

# La IA y los superchips

---

## AI and superchips

**Andrea Parra-2210062 , Christian Orduz-2152104, Milton Monsalve-2204004, Cristian Orduz-2211877, Oscar Mongui 2215104**

**Grupo A2**

**Junio de 2024**

### Resumen

Como ingenieros de sistemas es de suma importancia adaptarse e informarse sobre las nuevas tecnologías que entran al mercado y la academia. En vista del gran boom de las inteligencias artificiales y la creciente demanda de recursos de hardware para su entrenamiento, el presente artículo evaluará una posible solución a un problema relevante con su funcionamiento; haciendo uso de un superchip NVIDIA GH200 Grace-Hopper con cincuenta terabytes de almacenamiento, y analizando las necesidades en términos de procesamiento, almacenamiento, energía y demás, con los conocimientos discutidos en el curso.

**PALABRAS CLAVE:** Inteligencia Artificial, NVIDIA, CPU, GPU, Superchip

### Abstract

As systems engineers, it is of utmost importance to adapt and inform ourselves about the new technologies entering the market and academia. In view of the great boom of artificial intelligence and the growing demand of hardware resources for its training, this article will evaluate a possible solution to a problem relevant with its operation; making use of a superchip NVIDIA GH200 Grace-Hopper with fifty terabytes of storage, and analyzing the requirements in terms of processing, storage, energy, and more, using the knowledge discussed in the course

**KEYWORDS:** Artificial Intelligence, NVIDIA, CPU, GPU, Superchip

### I. INTRODUCCIÓN

La inteligencia artificial se ha tomado el mundo, aplicaciones como chatGPT, Copilot o Alexa forman parte de la cotidianidad de muchas personas, y sus posibles usos en el mercado no han sido ignorados, véase los reconocedores faciales o los carros autónomos. Sin embargo, lo que la persona promedio no conoce es la exigencia tecnológica, y en consecuencia energética, de estos sistemas. El procesamiento y gestión de grandes cantidades de datos a velocidades descomunales requiere de unidades capaces para la tarea, y puesto a que los modelos cada vez mejoran más, la industria de hardware se ha vuelto muy lucrativa; véase por ejemplo el caso de NVIDIA, que se convierte en la empresa más valiosa del mundo tras un impresionante repunte en sus acciones, destacándose el papel que desempeñó la demanda de sus chips, el estándar de oro en los espacios de IA (*AI Fever Drives Nvidia's Rise To World's Most Valuable Company, s. f.*).

Consecuentemente, el entendimiento y correcto funcionamiento de estos productos nos concierne en el ámbito de la arquitectura de computadores, en palabras de Carlos Barrios el docente de curso: *“el propósito fundamental de la asignatura es establecer un estado del arte de conocimientos fundamentales en arquitectura de computadores, que permita manejar el lenguaje técnico asociado, ubicar temporalmente el desarrollo tecnológico, conociendo tendencias y fundamentar conocimientos que permitan el auto-aprendizaje y profundización en el área, además de la interacción en equipos interdisciplinarios que requieran competencias en arquitectura de sistemas computacionales.”*. Esto se traduce en la solución de problemas complejos relacionados con la operación de nuevos sistemas computacionales, como lo vendrían siendo los nuevos chips desarrollados por NVIDIA para IA.

En este contexto, el desafío que nos concierne como grupo e ingenieros es proporcionar 50 terabytes de capacidad operativa utilizando la solución Nvidia Grace Hopper, todo mientras se minimiza el consumo de energía. Ello incluye recomendar un sistema de copia de seguridad y archivado de datos eficiente, evaluar costos y fabricantes, y en general hacer una investigación exhaustiva para dar una respuesta completa y competente. A continuación nuestra respuesta

### II. SOLUCIÓN

Ya siendo presentado el problema a solucionar procedemos a explicar la arquitectura del super chip NVIDIA GH200 Grace Hopper, donde con sus especificaciones de diseño para atacar problemas de IA así mismo como computación de alto rendimiento (HPC), siendo un diseño eficiente que puede ofrecer hasta 10 veces un rendimiento mayor para aplicaciones con una gran cantidad de flujo de datos(terabytes), permitiendo así solventar problemas más complejos.

