DAORDE

EL PEAJE DE CUDA

Nvidia domina el cálculo de la IA porque vende una máquina completa: procesadores, red, bibliotecas y años de trabajo ajeno convertido en dependencia. Sus cuentas permiten medir ese poder y descartar una cifra mágica de renta.

Colectivo Daorde · 12.07.2026

Nvidia cerró su ejercicio fiscal de 2026 con 215.938 millones de dólares de ingresos. Después de pagar fabricación, memoria, encapsulado, transporte, garantías y el resto de costes que contabiliza como coste de ventas, conservó un margen bruto del 71,1 %. El margen operativo quedó en el 60,4 %. El beneficio neto fue de 120.067 millones.

Son cifras extrañas para una empresa que suele describirse como fabricante de chips y que ni siquiera fabrica sus propias obleas. Nvidia diseña; TSMC y Samsung producen las obleas; SK Hynix, Micron y Samsung suministran memoria; Foxconn, Wistron y otras contratistas montan, prueban y encapsulan. La propia empresa explica esta división del trabajo en su informe anual de 2026. Su parte consiste en controlar el diseño del acelerador, la interconexión, la arquitectura del sistema y el entorno de software que hace útil el conjunto.

Nvidia vende silicio y cobra por una relación de dependencia construida durante veinte años.

UNA CUENTA DE RESULTADOS FUERA DE ESCALA

Los centros de datos aportaron 193.737 millones de dólares, casi el 90 % de los ingresos del ejercicio. Dos clientes directos concentraron el 22 % y el 14 % de las ventas totales. Son compradores con recursos para diseñar sus propios aceleradores, financiar centros de datos y negociar contratos gigantescos. Aun así, Nvidia mantuvo su margen.

Ejercicio fiscal 2026	Millones de dólares	Porcentaje de ingresos
Ingresos	215.938	100 %
Beneficio bruto	153.532	71,1 %
Investigación y desarrollo	18.497	8,6 %
Beneficio operativo	130.387	60,4 %
Beneficio neto	120.067	55,6 %

El beneficio bruto está calculado a partir del margen publicado; las demás cifras proceden del 10-K. La comparación descarta una explicación cómoda. Nvidia gasta mucho en ingeniería: 18.497 millones en un año y 31.000 de sus 42.000 trabajadores figuran como personal de investigación y desarrollo. Pero el gasto en I+D no absorbe una parte cercana al excedente que la empresa retiene. Tampoco lo hacen los riesgos reales de una cadena fabless. En 2026 cargó 4.500 millones por inventario y compromisos del H20 afectados por los controles de exportación de Estados Unidos. Tras ese golpe, el margen operativo siguió por encima del 60 %.

Un margen no demuestra por sí mismo una renta monopolista. Puede reunir ventaja técnica, economías de escala, escasez temporal, poder de negociación y rentas de propiedad intelectual. Las cuentas públicas no separan cada componente. Afirmar que todo

el beneficio de Nvidia es renta daría una cifra contundente y una investigación bastante mala.

Las cuentas muestran que la empresa fija condiciones gracias a algo más que el chip aislado.

LA MÁQUINA COMPLETA

Una GPU H100 es un procesador paralelo de gran capacidad. CUDA es la plataforma que permite programarlo y usar bibliotecas preparadas para álgebra lineal, redes neuronales, simulación, tratamiento de imagen y otras cargas. Alrededor aparecen cuDNN, NCCL, TensorRT, compiladores, controladores, contenedores, documentación y herramientas de perfilado. Para escalar miles de aceleradores entran NVLink, InfiniBand o Ethernet, adaptadores, conmutadores y una arquitectura de rack diseñada como una unidad.

Nvidia ya no presenta el centro de datos como un edificio que contiene ordenadores. Lo presenta como el ordenador. En su informe anual llama «plataforma» al conjunto de GPU, CPU, DPU, interconexiones, red, bibliotecas, modelos, datos de entrenamiento, API y marcos de aplicación. Es una descripción interesada, pero correcta.

El cliente no elige entre chips desnudos que ejecutan el mismo programa. Elige entre sistemas cuyo rendimiento depende de años de optimización del software, disponibilidad de personal, compatibilidad de modelos, herramientas de operación y posibilidad de conseguir miles de unidades a tiempo. El acelerador puede ser más rápido y, al mismo tiempo, beneficiarse de que salir resulte caro. Las dos cosas caben en la misma factura.

CUDA ACUMULA TRABAJO AJENO

Nvidia cifra en más de 7,5 millones las personas que usan CUDA y sus otras herramientas. Ese número incluye trabajadores de la propia empresa, universidades, laboratorios públicos, compañías de nube, proyectos libres y equipos que escriben aplicaciones para sus empleadores. Cada nueva biblioteca compatible aumenta el valor práctico de la plataforma. Una parte de ese trabajo la paga Nvidia. Otra la pagan Estados, empresas y trabajadores que no reciben derechos sobre la plataforma que valorizan.

La documentación de CUDA garantiza que aplicaciones compiladas con versiones antiguas sigan funcionando sobre controladores nuevos y ofrece rutas limitadas para ejecutar software reciente en instalaciones anteriores. Gracias a esa virtud técnica, una organización conserva código, imágenes de contenedor, procedimientos, conocimiento operativo y resultados ya validados cuando vuelve a comprar Nvidia.

Migrar significa reescribir extensiones, sustituir bibliotecas, ajustar kernels, rehacer pruebas numéricas, formar a la plantilla y aceptar un periodo de menor rendimiento. El precio del cambio aparece como horas de ingeniería y riesgo de producción, fuera de una línea llamada «peaje de CUDA», y muchas comparaciones de coste lo omiten.

ROCm ofrece una vía competidora con soporte para PyTorch y una gama creciente de aceleradores AMD; MLPerf Training 5.1 recibió resultados sobre doce tipos de acelerador; en 2026 AMD publicó entrenamiento distribuido sobre 512 MI300X. Los grandes proveedores diseñan chips propios. Los marcos de aprendizaje automático abstraen parte del hardware. Todo eso reduce dependencia.

Pero un benchmark prueba una carga y una configuración. No demuestra que una organización pueda migrar su catálogo entero, ejecutar sus extensiones, encontrar personal con experiencia y comprar capacidad suficiente en la región que necesita. Los competidores reducen el encierro sin eliminarlo.

EL PRECIO DE UNA HORA

La capacidad de cálculo llega al usuario a través de otra capa de propiedad. En julio de 2026, AWS anunciaba bloques de capacidad H100 en Londres a 3,933 dólares por GPU y hora. Google publicaba su máquina A3 con ocho H100 a 88,49 dólares por hora bajo tarifa bajo demanda, algo más de 11 dólares por GPU y hora. No son productos equivalentes: cambian compromiso, CPU, memoria, red, almacenamiento, soporte y disponibilidad. La diferencia impide presentar el precio más bajo como una comparación completa.

El precio de nube mezcla el margen de Nvidia con la amortización del servidor y el edificio, electricidad, refrigeración, red, personal, financiación y margen del hiperescalador. No sabemos qué parte de una hora alquilada termina en Nvidia. Podemos observar que todos los grandes proveedores ofrecen la misma plataforma porque sus clientes la piden, y que la oferta de CUDA refuerza la demanda de hardware Nvidia mientras la presencia del hardware refuerza la oferta de software CUDA.

Esta retroalimentación es la forma concreta del peaje. La organización paga por el equipo actual y carga además con el coste acumulado de dejarlo.

BENEFICIO INDUSTRIAL, ESCASEZ Y RENTA

Nvidia ha desarrollado procesadores excelentes y ha coordinado una transición técnica desde GPU sueltas hasta sistemas completos refrigerados por líquido. Esa capacidad productiva explica parte del beneficio. La demanda de IA creció más rápido que la oferta de memoria avanzada, encapsulado y electricidad; la escasez explica otra parte. La empresa

comprometió 95.200 millones de dólares en fabricación, suministros y capacidad para asegurar la producción futura. Su beneficio exige capacidad industrial además de propiedad intelectual.

La renta aparece donde el control exclusivo de una condición de producción permite apropiarse de una porción del excedente producido en otros lugares. CUDA, las bibliotecas optimizadas, las interconexiones y la escala instalada cumplen parte de esa función. Un laboratorio puede producir ciencia, una empresa puede vender inferencia y un Estado puede entrenar un modelo lingüístico. Nvidia captura valor antes de saber si esos proyectos llegarán a servir para algo. Controla un acceso escaso y difícil de sustituir.

No podemos calcular con información pública una «tasa de renta de CUDA». Podemos documentar sus condiciones: márgenes extraordinarios persistentes; una plataforma propietaria con millones de usuarios; costes de cambio pagados por el cliente; integración de hardware, red y software; y capacidad para conservar precios altos incluso ante compradores concentrados. Eso basta para dejar de hablar de Nvidia como un fabricante más.

MÁS ALLÁ DE OTRO CAMPEÓN SUBVENCIONADO

La respuesta liberal es aumentar la competencia. La Unión Europea subvenciona fábricas, los hiperescaladores diseñan ASIC y AMD mejora ROCm. Una segunda plataforma propietaria puede bajar precios. No convierte el cálculo en un bien común.

Una política inmediata tendría que pagar el trabajo que el mercado evita: mantener aplicaciones en varias arquitecturas, financiar compiladores y bibliotecas libres, exigir formatos abiertos en la compra pública, publicar pruebas de portabilidad y conservar equipos técnicos capaces de migrar. Las subvenciones a cómputo público deberían obligar a devolver código, resultados y mejoras de herramientas cuando no exista una razón de seguridad o privacidad para impedirlo. El Estado compra demasiadas veces «soberanía» en forma de un contrato que solo sabe operar el proveedor.

La salida socialista va más lejos. El diseño de chips, la fabricación, la memoria, la energía, los centros de datos y el software forman una cadena internacional. Socializar solo una empresa nacional dejaría casi todas las dependencias intactas. Hace falta propiedad social sobre la infraestructura decisiva, cooperación entre países y control de trabajadores y usuarios sobre la asignación de capacidad. También hace falta decidir qué cálculos merecen consumir trabajo y electricidad. Ejecutar publicidad personalizada diez veces más rápido sigue siendo ejecutar publicidad personalizada.

Nvidia ha demostrado algo que el discurso tecnológico suele ocultar: el software organiza la propiedad de una máquina mucho después de venderla. CUDA es trabajo acumulado, conocimiento y coordinación. Nvidia construyó una buena plataforma apropiándose de esa

fuerza productiva común y convirtió la coordinación en una barrera de pago.

FUENTES Y MÉTODO

Las cifras financieras proceden del informe 10-K de Nvidia para el ejercicio terminado el 25 de enero de 2026. Las condiciones de compatibilidad se comprobaron en la documentación de CUDA y ROCm. Los precios son una captura del 13 de julio de 2026 de las páginas públicas de AWS Capacity Blocks y Google Cloud. No son presupuestos negociados ni permiten inferir el ingreso de Nvidia por GPU. Los cálculos derivados están redondeados y pueden reconstruirse a partir de esas fuentes.