AFF A400 de NetApp con Lenovo ThinkSystem SR670 V2 para formación de modelos de IA y ML

NetApp Solutions

NetApp
April 25, 2024

This PDF was generated from https://docs.netapp.com/es-es/netapp-solutions/ai/a400-thinksystem-introduction.html on April 25, 2024. Always check docs.netapp.com for the latest.

Tabla de contenidos

AFF A400 de NetApp con Lenovo ThinkSystem SR670 V2 para formación de modelos de IA y ML 1
TR-4810: NetApp AFF A400 con Lenovo ThinkSystem SR670 V2 para formación de modelos AI y ML 1
Información general de la tecnología 2
Plan de pruebas 7
Resultados de la prueba 7
Configuración de prueba 8
Procedimiento de prueba y resultados detallados 9
Ajustes de arquitectura 12
Conclusión 12

AFF A400 de NetApp con Lenovo ThinkSystem SR670 V2 para formación de modelos de IA y ML

TR-4810: NetApp AFF A400 con Lenovo ThinkSystem SR670 V2 para formación de modelos AI y ML

Sathish Thyagarajan, David Arnette, NetApp Mircea Troaca, Lenovo

Esta solución presenta una arquitectura de clúster de gama media con almacenamiento de NetApp y servidores Lenovo optimizados para cargas de trabajo de inteligencia artificial (IA). Está destinado a empresas pequeñas y medianas para las que la mayoría de tareas informáticas sean de un único nodo (una única o varias GPU) o distribuidas en varios nodos informáticos. Esta solución se alinea con la mayoría de los trabajos de formación de IA diarios para muchas empresas.

En este documento se tratan las pruebas y la validación de una configuración de computación y almacenamiento compuesta por ocho servidores Lenovo SR670V2 de la GPU, un sistema de almacenamiento AFF A400 de NetApp de gama media y un switch de interconexión de 100 GbE. Para medir el rendimiento, utilizamos ResNet50 con el conjunto de datos ImageNET, un tamaño de lote de 408, media precisión, CUDA y cuDNN. Esta arquitectura proporciona una solución eficiente y rentable para organizaciones pequeñas y medianas con el inicio de iniciativas de IA que requieren las funcionalidades de clase empresarial del almacenamiento de datos conectado al cloud ONTAP de NetApp.

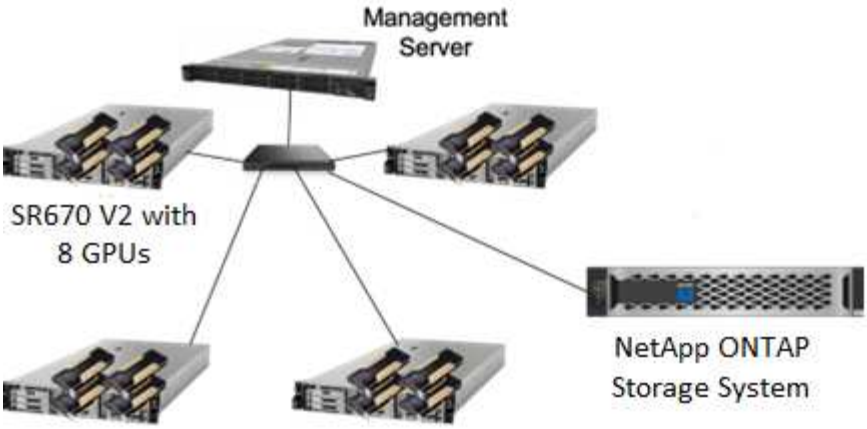
Público objetivo

Este documento está dirigido a los siguientes destinatarios:

- Científicos de datos, ingenieros de datos, administradores de datos y desarrolladores de sistemas de IA
- Arquitectos empresariales que diseñan soluciones para el desarrollo de modelos de IA
- Científicos e ingenieros de datos que buscan formas eficientes de alcanzar objetivos de desarrollo DE aprendizaje profundo (DL) y aprendizaje automático (ML)
- Líderes empresariales y responsables DE la toma de decisiones de OT/TECNOLOGÍA que desean conseguir el plazo de comercialización más rápido posible para iniciativas de IA

Arquitectura de la solución

Esta solución con los servidores Lenovo ThinkSystem y el sistema ONTAP de NetApp con almacenamiento AFF ha sido diseñado para gestionar el entrenamiento de IA en grandes conjuntos de datos utilizando la potencia de procesamiento de GPU junto con CPU tradicionales. Esta validación demuestra un alto rendimiento y una gestión de datos óptima con una arquitectura de escalado horizontal que utiliza uno, dos o cuatro servidores Lenovo SR670 V2 junto con un único sistema de almacenamiento AFF A400 de NetApp. En la siguiente figura, se proporciona una descripción general de la arquitectura.



Esta solución de NetApp y Lenovo ofrece las siguientes ventajas clave:

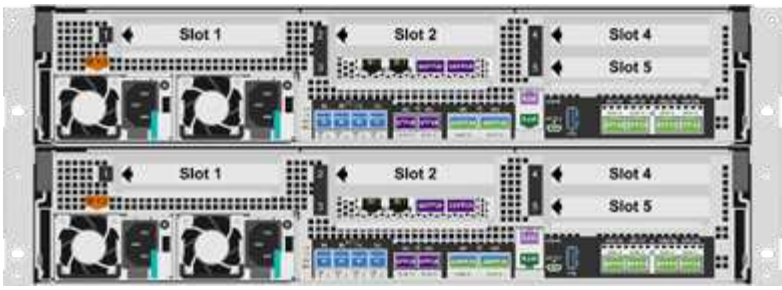
- Rendimiento muy eficaz y rentable al ejecutar varias tareas de formación en paralelo
- Rendimiento escalable basado en diferentes números de servidores Lenovo y diferentes modelos de controladoras de almacenamiento de NetApp
- Protección de datos sólida para cumplir con los objetivos de punto de recuperación (RPO) y los objetivos de tiempo de recuperación (RTO) bajos sin pérdida de datos
- Gestión de datos optimizada con copias Snapshot y clones para optimizar los flujos de trabajo de desarrollo

Información general de la tecnología

En esta sección se presentan con mayor detalle los principales componentes de esta solución.

