La arquitectura de referencia de Pure Storage para la fábrica de AI empresarial de NVIDIA acelera la inteligencia a escala

La fábrica de AI de Pure Storage con NVIDIA es una plataforma de infraestructura de AI validada diseñada para acelerar los proyectos de AI empresarial. A medida que los modelos de AI y los sistemas de razonamiento se vuelven más complejos, esta solución apunta a acelerar los resultados y simplificar las operaciones.

Diseñado específicamente para la AI empresarial moderna

TI tradicional tiene dificultades para satisfacer las demandas de las cargas de trabajo de AI modernas, lo que garantiza continuamente que los equipos informáticos y de AI sean productivos. Por otro lado, el almacenamiento de AI de alto rendimiento necesita incorporar capacidades empresariales como tiempo de actividad, agilidad y resistencia garantizados. La solución de Pure Storage para el diseño validado de NVIDIA Enterprise AI Factory aborda estos problemas con una arquitectura fácil de operar, integrada, de alto rendimiento y resistente. Admite cargas de trabajo avanzadas como AI generativa y la visión de computación al combinar las plataformas de software y computación acelerada más recientes de NVIDIA con el almacenamiento FlashBlade//S™ de Pure Storage®, lo que ofrece el rendimiento y la escalabilidad necesarios para el entrenamiento y la inferencia de modelos a gran escala. Flashblade//S de Pure Storage es un sistema de almacenamiento certificado por NVIDIA que puede escalar para admitir hasta 1024 GPU NVIDIA y varios petabytes de datos, con redes de alta velocidad que integran estrechamente la computación y el almacenamiento.

Beneficios clave

- **Actualizaciones sin interrupciones:** FlashBlade permite actualizaciones de datos en el lugar, lo que permite la expansión o actualización sin tiempo de inactividad, a diferencia del almacenamiento tradicional que a menudo requiere interrupciones planificadas.
- **Facilidad de uso:** Una plataforma unificada simplifica el proceso de datos de AI desde la ingestión hasta la inferencia, diseñada para que los

desarrolladores empresariales de TI e AI operen con mínimas complicaciones.

- **Rendimiento consistente:** La arquitectura ofrece un rendimiento confiable a cualquier escala, lo que permite a las organizaciones aumentar las implementaciones sin tiempo de inactividad ni pérdida de rendimiento.
- **Agilidad y preparación para el futuro:** El diseño abierto y modular admite tecnologías de AI en evolución y evita el bloqueo de proveedores, maximizando la flexibilidad y la interoperabilidad.
- **Eficiencia:** La alta densidad y el menor consumo de energía de FlashBlade reducen la huella del centro de datos y los costos operativos, mientras que la pila unificada simplifica el mantenimiento.

Configuraciones validadas

Juntos, Pure Storage, NVIDIA y otros socios del sistema están desarrollando productos, software y servicios para acelerar el cambio de TI empresarial a una fábrica de AI. Después del diseño validado por la fábrica de AI empresarial de NVIDIA, Pure lidera la tarea de crear una nueva clase de infraestructura en las instalaciones, que incluye servidores NVIDIA certificado RTX™ PRO y HGX Blackwell, redes Ethernet de NVIDIA Spectrum™-X, DPU NVIDIA BlueField® y software NVIDIA AI Enterprise junto con los sistemas de almacenamiento certificado por NVIDIA de Pure.

Las soluciones de diseño validadas por NVIDIA Enterprise AI Factory se basan en configuraciones de hardware recomendadas de las arquitecturas de referencia empresarial de NVIDIA (RA empresarial), adaptadas para implementaciones de clase empresarial que van de 4 a 32 nodos con 16 a 256 GPU. Cada RA empresarial sigue un patrón de diseño prescriptivo, llamado configuración de referencia, diseñado en torno a un servidor certificado por NVIDIA para garantizar un rendimiento óptimo en un clúster. Estas configuraciones de referencia estandarizan la descripción de los nodos de procesamiento en función de las especificaciones de CPU, GPU, redes y ancho de banda. La nomenclatura C-G-N-B simplifica la selección del sistema al indicar claramente la potencia de procesamiento, las capacidades de red y el rendimiento del ancho de banda. Cada dígito (p. ej., 2-8-5-200) se refiere a la cantidad de sockets de CPU, la cantidad de GPU, la cantidad de adaptadores de red (NIC, SuperNIC o DPU) y el ancho de banda medio este/oeste por GPU (GbE), respectivamente.