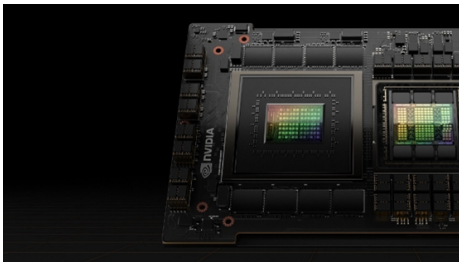


Ilustración 1. Superchip NVIDIA Grace Hopper GH200

Este superchip combina arquitecturas NVIDIA Grace y NVIDIA Hopper. NVIDIA Grace con 72 núcleos ARM Neoverse V2 es un tipo de CPU Arm(Advanced RISC Machine/máquina avanzada de RISC) conjunto de instrucciones reducido, lo que significa que cada instrucción realiza una operación simple, este diseño surgió para optimizar los servidores que están orientados para la nube, mejorando la seguridad y la velocidad de flujo de datos. Estas CPUs están diseñadas como System-on-a-Chip (SoC), de manera que tanto CPU y GPU están conectadas en el mismo hardware, para lograr un rendimiento uniforme y eficiente. Esta CPU es conectada a una GPU NVIDIA Hopper usando NVIDIA NVLink-C2C. NVIDIA, siendo este tipo de conexión diseñado para maximizar el rendimiento y la eficiencia de la comunicación entre chips. La GPU presente en el superchip es la NVIDIA H100, que con su estructura de alrededor de 80 mil millones de transistores entrega resultados óptimos para ejecutar las diversas aplicaciones de inteligencia artificial.

Especificación	Detalles
GH200 Grace Hopper Superchip	
CPU Core Architecture	Armv9-A Neoverse V2 Cores with 4x128b SVE2
CPU Core Count	72
CPU Cache	L1: 64KB I-cache + 64KB D-cache per core
	L2: 1MB per core
	L3: 117MB
CPU Memory Technology	LPDDR5X with ECC in the same package.
CPU Raw Memory BW	Up to 500 GB/s
CPU Memory Size	Up to 480GB
GPU Multi-Processor Architecture	Hopper SM compute capability 9.0
GPU Multi-Processor Count	132
GPU Memory Technology	High-Bandwidth Memory HBM3
	High-Bandwidth Memory HBM3e
GPU Memory Size	96GB HBM3
	144GB HBM3e
Power	550-1000W TDP with Memory, 12V Supply

Ilustración 2. Especificaciones Superchip NVIDIA Grace Hopper GH200

A medida que la complejidad de los modelos de IA se ha incrementado, la tecnología ha ido evolucionando para adaptarse a estos modelos. Cada vez son más los recursos necesarios de hardware necesarios para ofrecer soluciones a los distintos modelos. Donde también es necesario el uso de enormes capacidades de almacenamiento para manejar la gran cantidad de información guardada.

Se decide escoger un sistema de almacenamiento red (Network Attached Storage, NAS), ya que con su fácil acceso multiusuario permite que distintos usuarios pueda acceder a los datos simultáneamente, así mismo facilita la administración de datos desde un único punto de control, con el uso de SSDs NVMe de 2 TB cada uno (capacidad real aproximada: 1.86 TB por unidad debido al sistema de numeración binario o. decimal), usando la configuración RAID 6, porque ofrece una buena combinación de lectura y escritura, así como el procesamiento ágil en la distribución de datos, también tiene una gran tolerancia a fallos ofreciendo un sistema robusto en la recuperación de datos.

Se calculó la cantidad necesaria de SSDs NVMe requeridos usando (N-2)\*capacidad real por disco, donde N es el número total de discos. Ya que son requeridos 50TB de almacenamiento utilizable, se procede a hacer el cálculo.

