DAORDE

THE CUDA TOLL

Nvidia dominates AI compute because it sells a complete machine: processors, networking, libraries and years of other people's work turned into dependence. Its accounts let us measure that power and rule out a magic rent figure.

Daorde Collective · 12.07.2026

Nvidia ended fiscal year 2026 with $215.938 billion in revenue. After paying for fabrication, memory, packaging, shipping, warranties and the other expenses it records as cost of revenue, it retained a gross margin of 71.1 percent. Its operating margin was 60.4 percent. Net income reached $120.067 billion.

These are strange figures for a company commonly described as a chipmaker, especially one that does not manufacture its own wafers. Nvidia designs; TSMC and Samsung fabricate the wafers; SK Hynix, Micron and Samsung supply memory; Foxconn, Wistron and other contractors assemble, test and package the products. The company describes this division of labour in its 2026 annual report. Its position is to control the accelerator design, interconnect, system architecture and software environment that makes the whole arrangement useful.

Nvidia sells silicon and charges for a relationship of dependence built over twenty years.

AN INCOME STATEMENT OUT OF SCALE

Data centres brought in $193.737 billion, almost 90 percent of annual revenue. Two direct customers accounted for 22 and 14 percent of total sales. These buyers can design their own accelerators, finance data centres and negotiate enormous contracts. Nvidia kept its margin anyway.

Fiscal year 2026	USD millions	Share of revenue
Revenue	215,938	100%
Gross profit	153,532	71.1%
Research and development	18,497	8.6%
Operating income	130,387	60.4%
Net income	120,067	55.6%

Gross profit is calculated from the reported margin; the other figures come directly from the 10-K. The comparison rules out one easy explanation. Nvidia spends heavily on engineering: $18.497 billion in one year, and 31,000 of its 42,000 employees are listed as research and development staff. R&D nevertheless consumes nowhere near the surplus retained by the firm. Neither do the real risks of a fabless supply chain. In 2026 Nvidia recorded a $4.5 billion charge for H20 inventory and purchase commitments affected by US export controls. Its operating margin still exceeded 60 percent.

A margin does not prove monopoly rent by itself. It can combine technical advantage, economies of scale, temporary scarcity, bargaining power and intellectual-property rents. Public accounts do not separate each component. Calling all Nvidia profit rent would produce a forceful number and a poor investigation.

The accounts show that the company sets terms through something more than an isolated chip.

THE COMPLETE MACHINE

An H100 GPU is a high-capacity parallel processor. CUDA is the platform used to program it and to access prepared libraries for linear algebra, neural networks, simulation, image processing and other workloads. Around it sit cuDNN, NCCL, TensorRT, compilers, drivers, containers, documentation and profiling tools. At a scale of thousands of accelerators, NVLink, InfiniBand or Ethernet enter the picture, alongside adapters, switches and a rack architecture designed as one unit.

Nvidia no longer describes the data centre as a building that contains computers. It describes the data centre as the computer. In its annual report, the “platform” includes GPUs, CPUs, DPUs, interconnects, networks, libraries, models, training data, APIs and application frameworks. The description serves Nvidia’s interests, but it is accurate.

Customers do not choose between bare chips running the same program. They choose between systems whose performance depends on years of software optimisation, available staff, model compatibility, operational tools and the ability to obtain thousands of units on time. An accelerator can be faster while also benefiting from an expensive exit. Both facts fit on the same invoice.

CUDA ACCUMULATES OTHER PEOPLE’S WORK

Nvidia reports more than 7.5 million people using CUDA and its other software tools. They include Nvidia employees, university researchers, public laboratories, cloud companies, free-software projects and teams writing applications for their employers. Every compatible library increases the platform’s practical value. Nvidia pays for some of that work. States, firms and workers pay for another portion without receiving rights over the platform they make more valuable.

The CUDA documentation guarantees that applications compiled with older toolkits will continue to run on newer drivers and provides limited routes for recent software to run on older installations. Thanks to this good engineering, an organisation retains code, container images, procedures, operational knowledge and validated results when it buys Nvidia again.

Migration means rewriting extensions, replacing libraries, tuning kernels, repeating numerical validation, training staff and accepting a period of lower performance. Its price appears as engineering time and production risk, outside an invoice line named “CUDA toll”, and many cost comparisons omit it.

ROCm offers a competing path with support for PyTorch and a growing range of AMD accelerators; MLPerf Training 5.1 received submissions using twelve kinds of accelerator; in 2026 AMD published distributed training results across 512 MI300X GPUs. Hyperscalers design their own chips. Machine-learning frameworks abstract away part of the hardware. All of this reduces dependence.

But a benchmark proves one workload on one configuration. It does not show that an organisation can migrate its full catalogue, run its extensions, find experienced staff and purchase enough capacity in the required region. Competitors reduce lock-in without eliminating it.

THE PRICE OF AN HOUR

Compute capacity reaches the user through another layer of ownership. In July 2026, AWS advertised H100 capacity blocks in London at $3.933 per GPU-hour. Google listed an A3 machine with eight H100 GPUs at $88.49 per on-demand hour, slightly over $11 per GPU-hour. These are not equivalent products: commitment, CPU, memory, networking, storage, support and availability differ. The gap prevents the lowest price from posing as a complete comparison.

A cloud price combines Nvidia’s margin with depreciation on the server and building, electricity, cooling, networking, labour, financing and the hyperscaler’s margin. We do not know how much of a rented hour ends up at Nvidia. We can observe that every major provider offers the same platform because customers demand it, and that CUDA software reinforces demand for Nvidia hardware while installed hardware reinforces the supply of CUDA software.

This feedback loop is the concrete form of the toll. An organisation pays for the current equipment and carries the accumulated cost of leaving it.

INDUSTRIAL PROFIT, SCARCITY AND RENT

Nvidia has developed excellent processors and coordinated a technical transition from individual GPUs to liquid-cooled complete systems. This productive capacity explains part of its profit. Demand for AI grew faster than the supply of advanced memory, packaging and electricity; scarcity explains another part. The company held $95.2 billion in manufacturing, supply and capacity commitments at the end of the year. Its profit requires industrial capacity alongside intellectual property.

Rent appears where exclusive control over a condition of production allows its owner to appropriate part of the surplus produced elsewhere. CUDA, optimised libraries,

interconnects and installed scale perform some of that function. A laboratory may produce science, a company may sell inference and a state may train a language model. Nvidia captures value before it knows whether any of those projects will be useful. It controls access to a scarce resource that is difficult to replace.

Public information cannot produce a “CUDA rent rate”. It can document the conditions: persistently extraordinary margins; a proprietary platform with millions of users; switching costs paid by the customer; hardware, network and software integration; and the ability to maintain high prices even when facing concentrated buyers. That is enough to stop treating Nvidia as one manufacturer among many.

BEYOND ANOTHER SUBSIDISED CHAMPION

The liberal answer is more competition. The European Union subsidises factories, hyperscalers design ASICs and AMD improves ROCm. A second proprietary platform may lower prices. It does not turn compute into a commons.

An immediate policy would have to pay for the work the market avoids: maintaining applications across architectures, funding free compilers and libraries, requiring open formats in public procurement, publishing portability tests and keeping technical teams capable of migration. Public compute subsidies should require recipients to return code, results and tool improvements when no concrete security or privacy reason prevents it. States purchase “sovereignty” too often in the form of a contract only the supplier knows how to operate.

A socialist answer goes further. Chip design, fabrication, memory, energy, data centres and software form an international chain. Socialising one national company would leave nearly every dependency intact. What is required is social ownership of decisive infrastructure, cooperation among countries and control by workers and users over the allocation of capacity. Society would still need to decide which computations deserve labour and electricity. Running personalised advertising ten times faster is still running personalised advertising.

Nvidia has demonstrated something technology discourse routinely hides: software organises ownership of a machine long after the machine is sold. CUDA is accumulated labour, knowledge and coordination. Nvidia built a good platform by appropriating that common productive force and turned coordination into a barrier that charges admission.

SOURCES AND METHOD

Financial figures come from Nvidia’s Form 10-K for the fiscal year ended 25 January 2026. Compatibility claims were checked against the CUDA and ROCm documentation. Prices are a 13 July 2026 snapshot of the public AWS Capacity Blocks and Google Cloud pages. They are not negotiated quotes and cannot reveal Nvidia revenue per GPU. Derived calculations are rounded and can be reconstructed from the cited sources.