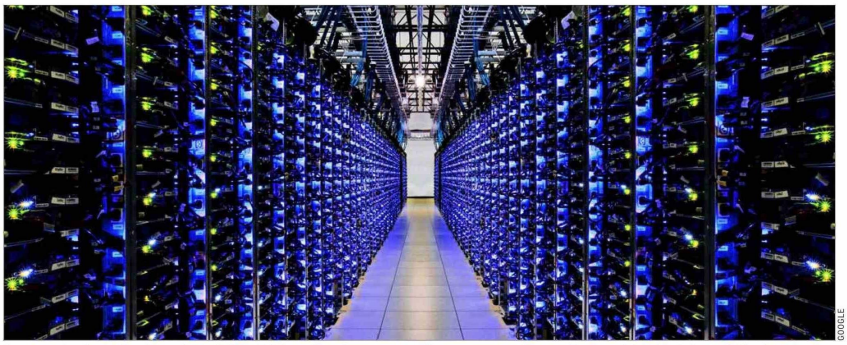
LITORALPRESS

Fecha: 16-04-2026
Medio: El Mercurio
Supl. : El Mercurio - Cuerpo B
Tipo: Noticia general
Título: La IA está utilizando tanta energía que la capacidad computacional se está agotando

Pág. : 8
Cm2: 815,2

Tiraje: 126.654
Lectoría: 320.543
Favorabilidad: No Definida



GOOGLE

Los precios de arriendo por hora de las GPU, los microchips que se utilizan para capacitar y activar modelos de IA, han subido abruptamente desde el último trimestre del año pasado.

En medio de feroz competencia por captar usuarios:

La IA está utilizando tanta energía que la capacidad computacional se está agotando

Las compañías de IA están racionando sus ofertas y productos, lo que irrita a los usuarios; una señal de alerta para un auge que depende de la rápida adopción.

ANGEL AU-YEUNG Y ROBBIE WHELAN
The Wall Street Journal

La fiebre del oro de la inteligencia artificial (IA) está agotando rápidamente el suministro del único recurso del que los creadores de IA no pueden prescindir: potencia computacional.

La grave escasez de capacidad ha causado desconcierto entre los usuarios de energía, ha obligado a las compañías a descartar productos y ha llevado a problemas de fiabilidad. Esta situación es una señal de alerta para el auge de la IA, puesto que puede limitar la utilidad de las nuevas y potentes herramientas de IA justo cuando enormes cantidades de usuarios han empezado a depender de ellas para aumentar la productividad.

Durante los últimos meses, se ha disparado la demanda de IA "agencial", herramientas autónomas que emplean la tecnología para realizar tareas en forma independiente, desde escribir código de *software* hasta programar visitas a casas para corredores de propiedades. Las empresas han estado haciendo grandes esfuerzos para asegurar la disponibilidad de la capacidad computacional necesaria para atender a una creciente base de clientes que también están aumentando en forma significativa su uso de IA.

"Todos están hablando de petróleo, pero creo que de lo que el mundo está escaso principalmente es de *tokens*", manifestó Ben Pouladian, ingeniero e inversionista tecnológico residente en Los Angeles. Un *token* es una unidad de medida en IA para hacer un seguimiento de cuántos recursos computacionales se están utilizando para una tarea. "En este punto, la IA ya no es solo un chatbot al que le pedimos una receta mientras estamos frente al refrigerador. Está organizando tareas, se está volviendo más inteligente", explicó Pouladian.

Todo esto apunta a un problema clásico que ha surgido en los auges tecnológicos a través de la historia, desde la expansión del ferrocarril en el siglo XIX hasta la irrupción de las telecomunicaciones e internet a principios de la década de 2000. La demanda está creciendo mucho más rápido que la velocidad con que las empresas pueden acceder a recursos y desarrollar infraestructura. Históricamente, los aumentos de precios han estado entre las únicas formas de abordar una escasez de oferta, pero esa medida podría ser peligrosa para las empresas de IA de vanguardia, las que están en una feroz competencia por captar usuarios.

COSTOS
Arrendar uno de los chips Blackwell, la generación más avanzada de Nvidia, por una hora cuesta US$ 4,08, un alza del 48% en relación con los US$ 2,75 que costaba hace dos meses.

Los precios de arriendo por hora de las GPU, los microchips que se utilizan para capacitar y activar modelos de IA, han subido abruptamente desde el último trimestre del año pasado. Anthropic, la compañía que creó el popular chatbot Claude y la viral aplicación de codificación Claude Code, ha estado plagada recientemente de interrupciones frecuentes. La compañía ha empezado a dosificar el suministro computacional a los usuarios durante las horas de mayor congestión, pero el despliegue se ha visto afectado por los clientes que se han quejado de que alcanzan el límite en forma muy rápida.

OpenAI descartó su aplicación de generación de videos Sora, en parte, para liberar recursos computacionales que se utilizarían en productos de codificación y empresariales que funcionarían en un nuevo modelo de IA, cuyo nombre en clave es Spud, según informó The Wall Street Journal.

