DAORDE

A PEAXE DE CUDA

Nvidia domina o cálculo da IA porque vende unha máquina completa: procesadores, rede, bibliotecas e anos de traballo alleo convertido en dependencia. As súas contas permiten medir ese poder e descartar unha cifra máxica de renda.

Daorde Colectivo · 12.07.2026

Nvidia pechou o seu exercicio fiscal de 2026 con 215.938 millóns de dólares de ingresos. Despois de pagar fabricación, memoria, encapsulado, transporte, garantías e o resto de custos que contabiliza como custo de vendas, conservou unha marxe bruta do 71,1 %. A marxe operativa quedou no 60,4 %. O beneficio neto foi de 120.067 millóns.

Son cifras estrañas para unha empresa que adoita describirse como fabricante de chips e que nin sequera fabrica as súas propias obleas. Nvidia deseña; TSMC e Samsung producen as obleas; SK Hynix, Micron e Samsung fornecen memoria; Foxconn, Wistron e outras contratistas montan, proban e encapsulan. A propia empresa explica esta división do traballo no seu informe anual de 2026. A súa parte consiste en controlar o deseño do acelerador, a interconexión, a arquitectura do sistema e a contorna de software que fai útil o conxunto.

Nvidia vende silicio e cobra por unha relación de dependencia construída durante vinte anos.

UNHA CONTA DE RESULTADOS FÓRA DE ESCALA

Os centros de datos achegaron 193.737 millóns de dólares, case o 90 % dos ingresos do exercicio. Dous clientes directos concentraron o 22 % e o 14 % das vendas totais. Son compradores con recursos para deseñar os seus propios aceleradores, financiar centros de datos e negociar contratos xigantescos. Aínda así, Nvidia mantivo a súa marxe.

Exercicio fiscal 2026	Millóns de dólares	Porcentaxe de ingresos
Ingresos	215.938	100 %
Beneficio bruto	153.532	71,1 %
Investigación e desenvolvemento	18.497	8,6 %
Beneficio operativo	130.387	60,4 %
Beneficio neto	120.067	55,6 %

O beneficio bruto está calculado a partir da marxe publicada; as demais cifras proceden do 10-K. A comparación descarta unha explicación cómoda. Nvidia gasta moito en enxeñaría: 18.497 millóns nun ano e 31.000 dos seus 42.000 traballadores figuran como persoal de investigación e desenvolvemento. Pero o gasto en I+D non absorbe unha parte próxima ao excedente que a empresa retén. Tampouco o fan os riscos reais dunha cadea *fabless*. En 2026 cargou 4.500 millóns por inventario e compromisos do H20 afectados polos controis de exportación dos Estados Unidos. Tras ese golpe, a marxe operativa seguiu por riba do 60 %.

Unha marxe non demostra por si mesma unha renda monopolista. Pode reunir vantaxe técnica, economías de escala, escaseza temporal, poder de negociación e rendas de

propiedade intelectual. As contas públicas non separan cada compoñente. Afirmar que todo o beneficio de Nvidia é renda daría unha cifra contundente e unha investigación bastante mala.

As contas amosan que a empresa fixa condicións grazas a algo máis ca o chip illado.

A MÁQUINA COMPLETA

Unha GPU H100 é un procesador paralelo de gran capacidade. CUDA é a plataforma que permite programalo e usar bibliotecas preparadas para álxebra lineal, redes neuronais, simulación, tratamento de imaxe e outras cargas. Ao redor aparecen cuDNN, NCCL, TensorRT, compiladores, controladores, contedores, documentación e ferramentas de perfilado. Para escalar miles de aceleradores entran NVLink, InfiniBand ou Ethernet, adaptadores, conmutadores e unha arquitectura de rack deseñada como unha unidade.

Nvidia xa non presenta o centro de datos como un edificio que contén computadores. Preséntao como o computador. No seu informe anual chama «plataforma» ao conxunto de GPU, CPU, DPU, interconexións, rede, bibliotecas, modelos, datos de adestramento, API e marcos de aplicación. É unha descrición interesada, pero correcta.

O cliente non escolle entre chips espidos que executan o mesmo programa. Escolle entre sistemas cuxo rendemento depende de anos de optimización do software, dispoñibilidade de persoal, compatibilidade de modelos, ferramentas de operación e posibilidade de conseguir miles de unidades a tempo. O acelerador pode ser máis rápido e, á vez, beneficiarse de que saír resulte caro. As dúas cousas caben na mesma factura.

CUDA ACUMULA TRABALLO ALLEO

Nvidia cifra en máis de 7,5 millóns as persoas que usan CUDA e as súas outras ferramentas. Ese número inclúe traballadores da propia empresa, universidades, laboratorios públicos, compañías de nube, proxectos libres e equipos que escriben aplicacións para os seus empregadores. Cada nova biblioteca compatible aumenta o valor práctico da plataforma. Unha parte dese traballo págaa Nvidia. Outra págaa Estados, empresas e traballadores que non reciben dereitos sobre a plataforma que valorizan.

A documentación de CUDA garante que aplicacións compiladas con versións antigas sigan funcionando sobre controladores novos e ofrece vías limitadas para executar software recente en instalacións anteriores. Grazas a esa virtude técnica, unha organización conserva código, imaxes de contedor, procedementos, coñecemento operativo e resultados xa validados cando volve mercar Nvidia.

Migrar significa reescribir extensións, substituír bibliotecas, axustar *kernels*, repetir probas numéricas, formar o cadro de persoal e aceptar un período de menor rendemento. O prezo do cambio aparece como horas de enxeñaría e risco de produción, fóra dunha liña chamada «peaxe de CUDA», e moitas comparacións de custo omíteno.

ROCm ofrece unha vía competidora con soporte para PyTorch e unha gama crecente de aceleradores AMD; MLPerf Training 5.1 recibiu resultados sobre doce tipos de acelerador; en 2026 AMD publicou adestramento distribuído sobre 512 MI300X. Os grandes provedores deseñan chips propios. Os marcos de aprendizaxe automática abstraen unha parte do hardware. Todo iso reduce dependencia.

Pero un *benchmark* proba unha carga e unha configuración. Non demostra que unha organización poida migrar o seu catálogo enteiro, executar as súas extensións, atopar persoal con experiencia e mercar capacidade suficiente na rexión que precisa. Os competidores reducen o peche sen eliminalo.

O PREZO DUNHA HORA

A capacidade de cálculo chega ao usuario por outra capa de propiedade. En xullo de 2026, AWS anunciaba bloques de capacidade H100 en Londres a 3,933 dólares por GPU e hora. Google publicaba a súa máquina A3 con oito H100 a 88,49 dólares por hora baixo tarifa baixo demanda, algo máis de 11 dólares por GPU e hora. Non son produtos equivalentes: cambian compromiso, CPU, memoria, rede, almacenamento, soporte e dispoñibilidade. A diferenza impide presentar o prezo máis baixo como unha comparación completa.

O prezo de nube mestura a marxe de Nvidia coa amortización do servidor e do edificio, electricidade, refrixeración, rede, persoal, financiamento e marxe do hiperescalador. Non sabemos que parte dunha hora alugada remata en Nvidia. Podemos observar que todos os grandes provedores ofrecen a mesma plataforma porque os seus clientes a piden, e que a oferta de CUDA reforza a demanda de hardware Nvidia mentres a presenza do hardware reforza a oferta de software CUDA.

Esta retroalimentación é a forma concreta da peaxe. A organización paga polo equipo actual e carga ademais co custo acumulado de deixalo.

BENEFICIO INDUSTRIAL, ESCASEZA E RENDA

Nvidia desenvolveu procesadores excelentes e coordinou unha transición técnica desde GPU soltas ata sistemas completos refrixerados por líquido. Esa capacidade produtiva explica unha parte do beneficio. A demanda de IA medrou máis rápido que a oferta de memoria avanzada, encapsulado e electricidade; a escaseza explica outra parte. A empresa

comprometeu 95.200 millóns de dólares en fabricación, fornecementos e capacidade para asegurar a produción futura. O seu beneficio esixe capacidade industrial ademais de propiedade intelectual.

A renda aparece onde o control exclusivo dunha condición de produción permite apropiarse dunha porción do excedente producido noutros lugares. CUDA, as bibliotecas optimizadas, as interconexións e a escala instalada cumpren unha parte desa función. Un laboratorio pode producir ciencia, unha empresa pode vender inferencia e un Estado pode adestrar un modelo lingüístico. Nvidia captura valor antes de saber se eses proxectos chegarán a servir para algo. Controla un acceso escaso e difícil de substituír.

Non podemos calcular con información pública unha «taxa de renda de CUDA». Podemos documentar as súas condicións: marxes extraordinarias persistentes; unha plataforma propietaria con millóns de usuarios; custos de cambio pagados polo cliente; integración de hardware, rede e software; e capacidade para conservar prezos altos incluso ante compradores concentrados. Iso abonda para deixar de falar de Nvidia como un fabricante máis.

MÁIS ALÁ DOUTRO CAMPIÓN SUBVENCIONADO

A resposta liberal é aumentar a competencia. A Unión Europea subvenciona fábricas, os hiperescaladores deseñan ASIC e AMD mellora ROCm. Unha segunda plataforma propietaria pode baixar prezos. Non converte o cálculo nun ben común.

Unha política inmediata tería que pagar o traballo que o mercado evita: manter aplicacións en varias arquitecturas, financiar compiladores e bibliotecas libres, esixir formatos abertos na compra pública, publicar probas de portabilidade e conservar equipos técnicos capaces de migrar. As subvencións a cálculo público deberían obrigar a devolver código, resultados e melloras de ferramentas cando non exista unha razón de seguridade ou privacidade que o impida. O Estado merca demasiadas veces «soberanía» en forma dun contrato que só sabe operar o provedor.

A saída socialista vai máis lonxe. O deseño de chips, a fabricación, a memoria, a enerxía, os centros de datos e o software forman unha cadea internacional. Socializar só unha empresa nacional deixaría case todas as dependencias intactas. Cómpre propiedade social sobre a infraestrutura decisiva, cooperación entre países e control de traballadores e usuarios sobre a asignación de capacidade. Tamén hai que decidir que cálculos merecen consumir traballo e electricidade. Executar publicidade personalizada dez veces máis rápido segue sendo executar publicidade personalizada.

Nvidia demostrou algo que o discurso tecnolóxico adoita ocultar: o software organiza a propiedade dunha máquina moito despois de vendela. CUDA é traballo acumulado, coñecemento e coordinación. Nvidia construíu unha boa plataforma apropiándose desa

forza produtiva común e converteu a coordinación nunha barreira de pagamento.

FONTES E MÉTODO

As cifras financeiras proceden do informe 10-K de Nvidia para o exercicio rematado o 25 de xaneiro de 2026. As condicións de compatibilidade comprobáronse na documentación de CUDA e ROCm. Os prezos son unha captura do 13 de xullo de 2026 das páxinas públicas de AWS Capacity Blocks e Google Cloud. Non son orzamentos negociados nin permiten inferir o ingreso de Nvidia por GPU. Os cálculos derivados están arredondados e poden reconstruírse a partir desas fontes.