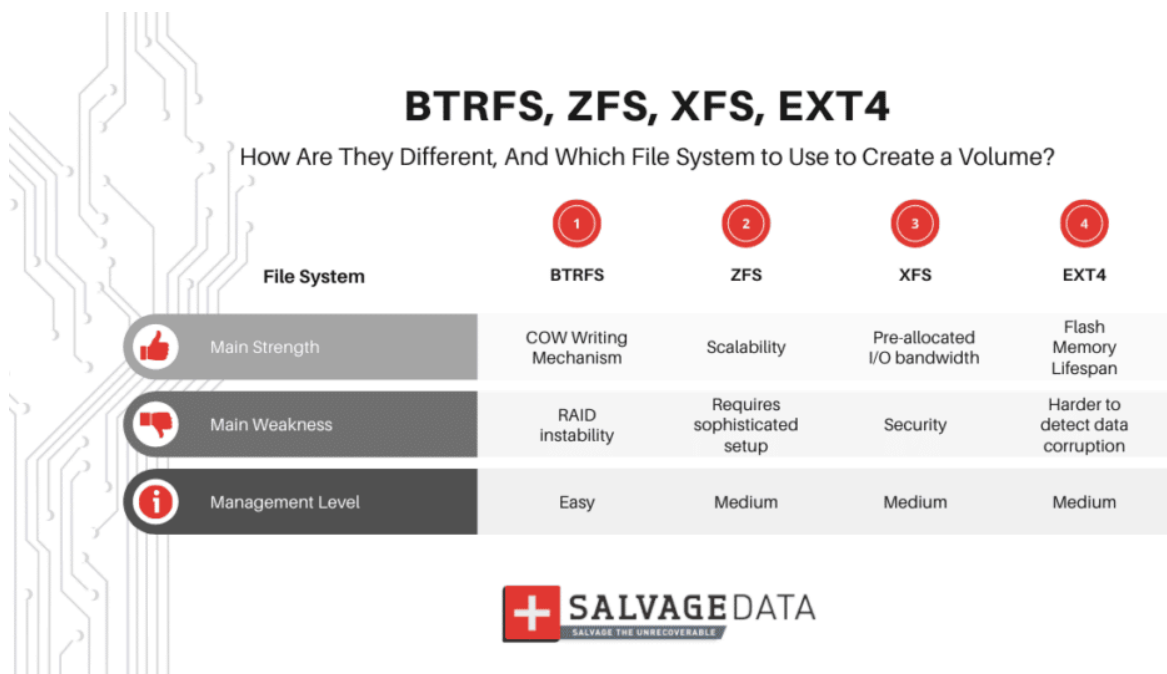
- $50 \text{ TB} / 1.86 \text{ TB} \approx 26.88 + 2$  (para paridad)  $\approx 29$  discos

Procediendo así con la configuración final del almacenamiento, usando 30 SSDs NVMe de 2TB cada uno en configuración RAID 6, dando una capacidad total bruta de 60TB, cuya capacidad operacional es de

- $(30 - 2) * 1.86 \text{ TB} \approx 52.08 \text{ TB}$

B. Ahora bien, si tomamos en cuenta la respuesta anterior, existen varios sistemas de copia de seguridad y archivado de datos óptimos y de calidad, por lo que es necesaria una comparación entre estas posibilidades. En primera instancia, los sistemas de archivado no funcionan meramente como una forma de almacenar datos; trabajan similar a una tabla de contenido en el sentido de que nombran, guardan y recuperan archivos de un dispositivo de almacenamiento. *Reza Lavarian* lo ejemplifica con las computadoras, pues cada vez que abres o modificas un archivo en tu dispositivo un sistema de archivos también debe verse involucrado.

