricas de rendimiento de computación y almacenamiento. En el apartado 7.2, Resultados de la prueba de validación, se ofrece un resumen de los datos.

Para demostrar las tasas de entrenamiento se utilizaron los siguientes modelos de red neuronal de convolución (CNN), con diversos grados de complejidad de computación y almacenamiento:

- **ResNet-152** suele considerarse el modelo de entrenamiento más preciso.
- **ResNet-50** ofrece mayor precisión que AlexNet, con un tiempo de procesamiento más rápido.
- **VGG16** produce la mayor comunicación entre GPU.
- **Inception-v3** es otro modelo habitual para TensorFlow.

Se puso a prueba cada uno de estos modelos con distintas configuraciones de hardware y software con el fin de estudiar el efecto de cada opción sobre el rendimiento:

- Probamos cada modelo tanto con datos sintéticos como con el conjunto de datos ImageNet de referencia. Otras pruebas con GPU adicionales, tanto internas en un DGX-1 como entre varios servidores DGX-1, nos ayudaron a evaluar la escalabilidad del clúster de computación y el rendimiento del acceso al almacenamiento.
- Utilizamos datos ImageNet con la distorsión desactivada para reducir la sobrecarga de procesamiento de la CPU antes de copiar los datos en la memoria de la GPU.
- Probamos cada modelo con núcleos tensores y núcleos CUDA para demostrar las mejoras de rendimiento que se obtienen con los primeros.
- El aumento del rendimiento de la GPU también tuvo el efecto de aumentar los requisitos de acceso al almacenamiento y demostró la capacidad del sistema AFF A800 de satisfacer sin problemas dichos requisitos.
- Probamos cada modelo de aprendizaje profundo con distintos tamaños de lote. El aumento del tamaño del lote tiene diversos efectos en el sistema. El resultado final es una mayor tasa global de entrenamiento, una reducción en los requisitos de comunicación entre GPU y un aumento en los requisitos de ancho de banda del almacenamiento. Probamos cada modelo con los siguientes tamaños de lote:
  - 64, 128 y 256 para ResNet-50.
  - 64 y 128 para los demás modelos.
- Cada modelo se probó con uno, dos y cuatro servidores DGX-1 para demostrar la escalabilidad de los modelos con varias GPU que utilicen RoCE para su interconexión (mediante Horovod).
- La inferencia se ejecutó utilizando todos los modelos con su tamaño de lote máximo (256 para ResNet-50 y 128 para todos los demás), con 32 GPU (núcleos tensores y núcleos CUDA) y el conjunto de datos ImageNet.
- Todas las métricas de rendimiento se obtuvieron pasadas al menos dos épocas. Observamos unos resultados de rendimiento ligeramente superiores al ejecutar el entrenamiento a lo largo de varias épocas. Cada prueba se realizó cinco veces y se anotó la media de las métricas de rendimiento observadas.

7.2 Resultados de la prueba de validación

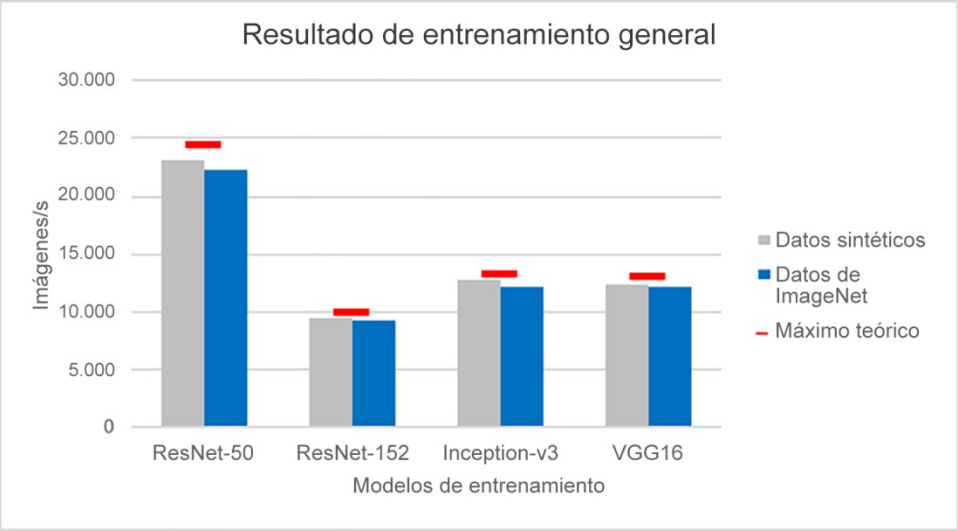
Como se ha descrito, realizamos distintas pruebas para valorar el funcionamiento general y el rendimiento de la solución. Este apartado contiene el resumen de los datos de rendimiento de computación y almacenamiento obtenidos durante las pruebas. El apéndice incluye los resultados completos y detallados. Deben tenerse en cuenta los siguientes detalles sobre los datos que se presentan en los siguientes subapartados de este informe:

- El rendimiento del entrenamiento de los modelos se mide en imágenes por segundo.
- Para el rendimiento del almacenamiento se miden la salida (MB/s) y la latencia (µs). También se monitorizó el uso de CPU del sistema de almacenamiento para evaluar la capacidad de rendimiento restante en dicho sistema.
- Cada sistema se sometió a prueba con varios tamaños de lote. Los tamaños de lote mayores aumentan el resultado de general del entrenamiento. Aquí solo se muestra la prueba de cada modelo con el tamaño de lote mayor. En el apéndice se encuentran los datos correspondientes a todos los tamaños de lote utilizados en la prueba:
  - En las pruebas de ResNet-50 se utilizó un tamaño de lote de 256.
  - En las pruebas de ResNet-152, Inception-v3 y VGG16 se utilizó un tamaño de lote de 128.

Resultado de entrenamiento general

La Figura 10 muestra el número máximo de imágenes por segundo de entrenamiento que se obtuvo con cada modelo probado, utilizando núcleos tensores para lograr el máximo rendimiento. La Figura 10 compara el resultado de entrenamiento obtenido con 32 GPU, utilizando como base datos ImageNet y datos sintéticos. También muestra el máximo teórico que es posible alcanzar cuando todas las GPU entrenan datos sintéticos de forma independiente y sin actualizar sus mutuos parámetros. Como se muestra en la Figura 10, la salida obtenida para los datos de ImageNet es muy similar a la obtenida con datos sintéticos.

Figura 10) Resultado de entrenamiento para todos los modelos.

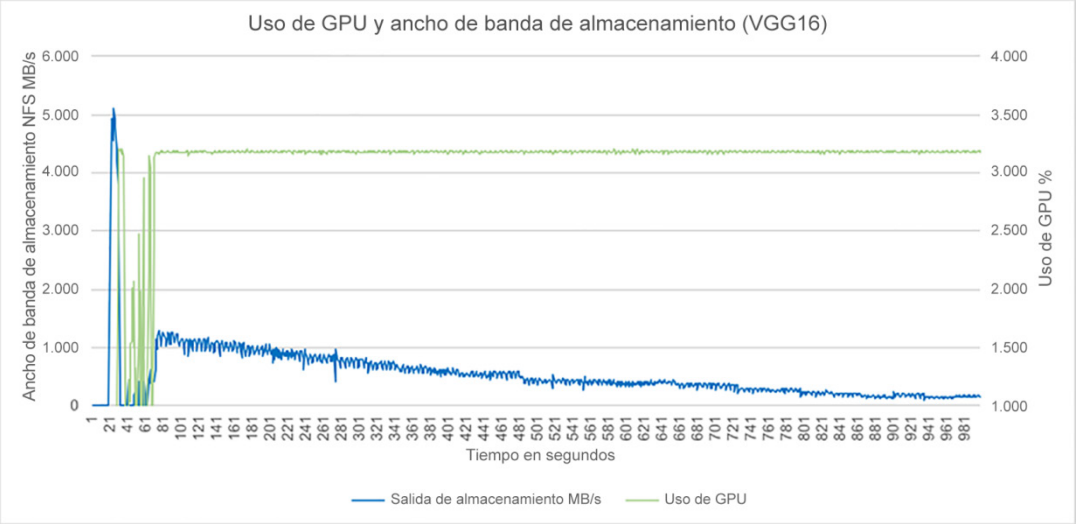


Rendimiento de carga de trabajo de la GPU

El siguiente conjunto de datos demuestra la capacidad del sistema de almacenamiento para satisfacer los requisitos del servidor DGX-1 con una carga completa. La Figura 11 muestra el uso que los servidores DGX-1 hacen de la GPU y el ancho de banda del almacenamiento que se genera cuando se ejecuta cada modelo utilizando 32 GPU. Como se aprecia en la gráfica, el ancho de banda del almacenamiento comienza muy alto cuando los datos iniciales se leen en el almacenamiento y pasan a la caché de canalización de TensorFlow, para luego caer gradualmente a medida que más partes del conjunto de datos van pasando a residir en la memoria local del servidor DGX-1.

Cuando todos los datos están en la memoria local, el acceso al almacenamiento se reduce hasta casi cero. Las GPU del servidor DGX-1 comienzan a procesar datos casi inmediatamente. El uso de GPU permanece constante durante toda la prueba. Esta gráfica muestra los resultados para el modelo VGG16 con un tamaño de lote de 128, lo que produjo el nivel más elevado de uso de GPU de la prueba. Las gráficas para los demás modelos están disponibles en el Apéndice. Debe tenerse en cuenta que la escala de uso de la GPU es la suma del uso de todas las GPU, por lo que en este caso en que se probaron 32 GPU, el uso máximo posible es del 3200 %.

Figura 11) Uso de la GPU y del ancho de banda del almacenamiento (VGG16).