El uso de *tokens* en API de OpenAI —una plataforma donde principalmente usuarios de empresa acceden a su *software*— subió de 6 mil millones por minuto en octubre a 15 mil millones por minuto a fines de marzo.

"Paso mucho tiempo tratando de encontrar capacidad de procesamiento disponible de último minuto", aseguró Sarah Friar, directora de finanzas de OpenAI, en una reciente entrevista pública en video con un inversionista. "En este momento, estamos tomando decisiones muy difíciles sobre cosas que no podemos realizar, porque no tenemos suficiente capacidad de procesamiento".

A fines del año pasado, Core Weave, una de las principales compañías de computación en la nube con IA que cotiza en bolsa, elevó los precios en más de un 20% y empezó a pedir a sus clientes más pequeños que firmaran contratos que los comprometen a utilizar los servicios de la compañía durante al menos tres años, en vez de un año como antes. Analistas de Bank of America restablecieron la cobertura de la compañía con una clasificación de "Compra" a fines del mes pasado, y señalaron que es probable que la demanda de sus servicios supere la oferta al menos hasta 2029.

Los precios de mercado al contado para acceder a las GPU (unidades de procesamiento gráfico) de Nvidia en las nubes de centros de datos han subido abruptamente en los últimos meses en toda la línea de productos de la compañía, según Ornn, un proveedor de datos con sede en Nueva York que publica datos de mercado y estructura productos financieros en torno al precio de las GPU.

Arrendar uno de los chips Blackwell, la generación más avanzada de Nvidia, por una hora cuesta US$ 4,08, un aumento del 48% en relación con los US$ 2,75 que costaba hace dos meses, según Ornn Compute Price Index.

"Hay una enorme escasez de capacidad que no había visto en los más de cinco años que he estado manejando este negocio", aseguró J. J. Kardwell, jefe ejecutivo de Vultr, una compañía de infraestructura en la nube. "La pregunta es, ¿por qué no desplegamos más equipos? Los tiempos de ejecución son demasiado largos. Los tiempos de construcción de centros de datos son largos, la energía que está disponible hasta 2026 ya está comprometida".

BLOOMBERG

Anthropic, dirigida por Dario Amodei, se ha visto afectada recientemente por frecuentes interrupciones del servicio.

Desde mediados de febrero, las interrupciones en los sistemas de Anthropic se han vuelto tan comunes que algunos de sus clientes de empresa se están cambiando a otros proveedores de modelos de IA.

David Hsu, fundador y director ejecutivo de la plataforma de desarrollo de *software* Retool, contó que prefiere utilizar el modelo Opus 4.6 de Anthropic para activar la herramienta de agente de IA de su compañía, porque cree que es el mejor modelo para empresa. Hace poco se cambió al modelo de OpenAI para activar el agente de su compañía. "Anthropic ha estado fallando todo el tiempo", dijo.

La fiabilidad de los servicios básicos en internet a menudo se mide en nueves. Cuatro nueves significa 99,99% de tiempo de actividad; un porcentaje habitual que una compañía de *software* se compromete a ofrecer a los clientes. El 8 de abril, Claude API de Anthropic tenía una tasa de tiempo de actividad de 98,95% en los últimos 90 días.

"Eso no es normal", comentó Amir Haghighat, cofundador y jefe de tecnología de Baseten, un emprendimiento de inferencia de IA. "Piense en AWS, las bases de datos, RDS o Stripe; estos tienen que ser muy resistentes con un tiempo de actividad muy alto. Pero ese no es el mundo en que vivimos cuando se trata de IA. Esa no es la calidad de servicio que espera de la compañía que está proporcionando la inteligencia para su aplicación".

Las interrupciones frecuentes en Anthropic suceden cuando el laboratorio de IA está experimentando un crecimiento explosivo. A fines de 2025, la compañía alcanzó los US$ 9 mil millones en tasa de proyección anual, lo que significa que la compañía estaba en vías de obtener esa cantidad de ingresos en los próximos 12 meses. En febrero, esa cifra se elevó a los US$ 14 mil millones. Dos meses más tarde, se duplicó a US$ 30 mil millones.

A fines de marzo, Anthropic anunció en forma repentina que limitaría la cantidad de *tokens* que los usuarios podían gastar durante las horas de mayor congestión, de 05:00 a 11:00, hora del Pacífico, en los días hábiles. Los clientes han recurrido a las redes sociales para quejarse del cambio. "No había alcanzado mi límite terminal de Claude Code en semanas, pero esta semana lo alcancé en unos 45 minutos", escribió un usuario en X.

"Hemos estado haciendo grandes esfuerzos para satisfacer el aumento en la demanda de Claude", escribió Boris Cherny, creador y jefe de Claude Code, en X. "La capacidad es un recurso que manejamos cuidadosamente y estamos dando prioridad a nuestros clientes que utilizan nuestros productos y API".

Artículo traducido del inglés por "El Mercurio".