Hecha esta salvedad, escoger un sistema depende de varios factores subjetivos acorde a lo que se busque; en la figura 1, hecha por *SalvageData*, puede verse una comparación resumida entre sistemas ZFS, BTRFS, XFS y EXT4. En todo caso, véase la siguiente lista más detallada de las ventajas y/o funcionalidades de cada sistema (Pompeu, 2022b):



File System	1 BTRFS	2 ZFS	3 XFS	4 EXT4
Main Strength	COW Writing Mechanism	Scalability	Pre-allocated I/O bandwidth	Flash Memory Lifespan
Main Weakness	RAID instability	Requires sophisticated setup	Security	Harder to detect data corruption
Management Level	Easy	Medium	Medium	Medium

**SALVAGEDATA**  
SALVAGE THE UNRECOVERABLE

Ilustración 3

#### Ext4

- EXT4 es la mejor opción para las necesidades de SOHO (Small Office/Home Office) y proyectos que requieren un rendimiento estable.
- EXT fue el primero en usar un interruptor de sistema de archivos virtual (VFS), por lo que permitió que Linux soportara múltiples sistemas de archivos al mismo tiempo en el mismo sistema.
- Extiende la vida útil de la memoria a través de la asignación diferida, lo que a su vez mejora el rendimiento y reduce la fragmentación al asignar efectivamente mayores cantidades de datos a la vez.
- Número ilimitado de subdirectorios y verificación del sistema de archivos más rápida.
- Si se tienen múltiples discos — y por lo tanto paridad o redundancia de la cual se puede recuperar teóricamente datos corruptos — EXT4 no tiene manera de saber eso y usarlo a favor de uno

#### Btrfs

- Ideal para grandes empresas que necesitan un sistema de archivos fácil de gestionar; adecuado para tecnologías y proyectos donde no se requiere alta tolerancia a fallos.
- Diseñado originalmente para abordar la falta de agrupación, checksums, snapshots y expansión integral de múltiples dispositivos en los sistemas de archivos de Linux.

- Sistema de archivos basado en el mecanismo de copy-on-write (COW), es decir, al modificar un archivo, el sistema de archivos no sobrescribe los datos existentes en el disco con la nueva información, sino que los datos nuevos se escriben en otra ubicación
- Se enfoca en funcionalidades avanzadas en tolerancia a fallos y reparación como subvolúmenes, autorreparación, crecimiento y reducción de volumen en línea, compresión de archivos, desfragmentación, deduplicación

## ZFS

- Está diseñado para funcionar con productos Oracle, como mainframes, entornos de servidores en clúster y supercomputadoras. Por lo tanto, algunos de los beneficios de ZFS no son aplicables a pequeñas empresas y usuarios privados.
- Va más allá de la funcionalidad básica de un sistema de archivos, pudiendo servir tanto como LVM o RAID
- Permite agregar dispositivos de almacenamiento adicionales al sistema actual y obtener de inmediato nuevo espacio en todos los sistemas de archivos existentes en ese conjunto
- Soporta una capacidad de almacenamiento de datos y metadatos casi ilimitada (hasta 1 billón de terabytes).
- Protección extensa contra la corrupción de datos comparado con otros sistemas de archivos
- Compresión de datos eficiente, snapshots y clones copy-on-write.
- Muchos de sus procesos dependen de la RAM, por lo que ZFS consume mucha memoria. En general necesita recursos computacionales o de servidor muy potentes para funcionar a una velocidad adecuada, por lo que no es la mejor opción para arquitecturas de microservicios y hardware débil.

## XFS

- Útil cuando se manejan archivos grandes, enormes almacenes de datos, proyectos científicos a gran escala o proyectos empresariales de gran tamaño.
- Proporciona una excelente escalabilidad de los hilos de I/O, el ancho de banda del sistema de archivos y el tamaño de los archivos y del sistema de archivos en sí cuando abarca múltiples dispositivos de almacenamiento físico.
- Garantiza la consistencia de los datos mediante el registro de metadatos y el mantenimiento de barreras de escritura.
- Deficiente contra la 'pudrición de bits', lo que causa una casi total incapacidad para recuperar archivos en caso de pérdida de datos.
- La asignación diferida ayuda a prevenir la fragmentación del sistema de archivos, y también se admite la desfragmentación en línea.