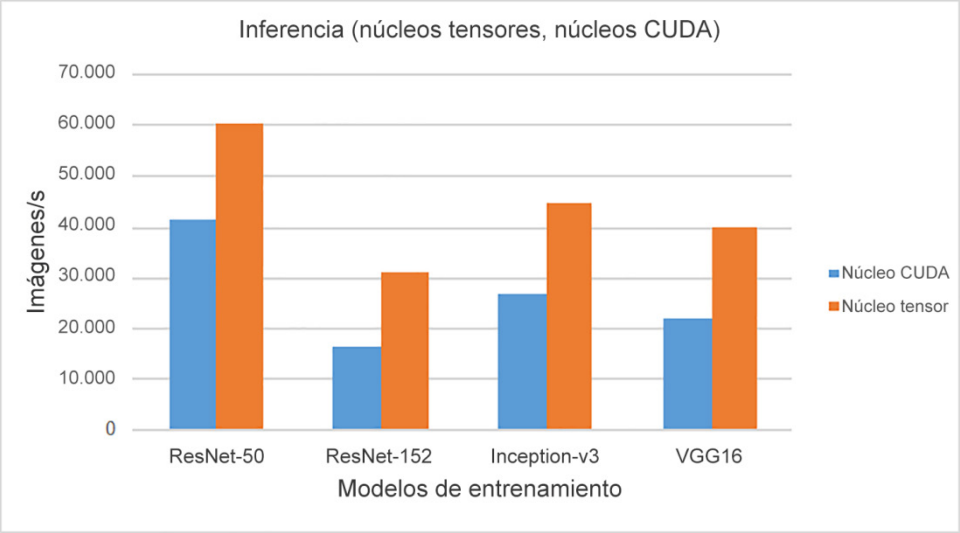
Como muestra la Figura 11, el uso de la GPU permanece por encima del 95 % para las 32 GPU. También es consistente sin importar cuántos datos lleguen del sistema de almacenamiento. El sistema de almacenamiento produce en un primer momento 5 GB/s de datos, para caer hasta los 2 GB/s y luego casi hasta cero durante el resto de la época de entrenamiento. Este resultado demuestra que el acceso al almacenamiento no es un cuello de botella para el rendimiento de la GPU con esta carga de trabajo. Con un conjunto de datos mayor que excediera la capacidad de memoria local, el rendimiento de acceso al almacenamiento permanecería en una tasa constante hasta las fases posteriores de la época de entrenamiento. Además, la Figura 11 compara el uso de la GPU en función del ancho de banda del almacenamiento. No refleja el tiempo necesario para la fase de entrenamiento completa, ya que el ancho de banda del almacenamiento se reduce gradualmente hasta casi cero a medida que avanza dicha fase.

Inferencia con las GPU

La inferencia es el proceso de implementar el modelo de aprendizaje profundo para valorar un nuevo conjunto de objetos y realizar predicciones con una precisión similar a la observada durante las fases de entrenamiento. En una aplicación con un conjunto de datos de imagen, el objetivo de la inferencia es clasificar las imágenes de entrada y responder a los solicitantes lo más rápido posible. Además de lograr una salida elevada, también es importante la minimización de la latencia.

Se utilizó ONTAP AI de NetApp para realizar la demostración de la inferencia y para medir la salida durante esta fase. La Figura 12 muestra el número de imágenes por segundo que es posible procesar durante la inferencia. Esta prueba compara la salida obtenida con 32 GPU utilizando datos ImageNet para cada uno de los modelos y empleando tanto núcleos tensores como núcleos CUDA. Gracias a la potencia de ONTAP AI de NetApp, se pueden utilizar núcleos tensores para clasificar un número significativo de imágenes de manera instantánea.

Figura 12) Inferencia para todos los modelos (núcleos tensores y núcleos CUDA).



Rendimiento del sistema AFF A800 con cargas de trabajo de entrenamiento de IA

Se capturó el ancho de banda del almacenamiento, la latencia y el margen de CPU para comprobar el rendimiento del sistema de almacenamiento con cada uno de los modelos probados. La Figura 13 y la Figura 15 muestran las métricas del sistema de almacenamiento para cada modelo probado con datos reales. Estas pruebas, centradas en el almacenamiento, se realizaron con tamaños de lote mayores para aumentar la carga de trabajo de almacenamiento y poner a prueba el escenario menos favorable.

Cabe señalar que, en todas las métricas, la carga de trabajo total generada en cada uno de los modelos con 32 GPU queda holgadamente dentro de los límites de capacidad del sistema AFF A800. Como marco de referencia para la carga de trabajo de entrenamiento, se generó una carga artificial utilizando E/S flexible (fio) con un perfil de E/S de lectura secuencial de 64K. Para la carga de trabajo generada con fio, la salida obtuvo un pico de 15 GB/s, la latencia de lectura se mantuvo bien por debajo de 1 ms y el uso de la CPU permaneció levemente por debajo del 50 %. Para conseguir la máxima salida posible con el número limitado de servidores DGX-1 disponibles, se utilizaron en los servidores montajes NFS adicionales y varios trabajos fio.

Nota: Se ha comprobado que una pareja de alta disponibilidad de AFF A800 de NetApp con cargas de trabajo NAS logran una salida de hasta 25 GB/s con una latencia inferior a 1 ms.

