HPE GreenLake

# Por qué las empresas necesitan los supercomputers para entrenar los modelos de IA más complicados

La complejidad de la IA generativa exige una potencia de computación superior, fuera del alcance de la mayoría de las empresas. Abran paso al supercomputer.

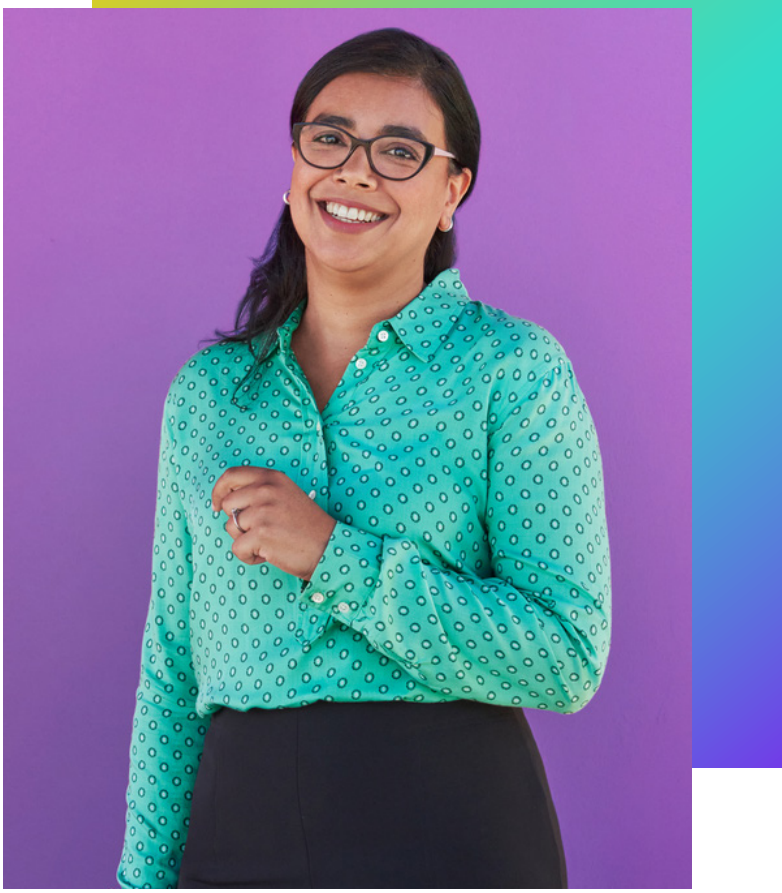
Desde hace ya casi 50 años, los supercomputers han estado intentando resolver los problemas más importantes y complejos del planeta: como predecir patrones meteorológicos a largo plazo, simular los efectos de la fusión nuclear, habilitar el descubrimiento de medicinas que salvan vidas y trazar con precisión los orígenes del universo.

Sin embargo, de forma creciente, estas máquinas con un potencial de computación inmenso están encontrando su lugar en las empresas. Y con la aparición de la inteligencia artificial (IA) como herramienta esencial para impulsar las decisiones empresariales, los supercomputers se preparan para desempeñar un papel mucho más prominente en el futuro.

Tal y como su nombre indica, los supercomputers suelen ser monstruos del tamaño de una habitación que se componen de decenas de miles de CPU y GPU, millones de gigabytes de memoria y almacenamiento y miles de nodos, cada uno con su propio PC inmensamente potente. Las versiones más avanzadas de estos, los llamados supercomputers a exaescala, pueden manejar 1000 millones x 1000 millones ($10^{18}$) de operaciones por segundo.

Los supercomputers se han construido para ejecutar aplicaciones masivas únicas que requieren comunicaciones rápidas entre miles de nodos que se ejecutan en paralelo, señala Paolo Faraboschi, vicepresidente del laboratorio de investigación sobre IA en Hewlett Packard Enterprise. Por este motivo, son ideales para entrenar grandes modelos de lenguaje (LLM) que sirven de base para las aplicaciones actuales de IA generativa.

«Entrenar un modelo de IA generativa se parece en cierto modo a simular patrones meteorológicos o crear un gemelo digital de un motor. En esencia, es una aplicación enorme con millones de parámetros. Es precisamente en las cargas de trabajo de entrenamiento de la IA donde el hardware de supercomputación brilla en lo que respecta al rendimiento», señala Faraboschi.



## Cómo superan los supercomputers a la nube en poder de computación bruto

Aunque el entrenamiento de grandes modelos de lenguaje sea, con toda probabilidad, la aplicación más visible de la computación de alto rendimiento (HPC), en absoluto es la única. Los supercomputers ya se están empleando a fondo para gestionar los sistemas de negociación de alta frecuencia de los agentes de bolsa de Wall Street, mantener gemelos digitales de redes de la cadena de suministro para los fabricantes y ayudar a las farmacéuticas a evaluar rápidamente millones de medicamentos candidatos en la carrera por descubrir vacunas más nuevas y eficaces.

Los factores por los cuales estos escenarios resultan ideales para las aplicaciones de la supercomputación —la necesidad de ejecutar millones de procesos estrechamente acoplados en paralelo— también los convierten en candidatos poco viables para las soluciones de nube a hiperescala, explica Faraboschi. La computación en la nube sobresale en la ejecución de aplicaciones diferenciadas para millones de usuarios, pero no está diseñada para las comunicaciones internas a alta velocidad necesarias para una aplicación única y altamente paralela.

«Estas aplicaciones necesitan funcionar al unísono. Pensemos en una simulación meteorológica. Cada nodo extrae un pequeño trozo del mundo y lo analiza. Sin embargo, también necesita comunicarse con sus nodos vecinos porque el clima de una zona afecta al clima de las demás. El intercambio rápido de datos entre los nodos es la clave para que todo funcione», afirma.

Las granjas de servidores en la nube no suelen tener la misma capacidad para intercambiar datos de forma rápida y predecible entre sí, explica Faraboschi. Si uno de esos servidores es más lento que los demás, el resto debe esperar hasta que se ponga al día.

«El autobús no puede salir hasta que el último pasajero se haya subido. Aquí es donde la arquitectura de la nube suele venirse abajo», señala.

Faraboschi afirma que es posible encargar un clúster de servidores estrechamente interconectados y acelerados por GPU en la nube, pero el coste podría ser prohibitivo. Estos servidores no abundan y, a diferencia de la mayoría de las arquitecturas de nube multiinquilino, suelen estar dedicados a un único cliente. Eso significa que los proveedores de nube pueden aplicarles un precio desorbitado.

«Muchas organizaciones empiezan utilizando la nube para realizar pequeños experimentos de entrenamiento de IA. Sin embargo, en cuanto se dan cuenta del coste que supondría escalar la infraestructura para la ejecución de tareas de entrenamiento completas, suelen recurrir a la tecnología interna. Si necesito 50 000 nodos por semana o mes para entrenar mi gran modelo de lenguaje, el modelo de negocio en la nube se desmorona», añade.

## La refrigeración líquida aumenta la sostenibilidad de los supercomputers

Otra sorprendente ventaja de los supercomputers respecto a la nube y la mayoría de centros de datos locales es su eficiencia energética. Un centro de datos empresarial medio tiene una calificación de eficacia de consumo energético (PUE) de 1,58 o superior. Es decir, que por cada kW consumido por recursos de computación, se gastan otros 580 vatios (58 %) en suministrar energía y refrigeración al centro de datos.[1]

Los proveedores de nube a hiperescala suelen tener un PUE aproximado de 1,1 o 1,2, comenta Faraboschi, por lo que son considerablemente más eficientes que los centros locales. En cambio, los supercomputers, especialmente los modelos con refrigeración líquida como el Frontier, el supercomputer a exaescala del Laboratorio Nacional Oak Ridge, tiene un PUE de 1,03, lo que significa que solo desperdicia un 3 % de la energía. Por tanto, son de tres a seis veces más eficientes que los centros de datos en la nube típicos.[2]

La refrigeración líquida es la responsable de la sostenibilidad superior de estos sistemas, apunta Faraboschi. Debido a que el agua disipa el calor con mayor eficacia que el aire, los nodos de computación pueden ejecutarse a una temperatura más elevada y colocarse más juntos. Esta mayor densidad supone un rendimiento significativamente superior en un espacio mucho más reducido. Y, a diferencia de los centros de datos refrigerados por aire, el calor residual se puede utilizar para calentar otras instalaciones o incluso para cultivar frutas y verduras en invernaderos en zonas con climas extremadamente fríos.

«La refrigeración líquida puede adoptar muchas formas. Si se hace bien, el fluido refrigerante llega a todos los elementos — memoria, GPU y CPU, e incluso condensadores y conversores de alimentación. Esto supone diseñar desde cero toda la infraestructura de las placas de refrigeración y los sistemas de distribución del líquido», señala Faraboschi.

## Las empresas necesitan desesperadamente especialización en IA

Las empresas necesitan tener en cuenta varios factores a la hora de decidir si necesitan el poder de computación de un supercomputer. Uno es la naturaleza y la escala de la carga de trabajo que desean ejecutar. Tal y como hemos mencionado anteriormente, los supercomputers destacan a la hora de abordar tareas de gran calado en las que intervengan muchos elementos de computación dispares. Si eres un gran distribuidor internacional que desea comprender y predecir el comportamiento de sus clientes, una empresa petrolífera que busca identificar reservas no explotadas de petróleo sin desembolsar el coste de perforar pozos de prospección o una empresa de ingeniería que diseña aeronaves comerciales y otra maquinaria compleja de gran tamaño, es probable que utilices un supercomputer para ello.

Las organizaciones que busquen crear ventaja competitiva a través de la construcción y el entrenamiento de sus propios modelos de IA de forma interna también querrán emplear un supercomputer para hacer el trabajo de manera más eficiente, afirma Faraboschi.

«Tienes que saber en qué nivel de la pirámide de los actores de la IA deseas estar. ¿Te puedes permitir estar en la cúspide de la pirámide, preentrenando un modelo desde cero? En ese caso, el problema es de supercomputación», señala.

Otro factor es cuánto tiempo vas a necesitar acceder a recursos de computación masivos. Si estás manteniendo un gemelo digital de un motor o una cadena de suministro, o necesitas realizar previsiones periódicas y regulares, necesitarás recurrir constantemente a la supercomputación y es probable que quieras contar con este recurso de forma local. Ahora bien, si lo que buscas es entrenar un gran modelo de lenguaje una sola vez (utilizando tus propios datos antes de implementarlo sobre el terreno), es probable que la HPC como servicio bajo demanda sea la forma más eficiente de proceder, en lugar de adquirir un supercomputer a la primera de cambio.

«La diferencia realmente radica en lo económico y en si dispones de un grupo de TI con experiencia en HPC e IA. De no ser así, y si solo necesitas entrenar tu gran modelo de lenguaje una sola vez o de forma cíclica, entonces es probable que quieras decantarte por la supercomputación como servicio», apunta Faraboschi.

No obstante, en todos los casos de uso, la experiencia es clave. Resulta complicado ejecutar cualquier aplicación durante un mes sin que se bloquee, en especial, una aplicación inmensamente compleja que se ejecuta en miles de nodos. En la actualidad, solo el 10 % de las empresas cuenta con la experiencia y los recursos internos para entrenar sus propios modelos de IA.[3] El resto utiliza grandes modelos de lenguaje disponibles de forma pública o recurren a la experiencia externa de integradores de sistemas.

«La complejidad de entrenar un gran modelo de IA es muy, muy elevada», advierte Faraboschi. En muchos casos, hemos visto a clientes avanzar hacia la IA y, a continuación, darse cuenta de que no es su negocio principal y de que su dinero está mejor invertido haciendo aquello que se supone que deberían estar haciendo, ya sea comercio minorista, finanzas o ingeniería.

[1] ¿Por qué la calificación PUE ha permanecido estable durante tanto tiempo después de años de avance?

[2] olcf.ornl.gov/wp-content/uploads/2022-OLCF-User-Meetig-Overview-of-Frontier-Whitt.pdf

[3] «2023: The State of Generative AI in the Enterprise (2023: estado de la IA generativa en la empresa)», Menlo Ventures, 13 de nov. de 2023

En ese momento es cuando suelen decidir confiarnos la experiencia a nosotros o a sus otros partners de integración de sistemas.

## Por qué el futuro de la IA va a ser «extraordinario»

Poco más de la mitad de las empresas usan alguna forma de IA y, de manera conjunta, prevén gastar en torno a 70 000 millones de dólares en servicios y tecnología de IA en 2024.[4] En la actualidad, la IA generativa representa una porción relativamente pequeña de esa inversión, pero es probable que crezca a un ritmo constante con el tiempo.

Las organizaciones que busquen capitalizar el potencial de la IA generativa primero deben ordenar sus almacenes de datos, señala Faraboschi. Sin unos datos limpios, fiables y de buena calidad, cualquier predicción que haga un modelo de IA no tendrá demasiado valor. A continuación, necesitan decidir la relevancia que va a tener la IA en sus impresas y cuál será su grado de compromiso con el aprovisionamiento del hardware y la experiencia necesaria.

«Las organizaciones acaban de empezar a darse cuenta de la complejidad del entorno de IA. Necesitas disponer de unos mecanismos de resiliencia eficaces, ya que algunos elementos de la infraestructura fallarán cada pocas horas y no podrás permitirte reiniciar una gran tarea de entrenamiento de IA desde cero. Necesitas aprovisionar las redes y el almacenamiento de forma que se saque el máximo partido a este costoso hardware. Todo el espacio de diseño del sistema es muy complejo, y ahí es donde los integradores de sistemas pueden ayudar», señala.

Faraboschi afirma que el tamaño y la escala de los modelos de IA actuales exigen soluciones de HPC (high performance computing). Hace diez o quince años, podías usar un supercomputer para entrenar un millón de los modelos de IA más grandes en una semana, señala. En la actualidad, supercomputers increíblemente potentes como el Frontier necesitarían una semana para entrenar solo uno de los grandes modelos de lenguaje de mayor tamaño. En otras palabras, tal y como afirma Faraboschi: «La gente necesita darse cuenta de que entrenar grandes modelos de IA es ahora un problema de supercomputación».

[4] Ibid

**Más información en**
HPE.com/AI

Visita **HPE GreenLake**

**Chatear ahora**

Hewlett Packard Enterprise



a50010080ESE