Así pues, basándonos en la solución previa y los componentes disponibles, ZFS viene a ser la mejor opción; pues soporta una gran capacidad de almacenamiento, ofrece mucha protección contra la corrupción de datos y permite un manejo eficiente y seguro de estos. Algo a destacar es su posibilidad de servir como RAID (Redundant Array of Independent Disks) con RAID-Z, lo que le permite la distribución de datos y paridad a través de múltiples discos dentro de un pool de almacenamiento (zpool), flexibilidad en cuanto a la cantidad de redundancia deseada según las necesidades y un buen rendimiento para entornos de gran escala como vendría siendo el entrenamiento de una IA.

### c. Evaluación de costos e idoneidad:

Parámetros para comparar:

#### 1. Rendimiento en operaciones de IA (FLOPS, latencia)

La NVIDIA GH200 Grace-Hopper debería ofrecer un alto rendimiento en términos de FLOPS debido a su arquitectura diseñada para cálculos intensivos. Específicamente, las tarjetas NVIDIA con núcleos tensoriales (como Tensor Cores en arquitecturas más recientes) están optimizadas para acelerar operaciones matemáticas fundamentales en IA.

#### 2. Consumo energético (W)

Se proporciona dos especificaciones claves: Consumo típico: Es la cantidad de energía bajo condiciones de carga de trabajo normales que va entre 450 W a 1000 W

Consumo Máximo: Es la cantidad máxima de energía que puede consumir en picos de carga máxima la cual puede ser entre 1000 W a 2000 W con el sistema de NVidia NVL2 activado.

#### 3. Costo total de propiedad (TCO)

El TCO comienza con el costo inicial de adquisición de la tarjeta gráfica, que en este caso es de 210,547,875 millones de pesos colombianos. Este es un costo considerable que impacta directamente en el presupuesto inicial del proyecto de inteligencia artificial.

Costos operativos varían según su consumo de energía y pues tiene un consumo significativo que puede resultar en costos operativos elevados, también el mantenimiento y reparaciones de forma regular para mantener un rendimiento óptimo y aumentar la vida útil de la tarjeta.

Costo de infraestructura, una infraestructura necesaria para soportar y optimizar el rendimiento de la tarjeta, como sistemas de refrigeración avanzados o actualizaciones de la fuente de alimentación.

Costos de Actualización y Ciclo de Vida hay que planificar los costos asociados con futuras actualizaciones de hardware o tecnología, así como la depreciación del valor de la tarjeta a medida que pasa el tiempo y surgen nuevas tecnologías.

#### 4. Escalabilidad

Configuraciones Multi-Tarjeta esto permite combinar múltiples tarjetas para aumentar el rendimiento computacional total.

Gestión de recursos sea eficiente y efectiva para así minimizar los cuellos de botella y maximizar el rendimiento global del sistemas.

La escalabilidad crecerá de manera efectiva y eficiente a medida del tiempo, adaptándose a nuevas cargas de trabajo y requisitos sin comprometer el rendimiento ni la estabilidad del sistema.

La solución propuesta es adecuada para aplicaciones de IA por:

- Alto rendimiento de la GH200 para cargas de IA
- Almacenamiento de alta velocidad con SSDs NVMe
- Equilibrio entre rendimiento y redundancia con RAID 6
- Sistema de respaldo robusto para protección de datos.

**D. En el mercado hay varias opciones de fabricantes para elegir, para este caso decidimos realizar la comparación entre Lenovo y HPE.**

#### Lenovo

Centrada en una visión audaz de ofrecer tecnología más inteligente para todos, Lenovo ha aprovechado su éxito como la empresa de PC más grande del mundo al expandirse aún más hacia áreas de crecimiento que impulsan el avance de las 'nuevas tecnologías de TI' (cliente, borde, nube, red y inteligencia) incluyendo servidores, almacenamiento, dispositivos móviles, software, soluciones y servicios. Esta transformación, junto con la innovación revolucionaria de Lenovo, está construyendo un futuro más inclusivo, confiable y más inteligente para todos, en todas partes.