Para simplificar la adopción, Pure Storage ha alineado sus sistemas con las configuraciones de referencia utilizadas en el diseño validado por la fábrica de AI

empresarial de NVIDIA, lo que permite la implementación sin esfuerzo y la escalabilidad de sistemas equilibrados que evitan cuellos de botella y un rendimiento subóptimo. A continuación se muestran las configuraciones de referencia repetibles que Pure diseñó y probó.

Estos representan bloques de construcción validados que los clientes y los OEM asociados pueden usar para implementar la solución a diferentes escalas. Esto es lo que significan:

- **2-4-3-200 (PCIe):** Un nodo de computación con 2 CPU, 4 GPU, 3 NIC/DPU de alta velocidad y ancho de banda de red de 200 Gbps por GPU.It podría corresponder a un servidor certificado por NVIDIA con cuatro GPU PCIe (NVIDIA L40S o las GPU NVIDIA RTX PRO 6000 Blackwell Server Edition o NVIDIA H100 NVL para modelos grandes) y tres interfaces Ethernet de 200Gb (a menudo aprovechando las DPU NVIDIA BlueField-3 para descargas). El bloque de construcción 2-4-3-200 ofrece un punto de entrada equilibrado para las fábricas de AI, suficiente para tareas como el entrenamiento de modelos pequeños y el ajuste o el servicio de modelos medianos en la producción. Las empresas pueden comenzar con un clúster de estos servidores 4-GPU sabiendo que la arquitectura se escalará de manera predecible.
- **2-8-5-200 (PCIe):** Un nodo más grande basado en PCIe con 2 CPU, 8 GPU, 5 NIC/DPU y ancho de banda de red de 200 Gbps por GPU. Con ocho GPU en un solo nodo, esta configuración incluye más procesamiento de AI por servidor, adecuado para trabajos de entrenamiento más pesados o granjas de inferencia consolidada. Esta configuración se alinea con el recientemente anunciado NVIDIA RTX PRO Server, con hasta ocho GPU RTX PRO 6000 Blackwell Server Edition. El patrón de diseño “2-8-5” garantiza aún más ancho de banda de red y recuento de adaptadores para manejar el aumento del recuento de GPU, manteniendo un proceso completo de 200 Gb/s para cada GPU. Los casos de uso incluyen la capacitación de modelos medianos y grandes (como modelos de NLP moderados o modelos de visión) y la inferencia de alto rendimiento para muchos usuarios simultáneos. Esta configuración se puede proporcionar a través de servidores como los sistemas PCIe HGX A100/H100 certificados por NVIDIA o los diseños OEM personalizados que cumplen con las especificaciones. Proporciona una excelente opción de escalabilidad horizontal y, al mismo tiempo, utiliza la estructura Ethernet estándar.
- **2-8-9-400 (optimizado por HGX):** Una configuración de alta gama con 2 CPU, 8 GPU (en un factor de forma NVIDIA HGX), 9 adaptadores de red y

redes de 400 Gbps por GPU. Esto corresponde a una plataforma NVIDIA HGX (p. ej., HGX H100 o próxima HGX H200) donde se conectan 8 GPU SXM a través de NVIDIA NVLink con un ancho de banda de GPU a GPU ultra alto (en el orden de 900 GB/s internamente). Las interfaces de red “9” (como las DPU BlueField-3 o las NIC NVIDIA ConnectX-7) proporcionan un rendimiento agregado masivo, 8x 400Gb para las GPU, además de extras para el almacenamiento o la comunicación entre grupos, lo que proporciona de manera eficaz 400Gb/s a cada GPU para las cargas de trabajo más intensivas de I/O. Este patrón 2-8-9-400 está dirigido a los escenarios más exigentes: entrenamiento de modelos grandes (modelos multimillonarios de parámetros), inferencia intensa en entradas de contexto grandes o cualquier carga de trabajo de AI donde el rendimiento máximo y la latencia baja son primordiales. Dada su potencia, naturalmente se adapta a casos de uso como AI generativa avanzada, la simulación compleja o la inferencia a gran escala.

Estas configuraciones de referencia se pueden escalar en todo el clúster. Por ejemplo, una fábrica de AI puede comenzar con (4) nodos 2-8-5-200 y luego expandirse hasta 32nodos por clúster: el almacenamiento FlashBlade y las redes Ethernet NVIDIA Spectrum-X escalarán en consecuencia (se proporciona orientación sobre cuántos chasis y enlaces ascendentes de FlashBlade se necesitan a medida que uno crece, lo que garantiza la escalabilidad lineal). Esto les brinda a las empresas un camino de crecimiento claro y modular: Comience con un bloque de construcción comprobado y agregue más a medida que crece la demanda de AI.

Para garantizar una accesibilidad amplia, NVIDIA y Pure Storage validan soluciones con los principales OEM de servidores para llevar al mercado estos componentes básicos de AI Factory. Los servidores Dell Technologies, HPE y Supermicro se probarán con las plataformas NVIDIA Blackwell combinadas con Pure Storage FlashBlade como soluciones completas de AI Factory. Esto acelerará el diseño y las implementaciones de NVIDIA AI Factory con el proveedor elegido por el cliente y proporcionará la confianza de que las soluciones están completamente validadas según las especificaciones de NVIDIA-Pure Storage.

La fábrica de AI no es un solo dispositivo de caja negra, sino una arquitectura de referencia abierta sobre la que pueden desarrollarse varios proveedores. Las empresas pueden adoptar la solución a través del proveedor de hardware que prefieran o incluso integrar componentes ellos mismos siguiendo el diseño de referencia. Esta franqueza contrasta con algunos enfoques de la competencia que

vinculan a los clientes con el ecosistema de un proveedor. En todos los casos, el cliente se beneficia del diseño validado conjuntamente y del trabajo de integración estricto realizado por los equipos de ingeniería.

Pila de software unificada: NVIDIA Base Command Manager, Run:ai y Portworx de Pure Storage

Una característica destacada de NVIDIA-Pure Storage AI Factory es su pila de software unificada para la organización de datos y computación. Tres componentes clave del software desempeñan un papel aquí: NVIDIA Base Command Manager, Portworx® de Pure Storage y NVIDIA Run:AI. Juntos, forman una capa cohesiva que simplifica la ejecución de la AI a escala:

- **NVIDIA Base Command Manager** es una solución de administración de clústeres que ofrece una implementación rápida y automatiza el aprovisionamiento y la administración de clústeres de cualquier tamaño. En la fábrica de AI, Base Command Manager proporciona capacidades de aprovisionamiento, monitoreo y administración en una sola herramienta que abarca todo el ciclo de vida del clúster. Al integrar Base Command Manager, AI AI Factory garantiza que las organizaciones optimicen el uso de sus recursos de clúster. El software incluye paneles para mostrar la utilización de clústeres y datos de salud para ayudar a ajustar el rendimiento o detectar problemas (por lo tanto, se crean *"herramientas de optimización de infraestructura*"). En resumen, NVIDIA Base Command Manager une el hardware en una sola supercomputadora de AI desde el punto de vista del usuario, lo que elimina la complejidad del clúster subyacente.
- **Portworx de Pure Storage** es una plataforma de datos nativa de Kubernetes que se destaca en la administración del almacenamiento persistente para aplicaciones en contenedores. Dentro de la fábrica de AI, Portworx puede aprovecharse para organizar datos en FlashBlade para cargas de trabajo de AI en contenedores. A medida que el desarrollo de AI utiliza cada vez más procesos basados en contenedores y Kubernetes, Portworx garantiza que los datos estén disponibles para los contenedores dondequiera que se ejecuten, con características como aprovisionamiento dinámico de volumen, snapshots de almacenamiento y replicación entre clústeres. Esto es especialmente importante en escenarios de nube híbrida; por ejemplo, si algunos microservicios de AI se ejecutan en la nube mientras la capacitación principal se realiza en las instalaciones, Portworx puede ayudar a mover o sincronizar datos sin problemas entre entornos.

Proporciona de manera efectiva una capa lista para la nube sobre el almacenamiento rápido de FlashBlade. La integración de Portworx significa que los desarrolladores no tienen que manejar manualmente la logística de datos; pueden solicitar volúmenes de datos a través de Kubernetes y confiar en que la plataforma proporcionará los datos con el rendimiento necesario. En combinación con la velocidad bruta de FlashBlade, esto garantiza que las tareas de AI en contenedores (por ejemplo, servicios de inferencia de modelos o procesos de preparación de datos) obtengan acceso a datos de baja latencia y alta tasa de transferencia según demanda.

- **NVIDIA Run:ai** es una plataforma especializada de organización de cargas de trabajo de AI que maximiza la utilización de GPU y simplifica la programación de trabajos en la fábrica de AI. Run:ai extiende Kubernetes con capacidades de programación inteligentes y basadas en políticas que asignan fracciones de GPU de forma dinámica, trabajos en cola y aplican cuotas que permiten compartir de manera eficiente la infraestructura de GPU entre varios usuarios y equipos. Por ejemplo, si un experimento solo usa una parte de la memoria de una GPU, Run:ai puede programar cargas de trabajo adicionales en la misma GPU, lo que aumenta la utilización de maneras en que los programadores tradicionales no pueden hacerlo. El resultado es un menor tiempo de inactividad y un mayor rendimiento de la carga de trabajo dentro de la fábrica de AI. Juntos, Run:ai y Portworx básicamente crean un *único plano unificado de administración de datos y recursos en GPU y almacenamiento,* lo que proporciona un acceso a datos de latencia más baja y una mejor utilización de recursos para trabajos de AI. Esta unificación cambia las reglas del juego para la productividad: Los recursos informáticos y de datos se organizan en conjunto, automáticamente.

Al aprovechar Base Command Manager, NVIDIA Run:ai y Portworx, la solución AI Factory de NVIDIA-Pure Storage ofrece un enfoque de pila completa. No se trata solo de hardware sin procesar, sino de hardware más software inteligente. Esto significa que las empresas obtienen una plataforma lista para usar que no solo tiene un rendimiento de clase mundial, sino también la automatización inteligente y la integración necesarias para usar ese rendimiento de manera efectiva. Los desarrolladores pueden iniciar experimentos de AI sin preocuparse por montar sistemas de archivos o encontrar dónde residen sus datos, la plataforma los maneja. Los operadores de TI pueden garantizar la gobernanza y la eficiencia con políticas de control detalladas. La pila unificada convierte de manera efectiva la

compleja matriz de GPU, redes y almacenamiento en una experiencia cohesiva para los usuarios finales dentro de la empresa.

Impulse la productividad de los desarrolladores y la eficiencia de los procesos de AI

Uno de los aspectos más emocionantes de la fábrica de AI de NVIDIA-Pure Storage es cómo puede acelerar el trabajo de los desarrolladores de AI y los científicos de datos. Al eliminar los cuellos de botella y la complejidad de la infraestructura, permite que los equipos se enfoquen en crear modelos y soluciones en lugar de lidiar con hardware y plomería de datos.

Así es como la fábrica de AI mejora la productividad en todo el proceso de AI:

- **Acceso a datos simplificado:** En muchas empresas, los datos están aislados en diferentes sistemas de almacenamiento, y obtener los datos correctos para alimentarse en experimentos de AI es un proceso lento y manual. Con el almacenamiento FlashBlade unificado de AI Factory, todas las etapas de los datos de AI, desde la ingesta sin procesar hasta las características refinadas, residen en un repositorio de alta velocidad. Esto significa que los clústeres de capacitación y los servidores de inferencia pueden acceder a conjuntos de datos directamente, sin pasos tediosos de copia o transferencia. Además, la capacidad de FlashBlade para atender cargas de trabajo de archivos y objetos significa compatibilidad con una amplia gama de marcos de trabajo de AI y formatos de datos. Los desarrolladores pueden usar protocolos estándar (como NFS o S3) para leer/escribir datos a velocidades de GPU. La integración de las redes RoCE garantiza aún más que las GPU puedan extraer datos del almacenamiento con latencia mínima y sobrecarga de CPU. En la práctica, esto se traduce en tiempos de carga de datos más cortos y la capacidad de iterar más rápido. Un trabajo de capacitación que podría haberse estancado esperando I/O ahora continúa sin obstáculos, manteniendo los costosos recursos de GPU completamente utilizados.
- **Ajuste mínimo de la infraestructura:** Sintonizar la infraestructura para la AI puede consumir innumerables horas de esfuerzo de TI. La solución Pure Storage con NVIDIA Enterprise AI Factory reduce en gran medida esta carga. Esta arquitectura de referencia está validada previamente y equilibrada para las cargas de trabajo de AI, lo que significa que el rendimiento listo para usar ya está optimizado. El software NVIDIA Base Command (incluido en la pila) proporciona herramientas para optimizar la

utilización de GPU y manejar la programación, para que los administradores no tengan que escribir sus propias soluciones para administrar clústeres de varios usuarios. En general, las opciones de diseño de la plataforma, como el uso de Ethernet y RDMA estándar, en lugar de las interconexiones exclusivas ocultas, también significan que se conecta fácilmente a entornos existentes. Los equipos de TI pasan menos tiempo trabajando y combatiendo incendios, y los desarrolladores de AI pasan menos tiempo esperando. Como resultado, las organizaciones pueden lograr ciclos de iteración más rápidos, más experimentos, ajustes de modelos y pruebas en la misma cantidad de tiempo, acelerando el camino de la idea a la comprensión.

- **Integración integral del proceso:** La fábrica AI de NVIDIA con Pure Storage no se trata solo del entrenamiento, sino de un enfoque holístico que cubre la preparación, el entrenamiento, la validación y la implementación de datos. Por ejemplo, considere un flujo de trabajo de AI empresarial típico: Los ingenieros de datos ingieren y seleccionan grandes conjuntos de datos, los científicos de datos capacitan modelos en clústeres de GPU y los ingenieros de ML implementan esos modelos para inferencia en la producción. Tradicionalmente, cada etapa puede ocurrir en sistemas separados, lo que requiere que los datos se muevan y los entornos se administren por separado. En esta solución, se puede unificar todo el proceso. Un solo FlashBlade puede alojar los datos sin procesar, los datos preprocesados intermedios, los puntos de verificación del modelo y los archivos del modelo final, además de las incrustaciones RAG y el contexto de caché KV para la inferencia. La misma infraestructura de GPU se puede particionar (con la programación adecuada) para manejar trabajos de capacitación y servicios de inferencia. Esta cohesión significa que tan pronto como se entrena un modelo, se puede probar y servir a los usuarios en la misma plataforma, lo que reduce drásticamente el tiempo de implementación. Además, el entorno consistente reduce los errores al hacer la transición de los modelos de desarrollo a producción. Todo esto lleva a un proceso de desarrollo de AI ágil en el que las ideas progresan rápidamente a implementaciones de producción, sin la fricción tradicional entre equipos y sistemas en silos.

En última instancia, al ofrecer rendimiento con sencillez, la fábrica de AI permite que las empresas logren lo que antes era muy difícil: una plataforma de AI que es increíblemente rápida y operativamente eficiente. Esta combinación produce beneficios comerciales tangibles: resultados más rápidos a partir de los datos, la

capacidad de iterar en los modelos de AI con más frecuencia y un equipo de desarrollo más productivo que no está abrumado por preocupaciones de infraestructura.