Sistemas AFF de NetApp

Los sistemas de almacenamiento AFF de NetApp permiten a las empresas cumplir sus requisitos de almacenamiento con un rendimiento líder del sector, una flexibilidad superior, integración con el cloud y la mejor gestión de datos. Los sistemas AFF han sido diseñados específicamente para flash y ayudan a acelerar, gestionar y proteger los datos esenciales para la empresa.



NetApp AFF A400 es un sistema de almacenamiento flash NVMe de gama media que incluye las siguientes funciones:

- Capacidad efectiva máxima: ~20PB
- Escalado horizontal máximo: 2-24 nodos (12 parejas de alta disponibilidad)
- Compatibilidad con host de FC de 25 GbE y 16 GB
- Conectividad RDMA over Converged Ethernet (roce) de 100 GbE con bandejas de almacenamiento de ampliación NVMe
- Los puertos roce de 100 GbE se pueden utilizar para la conexión adjunta de red de host si no se conectan las bandejas NVMe
- Bandejas de ampliación de almacenamiento con conectividad SAS completa de 12 Gbps
- Disponible en dos configuraciones:
  - Ethernet: 4 puertos Ethernet de 25 GB (SFP28)
  - Fibre Channel: 4 puertos FC de 16 GB (SFP+)
- 100 % lectura aleatoria de 8 KB @,4 ms 400 k IOPS

Las funciones AFF A250 de NetApp para puestas en marcha DE IA/ML de gama básica incluyen lo siguiente:

- Capacidad efectiva máxima: 35 PB
- Escalado horizontal máximo: 2-24 nodos (12 parejas de alta disponibilidad)
- 440 000 IOPS lecturas aleatorias a 1 ms
- Basado en la última versión de ONTAP de NetApp, ONTAP 9.8 o posterior
- Dos puertos Ethernet de 25 GB para alta disponibilidad e interconexión de clúster

NetApp también ofrece otros sistemas de almacenamiento, como A800 y AFF A700 de AFF, que proporcionan un mayor rendimiento y escalabilidad para puestas en marcha de IA/ML a gran escala.

ONTAP de NetApp

ONTAP 9, la última generación del software de gestión del almacenamiento de NetApp, permite a las empresas 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 permite la gestión y protección de los datos con un solo conjunto de herramientas, sin importar dónde residan. Los datos también pueden trasladarse libremente allí donde sea necesario: El perímetro, el núcleo o el 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 los datos

La gestión de los datos es crucial para las operaciones TECNOLÓGICAS empresariales, 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 reducir el coste total de las operaciones:

- **Compactación de datos inline y deduplicación expandida.** la compactación de datos reduce el espacio perdido dentro de los bloques de almacenamiento, y la deduplicación aumenta significativamente la capacidad efectiva. Esto es aplicable a los datos almacenados localmente y a los datos organizados en niveles en el cloud.
- **Calidad de servicio (QoS) mínima, máxima y adaptativa.** los controles granulares de QoS ayudan a mantener los niveles de rendimiento para aplicaciones críticas en entornos altamente compartidos.
- **ONTAP FabricPool**. Esta función organiza automáticamente en niveles los datos fríos en opciones de almacenamiento en cloud público y privado, incluido el almacenamiento de objetos Amazon Web Services (AWS), Azure y StorageGRID de NetApp.

Acelere y proteja sus datos

ONTAP no solo ofrece niveles de rendimiento y protección de datos superiores, sino que amplía estas capacidades de las siguientes maneras:

- **Rendimiento y menor latencia.** ONTAP ofrece el rendimiento más alto posible con la menor latencia posible.
- **Protección de datos.** ONTAP ofrece capacidades integradas de protección de datos con administración común en todas las plataformas.
- **Cifrado de volumen de NetApp.** ONTAP ofrece cifrado nativo a nivel de volumen con compatibilidad para la gestión de claves incorporada y externa.

Infraestructura preparada para futuros retos

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

- **Escalado sencillo y operaciones no disruptivas.** ONTAP admite la adición sin interrupciones de capacidad a las controladoras existentes y a los clústeres de escalado horizontal. Los clientes pueden empezar a utilizar tecnologías punteras como NVMe y FC 32 GB, sin necesidad de realizar costosas migraciones de datos y sin cortes.
- **Conexión en cloud.** ONTAP es el software de gestión del almacenamiento con mejor conexión en cloud, con opciones de almacenamiento definido por software (ONTAP Select) e instancias nativas del cloud (Cloud Volumes Service de NetApp) en todos los clouds públicos.
- **Integración con aplicaciones emergentes.** ONTAP ofrece servicios de datos de clase empresarial para plataformas y aplicaciones de próxima generación como OpenStack, Hadoop y MongoDB usando la misma infraestructura que soporta las aplicaciones empresariales existentes.

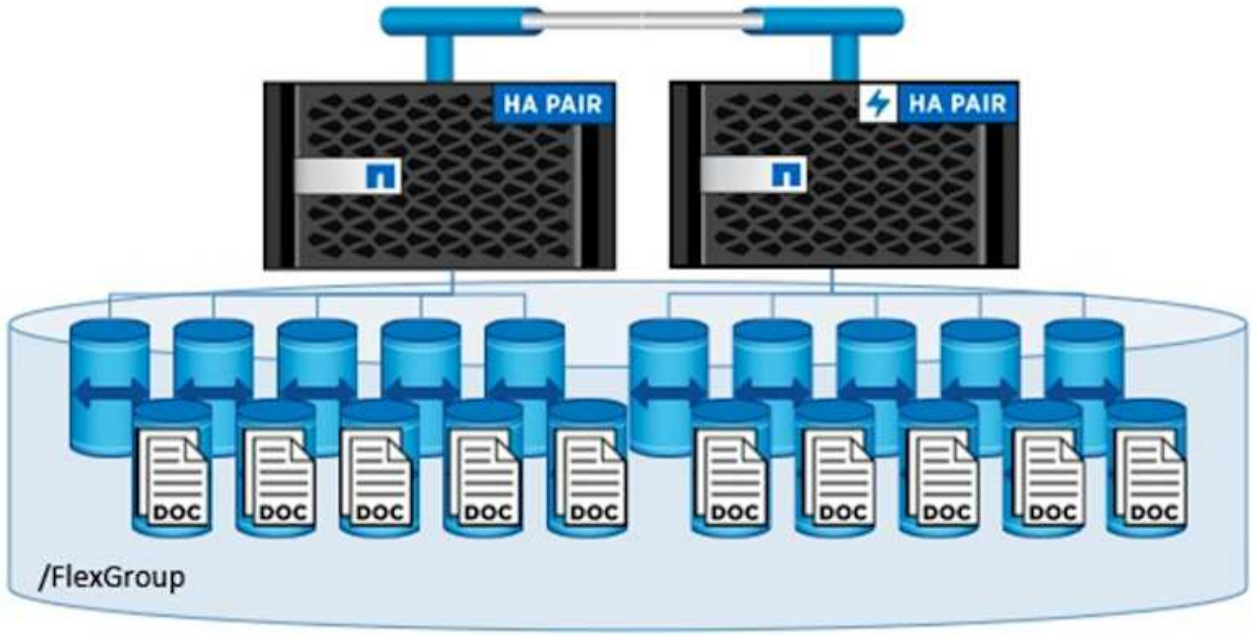
FlexGroup Volumes de NetApp

Los conjuntos de datos de entrenamiento suelen ser una colección de potencialmente miles de millones de archivos. 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 almacenar muchos archivos pequeños y debe leerlos en paralelo, con una entrada y salida secuencial o aleatoria

Un volumen FlexGroup (figura a continuación) es un espacio de nombres único compuesto por múltiples

volúmenes constituyentes que se gestiona y actúa como un volumen FlexVol de NetApp entre 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:

- Hasta 20 petabytes de capacidad y baja latencia predecible para cargas de trabajo con una gran cantidad de metadatos
- Hasta 400 000 millones de archivos en un mismo espacio de nombres
- Operaciones en paralelo para cargas de trabajo NAS entre varias CPU, nodos, agregados y volúmenes FlexVol constituyentes



"

Gama Lenovo ThinkSystem

Los servidores Lenovo ThinkSystem incluyen hardware, software y servicios innovadores que resuelven los desafíos actuales de los clientes y ofrecen un enfoque de diseño modular, evolutivo y adecuado para su propósito para afrontar los desafíos del futuro. Estos servidores se capitalizan en las mejores tecnologías estándar del sector, junto con innovaciones diferenciadas de Lenovo, para proporcionar la mayor flexibilidad posible en servidores x86.

Entre las ventajas clave de la implementación de servidores Lenovo ThinkSystem se incluyen las siguientes:

- Diseños modulares y de gran escalabilidad que crecen con el negocio
- Resiliencia líder en el sector para ahorrar horas de costosos tiempos de inactividad no programados
- Tecnologías flash rápidas para reducir las latencias, acelerar los tiempos de respuesta y gestionar los datos de forma más inteligente en tiempo real

En el ámbito de la IA, Lenovo está adoptando un enfoque práctico para ayudar a las empresas a comprender y adoptar las ventajas DEL APRENDIZAJE AUTOMÁTICO y la IA para sus cargas de trabajo. Los clientes de Lenovo pueden explorar y evaluar las ofertas de IA de Lenovo en los centros de innovación de IA de Lenovo para comprender por completo el valor de su caso de uso en particular. Con el fin de mejorar la rentabilidad de la inversión, este enfoque centrado en el cliente ofrece a los clientes pruebas de concepto para plataformas de

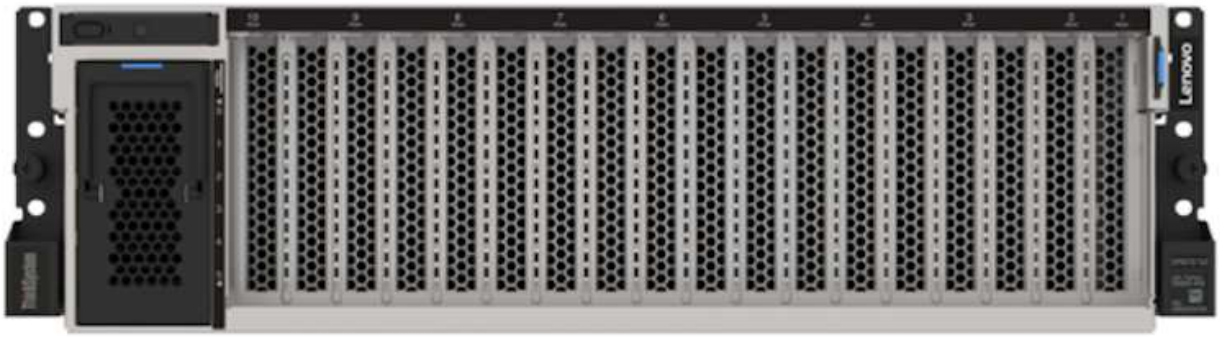
desarrollo de soluciones listas para usar y optimizadas para la IA.

Lenovo SR670 V2

El servidor en rack Lenovo ThinkSystem SR670 V2 ofrece un rendimiento óptimo para una IA acelerada y una informática de alto rendimiento (HPC). Con soporte para hasta ocho GPU, la SR670 V2 es adecuada para los requisitos de cargas de trabajo informáticas intensivas DE ML, DL e inferencia.

4x SXM GPUs with 8x 2.5-inch HS drives and 2x PCIe I/O slots

4x double-wide or 8x single-wide GPU slots and 2x PCIe I/O slots with 8x 2.5-inch or 4x 3.5-inch HS drives



8x double-wide GPU slots with 6x EDSFF HS drives and 2x PCIe I/O slots

Con las CPU Intel Xeon más recientes y escalables que admiten GPU de gama alta (incluida la GPU 8x PCIe NVIDIA A100 de 80 GB), ThinkSystem SR670 V2 ofrece un rendimiento optimizado y acelerado para cargas de trabajo de IA y computación de alto rendimiento.

Puesto que en más cargas de trabajo se utiliza el rendimiento de los aceleradores, ha aumentado la demanda de densidad de GPU. Sectores como el comercio minorista, los servicios financieros, la energía y la sanidad utilizan GPU para obtener una mayor información e impulsar la innovación con APRENDIZAJE

AUTOMÁTICO, DL y técnicas de inferencia.

ThinkSystem SR670 V2 es una solución empresarial optimizada para poner en marcha cargas de trabajo aceleradas de HPC e IA en la producción, maximizando el rendimiento del sistema a la vez que mantiene la densidad del centro de datos para los clústeres de supercomputación con plataformas de última generación.

Entre otras funciones se incluyen las siguientes:

- Compatibilidad con I/o RDMA directa de GPU en la que los adaptadores de red de alta velocidad están conectados directamente a las GPU para maximizar el rendimiento de I/O.
- Compatibilidad con almacenamiento directo de GPU en el que las unidades NVMe están conectadas directamente a las GPU para maximizar el rendimiento del almacenamiento.

Rendim. MLPerf

MLPerf es el conjunto de pruebas de rendimiento líder del sector para evaluar el rendimiento de la IA. En esta validación, utilizamos su punto de referencia de clasificación de imágenes con MXNet, uno de los marcos de IA más populares. El script de formación MXNet_Benchmark se utilizó para impulsar la formación de IA. El script contiene implementaciones de varios modelos convencionales populares y está diseñado para ser lo más rápido posible. Puede ejecutarse en una sola máquina o ejecutarse en modo distribuido entre varios hosts.

Plan de pruebas

En esta validación, realizamos el entrenamiento de reconocimiento de imágenes según lo especificado por MLPerf v2.0. Específicamente, entrenamos el modelo ResNet v2.0 con el conjunto de datos ImageNET hasta que alcanza una precisión del 76.1 %. La métrica principal es el tiempo para alcanzar la precisión deseada. También reportamos formación de ancho de banda en imágenes por segundo para valorar mejor la eficiencia del escalado horizontal.

El caso de prueba principal evaluó la ejecución simultánea de varios procesos de entrenamiento independientes (uno por nodo). Simula el caso práctico principal, un sistema compartido utilizado por varios científicos de datos. En el segundo caso de prueba se evaluó la eficiencia del escalado horizontal.

Resultados de la prueba

La tabla siguiente resume los resultados de todas las pruebas realizadas para esta solución.

Descripción de la prueba	Resumen de resultados
Formación de reconocimiento de imágenes: Varios trabajos simultáneos	Rendimiento muy eficiente. Todos los trabajos se ejecutaron a toda velocidad incluso cuando se utilizaba al máximo el clúster. Los sistemas de almacenamiento de NetApp proporcionaron un rendimiento de formación comparable al almacenamiento SSD local a la vez que posibilitan un uso compartido sencillo de los datos entre servidores.

Descripción de la prueba	Resumen de resultados
Formación para el reconocimiento de imágenes: Escalado horizontal	Altamente eficiente para un máximo de cuatro nodos. En ese momento, la escalabilidad horizontal era menos eficiente pero todavía factible. El uso de una red computacional de mayor velocidad mejora la escalabilidad. El sistema de almacenamiento de NetApp proporcionó un rendimiento de formación comparable al almacenamiento en unidades de estado sólido locales a la vez que permitió compartir fácilmente los datos entre servidores.

Configuración de prueba

En este apartado se describen las configuraciones probadas, la infraestructura de red, el servidor SR670 V2 y los detalles sobre el aprovisionamiento de almacenamiento NetApp.

Arquitectura de la solución

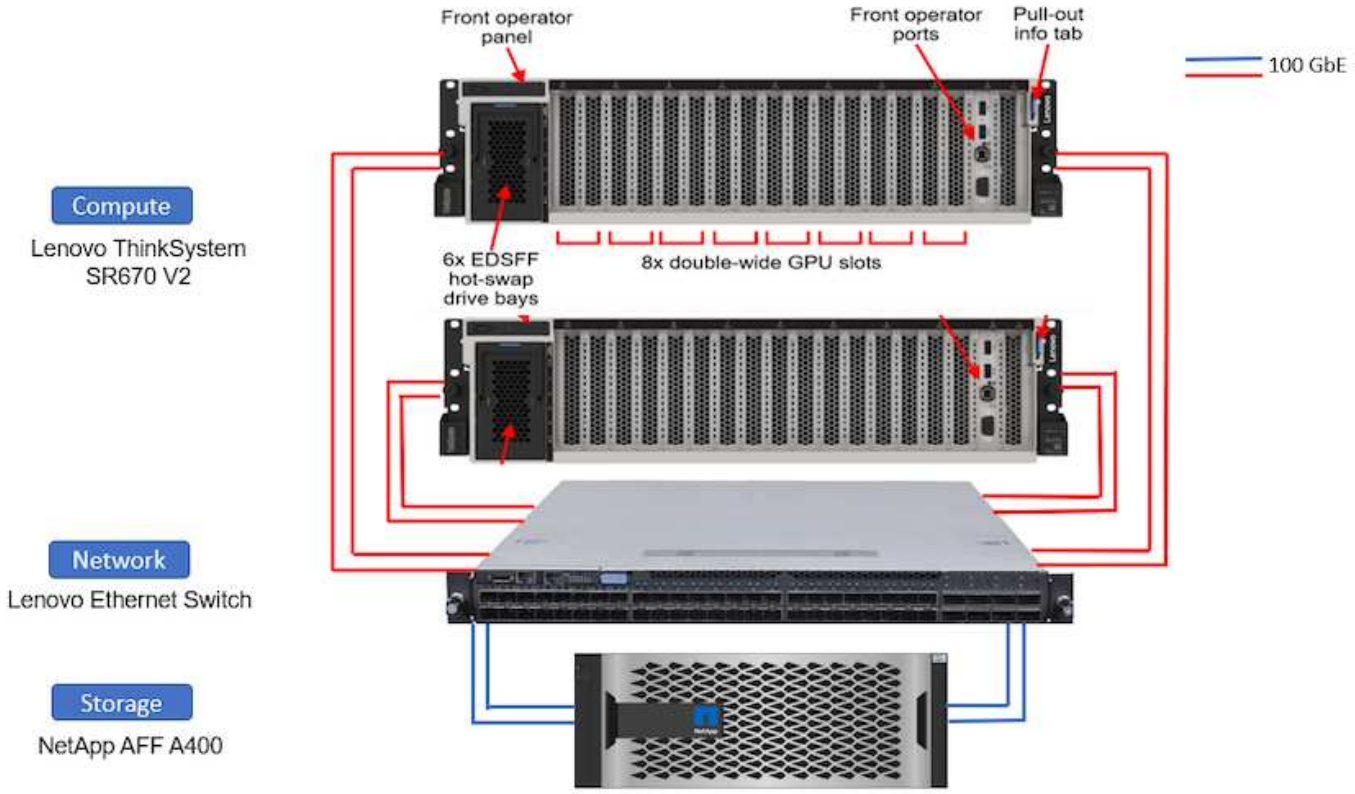
Hemos utilizado los componentes de la solución enumerados en la siguiente tabla para esta validación.

Componentes de la solución	Detalles
Servidores Lenovo ThinkSystem	• Dos servidores SR670 V2 cada uno con ocho tarjetas GPU NVIDIA A100 de 80 GB • Cada servidor contiene 2 CPU Intel Xeon Platinum 8360Y (28 núcleos físicos) y 1 TB de RAM
Linux (Ubuntu – 20.04 con CUDA 11.8)	
Sistema de almacenamiento AFF de NetApp (pareja de alta disponibilidad)	• Software ONTAP 9.10.1 de NetApp • 24 unidades SSD de 960 GB • Protocolo NFS • 1 grupo de interfaces (ifgrp) por controladora, con cuatro direcciones IP lógicas para los puntos de montaje

En esta validación, utilizamos ResNet v2.0 con la base ImageNET establecida como lo especifica MLPerf v2.0. El conjunto de datos se almacena en un sistema de almacenamiento AFF de NetApp con el protocolo NFS. Los SR670 se conectaron al sistema de almacenamiento AFF A400 de NetApp a través de un switch de 100 GbE.

ImageNET es un conjunto de datos de imágenes utilizado con frecuencia. Contiene casi 1.3 millones de imágenes por un tamaño de 144 GB. El tamaño medio de la imagen es de 108 KB.

En la siguiente figura, se muestra la topología de red de la configuración probada.



Controladora de almacenamiento

La siguiente tabla enumera la configuración de almacenamiento.

Controladora	Agregado	Volumen FlexGroup	Tamaño del agregado	Tamaño del volumen	Punto de montaje del sistema operativo
Control1	Agr1	/a400-100g	9,9 TB	19 TB	/a400-100g
Control2	Agr2	/a400-100g	9,9 TB		/a400-100g

La carpeta /a400-100g contiene el conjunto de datos utilizado para la validación ResNet.

Procedimiento de prueba y resultados detallados

En esta sección se describen los resultados detallados del procedimiento de prueba.

El entrenamiento de reconocimiento de imágenes mediante ResNet en ONTAP

Ejecutamos la prueba de rendimiento ResNet50 con uno y dos servidores SR670 V2. Esta prueba utilizó el contenedor MXNet 22.04-py3 NGC para realizar el entrenamiento.

Hemos utilizado el siguiente procedimiento de prueba en esta validación:

1. Se borró la caché del host antes de ejecutar el script para garantizar que los datos no se almacenaban en caché:

```
sync ; sudo /sbin/sysctl vm.drop_caches=3
```

2. Ejecutamos el script de pruebas de rendimiento con el conjunto de datos ImageNET en el almacenamiento de servidores (almacenamiento SSD local) y en el sistema de almacenamiento AFF de NetApp.
3. Validamos el rendimiento del almacenamiento local y en red con el `dd` comando.
4. Para la ejecución de un solo nodo, utilizamos el siguiente comando:

```
python train_imagenet.py --gpus 0,1,2,3,4,5,6,7 --batch-size 408 --kv
-store horovod --lr 10.5 --mom 0.9 --lr-step-epochs pow2 --lars-eta
0.001 --label-smoothing 0.1 --wd 5.0e-05 --warmup-epochs 2 --eval-period
4 --eval-offset 2 --optimizer sgdwfastlars --network resnet-v1b-stats-fl
--num-layers 50 --num-epochs 37 --accuracy-threshold 0.759 --seed 27081
--dtype float16 --disp-batches 20 --image-shape 4,224,224 --fuse-bn-relu
1 --fuse-bn-add-relu 1 --bn-group 1 --min-random-area 0.05 --max-random
-area 1.0 --conv-algo 1 --force-tensor-core 1 --input-layout NHWC --conv
-layout NHWC --batchnorm-layout NHWC --pooling-layout NHWC --batchnorm
-mom 0.9 --batchnorm-eps 1e-5 --data-train /data/train.rec --data-train
-idx /data/train.idx --data-val /data/val.rec --data-val-idx
/data/val.idx --dali-dont-use-mmap 0 --dali-hw-decoder-load 0 --dali
-prefetch-queue 5 --dali-nvjpeg-memory-padding 256 --input-batch
-multiplier 1 --dali- threads 6 --dali-cache-size 0 --dali-roi-decode 1
--dali-preallocate-width 5980 --dali-preallocate-height 6430 --dali-tmp
-buffer-hint 355568328 --dali-decoder-buffer-hint 1315942 --dali-crop
-buffer-hint 165581 --dali-normalize-buffer-hint 441549 --profile 0
--e2e-cuda-graphs 0 --use-dali
```

5. Para las ejecuciones distribuidas, utilizamos el modelo de paralelización del servidor de parámetros. Se utilizaron dos servidores de parámetros por nodo y se establece el número de épocas igual que para la ejecución de un único nodo. Lo hicimos porque el entrenamiento distribuido a menudo toma más épocas debido a la sincronización imperfecta entre procesos. El número diferente de épocas puede sesgar comparaciones entre casos de un solo nodo y distribuidos.

Velocidad de lectura de los datos: Local frente a almacenamiento en red

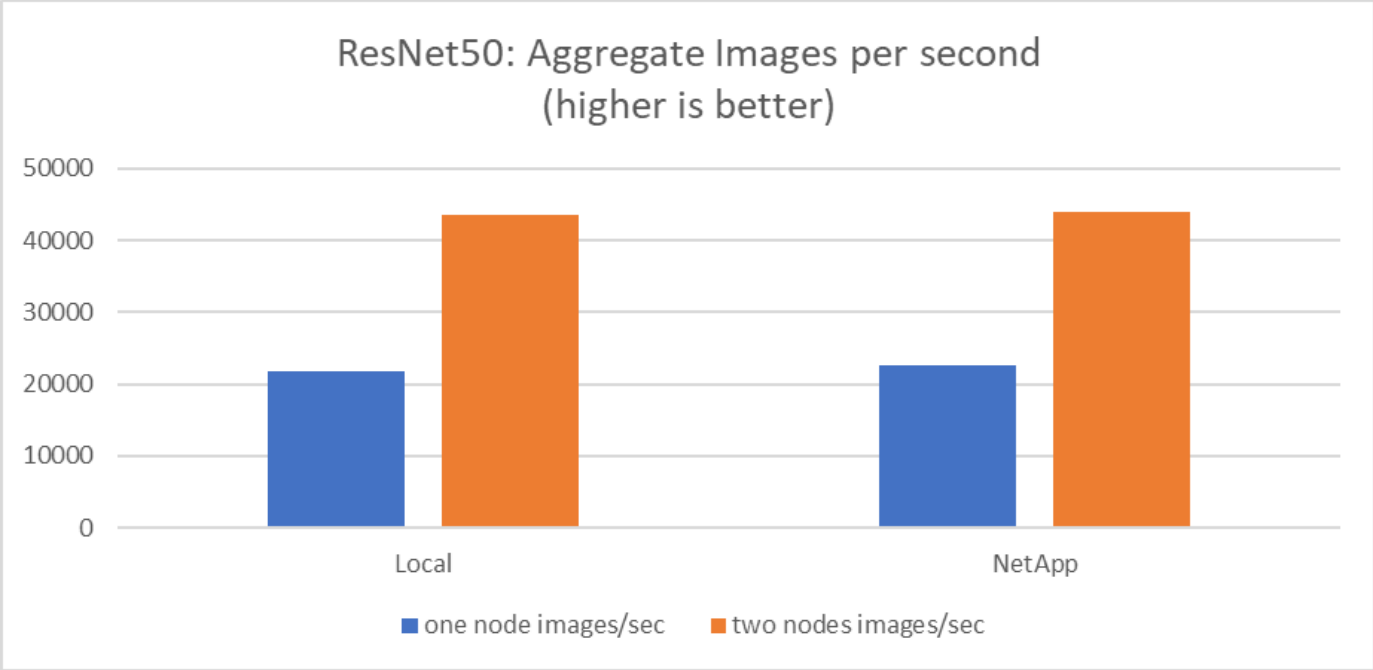
La velocidad de lectura se probó con el `dd` Comando en uno de los archivos para el conjunto de datos ImageNET. Específicamente, ejecutamos los siguientes comandos para datos locales y de red:

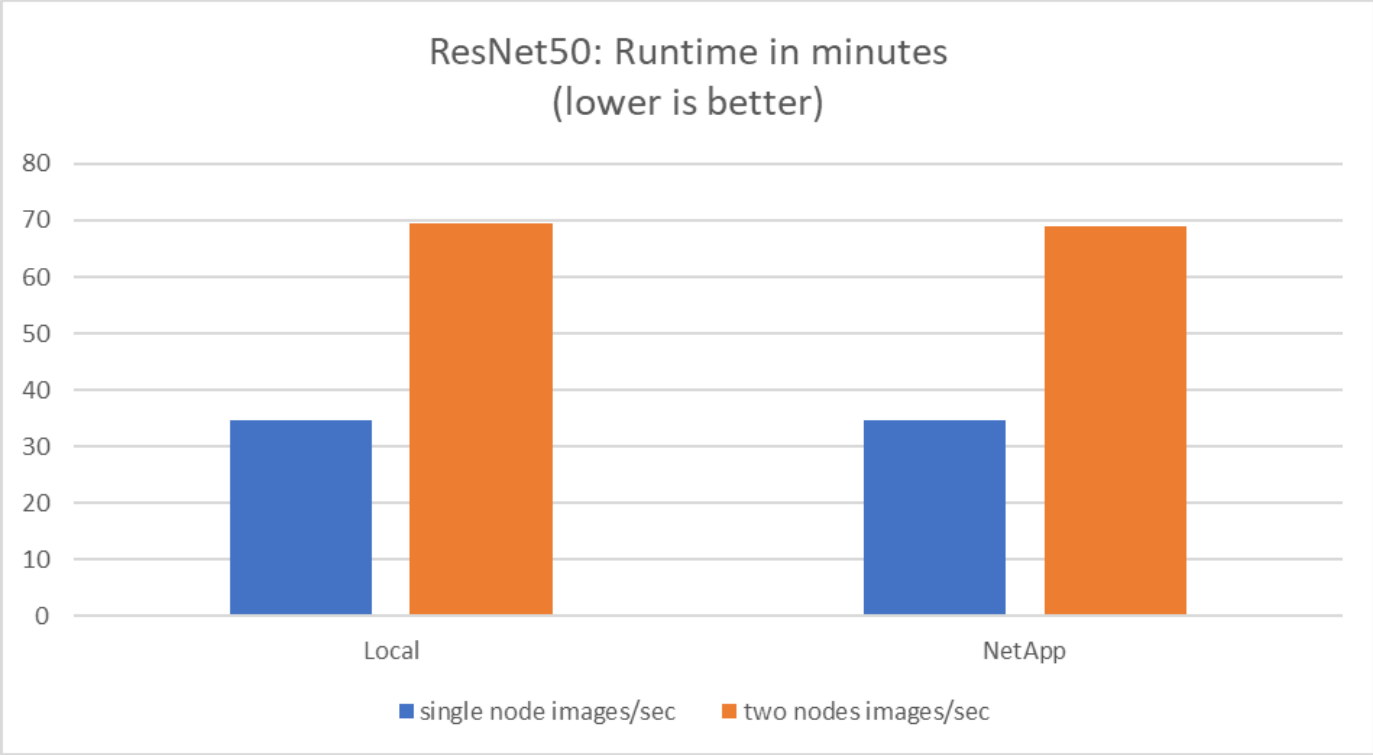
```
sync ; sudo /sbin/sysctl vm.drop_caches=3dd if=/a400-100g/netapp-
ra/resnet/data/preprocessed_data/train.rec of=/dev/null bs=512k
count=2048Results (average of 5 runs):
Local storage: 1.7 GB/s Network storage: 1.5 GB/s.
```

Ambos valores son similares, lo que demuestra que el almacenamiento en red puede ofrecer datos a un ritmo similar al del almacenamiento local.

Caso de uso compartido: Trabajos múltiples, independientes y simultáneos

Esta prueba simuló el caso de uso esperado para esta solución: Formación de IA para varios trabajos y varios usuarios. Cada nodo ejecutó su propia formación mientras utiliza el almacenamiento de red compartida. Los resultados se muestran en la siguiente figura, que muestra que el caso de la solución proporcionó un rendimiento excelente con todos los trabajos que se ejecutan básicamente a la misma velocidad que los trabajos individuales. El rendimiento total se escaló de forma lineal con el número de nodos.





Estos gráficos presentan el tiempo de ejecución en minutos y las imágenes agregadas por segundo para los nodos de computación que utilizaron ocho GPU de cada servidor en una red de cliente de 100 GbE, combinando tanto el modelo de entrenamiento simultáneo como el único modelo de entrenamiento. El tiempo

de ejecución medio del modelo de entrenamiento fue de 35 minutos y 9 segundos. Los tiempos de ejecución individuales fueron de 34 minutos y 32 segundos, 36 minutos y 21 segundos, 34 minutos y 37 segundos, 35 minutos y 25 segundos, y 34 minutos y 31 segundos. El promedio de imágenes por segundo para el modelo de entrenamiento fue de 22,573, y las imágenes individuales por segundo fueron de 21,764; 23,438; 22,556; 22,564; y 22,547.

Según nuestra validación, un modelo de entrenamiento independiente con tiempo de ejecución de datos de NetApp fue de 34 minutos y 54 segundos con 22,231 imágenes/s. Un modelo de entrenamiento independiente con un tiempo de ejecución de datos locales (DAS) era de 34 minutos y 21 segundos con 22,102 imágenes/s. Durante estas operaciones, el uso medio de la GPU era del 96 %, como se observó en nvidia-smi. Tenga en cuenta que esta media incluye la fase de prueba, durante la cual no se utilizaron las GPU, mientras que el uso de CPU fue del 40 % medido por mpstat. Esto demuestra que la tasa de entrega de datos es suficiente en cada caso.

Ajustes de arquitectura

La configuración utilizada para esta validación se puede ajustar para ajustarse a otros casos prácticos.

Ajustes de CPU

Utilizamos un procesador Intel Xeon Platinum 8360Y Skylake para esta validación, según lo recomendado por Lenovo. Esperamos que la CPU equivalente de Cascade Lake, un procesador Intel Xeon Gold 6330, proporcione un rendimiento similar porque esta carga de trabajo no está vinculada a la CPU.

Aumente la capacidad de almacenamiento

En función de sus necesidades de capacidad de almacenamiento, puede aumentar el espacio de almacenamiento compartido (volumen NFS) bajo demanda siempre que disponga de las bandejas de discos y los modelos de controladora adicionales. Puede hacerlo desde la CLI o desde la interfaz web de NetApp del controlador de almacenamiento como usuario administrador.

Conclusión

La solución de NetApp y Lenovo validada aquí es una arquitectura flexible de escalado horizontal ideal para IA empresarial de gama media.

El almacenamiento de NetApp proporciona el mismo rendimiento o mejor que el almacenamiento SSD local y ofrece las siguientes ventajas a los científicos de datos, ingenieros de datos y responsables de la toma DE decisiones TECNOLÓGICAS:

- Uso compartido de datos sin esfuerzo entre sistemas de IA, análisis y otros sistemas de negocio cruciales. Este uso compartido de datos reduce la sobrecarga de la infraestructura, mejora el rendimiento y optimiza la gestión de datos en toda la empresa.
- Almacenamiento y cálculo escalables independientemente para minimizar los costes y mejorar la utilización de recursos.
- Flujos de trabajo de desarrollo y puesta en marcha optimizados que utilizan copias Snapshot y clones integrados para espacios de trabajo de usuario instantáneos con gestión eficiente del espacio, control de versiones integrado y puesta en marcha automatizada.
- Protección de datos de clase empresarial para recuperación ante desastres y continuidad empresarial.

Reconocimientos

- Karthikeyan Nagalingam, Ingeniero Técnico de Marketing de NetApp
- Jarrett Upton, Admin, AI Lab Systems, Lenovo

Dónde encontrar información adicional

Para obtener más información sobre la información descrita en este documento, consulte los siguientes documentos y/o sitios web:

- Página de producto de cabinas all-flash de NetApp

  "https://www.netapp.com/us/products/storage-systems/all-flash-array/aff-a-series.aspx"

- Página AFF A400 de NetApp

  "https://docs.netapp.com/us-en/ontap-systems/a400/index.html"

- Página de producto del software de gestión de datos ONTAP de NetApp

  "http://www.netapp.com/us/products/data-management-software/ontap.aspx"

- Rendim. MLPerf

  "https://mlperf.org"

- Prueba de rendimiento de TensorFlow

  "https://github.com/tensorflow/benchmarks"

- SMI de NVIDIA (nvidia-smi)

  "https://developer.nvidia.com/nvidia-system-management-interface"

Información de copyright

Copyright © 2024 NetApp, Inc. Todos los derechos reservados. Imprimido en EE. UU. No se puede reproducir este documento protegido por copyright ni parte del mismo de ninguna forma ni por ningún medio (gráfico, electrónico o mecánico, incluidas fotocopias, grabaciones o almacenamiento en un sistema de recuperación electrónico) sin la autorización previa y por escrito del propietario del copyright.

El software derivado del material de NetApp con copyright está sujeto a la siguiente licencia y exención de responsabilidad:

ESTE SOFTWARE LO PROPORCIONA NETAPP «TAL CUAL» Y SIN NINGUNA GARANTÍA EXPRESA O IMPLÍCITA, INCLUYENDO, SIN LIMITAR, LAS GARANTÍAS IMPLÍCITAS DE COMERCIALIZACIÓN O IDONEIDAD PARA UN FIN CONCRETO, CUYA RESPONSABILIDAD QUEDA EXIMIDA POR EL PRESENTE DOCUMENTO. EN NINGÚN CASO NETAPP SERÁ RESPONSABLE DE NINGÚN DAÑO DIRECTO, INDIRECTO, ESPECIAL, EJEMPLAR O RESULTANTE (INCLUYENDO, ENTRE OTROS, LA OBTENCIÓN DE BIENES O SERVICIOS SUSTITUTIVOS, PÉRDIDA DE USO, DE DATOS O DE BENEFICIOS, O INTERRUPCIÓN DE LA ACTIVIDAD EMPRESARIAL) CUALQUIERA SEA EL MODO EN EL QUE SE PRODUJERON Y LA TEORÍA DE RESPONSABILIDAD QUE SE APLIQUE, YA SEA EN CONTRATO, RESPONSABILIDAD OBJETIVA O AGRAVIO (INCLUIDA LA NEGLIGENCIA U OTRO TIPO), QUE SURJAN DE ALGÚN MODO DEL USO DE ESTE SOFTWARE, INCLUSO SI HUBIEREN SIDO ADVERTIDOS DE LA POSIBILIDAD DE TALES DAÑOS.

NetApp se reserva el derecho de modificar cualquiera de los productos aquí descritos en cualquier momento y sin aviso previo. NetApp no asume ningún tipo de responsabilidad que surja del uso de los productos aquí descritos, excepto aquello expresamente acordado por escrito por parte de NetApp. El uso o adquisición de este producto no lleva implícita ninguna licencia con derechos de patente, de marcas comerciales o cualquier otro derecho de propiedad intelectual de NetApp.

Es posible que el producto que se describe en este manual esté protegido por una o más patentes de EE. UU., patentes extranjeras o solicitudes pendientes.

LEYENDA DE DERECHOS LIMITADOS: el uso, la copia o la divulgación por parte del gobierno están sujetos a las restricciones establecidas en el subpárrafo (b)(3) de los derechos de datos técnicos y productos no comerciales de DFARS 252.227-7013 (FEB de 2014) y FAR 52.227-19 (DIC de 2007).

Los datos aquí contenidos pertenecen a un producto comercial o servicio comercial (como se define en FAR 2.101) y son propiedad de NetApp, Inc. Todos los datos técnicos y el software informático de NetApp que se proporcionan en este Acuerdo tienen una naturaleza comercial y se han desarrollado exclusivamente con fondos privados. El Gobierno de EE. UU. tiene una licencia limitada, irrevocable, no exclusiva, no transferible, no sublicenciable y de alcance mundial para utilizar los Datos en relación con el contrato del Gobierno de los Estados Unidos bajo el cual se proporcionaron los Datos. Excepto que aquí se disponga lo contrario, los Datos no se pueden utilizar, desvelar, reproducir, modificar, interpretar o mostrar sin la previa aprobación por escrito de NetApp, Inc. Los derechos de licencia del Gobierno de los Estados Unidos de América y su Departamento de Defensa se limitan a los derechos identificados en la cláusula 252.227-7015(b) de la sección DFARS (FEB de 2014).

Información de la marca comercial

NETAPP, el logotipo de NETAPP y las marcas que constan en http://www.netapp.com/TM son marcas comerciales de NetApp, Inc. El resto de nombres de empresa y de producto pueden ser marcas comerciales de sus respectivos propietarios.