- **Diseño Modular:** Los servidores ThinkSystem están diseñados para ser modulares y flexibles, permitiendo configuraciones personalizadas y expansiones fáciles.
- **Interconexión:** Uso de tecnología PCIe Gen 4 para alta velocidad de interconexión entre componentes internos.
- **Eficiencia Energética:** Diseño optimizado para eficiencia energética y reducción de costos operativos.

#### Ventajas:

- **Flexibilidad:** Alta modularidad y opciones de configuración.
- **Costo-efectivo:** Generalmente más accesible en términos de costo total de propiedad (TCO).

#### Desventajas:

- **Menor innovación:** Lenovo se queda un poco atrás en temas de innovación (como energética o de gestion)

#### Ejemplo de lenovo:

**Lenovo HG650N** quien permite implementaciones de NVIDIA GH200 Grace Hopper

#### HPE

Hewlett Packard Enterprise (NYSE: HPE) es la compañía global edge-to-cloud que ayuda a las organizaciones a acelerar sus resultados de negocio, maximizando el valor de todos sus datos en cualquier lugar. Tras décadas reimaginando el futuro e innovando para mejorar la forma en que las personas viven y trabajan, HPE ofrece, como servicio, soluciones tecnológicas únicas, abiertas e inteligentes. Con una oferta que abarca desde los servicios cloud a la computación de alto rendimiento y la Inteligencia Artificial, pasando por el extremo inteligente, el software o el almacenamiento, HPE proporciona una experiencia

sólida y estable en todas las nubes y en todos los extremos para ayudar a los clientes a desarrollar nuevos modelos de negocio, impulsar su transformación y aumentar el rendimiento operativo

- **Innovación:** Líder en tecnologías avanzadas como infraestructura componible y memoria persistente.
- **Escalabilidad:** Alta escalabilidad y capacidad de integración con otras soluciones HPE.
- **Seguridad:** Tecnologías avanzadas de seguridad integradas en la arquitectura del hardware.

#### Desventajas:

- **Costo:** más caro en términos de inversión inicial y TCO.
- **Complejidad:** Mayor complejidad en la implementación y gestión.

#### Ejemplo de HPE:

**Ex254n** con una de sus características principales los 2 nodos de superchip de la GH200 Grace Hopper

#### Sistemas de enfriamiento:

##### Enfriamiento por Aire

**Descripción:** El enfriamiento por aire utiliza ventiladores y disipadores de calor para disipar el calor generado por los componentes del sistema.

#### Ejemplo de Sistema de Enfriamiento por Aire:

- **Noctua NH-D15:** Uno de los disipadores de aire más eficientes del mercado, adecuado para CPU de alto rendimiento.
- - **Precio:** Aproximadamente \$90 USD.
- **Cooler Master Hyper 212 EVO:** Otro disipador popular y eficiente para sistemas de alto rendimiento.
  - **Precio:** Aproximadamente \$40 USD.



1. **Costo Inicial Alto:** Los componentes y la instalación son significativamente más caros.
2. **Mantenimiento Especializado:** Requiere conocimientos especializados para instalación y mantenimiento. Además, existe el riesgo de fugas de agua.
3. **Riesgo de Daño:** Una fuga puede causar daños significativos a los componentes electrónicos.

#### Costos:

- **Costo inicial:** \$150 - \$300 USD por unidad de sistema de enfriamiento líquido todo en uno de alta gama.
- **Costo de operación:** Moderado a bajo, ya que es más eficiente en términos de consumo de energía (aproximadamente \$5 - \$15 USD por año por bomba de agua).
- **Costo de mantenimiento:** Alto, debido a la necesidad de inspecciones regulares y reemplazo de líquidos refrigerantes (aproximadamente \$30 - \$50 USD cada 2-3 años).