FlashBlade: Una estructura de almacenamiento optimizada para cargas de trabajo de AI

En el corazón de la fábrica de AI empresarial de NVIDIA con Pure Storage se encuentra FlashBlade//S, una plataforma de almacenamiento basada íntegramente en tecnología flash y con escalabilidad horizontal diseñada específicamente para el análisis moderno y las cargas de trabajo de AI. En esta solución, FlashBlade funciona como la plataforma de datos de alto rendimiento que respalda tanto el entrenamiento de AI (que requiere leer conjuntos de datos masivos de manera eficiente) como la inferencia de AI (que se beneficia del acceso rápido y de baja latencia a los datos de referencia y archivos de modelo). La arquitectura de FlashBlade es especialmente adecuada para estos desafíos. Proporciona rendimiento multidimensional y escalabilidad; baja latencia, acceso paralelo a datos; operaciones resistentes y sin interrupciones; y la capacidad de unificar el almacenamiento de archivos y objetos, lo que permite a los científicos e ingenieros de datos trabajar con datos en cualquier formato que sus herramientas esperen, sin recurrir al uso de silos de almacenamiento diferente.

FlashBlade de Pure Storage proporciona a una fábrica de AI empresarial de NVIDIA una base de datos de alto rendimiento y alta eficiencia, que ofrece capacidad multipetabyte, decenas de terabytes por segundo de rendimiento y latencia baja consistente a escala. Igual de importante, ofrece este rendimiento con las características empresariales (instantáneas, cifrado, replicación) y la facilidad de uso que TI empresarial espera, a diferencia de los sistemas de archivos HPC de nicho.

Conclusión: Acelerar la innovación de AI con confianza

El lanzamiento de la fábrica de AI empresarial de NVIDIA con Pure Storage marca un hito significativo en la evolución de la infraestructura de AI. Combina lo mejor de dos mundos: Las capacidades de computación y organización de AI de vanguardia de NVIDIA y la innovadora tecnología de almacenamiento flash de Pure Storage. Para los líderes de AI empresarial, esto significa que finalmente pueden implementar una plataforma de AI que ofrezca rendimiento a nivel de supercomputadora y confiabilidad/eficiencia de nivel empresarial en un solo

paquete. La capacidad de la plataforma para manejar cargas de trabajo de AI dinámicas y multimodales con facilidad mientras simplifica las operaciones y escala aumenta la agilidad organizacional en un dominio donde la velocidad importa.

Al aprovechar esta solución de ingeniería conjunta, las empresas pueden transformar sus centros de datos en verdaderas “fábricas de AI”, instalaciones que ingieren datos de forma continua, capacitan modelos de AI y producen resultados o servicios inteligentes que impulsan el negocio. La arquitectura de referencia de Pure Storage con NVIDIA Enterprise AI Factory ayuda a las organizaciones a eliminar las barreras históricas que ralentizaron estas iniciativas: El manejo de datos ya no será un obstáculo en la tasa de transferencia, la adición de capacidad ya no requerirá tiempo de inactividad y los desarrolladores ya no se verán obstaculizados por las peculiaridades de la infraestructura. En cambio, obtienen una plataforma optimizada y de alto octanaje para experimentar rápidamente y ofrecer resultados de AI.

A medida que el panorama de la AI continúa evolucionando (con modelos cada vez más grandes y aplicaciones emergentes en AI generativa, análisis en tiempo real y más), es clave contar con una base flexible y potente. La fábrica de AI proporciona esa base: rendimiento a escala, adaptabilidad para el futuro y elegancia operativa incorporada. Para cualquier organización que apunte a liderar en la era de la AI, la fábrica de AI de NVIDIA con Pure Storage ofrece un camino convincente hacia el futuro: Acelere sus proyectos de AI hoy, con la confianza de que su infraestructura satisfará las necesidades del futuro. Esta es una infraestructura de AI reinventada para agilidad, rendimiento y éxito.

Más información:

- Descargue Pure Storage con la arquitectura de referencia de fábrica AI empresarial de NVIDIA
- Soluciones de AI de Pure Storage
- FlashBlade//S

***Para más de 1024 entornos de GPU, explore FlashBlade//EXA™.**