

White paper

# Las cargas de trabajo de la IA y el futuro de la infraestructura de TI

La hoja de ruta adecuada para navegar  
en un futuro incierto





## Índice

|                                                                                         |    |
|-----------------------------------------------------------------------------------------|----|
| Los datos como el nuevo motor de la economía                                            | 3  |
| La infraestructura digital crítica que facilita la IA                                   | 4  |
| La eficiencia es vital, pero no será la fórmula mágica                                  | 5  |
| Dar sentido al futuro                                                                   | 6  |
| ¿La clave para adaptarse? Aprovechar la tecnología y forjar buenas asociaciones         | 11 |
| ¿Cuál será el futuro de la infraestructura de TI? Lo que dicen los principales expertos | 12 |
| ¿Cuál camino seguir?                                                                    | 16 |

# Los datos como el nuevo motor de la economía

Suele decirse que “los datos son el nuevo petróleo”, pero tal y como ocurre con el crudo, los datos necesitan refinarse.

Al igual que el crudo, los datos tienen el potencial de transformar la economía mundial. Sin embargo, tal y como fueron necesarios los avances en las tecnologías de refinamiento para convertir el crudo en el motor del crecimiento económico del siglo XX, la inteligencia artificial (IA) probablemente será clave para convertir los datos en la base de la economía del siglo XXI.

De acuerdo con McKinsey & Company, la IA generativa por sí sola podría agregar el equivalente de hasta \$4,4 trillones anuales a la economía global, con hasta \$100.000 millones en el sector de telecomunicaciones, \$130.000 millones en medios de comunicación y \$460.000 millones en alta tecnología<sup>1</sup>

Además, la alimentación de la IA no solo se medirá en dólares, euros, yuanes o yenes, sino también por medio de los avances que afectarán la calidad de la vida humana. Ya nos

encontramos en el umbral de importantes avances, como la conversión de grandes cantidades

## >\$25 trillones

de aumento potencial en el PIB global total, como resultado de la IA para 2030<sup>2</sup>

Sin embargo, liberar el potencial de la IA exigirá todo un ecosistema para soportar la tecnología. La arquitectura de centros de datos y la infraestructura digital crítica que la soporta deberán someterse a una transformación importante para satisfacer las necesidades de las cargas de trabajo de la IA. Además, el ecosistema deberá diseñarse para marcar la diferencia de forma duradera, no para el desuso.



<sup>1</sup> McKinsey Digital; “The economic potential of generative AI: The next productivity frontier,” publicado el 14 de junio de 2023; <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier>; Última visita el 30 de mayo de 2024.

<sup>2</sup> ibid.

## La infraestructura digital crítica que facilita la IA

Las tendencias actuales sugieren que el 19 % del consumo energético en los centros de datos estará asociado a la IA para 2028.<sup>3</sup> Entonces, ¿qué significa esto para los centros de datos, especialmente en lo relacionado con la arquitectura, el consumo energético y el enfriamiento?

Aunque no ocupa la primera plana, el crecimiento meteórico de la IA ya está alterando la arquitectura de TI y la infraestructura digital crítica del centro de datos, pero el efecto hasta este punto representa únicamente la punta del iceberg. Los aportes asociados a la IA generan una traba potencial en lo relacionado con el tamaño de los chips, el peso de los racks y la alimentación y el enfriamiento del centro de datos. El gasto

en infraestructura digital crítica para la IA generativa (GenAI) alcanzará los \$18.000 millones en 2024 y superará los \$48.000 millones para 2027.<sup>4</sup>

El nivel de inversión indica con cuánta profundidad afectará el aumento de la IA generativa a la infraestructura digital crítica. Necesariamente, la revolución de la IA llevará a una revolución en la manera en la cual operan los centros de datos. Por lo tanto, mientras la mayoría de personas se enfocan en los beneficios económicos de la IA, para obtener estos beneficios será necesario predecir, comprender y resolver los desafíos de infraestructura emergentes.



Cuando surgió la metáfora de que los datos eran el nuevo petróleo, esta no era del todo precisa. Hoy contamos con una increíble herramienta en IA que puede recopilar datos y liberar su valor oculto, pero esto también significa que se debe estar preparado para los cambios profundos en la infraestructura necesaria para soportar la IA”.



**Stephen Liang**

Director de Tecnología y Vicepresidente Ejecutivo, Vertiv



<sup>3</sup> Goldman Sachs. “AI is poised to drive 160% increase in data center power demand”. 14 de mayo de 2024. <https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand>. Última visita el 5 de agosto de 2024.

<sup>4</sup> Villars R. et al. “Worldwide Core IT Spending for GenAI Forecast, 2023–2027: GenAI Is Triggering Hyper-Expansion of AI Spending”. IDC #US51539723, diciembre de 2023.

# La eficiencia es vital, pero no será la fórmula mágica.

La IA se basa en muchos elementos, pero uno de los principales son los microprocesadores. Los chips utilizados para entrenar los modelos de IA exigen una cantidad considerable de potencia y generan una cantidad equivalente de calor.

Las GPU son el chip preferido de la actualidad para poner en marcha las cargas de trabajo de IA y usar las GPU en lugar de CPU para poner en marcha las cargas de trabajo de procesamiento informático en paralelo es 100 veces más eficiente en términos de consumo energético. Sin embargo, aunque las GPU son más eficientes, las necesidades informáticas seguirán aumentando exponencialmente y superarán en gran medida cualquier aumento en la eficiencia de los chips.<sup>5</sup> Por lo tanto, aunque las GPU sin duda producirán mucho más procesamiento informático con la misma cantidad

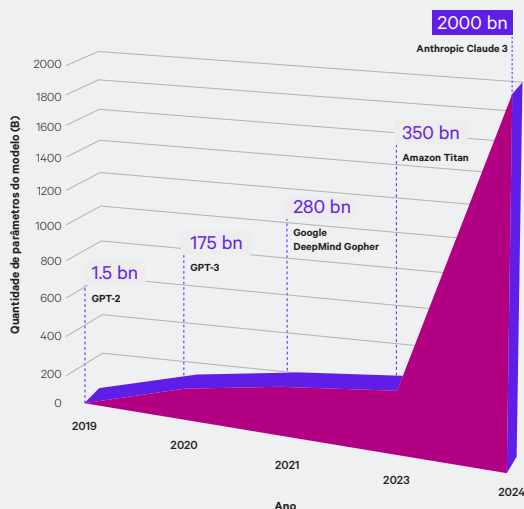
de potencia consumida, las demandas de la IA indican que la cantidad total de potencia utilizada seguirá aumentando.<sup>6</sup>

Probablemente, será necesaria una cantidad considerable de recursos para el desarrollo de GPU más eficientes. Con dicho nivel de inversión, sería insensato no esperar avances en el ámbito de la eficiencia.

Sin embargo, si la tendencia actual continua, los modelos que mejoran la IA generativa también continuarán aumentando de tamaño. La competencia entre los principales protagonistas de la industria como OpenAI, Anthropic, Google y Meta está siendo impulsada por el uso de modelos de mayor tamaño. Como resultado, los modelos han estado duplicando sus necesidades informáticas cada seis meses.<sup>7</sup>

## Más allá del crecimiento exponencial<sup>8</sup>

¿Predecirá el pasado el futuro del crecimiento de los modelos de lenguaje de gran tamaño (LLM)?