#### Ejemplo de Costos Totales para una Configuración Completa:

### Sistema con 2 unidades Corsair Hydro Series H150i PRO RGB:

- **Costo inicial:**  $2 \times \$160 = \$320$  USD
- **Costo de operación anual:**  $2 \times \$10 = \$20$  USD
- **Costo de mantenimiento cada 3 años:**  $2 \times \$50 = \$100$  USD

### Ventajas:

1. **Costo Inicial Más Bajo:** Los sistemas de enfriamiento por aire son generalmente más económicos de adquirir e instalar.
2. **Facilidad de Instalación y Mantenimiento:** Estos sistemas son más simples de instalar y mantener, no requieren de conocimientos especializados aparte la cantidad de técnicos con la capacidad de hacerle el respectivo mantenimiento es mucho mayor.
3. **Fiabilidad:** Son menos propensos a fallas catastróficas (por ejemplo, fugas de agua) y son bastante robustos.

### Desventajas:

1. **Eficiencia Limitada:** Menor capacidad para disipar grandes cantidades de calor, lo que puede ser insuficiente para sistemas de alta densidad y rendimiento.
2. **Ruido:** Los ventiladores pueden generar ruido significativo, especialmente en entornos con múltiples unidades lo cual el usuario puede volverse molesto.
3. **Requerimientos de Espacio:** Puede requerir más espacio físico debido a la necesidad de múltiples ventiladores y espacios de flujo de aire llevando a ocupar espacios que se pueden aprovechar para otras cosas igual o más importantes.

### Costos:

- **Costo inicial:** \$40 - \$100 USD por unidad de disipador de alta gama.
- **Costo de operación:** Moderado, debido al consumo constante de energía de los ventiladores (aproximadamente \$10 - \$20 USD por año por ventilador).
- **Costo de mantenimiento:** Bajo, requiere limpieza y reemplazo ocasional de ventiladores (aproximadamente \$10 USD por ventilador cada 2-3 años).

### Enfriamiento por Agua

**Descripción:** El enfriamiento por agua utiliza líquidos refrigerantes que circulan a través de bloques de agua montados en los componentes del sistema, transfiriendo el calor a un radiador donde se disipa al aire.

### Ejemplo de Sistema de Enfriamiento por Agua:

- **Corsair Hydro Series H150i PRO RGB:** Un sistema de enfriamiento por agua todo en uno adecuado para CPU de alto rendimiento.
  - **Precio:** Aproximadamente \$160 USD.



- **NZXT Kraken X62:** Otro sistema de enfriamiento por agua popular y eficiente.



- **Precio:** Aproximadamente \$150 USD.

#### Ventajas:

1. **Alta Eficiencia Térmica:** Mejor capacidad para manejar y disipar grandes cantidades de calor, ideal para sistemas de alto rendimiento y sistema de operación.
2. **Menor Ruido:** Los sistemas de enfriamiento por agua suelen ser más silenciosos que los de aire, ya que requieren menos ventiladores lo cual para el usuario va a resultar más agradable.
3. **Compacto:** Puede ser más eficiente en términos de espacio en configuraciones de alta densidad.

#### Desventajas:

1. **Costo Inicial Alto:** Los componentes y la instalación son significativamente más caros.
2. **Mantenimiento Especializado:** Requiere conocimientos especializados para instalación y mantenimiento. Además, existe el riesgo de fugas de agua.
3. **Riesgo de Daño:** Una fuga puede causar daños significativos a los componentes electrónicos.

#### Costos:

- **Costo inicial:** \$150 - \$300 USD por unidad de sistema de enfriamiento líquido todo en uno de alta gama.
- **Costo de operación:** Moderado a bajo, ya que es más eficiente en términos de consumo de energía (aproximadamente \$5 - \$15 USD por año por bomba de agua).
- **Costo de mantenimiento:** Alto, debido a la necesidad de inspecciones regulares y reemplazo de líquidos refrigerantes (aproximadamente \$30 - \$50 USD cada 2-3 años).

Comparación Técnica		
Detalle	Enfriamiento por aire	Enfriamiento por agua
Eficiencia energética	Menos eficiente, más energía gastada en mover aire para enfriar componentes.	Más eficiente, mejor transferencia de calor, menos energía necesaria para mantener temperaturas bajas.
Requerimiento de espacio	Requiere más espacio debido a los ventiladores y rutas de aire.	Más compacto, mejor para instalaciones de alta densidad.