Figura 13) Ancho de banda del almacenamiento para todos los modelos.

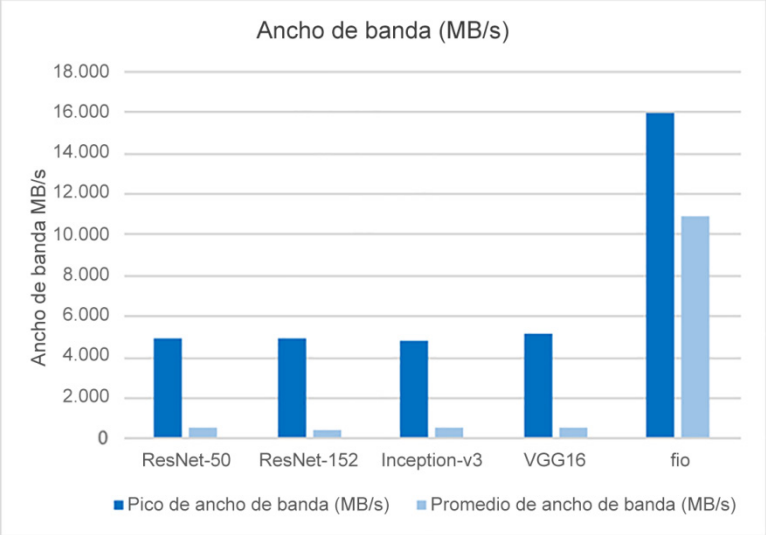


Figura 14) Latencia de almacenamiento para todos los modelos.

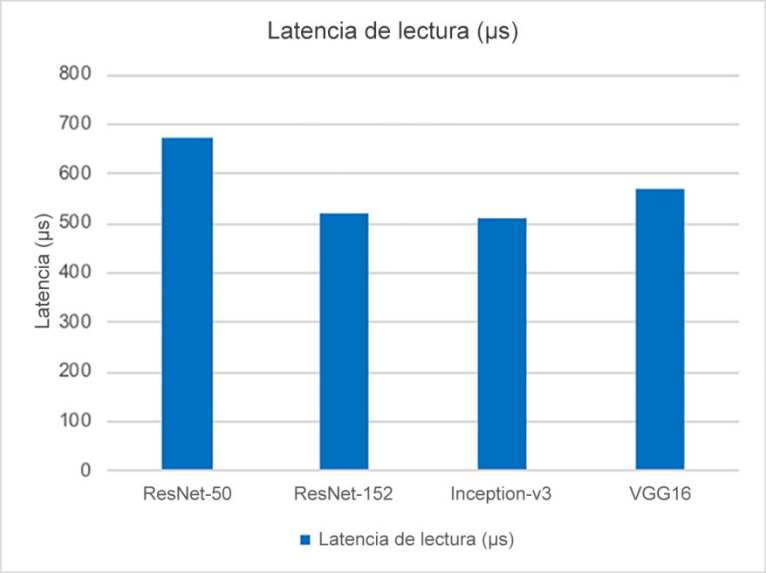
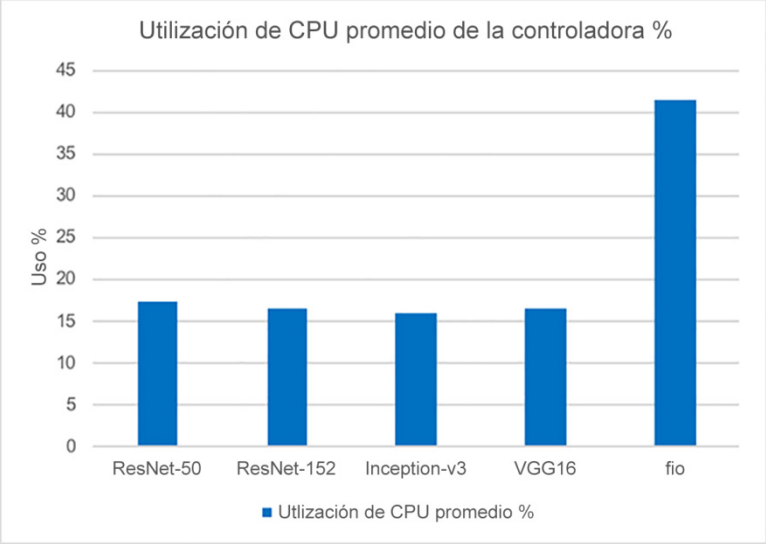


Figura 15) Uso de la CPU de almacenamiento para todos los modelos.



7.3 Guía de dimensionado de la solución

Esta arquitectura pretende servir como referencia para los clientes y asociados que quieran implementar una infraestructura de computación de alto rendimiento (HPC) con servidores NVIDIA DGX-1 y un sistema NetApp AFF.

Como queda demostrado en esta validación, el sistema AFF A800 admite sin problemas la carga de trabajo de entrenamiento de aprendizaje profundo generada por cuatro servidores DGX-1, con un margen restante aproximado del 70 % en el caso del par de alta disponibilidad. Por tanto, el sistema AFF A800 permite añadir servidores DGX-1 adicionales. Para implementaciones aún mayores y requisitos de rendimiento del almacenamiento superiores, es posible añadir sistemas AFF A800 adicionales al clúster ONTAP de NetApp. ONTAP 9 admite hasta 12 pares de alta disponibilidad (24 nodos) en un único clúster y, con la tecnología FlexGroup validada en esta solución, puede proporcionar hasta 20 PB en un solo volumen. El conjunto de datos empleado en esta validación era relativamente pequeño. Sin embargo, ONTAP 9 puede escalar su rendimiento de forma lineal hasta alcanzar una capacidad impresionante, ya que cada par de alta disponibilidad ofrece un rendimiento comparable al nivel validado en este documento.

Para clústeres de DGX-1 más pequeños, un sistema AFF A220 o AFF A300 proporciona un rendimiento suficiente a un precio más reducido. Como ONTAP 9 admite clústeres de modelo mixto, puede comenzar con una huella inicial pequeña e ir aumentando el sistema de almacenamiento a medida que crezcan sus requisitos de capacidad y rendimiento.

Desde el punto de vista de la red, la arquitectura validada consume solo 16 de los 32 puertos disponibles en cada switch Nexus 3232C. Cada switch admite hasta ocho servidores DGX-1 con puertos de acceso de almacenamiento adicionales para aumentar de forma significativa la potencia de computación sin necesidad de ampliar la red. Para implementaciones mayores, el modelo Cisco Nexus 7000 admite hasta 192 puertos con velocidad de cable de 100 GbE por switch. Como alternativa, puede adoptar una topología hoja-rama, con múltiples pares de switches Nexus 3000 conectados a un switch-rama central.

Según las pruebas de validación realizadas con esta carga de trabajo de entrenamiento de IA, cada DGX-1 requiere aproximadamente una salida de almacenamiento de 2 GB/s. Como el sistema AFF A800 tiene una capacidad probada de 25 GB/s con una carga de trabajo similar generada por otros medios, esta arquitectura admitiría nueve o más servidores DGX-1 por cada par AFF A800 de alta disponibilidad.

8 Conclusión

El servidor DGX-1 es una plataforma de aprendizaje profundo extremadamente potente que aprovecha una infraestructura de almacenamiento y de red igualmente potente para ofrecer el máximo valor. Combinando sistemas AFF de NetApp con switches Nexus de Cisco es posible implementar esta arquitectura validada a casi cualquier escala necesaria, desde un solo DGX-1 conectado a un sistema AFF A220 hasta un máximo de 96 servidores DGX-1 en un clúster AFF A800 de 12 nodos. Gracias a la superior integración con el cloud y las capacidades definidas por software de ONTAP de NetApp, AFF ofrece una completa gama de canalizaciones de datos que abarcan el perímetro, el núcleo y el cloud y que le permitirán llevar a buen puerto sus proyectos de DL.

Reconocimientos

Agradecemos y reconocemos las contribuciones que nuestros estimados colegas de NVIDIA, Darrin Johnson, Tony Paikeday, Robert Sohigian y James Mauro, han realizado a esta Arquitectura validada de NetApp. No hubiera sido posible completar este estudio sin el apoyo y la guía de dos miembros clave del equipo de NetApp, Robert Franz y Kesari Mishra.

Nuestro más sincero agradecimiento a todos aquellos que nos ofrecieron sus valiosos comentarios y su experiencia para la realización de este informe.

Dónde encontrar información adicional

Si quiere obtener más información sobre el contenido de este documento, consulte los siguientes recursos:

- Servidores DGX-1 de NVIDIA:
  - Servidores DGX-1 de NVIDIA: https://www.nvidia.com/en-us/data-center/dgx-1/
  - GPU de núcleo tensor NVIDIA Tesla V100: https://www.nvidia.com/en-us/data-center/tesla-v100/
  - GPU Cloud de NVIDIA : https://www.nvidia.com/en-us/gpu-cloud/
- Sistemas AFF de NetApp:
  - AFF: Especificaciones técnicas: https://www.netapp.com/es/media/ds-3582.pdf
  - Ventaja de NetApp Flash para AFF: https://www.netapp.com/es/media/ds-3733.pdf

  - Documentación de ONTAP 9.x:
    http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286
  - Informe técnico de NetApp FlexGroup:
    https://www.netapp.com/es/media/tr-4557.pdf
- Matriz de interoperabilidad de NetApp:
  - Herramienta de matriz de interoperabilidad de NetApp:
    http://support.netapp.com/matrix
- Redes Cisco Nexus:

  Los siguientes enlaces proporcionan más información sobre los switches de la serie Cisco Nexus 3232C:
  - Switches de la serie Cisco Nexus 3232C:
    https://www.cisco.com/c/en/us/products/switches/nexus-3232c-switch/index.html
  - Guía de configuración de Cisco Nexus 3232C :
    https://www.cisco.com/c/en/us/support/switches/nexus-3000-series-switches/products-installation-and-configuration-guides-list.html
  - Referencia de línea de comandos de Cisco Nexus 3232C:
    https://www.cisco.com/c/en/us/support/switches/nexus-3000-series-switches/products-command-reference-list.html
- Marco de aprendizaje automático:
  - TensorFlow: un marco de aprendizaje automático de código abierto para todos:
    https://www.tensorflow.org/
  - Horovod Marco de Uber para el aprendizaje automático distribuido de código abierto para TensorFlow:
    https://eng.uber.com/horovod/
  - Habilitación de GPU en el ecosistema Container Runtime:
    https://devblogs.nvidia.com/gpu-containers-runtime/
- Conjuntos de datos y pruebas:
  - ImageNet:
    http://www.image-net.org/
  - Pruebas de TensorFlow:
    https://www.tensorflow.org/performance/benchmarks

Apéndice

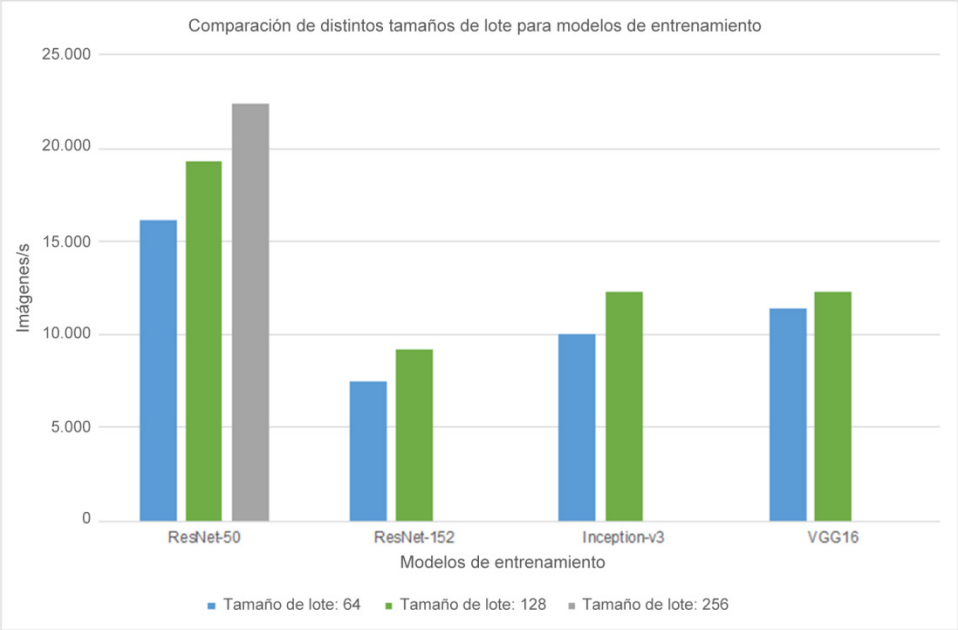
Este apartado contiene resultados adicionales de las pruebas realizadas sobre esta arquitectura.

Tasas de entrenamiento para distintos tamaños de lote y para cada modelo

La Figura 16 muestra una comparación de los distintos tamaños de lote y los distintos modelos de entrenamiento, utilizando los siguientes componentes:

- Número de GPU: 32 (4 servidores DGX-1).
- Núcleos: núcleos tensores.
- Tamaños de lote: 64,128 y 256 para ResNet-50; 64 y 128 para los demás modelos.

Figura 16) Comparación de distintos tamaños de lote para los modelos de entrenamiento.



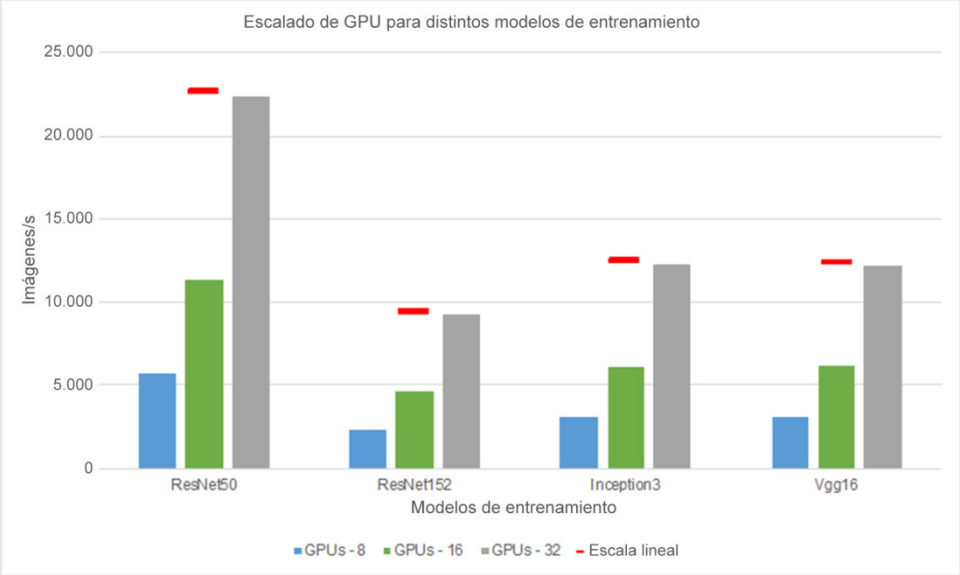
Conclusión: el rendimiento del entrenamiento aumenta cuando el tamaño de los lotes aumenta a 256 o 128.

Comparación de escalado de GPU para cada modelo

La Figura 17 muestra el escalado de GPU para los distintos modelos de entrenamiento, utilizando los siguientes componentes:

- Número de GPU: 8 (1 servidor DGX-1), 16 (2 servidores DGX-1) y 32 (4 servidores DGX-1).
- Núcleos: núcleos tensores.
- Tamaños de lote: 256 para ResNet-50; 128 para los demás modelos.

Figura 17) Escalado de GPU para diversos modelos de entrenamiento.



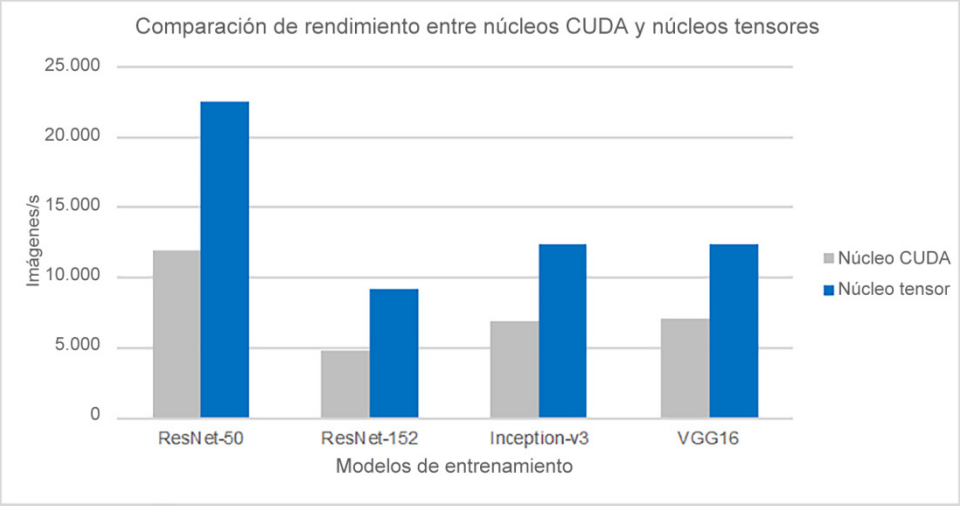
- Conclusión: se observa un escalado lineal de las GPU en todos los modelos de entrenamiento.

Comparación de núcleos tensores y núcleos CUDA

La Figura 18 muestra una comparación de rendimiento entre núcleos CUDA y núcleos tensores, utilizando los siguientes componentes:

- Número de GPU: 32 (4 servidores DGX-1).
- Núcleos: núcleos tensores y núcleos CUDA.
- Tamaños de lote: 256 para ResNet-50; 128 para los demás modelos.

Figura 18) Comparación de rendimiento entre núcleos CUDA y núcleos tensores.



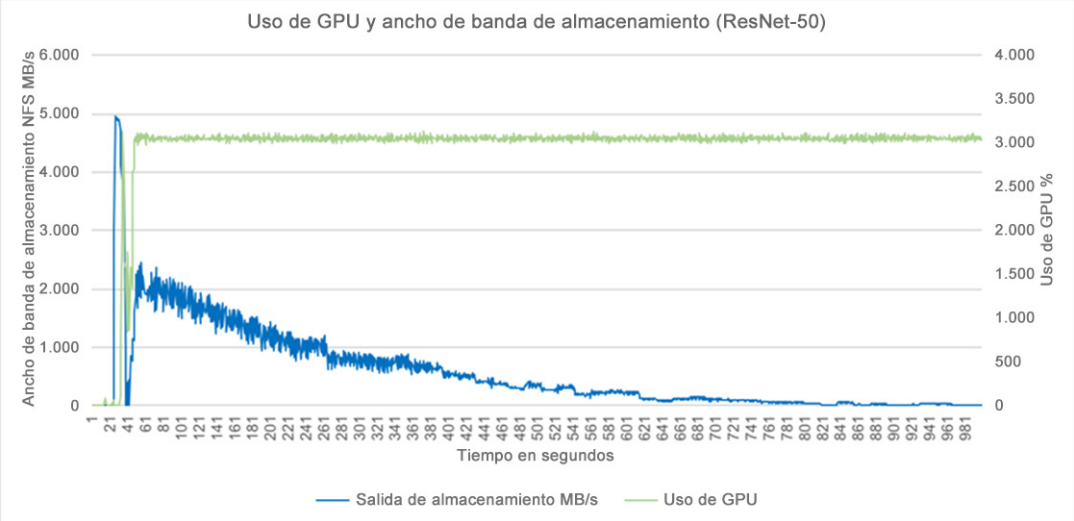
Conclusión: los núcleos tensores ofrecen mejor rendimiento que los núcleos CUDA.

Carga de trabajo de la GPU para todos los modelos

Las Figuras 19 a 21 muestran el uso de GPU y el ancho de banda para los modelos ResNet-50, ResNet-152 e Inception-v3 respectivamente, utilizando los siguientes componentes:

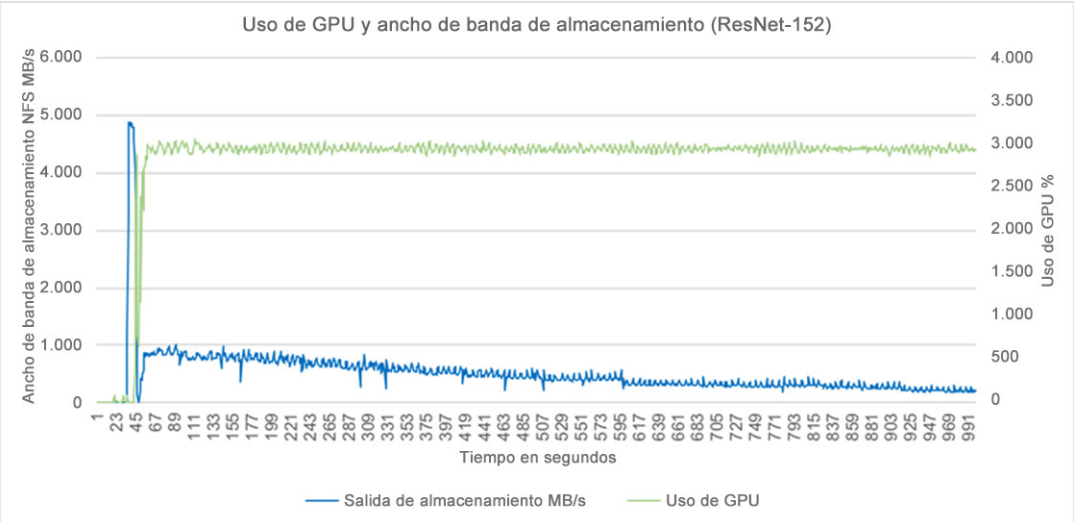
- Número de GPU: 32 (4 servidores DGX-1).
- Núcleos: núcleos tensores.
- Tamaños de lote: 256 para ResNet-50; 128 para los demás modelos.

Figura 19) Uso de GPU y ancho de banda del almacenamiento para ResNet-50.



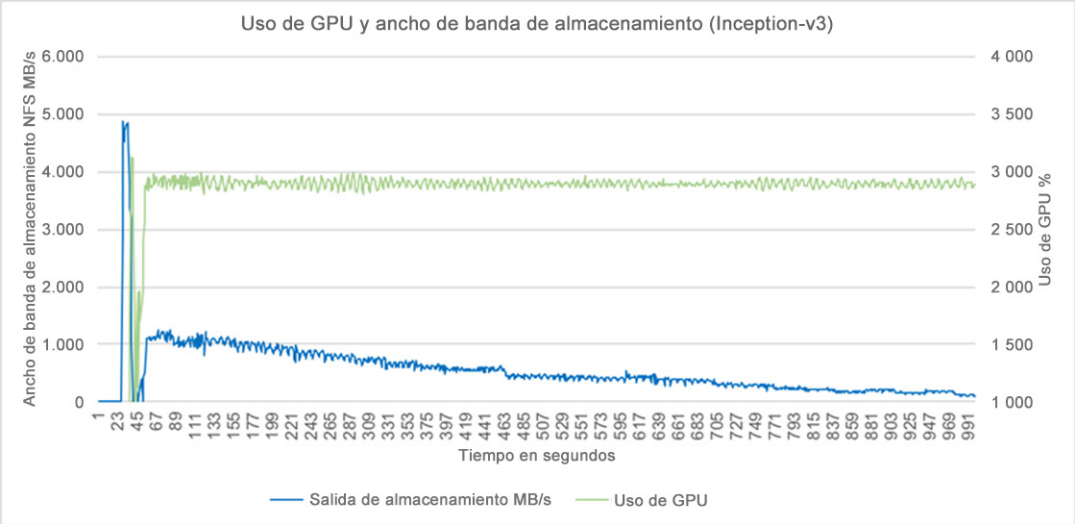
La Figura 20 muestra el uso de GPU y el ancho de banda para el modelo ResNet-152.

Figura 20) Uso de GPU y ancho de banda del almacenamiento para ResNet-152.



La Figura 21 muestra el uso de GPU y el ancho de banda para el modelo Inception-v3.

Figura 21) Uso de GPU y ancho de banda del almacenamiento para Inception-v3.



Consulte la matriz de interoperabilidad (IMT) en el sitio web de soporte de NetApp con el fin de 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.