Sin duda, el increíble rendimiento asociado a la IA justifica el aumento en la potencia de entrada, pero sigue siendo cierto que los avances en las capacidades de IA también parecen exigir modelos cada vez más grandes.

La demanda de precisión es parte del crecimiento continuo de estos modelos, pero el crecimiento va más allá de la simple precisión. Puede que la nueva funcionalidad requiera modelos aún más grandes para llevar a cabo labores como tareas de varias etapas en lugar de simplemente responder a instrucciones o determinar cuál respuesta algorítmicamente disponible satisface mejor las necesidades del usuario.

Aunque es cierto que los modelos más pequeños muestran resultados alentadores, la demanda entre los usuarios de una mayor precisión y una mayor funcionalidad indica que estos modelos más pequeños probablemente continuarán siendo un nicho.

Al mismo tiempo, la base de usuarios de IA crece rápidamente. Los consumidores individuales, las empresas de todos los tipos e incluso las organizaciones gubernamentales están impulsando una demanda cada vez mayor de IA.

<sup>5</sup> Newmark. "2023 U.S. Data Center Market Overview & Market Clusters". Enero de 2024; <https://www.nmrk.com/insights/market-report/2023-u-s-data-center-market-overview-market-clusters>. Última visita el 10 de julio de 2024.

<sup>6</sup> ibid.

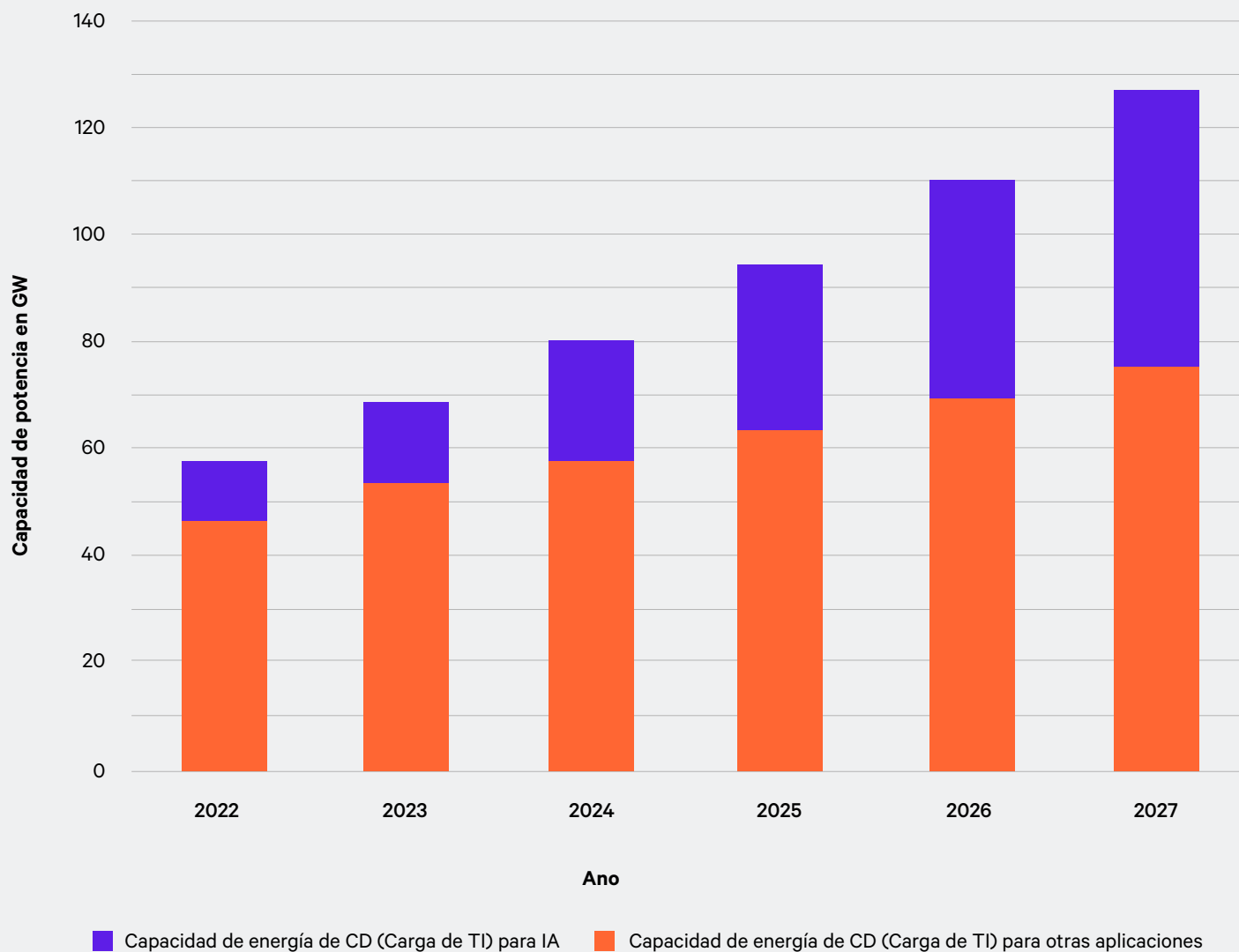
<sup>7</sup> Epoch AI. "Notable AI Models". 19 de junio de 2024. <https://epochai.org/data/notable-ai-models>. Última visita el 10 de julio de 2024.

<sup>8</sup> Adaptado de The Rise and Rise of A.I. Large Language Models (LLMs) & their associated bots like ChatGPT, by McCandless, D. Evans T. and Barton P, 2024 (<https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt>). En dominio público.

## Dar sentido al futuro

Según la paradoja de Jevons, un aumento en la eficiencia de un recurso generará un aumento en el consumo, en lugar de una disminución. Con base en los muchos factores que parecen impulsar tanto la demanda de IA como el uso de modelos más grandes, es difícil determinar cómo los aumentos en la eficiencia de los chips o los modelos contrarrestarán por completo los numerosos desarrollos con un uso intensivo de recursos informáticos en el ámbito de la IA. En otras palabras, el futuro podría ser más eficiente, pero la capacidad de carga de TI seguirá aumentando a lo largo de la década a pesar de los aumentos en eficiencia.

### La IA es una fuerza impulsora en el crecimiento de la capacidad de carga de TI<sup>9</sup>

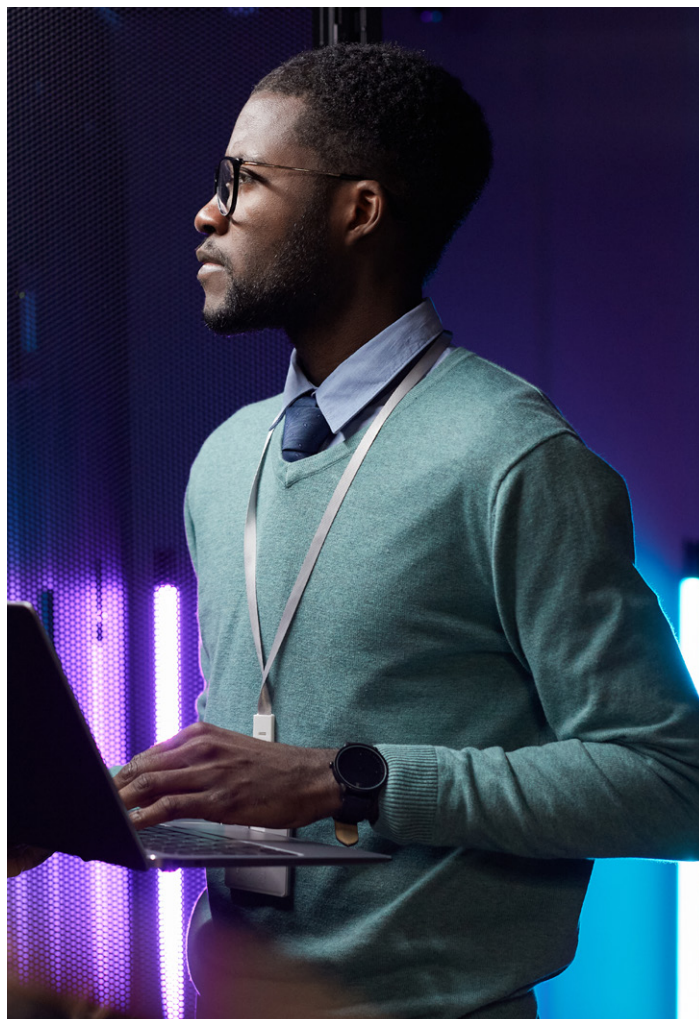


<sup>9</sup> Galabov, V and Sukumaran M. "Cloud and Data Center Market Snapshot – November 2023", Omdia.

Yogi Berra, jugador estadounidense de béisbol y pensador, dijo en cierta ocasión: “Es difícil hacer predicciones, especialmente del futuro”. Si bien esto puede parecer particularmente oportuno con tantas incógnitas conocidas en el desarrollo de infraestructura de IA, con base en las tendencias actuales y los probables desarrollos tecnológicos, es posible anticipar los mayores desafíos para la implementación y el uso de la IA, especialmente dentro del contexto del centro de datos. Entonces, ¿cuáles son los elementos más importantes a seguir si usted está interesado en preparar sus operaciones para el futuro?

### Específicamente, las siguientes tendencias deberían estar en el radar de todos.

- **Cambios en la capacidad de salida:** Necesitamos estar preparados para un mundo donde aumenta la capacidad de producción. Las cargas de los racks, PDU, UPS, conmutadores y CDU aumentarán en capacidad y tendrán una reacción en cadena equivalente en toda la infraestructura en general.
- **Cambios en la escala general:** Podría decirse que la reacción en cadena más evidente implicará el crecimiento de las instalaciones. Los actuales bloques de 3 MW serán los bloques de 20 MW del futuro.
- **La búsqueda de alimentación siempre disponible:** Un mayor tamaño también significa más potencia y más calor, y más calor significa que el enfriamiento líquido y el híbrido son el futuro. Debido a que las GPU del futuro deberán enfriarse continuamente, es imprescindible contar con una alimentación siempre disponible.
- **La IA se traduce en perfiles de carga no convencionales:** La IA está asociada a cargas de energía que pueden oscilar desde un 10 % de inactividad hasta una sobrecarga del 150 % en un instante.
- **La presión para modernizar o actualizar:** Los operadores de centros de datos deberán determinar la mejor manera de modernizar sus operaciones existentes para satisfacer las necesidades mencionadas anteriormente.



Creo que la arquitectura va a cambiar considerablemente y una parte vital para alcanzar el éxito es prever estos cambios. Por ejemplo, es importante preguntarse: “¿Cómo planificamos racks de 250 kilovatios o más?”. La manera en que se hace actualmente no es viable, así que ahora es el momento de pensar en el cambio”.



**Gregory Ratcliff**  
Chief Innovation Officer, Vertiv

# Se avecinan cambios importantes en la arquitectura, la alimentación y el enfriamiento

## Un resumen del mayor desarrollo de esta década

Las tendencias actuales indican que se avecinan cambios importantes en la infraestructura de TI en general, pero, ¿cuáles veremos y cuándo sucederán?

Las respuestas a estas preguntas comienzan en el nivel más básico. A medida que la arquitectura de GPU, arquitectura de rack y la la arquitectura de filas evolucionen, veremos

una evolución equivalente tanto en la alimentación como en el enfriamiento.

Aunque no es posible estar realmente seguros del futuro preciso de la arquitectura de TI, observar los motores básicos del cambio es la mejor manera de determinar cómo será dicho futuro.

|                                          | 2024                                                                                                                                                                                                  | 2025                                                                                                                                                                                                                                                                       | 2027**                                                                                                                                                                                                    | 2029+**                                                                                                                                                                                                                                                                        |
|------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                          | Actualmente se envían diseños de GPU con IA basados principalmente en enfriamiento por aire o enfriamiento líquido híbrido y enfriamiento por aire para una mayor eficiencia operativa. <sup>10</sup> | Diseños de GPU con IA de última versión basados principalmente en enfriamiento líquido monofásico al chip y enfriamiento por aire para las cargas de calor residual. <sup>11</sup>                                                                                         | Diseños de GPU con IA de última generación (no anunciados, pero proyectados) basados principalmente en un mayor flujo y una temperatura más baja para un mayor rendimiento. <sup>12</sup>                 | Diseños de GPU con IA de última generación (trayectoria proyectada) basados principalmente en enfriamiento líquido al chip o alguna forma de inmersión contenida. <sup>13</sup>                                                                                                |
| <b>GPU Arquitectura</b>                  | Hasta 32 núcleos de GPU NVIDIA Hopper H100 (o similares) por rack, configurados en clústeres de filas/POD de hasta 240 núcleos de GPU. <sup>14</sup>                                                  | Hasta 72 núcleos de GPU NVIDIA Blackwell B200 (o similares) por rack, configurados en clústeres de filas/POD de hasta 720x núcleos GPU. <sup>15</sup>                                                                                                                      | Núcleos de GPU de última generación (ciclos de lanzamiento de 12 meses), en configuraciones de POD/clústeres de filas similares.                                                                          |                                                                                                                                                                                                                                                                                |
| <b>Arquitectura en rack<sup>16</sup></b> | Hasta 50 kW por rack con distribución eléctrica de 400/230 VCA N+1. Enfrida con intercambiadores de calor líquido-aire en puerta trasera.                                                             | Hasta 140 kW por rack con distribución eléctrica de 400/230 VCA o 400 VCD N+1. Enfrida con placas frías de líquido a chip monofásicas (80 % de la carga térmica), aumentadas por intercambiadores de calor de líquido a aire en puerta trasera (20 % de la carga térmica). | Hasta 300kW por rack con distribución eléctrica de 400/230 VCA o 400 VCD N+1. Enfrida con placas frías de líquido a chip, aumentadas por intercambiadores de calor de líquido a aire en puerta trasera. * | Densidades de hasta 1 MW por rack. En el diseño se deben considerar densidades de 480/277 VCA, 600/347VCA o 800 VCD para 2031. Distribución eléctrica de 400/230VAC o 400VDC N+1. * En el diseño se debe considerar el enfriamiento de placa fría (100 % de la carga térmica). |

\*Probablemente, el factor de forma del rack cambiará en tamaño y configuración a partir de 2027 y no corresponderá con las versiones EIA y OCP utilizadas en la actualidad.

<sup>10</sup> NVIDIA; "NVIDIA H100 Tensor Core GPU Architecture;" <https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper?ncid=no-ncid>; Última visita el 12 de agosto de 2024.

<sup>11</sup> NVIDIA; "NVIDIA Blackwell Architecture Technical Brief;" <https://resources.nvidia.com/en-us-blackwell-architecture?ncid=no-ncid>; Última visita el 12 de agosto de 2024.

<sup>12</sup> Trueman, C. "Stack Infrastructure to support AI workloads requiring up to 300kW-per-rack will be achieved through closed-loop water cooling systems;" Data Center Dynamics; 8 de enero de 2024; <https://www.datacenterdynamics.com/en/news/stack-infrastructure-to-support-ai-workloads-requiring-up-to-300kw-per-rack/>; Última visita el 14 de agosto de 2024.

<sup>13</sup> ibid.

<sup>14</sup> NVIDIA; "NVIDIA H100 Tensor Core GPU Architecture;" <https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper?ncid=no-ncid>; Última visita el 12 de agosto de 2024.

<sup>15</sup> NVIDIA; "NVIDIA Blackwell Architecture Technical Brief;" <https://resources.nvidia.com/en-us-blackwell-architecture?ncid=no-ncid>; Última visita el 12 de agosto de 2024.

<sup>16</sup> Wang, S. "2024 Trends to Watch: Data Center Physical Infrastructure". Omdia; 22 de diciembre de 2023; <https://omdia.tech.informa.com/om033896/2024-trends-to-watch-data-center-physical-infrastructure>; Última visita el 12 de agosto de 2024.



|                                           | 2024                                                                                                                                                                                                                                                                                                                                                                                                                | 2025                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | 2027**                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | 2029+**                                                                                                                                                                                                                                                                                                                               |
|-------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Arquitectura en fila</b> <sup>17</sup> | Se recomienda un busway rígido suspendido de 400 A con cajas de conexiones, las cuales ofrecen una máxima flexibilidad para el reemplazo directo y la actualización de las GPU. Por su costo, se pueden considerar los circuitos exclusivos cableados en el sitio desde las PDU. Unidades de distribución de enfriamiento de 600 kW en fila y unidades de aire acondicionado perimetral para salas de computadoras. | Se recomienda un busway rígido suspendido de 2000 A con cajas de conexiones, las cuales ofrecen una máxima flexibilidad para el reemplazo directo y la actualización de las GPU. No se recomiendan los circuitos exclusivos cableados en el sitio desde la PDU debido al costo de instalación y al calibre inflexible del cable. Unidades de distribución de enfriamiento de 1350 kW en fila y unidades de aire acondicionado perimetral para salas de computadoras. | Se recomienda un busway rígido suspendido de 4000 A con cajas de conexiones, las cuales ofrecen una máxima flexibilidad para el reemplazo directo y la actualización de las GPU. No se recomiendan circuitos exclusivos cableados en el sitio desde las PDU. Unidades de distribución de enfriamiento de 3000 kW en fila y unidades de aire acondicionado perimetral para salas de computadoras.                                                                                                                                         | Se recomienda un busway rígido suspendido de 6000 A o más, con cajas de conexiones, las cuales ofrecen una máxima flexibilidad para el reemplazo directo y la actualización de las GPU. No se recomiendan circuitos exclusivos cableados en el sitio desde las PDU. Unidades de distribución de enfriamiento de múltiples megavatios. |
| <b>Gestión de potencia</b>                | Se debe considerar una alimentación de la red pública mínima de Nivel II o Nivel III respaldada por generación de energía en espera N+1 con Sistemas de Almacenamiento de Energía con Baterías (BESS) en el sitio o similares para la interacción, el arbitraje y la resiliencia de la red.                                                                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Se debe considerar una alimentación de la red pública mínima de Nivel II o Nivel III respaldada por generación de energía en espera N+1 con Recursos Energéticos Distribuidos (DER), Sistemas de Almacenamiento de Energía con Baterías (BESS) en el sitio o celdas de combustible para una interacción, un arbitraje y una resiliencia de la red totalmente dinámicos. Se requiere el monitoreo integrado de la alimentación de alta velocidad en la caja de conexiones (busway) con control térmico inteligente (patentado de Vertiv). |                                                                                                                                                                                                                                                                                                                                       |
| <b>Carga de trabajo de GPU</b>            | Hasta un 160%<br>We are assuming initial GPU deployments to be >20% of total workloads for enterprise up to 2025 and growing to 100% of total load by 2027.                                                                                                                                                                                                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Hasta un 160%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | Hasta un 160%                                                                                                                                                                                                                                                                                                                         |
| <b>Ciclo térmico</b> <sup>18</sup>        | Expansión directa (DX) y agua helada (CW).                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Circuito de agua helada a alta presión.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | Circuito de agua helada                                                                                                                                                                                                                                                                                                               |
| <b>Reutilización del calor</b>            | Utilización del ciclo orgánico de Rankine (ORC), la captura directa de aire (DAC) y plantas de enfriamiento de absorción.                                                                                                                                                                                                                                                                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | Más del 80 % de                                                                                                                                                                                                                                                                                                                       |
| <b>Control</b>                            | Captura de la telemetría operativa en el lago de datos e implementación del mantenimiento basado en condiciones (CBM).                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Implementación de gemelos digitales en todo el centro de datos y aprovechamiento de las entradas continuas de telemetría para optimizar la operación por medio de la IA.                                                                                                                                                                                                                                                                                                                                                                 | Optimización automática y funcionamiento robótico autónomo.                                                                                                                                                                                                                                                                           |

\*\*Los detalles técnicos y de gestión, las estimaciones y las proyecciones citadas aquí se basan en las investigaciones de Vertiv, experiencia interna e información actualmente disponible de parte de los socios de Vertiv en la industria. Los pronósticos analizados se basan en las tecnologías de infraestructura y diseño disponibles a la fecha de publicación de este artículo técnico. Los pronósticos se presentan únicamente con fines ilustrativos.

La trayectoria al alza en la densidad de los racks como consecuencia de la actual ola de IA no será algo pasajero. Para satisfacer la demanda cada vez mayor de IA, la arquitectura de los racks aumentará de las densidades superiores de 30 kW a los 300-600 kW en el corto plazo y posiblemente a 1 MW o más para 2030.

Naturalmente, seguirá existiendo un nivel de incertidumbre en lo relacionado con esta cifra, con base no solo en los avances tecnológicos, sino también en la demanda de IA. Sin embargo, es razonable esperar que el futuro se basará en la creciente densidad de los racks requerida por la IA y las mejoras tecnológicas en infraestructura crítica necesarias para satisfacer estas demandas.



Las cosas están cambiando más rápidamente ahora que en cualquier otro momento de los últimos 30 años. Para gestionar este cambio, debemos trabajar juntos como industria para establecer buenas prácticas y diseños de referencia. Es fundamental estar preparado para los cambios que ya están ocurriendo”.

**Peter Panfil**

Vicepresidente, Potencia Global, Vertiv

<sup>17</sup> ibid.

<sup>18</sup> Wang, S. “AI-driven Data Center Cooling Systems and Technologies,” Omdia; Agosto 2, 2024; <https://omdia.tech.informa.com/blogs/2024/aug/ai-driven-data-center-cooling-systems-and-technologies>; Última visita el 15 de agosto de 2024.

Aunque puede existir cierta incertidumbre en la rapidez con que evolucionará la arquitectura de los racks, especialmente a largo plazo, no hay duda de que los cambios en la arquitectura exigirán nuevos enfoques para el enfriamiento. Los actuales diseños de GPU con IA se basan principalmente en el enfriamiento por aire.

Sin embargo, el enorme crecimiento en el consumo energético indicará que varias formas de enfriamiento líquido e híbrido son el futuro. El enfriamiento líquido es considerablemente más eficiente que el enfriamiento por aire y esa eficiencia es sumamente importante para desarrollar la infraestructura necesaria para la IA. Para finales de la década, seremos testigos de que los centros de datos dependerán principalmente del enfriamiento líquido del chip, la inmersión independiente y el enfriamiento por aire para las cargas de calor residual.



El auge del enfriamiento líquido debería formular unas cuantas preguntas para quienes gestionan los centros de datos. Por ejemplo, '¿Cómo puedo asegurarme de que el circuito de líquido sea estable? ¿Cómo puedo asegurarme de que sea redundante? ¿Cómo coloco un rack en el sistema? ¿Cómo retiro un rack del sistema?' Estas son preguntas en las que la gente debe pensar con los cambios que se avecinan".



**Steve Madara**  
Vicepresidente,  
Enfriamiento Global,  
Vertiv



El enfriamiento no es la única preocupación. Los centros de datos del futuro también deberán ser lo suficientemente flexibles para hacer frente a las "cargas de pulso" o cargas periódicas con elevadas irrupciones energéticas en muy poco tiempo. Los clústeres de entrenamiento de IA son conocidos por mostrar picos de tensión de corta duración durante ciertos ciclos de procesamiento informático, por lo cual las disposiciones de sobrecarga se volverán obligatorias.

Sin embargo, es probable que los centros de datos sufran de un dolor de cabeza más grande que los simples picos de tensión en lo relacionado con la alimentación. En Estados Unidos, se espera que el aumento total en el consumo total de electricidad se basará casi en su totalidad en la demanda de los centros de datos. Según datos de Goldman Sachs, habrá una tasa de crecimiento anual compuesto del 15 % en la demanda energética de los centros de datos hasta 2030 y los centros de datos representarán el 8 % de la demanda energética total de EE. UU. para finales de la década, en comparación con el 3 % actual.



Uma de nossas funções hoje é tornar mais disponível a energia que existe e garantir que ela seja usada com mais eficiência no data center. Você precisa ter certeza de que usa a energia que tem de forma eficaz."



**Stephen Liang**  
Director de Tecnología y  
Vicepresidente Ejecutivo, Vertiv



Aunque es difícil señalar con exactitud todos los hitos tecnológicos que tendrán lugar en esta década y en el futuro, el futuro a grandes rasgos es cada vez más claro. Sabemos que la densidad de los racks aumentará y que los métodos y la gestión de enfriamiento necesitarán evolucionar. También, sabemos que la gestión de la alimentación y la carga se están convirtiendo rápidamente en desafíos importantes.

En la actualidad, necesitamos aplicar lo que sabemos para beneficiar el desarrollo de soluciones y productos que darán solución a las necesidades de los centros de datos tanto hoy como en el futuro. Además, debemos hacerlo rápidamente si queremos aprovechar todo el potencial de la IA.

Por ejemplo, deberíamos planificar mayores densidades de racks mediante el desarrollo de unidades de distribución de energía en racks de alta densidad para gestionar la distribución eléctrica de manera más efectiva en los centros de datos en rápida evolución. También deberíamos reexaminar las tecnologías existentes para el enfriamiento líquido y preguntarnos cómo podemos prepararnos para un mundo en el que el enfriamiento de placa fría de líquido a chip en dos fases sea una necesidad, y deberíamos observar los perfiles de carga de IA reales y desarrollar formas de optimizar el rendimiento en función de los perfiles que estamos viendo ahora y lo que anticipamos ver a medida que avanza la revolución de la IA.

## ¿La clave para adaptarse?

### Aprovechar la tecnología y forjar buenas asociaciones

Sin embargo, lograr todo esto no es necesariamente fácil. Aunque resulta vital prepararse para los cambios que traerá la IA, también es importante reconocer que cualquier intento de preparar sus operaciones para el futuro no se puede dar en el vacío.

Jensen Huang, director ejecutivo de NVIDIA, resumió eficazmente la situación cuando dijo: “La IA no es un problema de chips. Es un problema de reinversión de la computación. No se puede resolver esta nueva forma de computación con solo diseñar un chip. Cada aspecto de la computadora ha cambiado fundamentalmente”.

La manera en la cual soportamos y aprovechamos la IA también está cambiando. Ya somos testigos del surgimiento de la fábrica de IA, un hub centralizado que les permite a los expertos, a los científicos de datos y a los ingenieros colaborar en el desarrollo, la implementación y la ampliación de soluciones basadas en IA.

A medida que la IA crezca, también lo hará el modelo de fábrica. Con un funcionamiento similar a las fábricas tradicionales, pero con la creación de inteligencia procesable y nuevos conocimientos en lugar de productos físicos, el modelo necesita una infraestructura compleja y un diseño integral para funcionar.

En la práctica, esto significa que ningún protagonista en este ámbito tiene todas las respuestas. Por lo tanto, prosperar en un mundo dominado por la IA significará ir más allá de solo resolver un problema de alimentación, enfriamiento o chips. Más bien, los operadores inteligentes verán cómo los problemas a los que se enfrentan encajan en el ecosistema general.

Por esta razón, estar preparados para el futuro exigirá forjar relaciones sólidas con socios expertos que no sólo observen los cambios provocados por la IA, sino que también los impulsen.

Incluso para las empresas líderes en el sector de la IA, reunir la experiencia de los protagonistas más experimentados es una parte importante para maximizar los beneficios de la IA. Integrar las perspectivas de todos, desde los desarrolladores de chips hasta los gerentes de potencia, es fundamental para hacer predicciones inteligentes sobre las demandas de la IA y tener la capacidad de satisfacer dichas necesidades.

¿Cómo se ve esto en la práctica? Algunos elementos como los innovadores chips tendrán un efecto enorme en la arquitectura y el funcionamiento futuros de los centros de datos; por esta razón, Vertiv ha estado trabajando en estrecha colaboración con los principales fabricantes de chips de IA y sirviendo como consultor en [el rápido desarrollo de tecnología de chips](#) y como

arquitecto de la [infraestructura necesaria para alojar dichos chips](#). Como parte de varios programas globales que involucran a innovadores tecnológicos que ofrecen soluciones basadas en tecnologías o impulsadas por ellas, podemos contribuir a garantizar que la infraestructura digital crítica y las GPU funcionen conjuntamente para ofrecer un rendimiento eficiente.

Además, al ofrecer nuestra experiencia a redes que también incluyen a los principales proveedores de software, de servicios en la nube, de soluciones y a los integradores de sistemas, fortalecemos el trabajo que hacemos hoy y nos preparamos para los cambios en el futuro a través de [la integración del portafolio completo de soluciones de alimentación y enfriamiento de Vertiv en las últimas tecnologías](#) de todo el sector.

En el mercado actual, resulta muy importante contar con una base de conocimientos técnicos incomparables, pero dichos conocimientos deben trabajar en conjunto con otras partes del ecosistema de IA en desarrollo si realmente se quiere preparar las operaciones para el futuro.



Debido a los cambios sin precedentes que los centros de datos han comenzado a enfrentar con el auge de la IA, me dije a mí mismo: ‘Creo que debemos recurrir a los muchachos de los chips y específicamente a los muchachos de los chips disruptivos’. Necesitamos saber qué van a hacer y construir que vaya a alterar nuestro ecosistema”.

**Greg Stover**  
Director Global,  
Desarrollo de Alta  
Tecnología, Vertiv



# ¿Cuál será el futuro de la infraestructura de TI?

## Lo que dicen los principales expertos

La necesidad de soportar la IA ya está impulsando cambios importantes en la infraestructura del centro de datos. Sin embargo, los mayores desarrollos aún están por llegar. En lo que queda de esta década, la IA será la fuerza impulsora detrás de una serie de cambios, con la nueva arquitectura de GPU, la arquitectura de racks y la arquitectura de filas como contribuidores de cambios significativos en la infraestructura digital crítica y de TI.

¿Qué se necesitará para aprovechar las oportunidades y los cambios que presenta el auge de la IA?

### Alimentar el futuro significa mirar más allá de lo evidente



**Peter Panfil**

Vicepresidente, Potencia Global, Vertiv

Peter lidera el desarrollo estratégico de los clientes para el negocio de alimentación de Vertiv. Es experto en resolver los desafíos de los clientes con las últimas tecnologías de alimentación y control que ofrecen niveles de disponibilidad, escalabilidad y eficiencia para satisfacer las diversas necesidades de sostenibilidad y de los clientes. Peter tiene un profundo conocimiento en soluciones de alimentación para satisfacer las necesidades únicas de las aplicaciones de IA y es un defensor del enfoque de “Trae tu propia potencia” (BYOP) para resolver los problemas de dependencia de los servicios públicos. Con más de 30 años en el sector de infraestructura crítica, ha ocupado puestos como vicepresidente de ingeniería y vicepresidente/gerente general de energía de CA antes de sus actuales responsabilidades. Es un frecuente presentador y vocero en ferias comerciales de la industria, conferencias y medios informativos que dan servicio a la industria de ingeniería e instalaciones de TI, así como autor publicado gracias a su contribución en el libro de 2024, “Greener Data Vol. 2”.



Por mucho tiempo, hemos vivido felices en el mundo de los centros de datos con bloques de tres MW acoplados directamente a un generador de tres MW en un bus de 4000 amperios. Esos días pronto pasarán a la historia. Con el auge de la IA, es común empezar a hablar de bloques de 10 MW e incluso de 20 MW.

De manera que cuando se trata del consumo energético en los centros de datos, no es ningún secreto que nos encontramos ante un aumento masivo en tamaño y escala. La conclusión es la siguiente: los operadores tendrán que determinar de dónde vendrá esa alimentación, pero, con la misma relevancia, la IA también dará lugar a cargas elevadas asociadas al entrenamiento de la IA generativa y una demanda energética continua asociada a las necesidades de enfriamiento. Estos dos últimos factores son necesidades relativamente recientes y supondrán un desafío cada vez mayor en toda la industria en los próximos años.

Entonces, ¿cuál es el secreto para hacer frente a estos desafíos de alimentación? Creo que gran parte será anticipar los cambios más específicos en la alimentación. Es necesario conocer hacia dónde se dirigen los fabricantes de chips y saber qué piensan los expertos de la industria de TI.

En cuanto a las soluciones energéticas, los innovadores ya se hacen preguntas como ‘¿puedo simplificar esto e implementarlo fácilmente y a bajo costo?’ y ‘¿puedo estar seguro de que esta solución es confiable?’. Las respuestas en mi mente son definitivamente un ‘Sí’. Sin embargo, es importante comprender que para llegar a dicha respuesta afirmativa se debe trabajar con las personas adecuadas, quienes comprenden tanto los últimos chips como las últimas soluciones energéticas”.

## El uso de la resiliencia para proteger sus activos



**Stephen Liang**

Director de Tecnología y Vicepresidente Ejecutivo, Vertiv

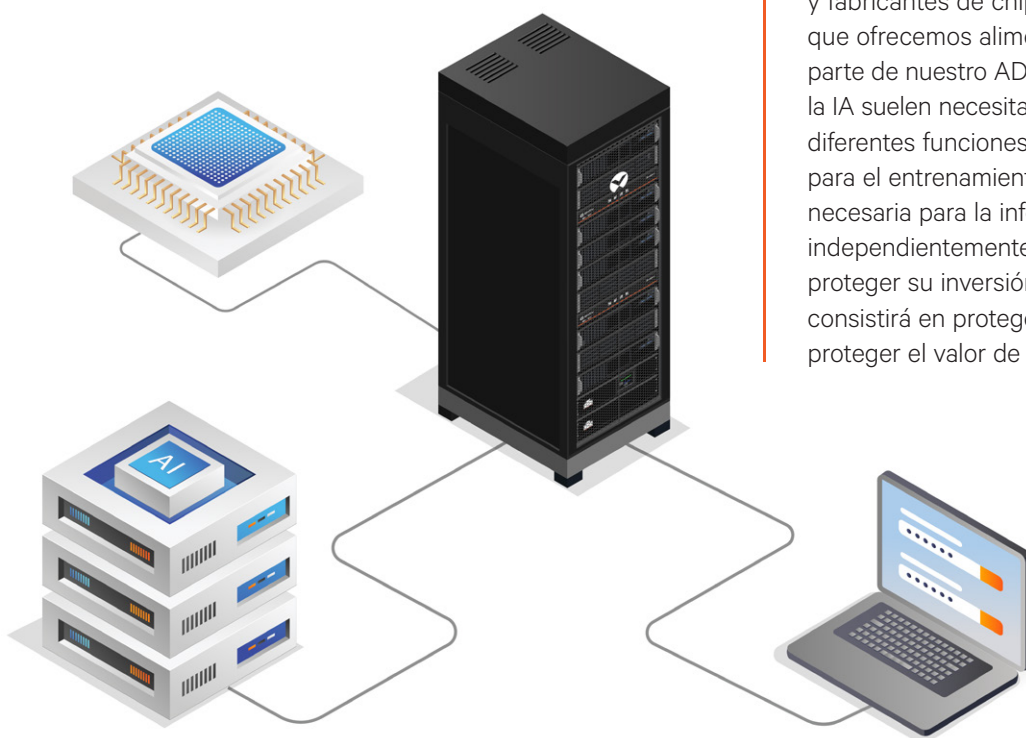
Stephen Liang alinea las estrategias y los recursos tecnológicos de Vertiv para dar prioridad a la voz del cliente en la investigación y el desarrollo de productos. Cuenta con tres décadas de experiencia en soluciones de alimentación, operaciones de fabricación e I+D. Con una Licenciatura y una Maestría en Ingeniería Mecánica del Instituto de Tecnología de Massachusetts, combina tanto su conocimiento técnico como su gran experiencia en su visión de la manera en la cual la tecnología está cambiando la infraestructura de TI.



Tengo una larga experiencia con infraestructuras TI y puedo decir sin temor a dudas que el cambio tecnológico está ocurriendo más rápido que nunca en el sector. Una razón es porque esperamos que la tecnología haga más que nunca y le pedimos a la infraestructura general que haga lo mismo.

Hemos sido testigos de una considerable inversión en los valiosos chips informáticos y sistemas de servidores de última generación. Esta inversión se relaciona en gran parte con la IA y significa que tenemos más en juego que nunca para mantener estos sistemas en funcionamiento. Para aprovechar al máximo estos activos, nuestros clientes necesitarán que funcionen las 24 horas del día, los 7 días de la semana. Por lo tanto, creo que una de las mayores consignas para el futuro es la resiliencia. Naturalmente, contar con una infraestructura de TI resiliente siempre ha sido una prioridad, pero en el futuro, la resiliencia será más importante y más compleja que nunca.

¿Qué quiero decir con esto? En Vertiv, tenemos un largo historial en la formación de asociaciones con desarrolladores y fabricantes de chips para obtener información, a la vez que ofrecemos alimentación y enfriamiento continuos. Es parte de nuestro ADN. Podemos ver cómo las demandas de la IA suelen necesitar diferentes niveles de resiliencia para diferentes funciones. Por ejemplo, la resiliencia necesaria para el entrenamiento es diferente de la resiliencia necesaria para la inferencia, pero debemos recordar que, independientemente de su objetivo, la clave siempre será proteger su inversión. En el futuro, la resiliencia no siempre consistirá en proteger la aplicación, sino también en proteger el valor de los activos”.



## Aprenda a aprovechar lo que tiene



**Martin T. Olsen**

Vicepresidente Ejecutivo, Estrategia de Productos Global, Vertiv

Martin Olsen es responsable de la estrategia y la planificación globales de los productos. Su trabajo impulsa la interoperabilidad de los productos en la hoja de ruta para ofrecer soluciones completas, la priorización, la evaluación y el alcance de las inversiones en todas las unidades de negocios globales de Vertiv. Cuenta con amplia experiencia en varios puestos de liderazgo en ingeniería, estrategia y planificación, y es un experto en infraestructura crítica global en la industria de centros de datos. Martin tiene varias patentes térmicas en EE. UU. y obtuvo su Licenciatura en Mercadeo y Análisis de Negocios en la Universidad de Dinamarca del Sur, en Kolding, Dinamarca. Además, completó el Programa de Gestión General como egresado de la Escuela de Negocios de Harvard.



Vivimos en un mundo en el cual el aumento de la IA significa que las densidades de los racks podrían ser de 1 MW por rack para finales de la década. Quienes ya han hecho importantes inversiones en infraestructura deben determinar cómo aprovechar al máximo estas inversiones sin dejar de estar preparados para un nuevo aumento de las demandas en el futuro.

En otras palabras, en un entorno que cambia rápidamente, ¿cómo puede aprovechar lo que ya se tiene sin correr el riesgo de quedar desactualizado?

Muchos operadores de centros de datos necesitarán evaluar la posibilidad de modernizar sus activos. Sin embargo, a medida que la densidad de los racks continúa creciendo a un ritmo acelerado, decidirse solo por una modernización no será suficiente. Más bien, será necesario determinar las densidades para las cuales realizará la modernización y las tecnologías de enfriamiento requeridas como parte de dicha modernización.

Implementar y utilizar la IA en los negocios resulta muy transformador, y los clientes conocen muy bien la frase “Disrupt or be Disrupted” de John Chambers, exdirector ejecutivo y presidente emérito de Cisco Systems. Los clientes quieren sobresalir sobre sus competidores. A la hora de pensar en una modernización, esto implica tomar una decisión después de ver todas las opciones disponibles y considerar cómo pueden funcionar juntas de forma inteligente.

Esto puede resultar ser una tarea difícil. Después de todo, ¿cómo conocer todas las opciones? ¿Cuál es la mejor manera de ver tanto los detalles como el panorama general al mismo tiempo? En mi opinión, el elemento que hará que un operador de centros de datos esté verdaderamente preparado para el futuro será encontrar socios que no solo tengan un amplio conocimiento de las tecnologías emergentes, sino también en la manera en la cual estas tecnologías pueden trabajar juntas para satisfacer las necesidades futuras. Sin un socio así, creo que los operadores de centros de datos corren el riesgo de ver el futuro pasar de lejos”.

## ¿Dónde buscar los productos más innovadores del futuro?



**Greg Ratcliff**

Director de Innovación, Vertiv

Greg Ratcliff es un veterano de más de 30 años en la industria con énfasis en la innovación basada en datos. Dirigió al grupo de análisis avanzado global para el equipo de servicios, con una especialización en datos en tiempo real y donde conectó a casi 1000 millones de productos operativos de Vertiv con la nube de Vertiv. Greg forja relaciones con socios tecnológicos, universidades y equipos de ingeniería de Vertiv para garantizar que la empresa esté enfocada en el futuro y preparada para soportar el cambio en la industria. Su formación académica incluye títulos universitarios en Matemáticas Aplicadas e Ingeniería Informática con estudios menores en Ingeniería Eléctrica de la Universidad Estatal de Ohio, una Maestría en Administración de Empresas de la Universidad de Phoenix y capacitación especializada en la gestión ágil de proyectos de IoT y Big Data de la Universidad Liberty.



Me gusta pensar en I+D en tres contextos: Lo inmediato, lo cercano y lo lejano. “Lo inmediato” es donde se encuentra los productos que se pueden usar en la actualidad y son esenciales para mantener la infraestructura de TI contemporánea.

Sin embargo, cualquier empresa que quiera mantenerse al ritmo de los rápidos cambios ya en marcha no puede darse el lujo de ignorar “lo cercano” y “lo lejano”. Las tecnologías esenciales para la infraestructura de TI serán considerablemente diferentes de las empleadas en la actualidad. Esto significa que mantener los ojos puestos en el proceso de desarrollo es parte esencial de la preparación para el futuro.

En mi mundo, una forma de prepararnos para el futuro es mantener relaciones activas con docenas de universidades. Algunas de estas relaciones son formales, como el Center for Energy- Smart Electronic Systems (ES2). ES2 es un Centro de Investigación Cooperativa de la Fundación Nacional de Ciencias/Industria/Universidad (I/UCRC) creado para desarrollar herramientas para maximizar la eficiencia energética para el funcionamiento de sistemas electrónicos, incluidos los centros de datos, “desde el chip hasta la torre de enfriamiento”. Otras relaciones son menos formalizadas y muy valiosas con el mundo académico, pero estas relaciones con las universidades no significan que vivamos en una urna de cristal. De hecho, los clientes finales de Vertiv suelen ser parte importante de estas relaciones. Me gusta pensar en ello como un taburete de tres patas que apoya la innovación práctica. Las innovaciones pueden venir de un laboratorio universitario, pero Vertiv y nuestros clientes desempeñan un papel esencial a la hora de convertir la innovación en un producto listo para satisfacer las necesidades de un mercado más amplio.

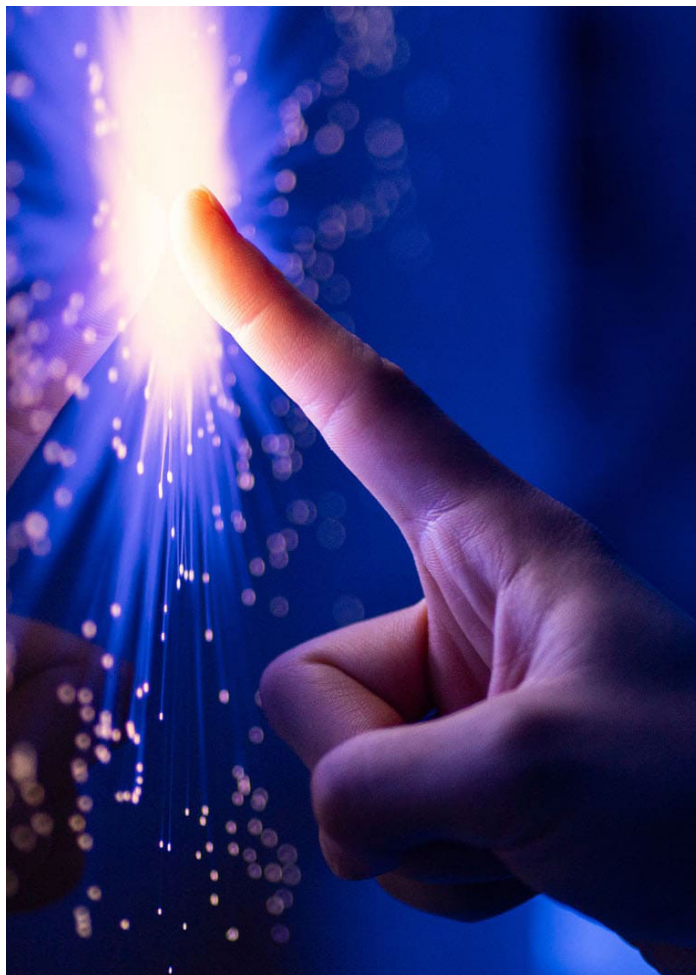
Además, las innovaciones que se conviertan en productos reales serán parte esencial de la evolución de la infraestructura digital crítica. Si se quiere estar preparado para el futuro, es necesario asociarse con una empresa que tenga vínculos profundos con los líderes e innovadores tecnológicos, porque en el entorno actual puede que uno esté enfocado en “lo inmediato”, pero también necesita estar atento a “cercano” y “lo lejano”.

## ¿Cuál camino seguir?

El auge de la IA ha hecho que este sea un momento realmente emocionante en el mundo de la computación, pero con la emoción y las posibilidades vienen las preguntas y la incertidumbre. Muchas entidades con centros de datos están luchando por agregar capacidad de IA a las operaciones existentes. Al mismo tiempo, quienes planean nuevos centros de datos se preguntan exactamente cuánta capacidad de IA deberían incluir en sus proyectos. ¿Y cómo será la arquitectura de los racks? ¿Y qué hay de las necesidades de alimentación y enfriamiento?

Responder este tipo de preguntas será difícil, pero una forma de hacerlo de manera más sencilla es asegurarse de diseñar su infraestructura digital crítica con líderes en el desarrollo de IA, no con seguidores, porque en un mundo donde el cambio es rápido y seguro, el enfoque incorrecto podría ser invertir en una solución que quede obsoleta poco después de implementarla. Al ser inteligente y prudente, puede evitar diseñar y caer en el desuso, y, por el contrario, crear algo más sostenible, confiable y duradero.

Prosperar en la revolución de la IA significa elegir sabiamente sus soluciones, pero significa ser aún más sabio acerca de quién las ofrece. No solo necesita la tecnología adecuada, sino también un socio con un amplio conocimiento del tren de potencia y la cadena térmica de principio a fin, y con expertos que puedan asesorarlo sobre cómo satisfacer sus necesidades de IA hoy y en el futuro.



Su socio ideal es aquel con una amplia cartera de tecnología de alimentación y enfriamiento, con una gran red de servicios y presencia global. ¿Por qué? Porque usted necesita un socio que le permita ser transformador, y para ser transformador en esta etapa del juego con la IA, necesita velocidad y escala”.



**Martin Olsen**

Vicepresidente Ejecutivo, Estrategia de Productos Global, Vertiv

## Referências

- Adaptado de The Rise and Rise of A.I. Large Language Models (LLMs) & their associated bots like ChatGPT, by McCandless, D. Evans T. and Barton P, 2024 (<https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt>). En dominio público.
- Epoch AI; "Notable AI Models;" 19 de junio de 2024; <https://epochai.org/data/notable-ai-models>. Última visita el 10 de julio de 2024
- Galabov, V and Sukumaran M; "Cloud and Data Center Market Snapshot – November 2023;" Omdia. Noviembre de 2023
- Goldman Sachs; "Generational growth: AI, data centers and the coming US power demand surge;" publicado el 28 de abril de 2024; <https://www.goldmansachs.com/intelligence/pages/gs-research/generational-growth-ai-data-centers-and-the-coming-us-power-surge/report.pdf>; Última visita el 30 de mayo de 2024.
- Goldman Sachs; "AI is poised to drive 160% increase in data center power demand;" 14 de mayo de 2024. <https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand>. Última visita el 5 de agosto de 2024.
- McKinsey Digital; "The economic potential of generative AI: The next productivity frontier;" publicado el 1 de junio de 2023; <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier>; Última visita el 30 de mayo de 2024
- Newmark; "2023 U.S. Data Center Market Overview & Market Clusters;" enero de 2024; <https://www.nmrk.com/insights/market-report/2023-u-s-data-center-market-overview-market-clusters>. Última visita el 10 de julio de 2024.
- NVIDIA; "NVIDIA H100 Tensor Core GPU Architecture;" <https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper?ncid=no-ncid>; Última visita el 12 de agosto de 2024.
- NVIDIA; "NVIDIA Blackwell Architecture Technical Brief;" <https://resources.nvidia.com/en-us-blackwell-architecture?ncid=no-ncid>; Última visita el 12 de agosto de 2024.
- Trueman, C. "Stack Infrastructure to support AI workloads requiring up to 300kW-per-rack will be achieved through closed-loop water cooling systems;" Data Center Dynamics; 8 de enero de 2024; <https://www.datacenterdynamics.com/en/news/stack-infrastructure-to-support-ai-workloads-requiring-up-to-300kw-per-rack/>; Última visita el 14 de agosto de 2024
- Villars R. et al. "Worldwide Core IT Spending for GenAI Forecast, 2023–2027: GenAI Is Triggering Hyper-Expansion of AI Spending;" IIDC #US51539723, diciembre de 2023.
- Wang, S. "2024 Trends to Watch: Data Center Physical Infrastructure;" Omdia; 22 de diciembre de 2023; <https://omdia.tech.informa.com/om033896/2024-trends-to-watch-data-center-physical-infrastructure>; Última visita el 12 de agosto de 2024.
- Wang, S. "AI-driven Data Center Cooling Systems and Technologies;" Omdia; 2 de agosto de 2024; <https://omdia.tech.informa.com/blogs/2024/aug/ai-driven-data-center-cooling-systems-and-technologies>; Última visita el 15 de agosto de 2024.



**Vertiv.com** | Sede da Vertiv, 505 N Cleveland Ave, Westerville, OH 43082, EE. UU.

© 2025 Vertiv Group Corp. Todos los derechos reservados. Vertiv™ y el logo de Vertiv son marcas o marcas registradas de Vertiv Group Corp. Todos los demás nombres y logos a los que se hace referencia son nombres comerciales, marcas, o marcas registradas de sus dueños respectivos. Aunque se tomaron todas las precauciones para asegurar que esta literatura esté completa y exacta, Vertiv Group Corp. no asume ninguna responsabilidad y renuncia a cualquier demanda por daños como resultado del uso de esta información o de cualquier error u omisión. Las especificaciones, los reembolsos y otras ofertas promocionales están sujetas a cambio a la entera discreción de Vertiv y mediante notificación.