Instalación y mantenimiento	Fácil instalación y mantenimiento, bajo costo.	Instalación y mantenimiento más complejos, costos más altos.
Ruido	Generalmente más ruidoso debido a los ventiladores.	Más silencioso, menos ventiladores necesarios.
Fiabilidad	Alta fiabilidad, menos componentes que pueden fallar catastróficamente.	Riesgo de fugas, mantenimiento crítico para evitar fallos.

### Ejemplo de Comparación de Costos y Eficiencia

#### Sistema de Enfriamiento por Aire (Noctua NH-D15):

- **Costo inicial:** \$90 USD
- **Costo de operación anual:** \$15 USD
- **Costo de mantenimiento cada 3 años:** \$10 USD
- **Costo total en 3 años:**  $90 + (3 \times 15) + 10 = 145$  USD

#### Sistema de Enfriamiento por Agua (Corsair Hydro Series H150i PRO RGB):

- **Costo inicial:** \$160 USD
- **Costo de operación anual:** \$10 USD
- **Costo de mantenimiento cada 3 años:** \$50 USD
- **Costo total en 3 años:**  $160 + (3 \times 10) + 50 = 210$  USD

Si se prioriza el costo inicial y la facilidad de mantenimiento, el enfriamiento por aire es la mejor opción.

Si se prioriza la eficiencia térmica y el ruido reducido en un entorno de alta densidad y alto rendimiento, el enfriamiento por agua sería más adecuado a pesar del mayor costo inicial y de mantenimiento.

### III. CONCLUSIONES

- Al integrar la potencia de la CPU ARM Neoverse V2 con 72 núcleos y la GPU NVIDIA Hopper mediante NVLink-C2C, el superchip NVIDIA GH200 Grace Hopper, ofrece una solución integral para el uso en aplicaciones exigentes de IA y HPC. El tipo de arquitectura otorga un rendimiento excepcional y eficiencia energética, también establece un estándar en la capacidad de procesamiento y comunicación entre chips.
- El uso de un sistema NAS con SSDs NVMe de 2TB entrega una solución correcta haciendo uso de estos en una configuración de RAID 6 para alcanzar la capacidad operacional requerida.
- Para el sistema con la NVIDIA GH200 Grace-Hopper y un almacenamiento de 50 TB, donde se espera un alto rendimiento y carga de trabajo intensiva, **la recomendación es optar por un sistema de enfriamiento por agua**, debido a su mayor eficiencia en la gestión térmica y la capacidad de soportar aplicaciones intensivas de inteligencia artificial.
- Para el entrenamiento de modelos de IA, ZFS es un muy buen candidato debido a las capacidades que le entrega RAID-Z. Su capacidad para escalar y manejar el almacenamiento de grandes cantidades de datos sin comprometer la seguridad lo convierte en una solución ideal para sistemas tan exigentes como estos

## REFERENCIAS

- [1] AI fever drives Nvidia's rise to world's most valuable company. (s. f.). Reuters.  
<https://www.reuters.com/technology/artificial-intelligence/ai-fever-drives-nvidias-rise-worlds-most-valuable-company-2024-06-18/>
- [2] Lavarian, R. (2022, 12 enero). *What Is a File System? Types of Computer File Systems and How they Work – Explained with Examples*. freeCodeCamp.org. <https://www.freecodecamp.org/news/file-systems-architecture-explained/>
- [3] Pompeu, O. (2022a, octubre 25). BTRFS, ZFS, XFS, EXT4: What Difference & Which File System to Use? *SalvageData*.  
<https://www.salvagedata.com/btrfs-zfs-xfs-ext4-how-are-they-different/>
- [4] NVIDIA Grace Hopper Superchip Data Sheet. (s. f.). NVIDIA.  
<https://resources.nvidia.com/en-us-grace-cpu/grace-hopper-superchip>
- [5] Nvidia. (s. f.). Nvidia Grace Performance Tuning Guide.  
<https://docs.nvidia.com/grace-performance-tuning-guide.pdf>
- [6] Nvidia. (s. f.). Nvidia Grace Hopper.  
<https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper>