The sleeping equity giant

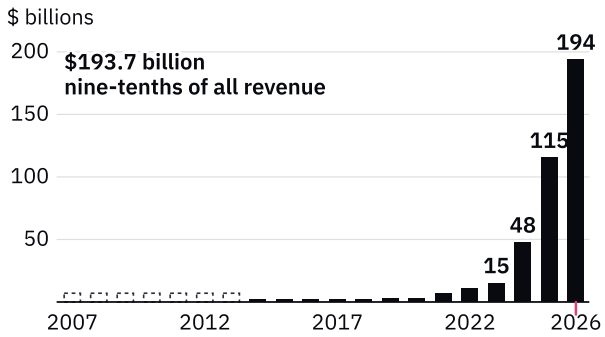
James E. Coleman

STRATEGIC DESIGN · WHITE PAPER SERIES · 2026

ABSTRACT In November 2006 Nvidia put a free programming layer called CUDA on a graphics card built for video games. For the next ten fiscal years the business it was meant to create was too small to round up to a billion dollars. In that time the company's research bill rose through two years of losses and its shares fell 85 percent. Jensen Huang calls a bet of that kind a zero-billion-dollar market. The title refers to that wait. It took ten years for the product to find its customer, and the report for fiscal 2016 is the first to give the business a line of its own. Last fiscal year the same business brought in $194 billion. This dispatch reads twenty annual reports to measure what the wait cost, who built the thing, and which part of it was one person's doing.

FIG. 1 — DATA-CENTER REVENUE BY FISCAL YEAR, 2007 TO 2026

STATE 7 OF 7



FISCAL 2007

Not yet a line

CUDA is announced in November 2006 with the GeForce 8800. The annual report describes it in a few sentences and reports no revenue for it.[2]

FISCAL 2009

Not yet a line

The first loss-making year. The report counts 25,000 developers and still has no revenue line for what they do.[3]

FISCAL 2013

Not yet a line

AlexNet wins the image contest in the autumn of 2012 on two Nvidia games cards. The annual report for the year says nothing about neural networks.[15,5]

FISCAL 2016

$0.34 billion

The first report to give data centers a line of their own: $199 million, $317 million and $339 million for fiscal 2014 to 2016. To the nearest billion, each of the three is zero.[5]

FISCAL 2017

$0.83 billion

The first year that rounds up to one billion. Huang had delivered the first DGX-1 to OpenAI that August.[5,19]

FISCAL 2023

$15.0 billion

More than half of all revenue for the first time. ChatGPT was released two months before the fiscal year ended.[5]

FISCAL 2026

$193.7 billion

About 970 times the fiscal 2014 figure, and 90 percent of everything Nvidia sold.[1]

Fiscal years end in late January, so fiscal 2007 is mostly calendar 2006. A hollow mark is a year in which the annual report gave data-center revenue no line of its own. The scale is linear, and the first ten years are as flat as they look.



10 years

from the launch in **November 2006** to the first fiscal year in which data-center revenue rounded up to a billion dollars.[2,5]



+31%

research spending from fiscal 2008 to 2010, while revenue fell **19 percent** and the company lost money twice.[6]



85.5%

fall in the share price, from **$39.67** in late 2007 to $5.75 a little over a year later.[3]



970×

data-center revenue in fiscal 2026 against fiscal 2014: **$193.7 billion** against $199 million.[1,5]

§ 01 The claim

The most valuable decision in the AI industry was a budget line. *It was held for ten years while the market for it rounded to zero.*

Jensen Huang has a name for a market that does not exist yet. He calls it a “zero-billion-dollar market,” and he has said these are the markets he prefers.[11,12] The idea is that a company should build for a use nobody is paying for yet, so that it is alone there when the customers arrive. CUDA is the largest test of that idea on record. It is the software that lets a programmer use an Nvidia graphics chip for general computation, and it was announced in November 2006 alongside a card designed for video games.[2,10]

This paper measures the test with the company’s own filings. Nvidia has filed an annual report every year since CUDA was launched. Those twenty reports contain the revenue, the research spending, the share price and the count of developers for each year, and read in order they show three things. First, the business CUDA was meant to create stayed small for a decade. The report for fiscal 2016 is the

first to give data-center revenue a line of its own, and the figure for that year, the tenth after the launch, is $339 million.[5] Second, the cost was carried in plain sight. Between fiscal 2008 and fiscal 2010 revenue fell 19 percent, the company lost money two years running, and research spending rose 31 percent.[6] Third, the payoff was out of all proportion to the first ten years. Data-center revenue in fiscal 2026 was $193.7 billion, which is about 970 times the fiscal 2014 figure and nine-tenths of everything Nvidia sold.[1] Fig. 1 walks through the twenty years.

What this paper claims, and what it does not. The claim is that Huang's contribution to CUDA was neither the invention nor a forecast of artificial intelligence. Other people wrote the language, and the record shows that nobody at Nvidia, Huang included, was building it for neural networks. His contribution was two decisions and the length of time he held them. He required every Nvidia graphics chip to carry CUDA, so that people buying cards for games paid for the installed base. Then he kept funding it for about ten years while the stock market told him to stop. This paper does not claim the decade was as bleak as he sometimes tells it. Section 08 sets two of his recollections against the filings, and one of them does not hold.

A disclosure. I am a customer. I run Nvidia graphics cards in a rig that mines Ethereum Classic, which is one more job the cards were never designed for. Section 05 comes back to that.

This is the third paper in this series built on Nvidia's filings. The Name on the Chip, No. 13, graded the reporting on diverted chips, and Key Person Risk, No. 15, covered Huang's resignation on November 11, 2026.[23,24,25] It is also the last of the three. The reason for writing it this week is that the days after a founder leaves are when a company's history gets rewritten, in both directions. The filings do not change.

§ 02 The thing itself

CUDA lets ordinary code run on a graphics chip. *It came out of a Stanford doctoral project called Brook.*

A graphics chip is built to do one simple calculation on millions of pixels at the same moment. A central processor does complicated calculations one after another. In the early 2000s a few scientists noticed that many of their problems looked more like the first kind of work than the second. They began disguising physics as pictures so that a games card would run it.[13,14] It worked, and it was miserable to program.

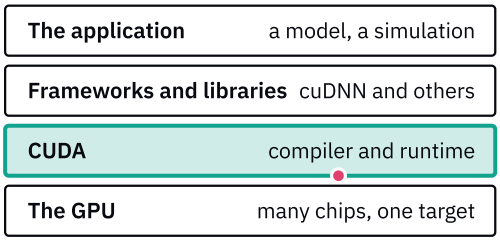
Ian Buck was a doctoral student at Stanford who made it less miserable. Brook, the system Buck and six colleagues presented in 2004, extended the C language so that a programmer could describe a calculation over a stream of data and let the graphics chip run it, without pretending it was a picture.[7] Nvidia hired Buck that year.[9] The paper that first set CUDA out for programmers, published in 2008, has four authors: John Nickolls, Ian Buck, Michael Garland and Kevin Skadron.[8]

Nvidia's annual report for fiscal 2007 is the first to mention it. The report says the company announced CUDA with the GeForce 8800, and that from November 2006 it began seeding the industry with GeForce 8 chips that could run programs written in C.[2] The 8800 GTX went on sale on November 8, 2006 for about $600, with 128 processing units on a chip of 681 million transistors.[10] The kit for programmers was made public the following February.[9] It was free to download then and

it is free now. It has never been open source, and Nvidia has only ever supported it on its own chips.[14] Section 09 returns to that.

FIG. 2 — WHERE CUDA SITS

STATE 3 OF 4



TOP LAYER

The application

A weather model, a drug simulation or a chatbot. The person who writes it wants an answer sooner, and does not want to learn how a graphics chip works.

SECOND LAYER

Frameworks and libraries

Shared code that turns the application's mathematics into operations a chip can run. Nvidia's annual report counts hundreds of its own libraries for particular fields.[1] cuDNN, the one for neural networks, arrived in 2014.[18]

THIRD LAYER

CUDA

The compiler, the runtime and the programming model. It takes a program written in C or one of its descendants and spreads the work across the chip's many processors. The same program runs on many different chips.[2,12]

BOTTOM LAYER

The GPU

The chip itself. The first to carry CUDA had 128 processing units.[10] Every Nvidia graphics chip since has been able to run the same programs.[12]

A diagram of the relation, not a chart of data. Four layers, with the application at the top and the chip at the bottom. CUDA is the third.

It is not a feature of the chip. The common mistake is to look at CUDA at the chip layer, as if it were a setting on the hardware. CUDA cannot do its job by looking at one chip. Its job is to hide the differences between chips, so that a program written once runs on many different Nvidia GPUs, from a games card to a rack in a data center, and goes on running on the generations that follow. It is better understood as the agreement between all four layers in Fig. 2, and the reason the layers can be improved separately without breaking each other. A scientist or a model builder works in the top two layers and rarely sees CUDA at all. The code they depend on sees nothing else. Nvidia now describes its product in the same terms. The latest annual report says more than half of the company's engi-

neers work on software, and that the data center, networking included, has become the unit of computing.[1] A rival can match a chip. To displace CUDA it has to match the whole picture.

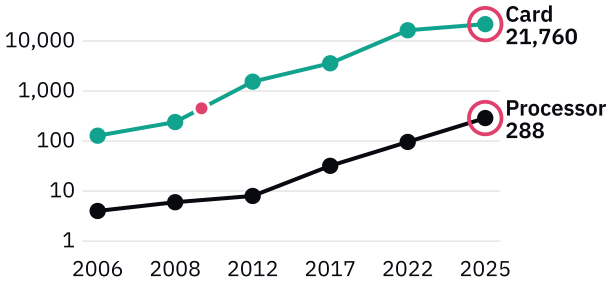
The chip holds many small processors, and CUDA puts all of them to work. An ordinary processor has a few powerful cores and does complicated work one step after another. A graphics chip has many small processing units that each do a simple calculation, and all of them run at the same moment. The first card to carry CUDA had 128 of them. The server processor Intel launched in the same month had four cores.[10,33] Before CUDA those units could be used only to draw pictures. CUDA let a program written in C hand them any calculation that can be split into many small pieces, and a weather model, a drug simulation and a neural network are all calculations of that kind.[2,8]

The first published measurements showed gains of ten to several hundred times. The four engineers who set CUDA out in 2008 reported programs running 10 to 100 times faster in molecular dynamics, 50 to 250 times faster in a simulation of bodies moving under gravity, and 263 times faster in reconstructing a medical scan. Each figure compares the graphics chip with the same work done one step at a time on an ordinary processor.[8] They are the authors' own measurements, for problems chosen because they divide well, and a calculation that cannot be divided gains nothing from the extra units.

The count of units then rose with each generation of card. Nvidia's leading games card carried 240 units in 2008, 1,536 in 2012, 3,584 in 2017, 16,384 in 2022 and 21,760 in 2025.[31,32] A program written for CUDA used the extra units without being rewritten, because every card ran the same code. Server processors gained cores over the same years, from 4 to 288 in the examples in Fig. 3, and in each pair the card has between 32 and 192 times as many units as the processor has cores.[33] The two counts do not measure the same thing. One unit on a graphics chip does far less than one core of a processor, so Fig. 3 compares how many there are and not what each can do.

FIG. 3 — PROCESSING UNITS PER CARD, CORES PER PROCESSOR

STATE 6 OF 6



21,760 units on the card, 288 cores on the processor

2006

128 against 4

The GeForce 8800 GTX, the first card to carry CUDA, had 128 processing units. The Xeon server processor Intel launched the same month had four cores.[10,33]

2008

240 against 6

The GeForce GTX 280 had 240 units. The Xeon launched that autumn had six cores.[32,33]

2012

1,536 against 8

The GeForce GTX 680 had 1,536. Nvidia's specification page calls them CUDA cores. The Xeon launched that spring had eight cores.[31,33]

2017

3,584 against 32

The GeForce GTX 1080 Ti had 3,584. AMD's first EPYC server processor, launched that June, had 32 cores.[32,33]

2022

16,384 against 96

The GeForce RTX 4090 had 16,384. AMD's EPYC 9654, launched that November, had 96 cores.[31,33]

2025

21,760 against 288

The GeForce RTX 5090 has 21,760. The Xeon generation Intel launched in June 2024 reaches 288 cores, the most of any processor in this figure.[31,33]

The vertical scale is logarithmic: each gridline is ten times the one below. The teal line counts the processing units on Nvidia's leading games card of the year. The dark line counts the cores on a server processor launched in the same year, except the last, which was launched in 2024. The processors are examples and are not claimed to be the largest of their year. A unit on a graphics chip does much less than a processor core, so the figure compares how many there are and not what each can do.

Over the same years the ordinary processor stopped getting faster at its old rate. John Hennessy and David Patterson charted how fast a single processor runs a standard test, year by year. In the best years it improved 52 percent a year, which doubles performance every year and a half. The rate fell to 23 percent a year when chips moved to several cores, then to 12 percent, and in the last period on their chart, which begins around 2015, to 3 percent a year, which doubles performance every twenty years.[30] Huang told Nvidia's conference in March 2024 that general-purpose computing had run out of steam.[37] Nvidia's own account of a 2023 talk by its chief scientist puts the gain in artificial-intelligence work on a single graphics chip at 1,000 times in ten years.[36]

By two public measures the graphics chip has overtaken the processor in the data center. The first is the list of the world's 500 fastest computers. In November 2010, 17 of the 500 used a graphics chip or another accelerator beside their processors, and 10 of the 17 used Nvidia's. In June 2026, 276 of the 500 did, and 169 of those ran on one of two recent generations of Nvidia chip.[34] The second measure is revenue. Intel's data-center division took in $14.4 billion in 2014. Nvidia's data-center revenue for almost the same twelve months, its fiscal 2015, was $317 million. Intel's division reported $26.1 billion for 2020, the highest figure in the series, and $16.9 billion for 2025. Nvidia's data-center revenue passed it in fiscal 2024, at $47.5 billion against Intel's $15.5 billion for 2023, and in fiscal 2026 it was $193.7 billion.[35,5,1] Intel has redrawn the division several times, so its figures are given here as first reported and are not strictly comparable from one year to the next.

The rule was that every card would run it. *People who bought the cards for games paid for a supercomputer they never used.*

A new computing platform has a familiar problem. Programmers will not write for it until many machines can run their programs, and nobody buys the machines until the programs exist. Most attempts to start a platform fail at this point. Nvidia had an unusual way around it, because it already sold graphics cards in very large numbers to people who wanted them for another reason.

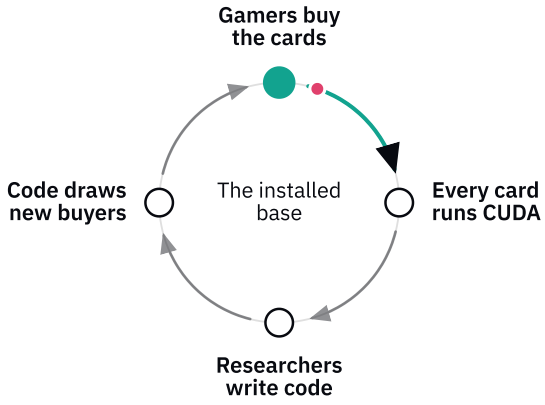
Huang explained the logic in a 2023 interview. The card was paid for by the buyer who wanted it for games and graphics, and CUDA rode along on it. So the installed base for the new platform was built with other people's money. The interviewers put it to him that he had insisted on this personally, and he agreed. In his account a computing platform is only a platform if every unit is compatible with every other, and that is the one rule at Nvidia that is not open to negotiation. By then, on his count, between 250 and 300 million CUDA-capable chips were in use.[12]

The cost of the rule fell on the chip. Circuitry that a game never used still had to be designed, tested and manufactured on every card. In a 2024 interview Huang said CUDA added a great deal of cost to the chips at a time when no customer would pay extra for it, and that the company's margins and its market value suffered as a result.[11] The next section checks that against the accounts.

The benefit fell on a student with no budget. A researcher who could not get time on a university supercomputer could buy a GeForce card at an electronics shop. For a few hundred dollars that was a parallel computer, and code written for it would still run on next year's card. One of Nvidia's founders has described graduate students doing exactly this, and Huang recalls a chemist in Taiwan with a cupboard full of games cards.[13] Fig. 4 shows the loop this set going.

FIG. 4 — WHO PAID FOR THE INSTALLED BASE

STATE 1 OF 4



STEP 1

Gamers buy the cards

The buyer wants better graphics and pays for the whole chip. In fiscal 2008 Nvidia had revenue of $4.1 billion, and no annual report of the period gives computing a revenue line of its own.[6,5]

STEP 2

Every card runs CUDA

The same chip can run general programs whether or not its owner ever does. The installed base grows with every card sold for any purpose.[12]

STEP 3

Researchers write code

A scientist can assume the machine exists, on a desk or in a shop. By early 2009 the annual report counted more than 25,000 developers, and more than 100 universities teaching it.[3]

STEP 4

Code draws new buyers

Each program that runs faster on the cards is a reason for a laboratory, and later a data center, to buy more of them.

A diagram of the relation, not a chart of data. Each state draws one stretch of the loop, and the dot travels it.

§ 04 The second decision

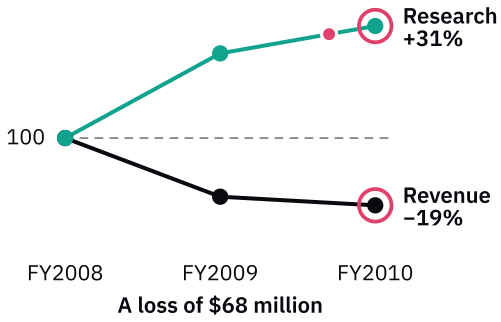
Revenue fell 19 percent and research spending rose 31 percent. *The accounts show what holding the bet cost, year by year.*

No filing gives a cost for CUDA alone. Nvidia has never reported one. The figure sometimes quoted, a billion dollars or more a year, is close to the company's whole research budget in those years.[14,6] From fiscal 2008 to fiscal 2016 that budget came to $9.5 billion, and it paid for games chips as well as CUDA. The filings do not divide it.

What the accounts do show is the shape of the decision. Fiscal 2008, which ended in January 2008, was a record year, with revenue of $4.10 billion and a profit of $798 million. Then the financial crisis arrived. Revenue fell to $3.43 billion and then to $3.33 billion, and the company lost $30 million and then $68 million.[6,4] A chip company in that position usually cuts research. Nvidia's research spending went from $692 million to $856 million to $909 million. By fiscal 2010 it was spending 27 cents of every dollar of revenue on research, up from 17 cents two years earlier.

FIG. 5 — REVENUE AND RESEARCH SPENDING, FISCAL 2008 = 100

STATE 3 OF 3



FISCAL 2008

A record year

Revenue of $4.10 billion and a profit of $798 million. Research spending was $692 million, or 17 percent of revenue.[6] CUDA had been on sale for a little over a year.

FISCAL 2009

The first loss

Revenue fell to $3.43 billion and the company lost $30 million. Research spending rose 24 percent to $856 million.[6] The report for this year is the first to describe CUDA at length: what it is, how many developers use it, and where it is taught.[3]

FISCAL 2010

The second loss

Revenue slipped again to $3.33 billion and the loss widened to $68 million. Research spending rose again, to $909 million, which was 27 percent of revenue.[6,4]

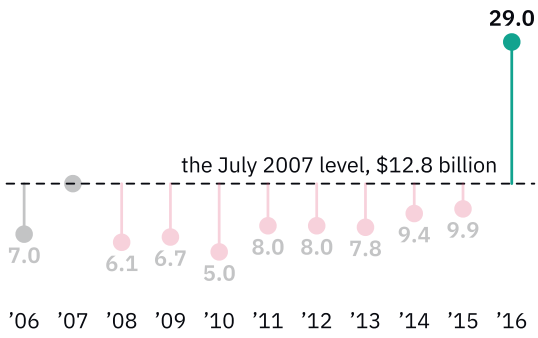
Both lines start at 100 in fiscal 2008. The dark line is revenue. The teal line is research and development spending for the whole company.

The stock market's view is in the filings too. Each annual report prints the highest and lowest price of the shares in each quarter. The high in the autumn of 2007 was $39.67. The low in the quarter that ended in January 2009 was $5.75, a fall of 85.5 percent. With 542 million shares outstanding, the company was worth about $3.1 billion at that price.[3] Much of the fall was the crisis, which took most technology shares down with it. What was particular to Nvidia was how long the recovery took.

The cover of each annual report states what the shares held by outside investors were worth on a day in late July of the year before. In July 2007 that figure was $12.8 billion. It was lower than that at the same point in each of the next eight years. It did not pass the 2007 level until July 2016, when it reached $29 billion.[3,5] Fig. 6 plots the eleven Julys.

FIG. 6 — WHAT OUTSIDE INVESTORS' SHARES WERE WORTH EACH JULY, $ BILLIONS

STATE 4 OF 4



JULY 2006 AND 2007

$12.8 billion

The shares held by outside investors were worth $7.0 billion in July 2006 and $12.8 billion a year later, eight months after CUDA went on sale.[5]

JULY 2008 TO 2010

Less than half

They were worth $6.1 billion, $6.7 billion and $5.0 billion. The low point on this measure, in 2010, was 61 percent below July 2007.[3,4,5]

JULY 2011 TO 2015

Five more years below

Between $7.8 billion and $9.9 billion. AlexNet was published in the middle of this stretch, and the figure did not move.[5]

JULY 2016

$29 billion

The first July above the 2007 level. The fiscal year then under way was the first in which data-center revenue rounded up to a billion dollars.[5]

Each dot is the figure printed on the cover of an annual report, for a day in late July. It counts shares held by outside investors and leaves out those held by directors and officers. Rose dots are below the July 2007 level.

Stephen Witt wrote a history of the company with Huang's cooperation. Witt says that in these years Wall Street did not believe in the project, that activist investors bought in and pressed for change, and that Huang was not well regarded.[14] The filings cannot confirm what investors said in private. They do confirm that anyone who bought at the high in the autumn of 2007 did not see that price again until 2016, more than eight years later. In those years the research budget fell only twice.[5,6]

One qualification. Huang has said that CUDA's cost crushed the company's gross margin.[11] The margin did fall, from 45.6 percent in fiscal 2008 to 34.3 percent in fiscal 2009. But the annual report for that year gives its own reasons. They are a $196 million charge for a batch of faulty notebook chips, larger write-downs of unsold stock, and the recession. In the first full year after CUDA's launch the margin had risen.[3] The cost of CUDA was real. In the accounts it cannot be separated from everything else that went wrong in 2008.

§ 05 The first users

The first customers were scientists with small budgets. *Artificial intelligence was not on the list.*

Huang's list of the early uses is the same from one interview to the next: molecular dynamics, seismic processing for oil companies, the reconstruction of medical scans, and image processing.[11,12,13] The annual reports of the time say the same in more commercial language. The report for fiscal 2009 lists molecular dynamics, medical imaging and the modelling of financial risk. It names Motorola, Chevron, GE Health Care and General Mills among the customers for the computing products built on CUDA.[3] Huang has said these uses did not make the project a success. They brought in just enough to buy it time.[11]

Witt is blunter about how thin the business was. One early medical scanner built around the technology, on Witt's account, cost a couple of million dollars and had two customers. Scientists who already

had working programs did not want to rewrite them. Nvidia responded the way it had with games developers, by sending its own engineers to work beside the users.[14] It also started a conference. The first GPU Technology Conference was held in San Jose in the autumn of 2009, in the middle of the second loss-making year, and about 1,500 people came.[20]

The numbers in the filings are small and specific. There were more than 25,000 developers at the start of 2009. More than 100 universities were teaching parallel programming with CUDA that year, and more than 300 a year later.[3,4] After that the annual report stops counting developers for nine years.

FIG. 7 — THE EARLY FOLLOWING

STATE 1 OF 3

ANNUAL REPORT, FISCAL 2009

A small following

The report says more than 25,000 developers around the world were using CUDA.[3] It is the only count the company published for the next nine years.

A YEAR LATER

Tripled in a year

More than 100 universities taught parallel programming with CUDA according to the report for fiscal 2009, and more than 300 according to the next one.[3,4]

FISCAL 2009 AND 2010

Built for science

The reports list molecular dynamics, image processing, medical imaging and the modelling of financial risk. Deep learning is first named as a use in the report for fiscal 2015.[3,4,5]

Three readings from the annual reports, one at a time.

Neural networks were not what any of this was for. Huang has said that before 2012 the company was trying to make CUDA work for computer vision and finding that the existing methods were a poor fit for it.[12] Witt describes a list of target markets with weather and oil at the top and computer vision near the bottom, and says the researchers who used the cards for neural networks were among the smallest customers Nvidia had.[14] The annual reports of those years list CUDA's uses in detail, and neural networks are not among them. Deep learning is first named as a use in the report for fiscal 2015.[5]

The same thing has happened since with other uses. I use the cards to mine Ethereum Classic. Nobody designed them for that either. For a while it mattered enough that the annual report for fiscal 2019 said demand from cryptocurrency miners had distorted the trend in the company's games revenue.[5] A general-purpose machine in enough hands will be put to uses its maker did not plan. The scientists were the first case and the miners were a later one. The largest is the subject of the next section.

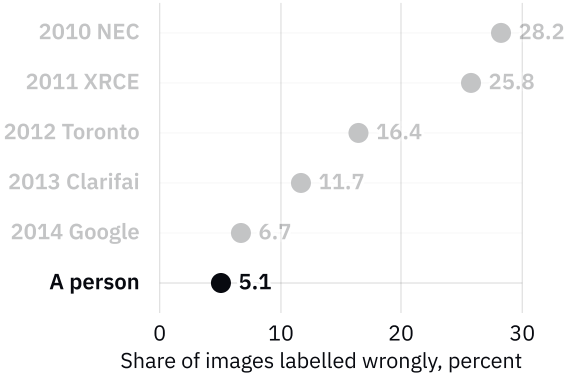
The proof took two games cards and six days. *It came from three researchers in Toronto, and Nvidia had no part in it.*

In the autumn of 2012 Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton of the University of Toronto entered an annual contest in which programs label photographs. Their entry was a neural network with 60 million adjustable values, trained on 1.2 million images. Their paper says the training took five to six days on two GeForce GTX 580 cards with three gigabytes of memory each.[15] These were consumer cards. Witt puts the cost of the pair at about a thousand dollars.[14]

The entry got 15.3 percent of the test images wrong, counting an answer as right if the correct label was among its five guesses. The next best entry got 26.2 percent wrong.[15] Contests of this kind are usually won by a point or two. The organizers' history of the contest gives the Toronto result on the standard training data as 16.4 percent, and shows what happened next. Within two years every leading entry was a neural network and the error was 6.7 percent. A trained person scored 5.1 percent on the same task.[16]

FIG. 8 — THE WINNING ERROR RATE IN THE IMAGE CONTEST, 2010 TO 2014

STATE 6 OF 6



2010

28.2 percent

The first contest. The winner, from NEC, used features designed by hand and a conventional classifier.[16]

2011

25.8 percent

A team from Xerox's research centre in Europe won with a refinement of the same approach. The gain was 2.4 points in a year.[16]

2012

16.4 percent

Krizhevsky, Sutskever and Hinton, entered as SuperVision. With extra training data their entry reached 15.3 percent. The runner-up scored 26.2.[15,16]

2013

11.7 percent

The winner was Clarifai. Every leading entry was now a neural network trained on graphics chips.[16]

2014

6.7 percent

The winner was Google's entry, GoogLeNet. The error had fallen by three-quarters in four years.[16]

FOR COMPARISON

5.1 percent

The organizers had a trained annotator label 1,500 of the test images. The annotator got 5.1 percent wrong.[16]

An answer counts as right if the correct label is among the program's five guesses. Lower is better.

Two statements in the paper matter for this story. The authors write that the size of their network was limited mainly by the memory on the cards, and that their results could probably be improved just by waiting for faster ones.[15] The statements read like a purchase order. They told every laboratory in the field that the way to a better model was to buy more of one company's product.

It was not the first sign. Andrew Ng has said that students at Stanford were pressing for graphics chips from about 2008, and that the students were the ones who brought up CUDA.[13] Nvidia's own account is that a Google experiment which needed 2,000 processors to learn to recognize cats was matched with 12 of its graphics chips.[17] Huang says the laboratories of Hinton in Toronto and Yann LeCun in New York got in touch at about the same time.[13] But the 2012 contest was the public proof, and what made it possible was six years old by then. The cards were cheap because they were sold for games. The code ran because every card ran CUDA.

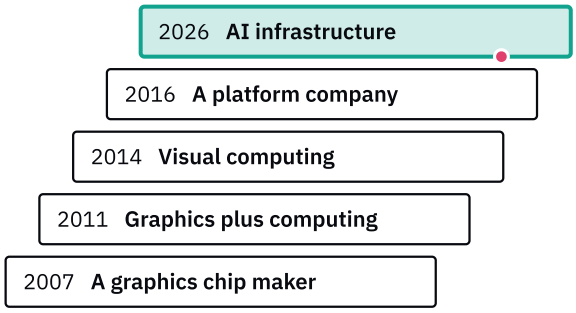
§ 07 The turn

Two years after the proof, Nvidia shipped a library for it. *Four years after, Huang carried the first machine to the laboratory himself.*

The usual telling has Huang seeing the 2012 result and turning the company at once. The filings suggest something slower. The annual report for fiscal 2014 was written more than a year after the contest. It still introduces Nvidia as a visual computing company, known for the chips that bring video games to life. The first product aimed at the field was cuDNN, a library of neural-network routines released in September 2014. It plugged into Caffe, a framework researchers were then using, and Nvidia's own test showed training running more than ten times faster than on a central processor.[18] The report filed the following spring is the first to name deep learning among the things researchers do with the chips. A year after that, the opening page says Nvidia has become a platform company and names artificial intelligence as a new area. The most recent report opens by calling Nvidia an AI infrastructure company.[5,1] Fig. 9 follows the company's description of itself.

FIG. 9 — HOW NVIDIA INTRODUCED ITSELF IN ITS ANNUAL REPORT

STATE 5 OF 5



ANNUAL REPORT, FISCAL 2007

A graphics chip maker

The report opens by calling Nvidia the leader in programmable graphics processors. CUDA, announced that year, is described further in, in a few sentences.[2]

FISCAL 2011

Graphics plus computing

The opening now says the company's work on programmable chips has made supercomputing cheaper and more widely available. High-performance computing is one of three areas it says it is investing in.[5]

FISCAL 2014

Visual computing

More than a year after AlexNet, the report still introduces Nvidia as a visual computing company, known for the chips that bring video games to life. Data centers appear in a list of target markets.[5]

FISCAL 2016

A platform company

The opening says Nvidia began as a maker of PC graphics chips and has become a platform company. It names artificial intelligence as a new area that its chips are powering.[5]

FISCAL 2026

AI infrastructure

The report now introduces Nvidia as an AI infrastructure company working at the scale of the data center. It names CUDA in its third sentence, as the foundation of the software stack.[1]

A diagram of the relation, not a chart of data. Each step paraphrases the opening lines of that year's annual report.

Then came the hardware. Nvidia announced the DGX-1 at its own conference on April 5, 2016. It was a single computer holding eight of the company's fastest graphics chips and built only for training neural networks, and the company said it did the work of 250 ordinary servers.[38] The list price was $129,000.[19] Huang said that August that the machine had taken 3,000 people three years to build, and that if only one had ever been shipped the project would have cost $2 billion.[39]

The first one went to OpenAI, and the record says it was a gift. OpenAI was then a nonprofit research laboratory less than a year old, and the press described it as Elon Musk's artificial intelligence project.[39,40] Huang took the machine to its office in San Francisco in the second week of August 2016

and handed it over himself. He wrote on the case that he was presenting the world's first DGX-1 to Elon and the OpenAI team, and dedicated it to the future of computing and humanity.[39,40] Musk thanked Nvidia and Huang in public for donating it, the business press reported a donation, and a later OpenAI research paper thanks Nvidia for the gift of a DGX-1.[40,41,42] The trade press at the time could not establish whether OpenAI had paid, and no source read for this paper shows that it did.[19]

Nvidia's account of the day quotes three people at OpenAI. Ilya Sutskever, one of the three Toronto researchers of Section 06, said the machine would let them work on problems they had not been able to explore. Andrej Karpathy said that more computing power was a simple way to make the models work better, and that training could go from a month of conversations on Reddit to whole years of them. Greg Brockman, the chief technology officer, said artificial intelligence could be the most positive technology people create.[39] By Huang's account Sutskever was in the room,[12,44] so one of the three people who had run the 2012 proof on two games cards was there four years later to receive the machine built for that work.

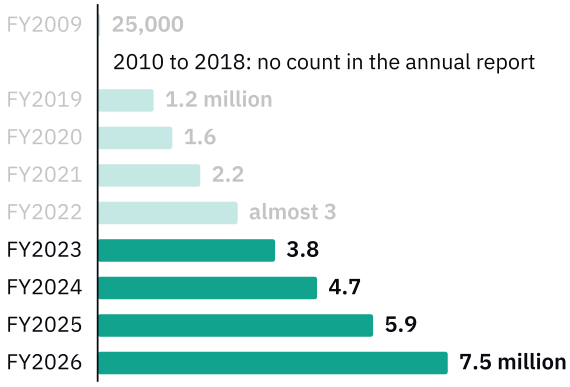
Huang has told the story of how it came about in several ways. In 2023 he said he knew the people at OpenAI and knew they needed the computer Nvidia was building.[12] Later that year he said Musk had heard about the machine at a conference and said he wanted one.[43] In December 2025 he gave the longest version. He said he announced the DGX-1 to a silent room and had no purchase order for it. He thought he had found his first customer when Musk said he had a company that could use one, and Musk then told him the company was a nonprofit. Huang said he boxed up a machine that had been built for Nvidia's own use and drove it to San Francisco.[44] In that telling the price was $300,000, which is not the $129,000 reported in 2016.[44,19] Reports published in 2026 say that emails between the two men from April 2016, released as exhibits in Musk's lawsuit against OpenAI's chief executive, show Musk asking whether OpenAI could buy one of the first units. The emails themselves were not read for this paper.[45] The versions agree on the delivery and differ on who asked first and on whether anyone else wanted the machine.

The delivery fits the pattern this paper describes. Nvidia had built a product ahead of its customer, and in Huang's telling the first taker was a laboratory that could not pay, so the company gave it the machine. He has since repeated the gesture twice. He delivered the first of a later model, the DGX H200, to OpenAI by hand in April 2024, and he took one of the first DGX Spark machines to Musk at the SpaceX site in Texas in October 2025.[46,47] In September 2025 the two companies signed a letter of intent for OpenAI to deploy at least 10 gigawatts of Nvidia systems, with Nvidia intending to invest up to $100 billion as they were built, and Brockman said the plan amounted to a billion times the computing power of that first machine.[48] The plan was not completed on those terms. Nvidia's finance chief said in December 2025 that no final agreement had been signed, and in February 2026 OpenAI announced a funding round that included $30 billion from Nvidia.[49,24]

The revenue followed in the order the loop in Fig. 4 predicts. Data-center revenue was $199 million in fiscal 2014, $317 million in fiscal 2015 and $339 million in fiscal 2016. Then it was $830 million, $1.9 billion and $2.9 billion.[5] By fiscal 2023, the year ChatGPT was released, it was $15.0 billion. In the three years after that it went to $47.5 billion, $115.2 billion and $193.7 billion.[5,1] The annual report began counting developers again in fiscal 2019, at 1.2 million. The latest count is more than 7.5 million.[5,1]

FIG. 10 — DEVELOPERS USING CUDA AND NVIDIA'S OTHER SOFTWARE TOOLS

STATE 3 OF 3



FISCAL 2009

25,000

The only count for a decade. On the scale of this chart it is a sliver.[3]

FISCAL 2019 TO 2022

1.2 million to 3 million

The report resumes the count at 1.2 million and adds about 600,000 a year.[5]

FISCAL 2023 TO 2026

3.8 million to 7.5 million

After ChatGPT the additions pass a million a year. The counts are 3.8, 4.7, 5.9 and 7.5 million.[5,1]

Each bar is the count given in that year's annual report. The early count is of CUDA developers. The later ones also include Nvidia's other software tools.

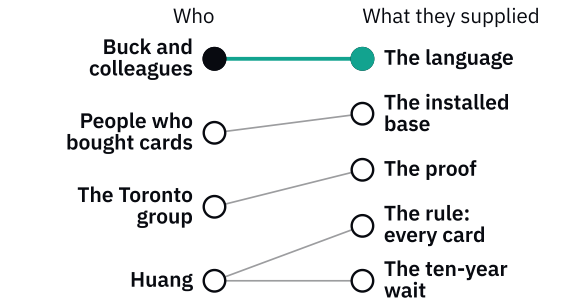
§ 08 The credit

He did not write CUDA and he did not predict AI. *He made two decisions and held them for ten years.*

The week after a founder leaves, it is tempting to credit the founder with everything or with nothing. The record supports neither. Fig. 11 sets out who supplied what.

FIG. 11 — WHO SUPPLIED WHAT

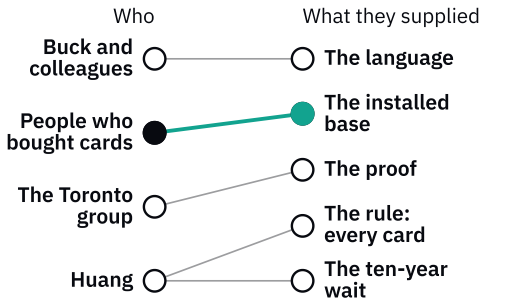
STATE 1 OF 4



STANFORD, THEN NVIDIA

The language

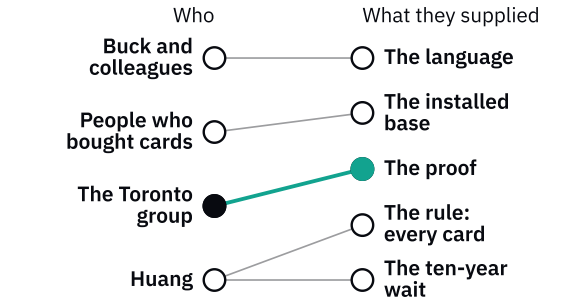
Ian Buck led Brook at Stanford and was hired in 2004. The 2008 paper that sets out CUDA has four authors, and Huang is not one of them.[7,8,9]



CUSTOMERS

The installed base

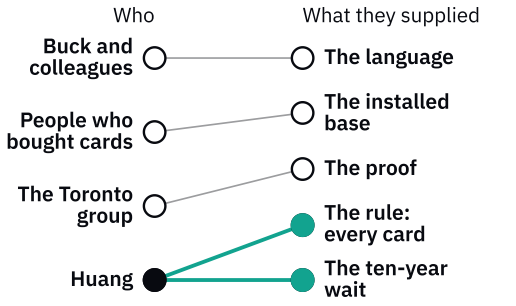
People who bought cards for games and graphics paid for chips that could also compute. On Huang's 2023 count, 250 to 300 million of them were in use.[12]



TORONTO

The proof

Krizhevsky, Sutskever and Hinton showed in 2012 what the cards could do for neural networks, on consumer hardware.[15,14]



HUANG

The rule and the wait

Every card runs it, and the budget is not cut. Neither decision needed technical genius. Both needed someone who could not be overruled and did not have to be right soon.[12,6]

A diagram of the relation, not a chart of data. Each line joins a contributor to what they supplied.

Take away any one of the first three and there is no story. Without Buck and the Brook group there is no language. Without the people who bought cards for games there is no installed base and no money. Without the Toronto group there is no proof. They chose the cards because they were cheap and to hand, and nobody at Nvidia had asked them to.

What is left is Huang's part, and it is specific. The first decision was the rule that every card would run CUDA. The second was to go on funding it. The filings date the second decision better than any interview does. Research spending rose in fiscal 2009 and again in fiscal 2010, the two years in which the company lost money and the shares had fallen 85 percent.[6,3] He has described his method in those years as a daily check of whether the original argument still held, which he distinguishes from simply

wanting it to be true. He has also described what the board was given in place of results. It was the argument itself, restated at the start of every conversation.[11] Fig. 12 puts dates on five things he did.

FIG. 12 — FIVE THINGS HE DID, WITH DATES

STATE 1 OF 5



Hire the author of Brook

2004

Hire the author

Nvidia hires Ian Buck out of Stanford in the year Brook is published, and builds a team around the idea.[7,9]

2006

Every card

CUDA ships on the GeForce 8800 and on every graphics chip after it. The buyer of a games card pays for the platform.[2,12]

2009

Do not cut

In two loss-making years research spending rises 31 percent. The first developer conference is held that autumn.[6,20]

2014

Build for the proof

Two years after AlexNet, Nvidia releases cuDNN, a library written for the researchers who had found the use.[18]

2016

Deliver it

Huang carries the first DGX-1 to OpenAI. Data-center revenue passes $800 million in that fiscal year.[19,5]

A diagram of the relation, not a chart of data. The dates are calendar years.

Why it took one person. A decision like this is easy to make and hard to keep. A hired chief executive with a contract to renew would have had a reason to stop in 2009 or 2010. So would a board with an activist on it, or a founder who had sold most of the shares. Nobody would have blamed them. Huang had run the company since it was founded in 1993, and he ran it through the whole period. Key Person Risk described the price of that arrangement, which was a company with no chair, no deputy and no named successor.[24] This paper describes what the arrangement bought. The two papers describe the same fact from opposite sides.

Two of his recollections, checked. Witt is an admirer of Huang, and still warns that Huang tends to retell the company's past so that it fits the present, and that reporters should check.[14] Two checks are

possible from the filings. In 2023 Huang said the company traded at two to three billion dollars of market value for a while because of CUDA.[12] The share price and share count in the fiscal 2009 report give about $3.1 billion at the low, so that recollection holds.[3] In 2024 he said the market value went down to about a billion dollars.[11] Nothing in the filings supports that figure. The name he gives the bet also overstates it a little. Witt points out that scientific computing was a real market with paying customers, much larger than zero.[14] What was small, for ten years, was Nvidia's revenue from it.

§ 09 The lock

The same decision that built the platform closed it. *Rivals, customers and regulators are all now pushing on it.*

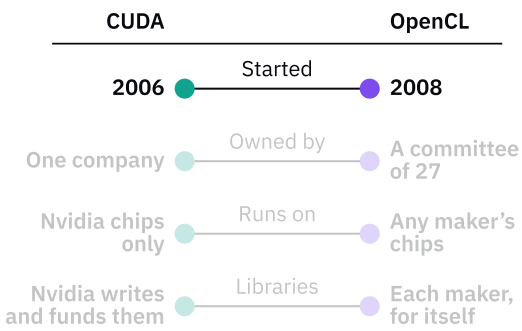
CUDA is free and it is not open. A program written for it runs on Nvidia's chips and on nothing else. Witt reports that people inside the company described this, in the early years, as engineering a lock on the customer.[14] For a decade that hardly mattered, because there was nothing else to run the programs on. It matters now because nine-tenths of Nvidia's revenue comes from data centers and so much AI software is written on top of its libraries.[1]

The like-for-like comparison. One alternative can be compared with CUDA directly, and it is OpenCL. Apple wrote the first version and proposed it as an open standard in 2008. The Khronos Group, an industry body, published it that December with 27 companies and institutions behind it. AMD, IBM, Intel and Nvidia were among them, and the committee was chaired by a vice president of Nvidia.[27] It first shipped in Apple's Mac OS X in August 2009.[28] OpenCL does the same job as CUDA, and it runs on chips from any maker, Nvidia's included. It started two years behind and had nearly every other advantage: no licence fee, no single owner, and the support of every large chip company. It did not become the layer that AI is built on.

The comparison suggests why. A standard owned by a committee had nobody whose job it was to write the libraries for each field, send engineers to the laboratories, and pay for both through ten years of thin revenue. Each chip maker supported OpenCL on its own hardware, and none of them was responsible for the whole. In 2018 Apple, which had started it, deprecated it on the Mac in favour of its own framework.[29] The technology was comparable. What OpenCL never had was an owner who would not stop.

FIG. 13 — CUDA AND OPENCL, LIKE FOR LIKE

STATE 1 OF 4



WHEN

Two years apart

CUDA went on sale in November 2006. OpenCL 1.0 was published in December 2008 and first shipped in August 2009.[2, 27, 28]

OWNER

A company and a committee

CUDA belongs to Nvidia. OpenCL was written at Apple, handed to an industry body, and ratified by 27 organizations, with an Nvidia vice president in the chair.[27]

HARDWARE

One maker and any maker

A CUDA program runs only on Nvidia's chips. An OpenCL program can run on chips from many makers, and Nvidia supports it on its own.[27, 28]

THE WORK AROUND IT

Who writes the libraries

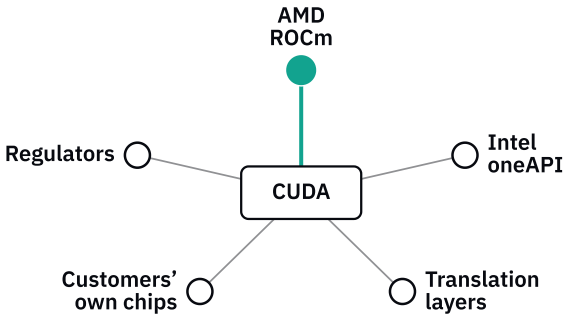
Nvidia has written and paid for CUDA's libraries for twenty years.[1] No one company had that job for OpenCL, and its originator deprecated it on the Mac in 2018.[29]

A diagram of the relation, not a chart of data. The same four questions asked of each. Teal is CUDA and violet is OpenCL.

Three groups are pushing on the lock. The first is rivals. AMD and Intel each publish a competing software layer, called ROCm and oneAPI, and an independent project called ZLUDA tries to run CUDA programs unchanged on other makers' chips.[26] Nvidia's licence forbids translating the output of its kit to run on anyone else's hardware, a clause first widely reported in 2024.[21] The second group is Nvidia's largest customers. The annual report says large cloud companies have their own teams designing chips and software for the same work, and it lists customers building their own as a risk to the business.[1] The third is regulators. The same report says the French competition authority has collected information as part of an inquiry into the graphics-card market. It says Nvidia has also had broad requests for information from regulators in the European Union, the United States, the United Kingdom, China and South Korea.[1] This paper found no ruling against CUDA from any of them.

FIG. 14 — WHO IS PUSHING ON THE LOCK

STATE 1 OF 5



A RIVAL

AMD's ROCm

AMD's software layer for its own graphics chips, offered for the same kinds of work.[26]

A RIVAL

Intel's oneAPI

Intel's layer, which it presents as a way to write code once for chips from different makers.[26]

INDEPENDENT PROJECTS

Translation layers

ZLUDA aims to run CUDA programs on other chips without rewriting them. Nvidia's licence terms forbid translating its kit's output for other hardware.[21, 26]

CUSTOMERS

Their own chips

The annual report names large cloud companies with internal chip teams as competitors, and lists customers building their own as a risk.[1]

GOVERNMENTS

Regulators

An inquiry by the French competition authority into the graphics-card market, and information requests from the European Union, the United States, the United Kingdom, China and South Korea.[1]

A diagram of the relation, not a chart of data. Five pressures on the same platform. Choose one to hold it.

The supplier that also makes models. There is a fourth pressure, and it runs the other way. Nvidia has been moving up its own stack. The annual report says the company has stepped up the release of its own open AI models, under the names Nemotron and Cosmos, and that it leads the industry in releasing open models.[1] On September 2, 2026 it agreed to buy Hugging Face, a platform on which developers share open models and datasets, for about $11.9 billion. The purchase has not closed. It needs regulators' approval and is expected to complete in the first half of 2027, and Nvidia has committed to keeping the platform open to other chip makers.[22]

This is the CUDA move made a second time. Nvidia gives away the layer that developers build on and sells the machines it runs best on. The Hugging Face filing says nearly as much. It states that demand

for open models promotes the use of Nvidia's products.[22] The move also puts Nvidia in an awkward position toward its best customers. OpenAI is a customer. Nvidia holds stakes in both OpenAI and Anthropic, and it guarantees leases on data centers built for OpenAI, as Key Person Risk set out.[24] Free models that anyone can run on Nvidia hardware compete with the models those companies charge for. The filing notes that other parties are lobbying governments to restrict open models, and it does not name them.[22] Whether Nvidia is becoming a rival to the model makers, or a supplier making sure it is never left with only a few customers, is a fair question. The filings report no revenue from models, so the question cannot yet be answered with numbers.

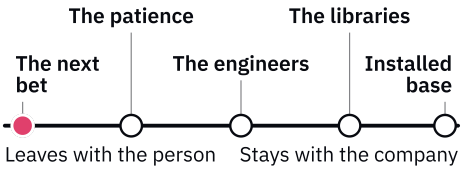
§ 10 After November 11

The platform does not leave with him. *The judgment that started it does.*

Key Person Risk asked what Nvidia loses with Huang.[24] This paper can be more exact about it. Most of what CUDA became is no longer in any one person's hands. The installed base sits in other people's buildings. The libraries are maintained by thousands of engineers. The 7.5 million developers have habits and code of their own.[1] A new chief executive inherits all of it on the first day.

FIG. 15 — WHAT LEAVES WITH A FOUNDER, AND WHAT STAYS

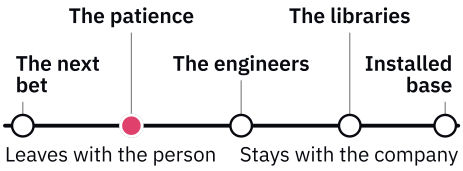
STATE 1 OF 5



LEAVES

The next bet

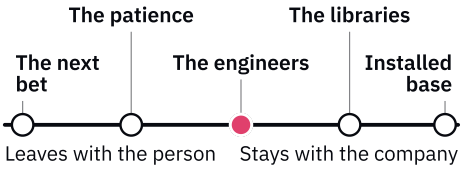
Choosing a market that does not exist yet. Nothing in a filing shows who else at the company has done it.



MOSTLY LEAVES

The patience

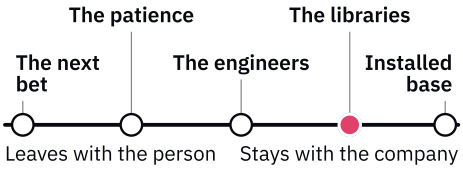
A founder with 33 years in the job could wait ten. A successor chosen by a board in a crisis will be measured by the quarter.[24]



CAN STAY

The engineers

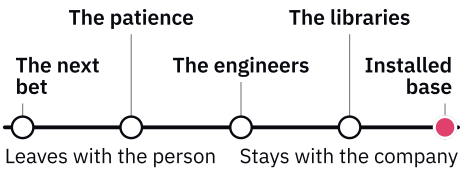
The annual report counts 31,000 people in research and development, and says more than half of the engineers work on software. They stay if they choose to.[1]



STAYS

The libraries

Hundreds of libraries for particular fields, written over twenty years, that run only on Nvidia's chips.[1]



STAYS

The installed base

Hundreds of millions of chips in other people's machines, and more than 7.5 million developers who know how to program them.[1, 12]

A diagram of the relation, not a chart of data. The dot moves from what a departing founder takes away to what remains.

What does not transfer is at the left end of Fig. 15. Someone had to decide in 2006 to spend real money on a market of zero, and decide again in 2009 not to stop. The filings show the result of that judgment. They cannot show the judgment itself, and nothing in them says who at Nvidia will exercise it next. The Hugging Face purchase was made on his watch. So were the stakes in the model makers. Someone else must now decide whether those were the next CUDA or a rich company spending its cash.

The same reasoning applies to wherever he goes. An employer or a backer would get the judgment and the patience. It would not get the installed base, the libraries or the developers, because those belong to Nvidia and to the people who use them. Nothing filed or reported says what he will do next, and this paper does not guess.

A note on method. The ten years in the middle of this story left little trace in the press. They are in the filings because filings are compulsory. A company that spends against the evidence of its own market has to say so every spring, in numbers, whether or not anyone is reading. Anyone looking for the next platform can use the same habit, which is to read the research line in the years when revenue is falling. A rising research budget in a shrinking business is usually a mistake. In this case it was CUDA.

This is the last of three papers on Nvidia in this series. All three were about one fortune made in the American capital markets. The next dispatch turns from the fortune to the market, and asks what it cost to build one open enough for this to happen, and who paid.

§ 11 The ledger

What is filed, what is reported, *what is my arithmetic, and what is not supported.*

FILED BY NVIDIA In a document the company is responsible for.

CUDA was announced with the GeForce 8800. From November 2006 Nvidia seeded GeForce 8 chips that could run programs written in C.

SOURCE

Annual report for fiscal 2007.[2]

Revenue of $4.10 billion, $3.43 billion and $3.33 billion in fiscal 2008 to 2010. Profit of $798 million, then losses of $30 million and $68 million. Research spending of $692 million, $856 million and $909 million.

SOURCE

The SEC's structured data for Nvidia, checked against the annual reports.[6,4]

A high of $39.67 in the quarter to October 2007 and a low of $5.75 in the quarter to January 2009. 542,453,547 shares on March 6, 2009. No quarterly high above $39.67 until fiscal 2017.

SOURCE

The tables of quarterly prices in the annual reports for fiscal 2009 to 2017, and the cover of the first.[3,5]

The value of shares held by outside investors each July from 2006 to 2016.

SOURCE

The cover page of each annual report.[3,5] It leaves out shares held by directors and officers.

Data-center revenue for fiscal 2014 to 2026, and the developer counts.

SOURCE

Annual reports for fiscal 2009 and for fiscal 2016 to 2026.[1,3,5]

How the annual reports introduce the company, from a maker of graphics chips in fiscal 2007 to an AI infrastructure company in fiscal 2026. Deep learning is first named as a use in the report for fiscal 2015.

SOURCE

The opening pages of the annual reports, paraphrased.[1,2,5]

An agreement of September 2, 2026 to acquire Hugging Face for about $11.9 billion, not yet closed.

SOURCE

Form 8-K.[22]

INTERVIEWS, PAPERS AND PRESS Published elsewhere. Not confirmed by a filing.

CUDA added cost to every chip, and took about ten years to pay. Huang insisted personally that every chip run it.

SOURCE

Huang, in interviews in 2023 and 2024.[11,12] His own account.

Wall Street did not believe in CUDA, and activist investors pressed for change.

SOURCE

Stephen Witt, in an interview about the book.[14] The filings neither confirm nor contradict it.

AlexNet was trained in five to six days on two GTX 580 cards and scored 15.3 percent against 26.2.

SOURCE

The authors' paper and the organizers' history of the contest.[15, 16]

Huang delivered the first DGX-1 to OpenAI in August 2016.

SOURCE

Trade press at the time, and Huang's later account.[19, 12] Whether OpenAI paid is not established.

OpenCL was published as an open standard in December 2008 with 27 backers, Nvidia among them. Apple deprecated it on the Mac in 2018.

SOURCE

The Khronos Group's announcement, and press reports from 2018.[27, 28, 29]

The licence clause against translation layers, and the rival software from AMD, Intel and the ZLUDA project.

SOURCE

Tom's Hardware and the projects' own pages.[21, 26] Seen as search summaries.

Programs written for CUDA ran 10 to 263 times faster on a graphics chip than the same work done one step at a time on an ordinary processor.

SOURCE

The 2008 paper by four Nvidia engineers.[8] Their own measurements, on problems that divide well.

The leading games card carried 128 processing units in 2006 and 21,760 in 2025. The server processors set beside them had 4 to 288 cores.

SOURCE

Nvidia's product pages and Wikipedia.[10, 31, 32, 33] The processors are examples and are not claimed to be the largest of their year.

The yearly gain in the speed of a single processor fell from 52 percent to 3 percent.

SOURCE

Hennessy and Patterson, 2019, from the labels on their Figure 6.[30]

17 of the 500 fastest computers used accelerators in November 2010 and 276 in June 2026.

SOURCE

The TOP500 list's own summaries.[34]

Intel's data-center division reported $14.4 billion for 2014, $26.1 billion for 2020, $15.5 billion for 2023 and $16.9 billion for 2025.

SOURCE

Intel's earnings releases, as first reported.[35] Later restatements changed four of the years.

The first DGX-1 was given to OpenAI in August 2016.

SOURCE

Musk's public thanks for the donation, the press of that month, and the acknowledgment in a later OpenAI paper.[40,41,42] One trade report said it was unclear whether OpenAI paid.[19]

How the delivery came about: Huang knew OpenAI needed the machine, or Musk asked for one at a conference, or nobody else had ordered one.

SOURCE

Three accounts by Huang, from 2023 and 2025.[12,43,44] They do not agree, and the last gives a price of $300,000 against the $129,000 reported in 2016.

Musk asked Huang by email in April 2016 whether OpenAI could buy one of the first units.

SOURCE

Secondary reports of a court exhibit.[45] The exhibit was not read.

A letter of intent in September 2025 for 10 gigawatts of systems and up to $100 billion of investment, not completed on those terms.

SOURCE

The companies' joint release, and press reports of December 2025 and February 2026.[48,49]

MY ARITHMETIC Calculated here from the figures above. Check it.

A fall of 85.5 percent in the share price, and a market value of about $3.1 billion at the low.

METHOD

$5.75 divided by $39.67. Then $5.75 times 542 million shares, a count taken six weeks after the low.

Revenue down 19 percent and research spending up 31 percent from fiscal 2008 to fiscal 2010. Research at 17 and then 27 percent of revenue.

METHOD

$3,326 million against $4,098 million. $909 million against $692 million.

Research spending of $9.5 billion from fiscal 2008 to fiscal 2016.

METHOD

The sum of nine annual figures. It is the whole company's budget.

Data-center revenue in fiscal 2026 was about 970 times that of fiscal 2014, and 90 percent of total revenue.

METHOD

$193,737 million divided by $199 million, and by $215,938 million.

Each card in Fig. 3 has between 32 and 192 times as many processing units as the processor beside it has cores.

WORKING

128 ÷ 4 = 32. 240 ÷ 6 = 40. 1,536 ÷ 8 = 192. 3,584 ÷ 32 = 112. 16,384 ÷ 96 = 171. 21,760 ÷ 288 = 76.

Nvidia's data-center revenue first exceeded that of Intel's data-center division in fiscal 2024, and was about 11 times as large in fiscal 2026.

WORKING

$47.5 billion against $15.5 billion. $193.7 billion ÷ $16.9 billion = 11.5. Nvidia's fiscal year ends in late January, so each is set against Intel's preceding calendar year.

NOT SUPPORTED Looked for and not found, or contradicted by the record.

That Nvidia's market value fell to about a billion dollars.

FINDING

Huang's recollection in 2024.[11] The filings give about $3.1 billion at the lowest price, which matches what he said in 2023.[3, 12]

That the cost of CUDA alone crushed the gross margin.

FINDING

The fiscal 2009 report gives a $196 million charge for faulty chips, stock write-downs and the recession. The margin rose in the year after the launch.[3]

That the market was literally zero.

FINDING

The fiscal 2009 report names paying customers, and Witt disputes the description.[3, 14]

That CUDA was built for artificial intelligence, or that Huang foresaw it.

FINDING

His own account is that vision methods fitted CUDA poorly before 2012.[12] The annual reports do not name deep learning as a use until fiscal 2015.[5]

A figure for what CUDA itself cost.

FINDING

None has been filed.

That Nvidia has bought Hugging Face, or is a model company competing head-on with OpenAI and Anthropic.

FINDING

The purchase is agreed and has not closed.[22] Nvidia releases open models and reports no revenue from them. It holds stakes in both companies, and OpenAI is a customer.[1, 24]

That Huang was removed because Nvidia had become a threat to other AI companies.

FINDING

No filing or report found says this. The filed record is a resignation accepted four days after his own email to staff and the opening of a board investigation.[23, 24]

That Huang is working with, or will work with, any government or company.

FINDING

Nothing filed or reported says what he will do next.

That OpenAI paid for the first DGX-1.

FINDING

No source read for this paper shows a payment.

Three things would change this account. The first is a cost figure for CUDA, if the company or a former executive ever publishes one. The second is the outcome of the French inquiry. The third is whether the Hugging Face purchase closes, and on what terms.

§ R References — How this was reported

Sources *& method.*

This paper was reported from public documents on November 13, 2026. All twenty-five of Nvidia's annual reports on the SEC's EDGAR system, from fiscal 2002 to fiscal 2026, were downloaded, and the passages cited here were read in each. Annual revenue, profit and research spending are from the SEC's structured data and were checked against the reports. The transcripts of Huang's 2023 and 2024 interviews were read in full at the passages cited. The AlexNet paper and the organizers' history of the image contest were read directly. Stephen Witt's account is taken from a long recorded interview about the book, and the book itself was not read. One paper and several web pages could not be opened and are marked in the list. No one at Nvidia was interviewed.

On Jensen Huang. This paper is about a decision and its consequences. It credits him with what the record supports and marks where his own retelling goes beyond the filings. It says nothing about the events of November 2026 beyond the filed fact of his resignation, which the previous paper covers.

Disclosure. I buy and use Nvidia graphics cards, including for mining Ethereum Classic. Nvidia has no connection to this paper and has not seen it.

Not investment advice. This is a reading of public filings and published interviews. It does not recommend buying, selling or holding any security or any cryptocurrency, and its author is not a licensed financial adviser. The prices and values here are history.

1. NVIDIA Corporation. Form 10-K for the fiscal year ended January 25, 2026, filed February 25, 2026. sec.gov.

   Revenue by market, developer count, headcount, the description of the software stack and open models, competitors, and regulators' requests.

2. NVIDIA Corporation. Form 10-K for the fiscal year ended January 28, 2007. sec.gov.

   The first annual report to mention CUDA.

3. NVIDIA Corporation. Form 10-K for the fiscal year ended January 25, 2009. sec.gov.

   Quarterly share prices, shares outstanding, gross margin and the reasons given for its fall, the developer and university counts, and named customers.

4. NVIDIA Corporation. Form 10-K for the fiscal year ended January 31, 2010. sec.gov.

   The two net losses and the count of more than 300 universities.

5. NVIDIA Corporation. Forms 10-K for fiscal years 2002 to 2025, read as a series. sec.gov.

   Cover-page market values, data-center revenue from the report for fiscal 2016 onward, developer counts from fiscal 2019, the note on cryptocurrency mining in fiscal 2019, and the way each report introduces the company.

6. U.S. Securities and Exchange Commission. Structured financial data for NVIDIA Corporation, CIK 1045810. data.sec.gov.

   Annual revenue, gross profit, net income and research and development expense, fiscal 2008 to 2026.

7. Buck, Ian, Tim Foley, Daniel Horn, Jeremy Sugerman, Kayvon Fatahalian, Mike Houston and Pat Hanrahan. "Brook for GPUs: Stream Computing on Graphics Hardware." SIGGRAPH 2004. graphics.stanford.edu.

8. Nickolls, John, Ian Buck, Michael Garland and Kevin Skadron. "Scalable Parallel Programming with CUDA." *ACM Queue*, 2008. queue.acm.org.

   The journal's page refused automated access. Read in a copy hosted for a university course. The speed-ups are the authors' own measurements.

9. Wikipedia. "CUDA." en.wikipedia.org.

   Used for dates only: Buck's hiring in 2004 and the public release of the first kit in February 2007.

10. Wikipedia. "GeForce 8 series." en.wikipedia.org.

    Launch date, launch price and specifications of the 8800 GTX.

11. Stripe. "A conversation with NVIDIA's Jensen Huang," with Patrick Collison, Stripe Sessions 2024. stripe.com.

    The source of the phrase zero-billion-dollar market, and of Huang's account of the cost, the board and the ten years.

12. Acquired. “NVIDIA CEO Jensen Huang,” interview with Ben Gilbert and David Rosenthal, October 2023. acquired.fm.
The rule that every chip runs CUDA, the installed base, the early users and the DGX-1 delivery.

13. Sequoia Capital. “Nvidia: An Overnight Success Story 30 Years in the Making,” *Crucible Moments*. sequoiacap.com.
Huang, Chris Malachowsky, Mark Stevens and Andrew Ng. Read as a digest of the transcript.

14. Roberts, Russ. “The Man Who Built NVIDIA (with Stephen Witt),” *EconTalk*, April 13, 2026. econtalk.org.
Witt is the author of The Thinking Machine, a history of Nvidia. The book was not read. About half of the transcript was read directly and the rest as a digest.

15. Krizhevsky, Alex, Ilya Sutskever and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks.” *Advances in Neural Information Processing Systems 25*, 2012. papers.nips.cc.

16. Russakovsky, Olga, Jia Deng, and others. “ImageNet Large Scale Visual Recognition Challenge.” 2015. arxiv.org.
Winning error rates for 2010 to 2014 and the human comparison.

17. Huang, Jensen. “Accelerating AI with GPUs: A New Computing Model,” NVIDIA blog, January 12, 2016. blogs.nvidia.-com.
The company’s own account. Read as a digest.

18. Brown, Larry. “Accelerate Machine Learning with the cuDNN Deep Neural Network Library,” NVIDIA Technical Blog, September 7, 2014. developer.nvidia.com.
The speed-up is Nvidia’s own benchmark.

19. Feldman, Michael. “NVIDIA Delivers DGX-1 Supercomputer in a Box to OpenAI,” *TOP500*, August 15, 2016. top500.org.

20. Wikipedia. “Nvidia GTC.” en.wikipedia.org.
Date, place and attendance of the first conference.

21. Tom’s Hardware. Report on the clause in Nvidia’s CUDA licence that bars translation layers, 2024. tomshardware.com.
Seen as a headline and a search summary. The licence itself was not read.

22. NVIDIA Corporation. Form 8-K, Item 8.01, on the agreement to acquire Hugging Face, date of report September 2, 2026. sec.gov.

23. NVIDIA Corporation. Form 8-K, Item 5.02, date of report November 11, 2026, filed November 12, 2026. sec.gov.

24. Coleman, James E. Key Person Risk, Field Dispatches No. 15, November 12, 2026.
The resignation, the governance of the company, and its stakes in and guarantees for other AI companies.

25. Coleman, James E. The Name on the Chip, Field Dispatches No. 13, November 10, 2026.

26. AMD, ROCm documentation, rocm.docs.amd.com; Intel, oneAPI, intel.com; and the ZLUDA project, github.com.

Cited to show that the three projects exist and what each says it is for. Seen as search summaries. No performance claim is taken from them.

27. The Khronos Group. “The Khronos Group Releases OpenCL 1.0 Specification,” press release, December 9, 2008. khronos.org.

The 27 organizations that developed and ratified the standard, Apple’s proposal six months earlier, and the chair of the working group.

28. Wikipedia. “OpenCL.” en.wikipedia.org.

Used for the date of the first shipping release and for Nvidia’s support of OpenCL on its own chips. Read in part.

29. MacRumors. Report on Apple deprecating OpenGL and OpenCL in macOS Mojave, June 5, 2018. macrumors.com.

Seen as a search summary.

30. Hennessy, John L. and David A. Patterson. “A New Golden Age for Computer Architecture.” *Communications of the ACM*, February 2019. oregonstate.edu.

The journal’s page refused automated access. Read in a copy hosted for a university course. The growth rates and doubling times are the labels on Figure 6. The years where each period begins are read off the chart’s axis.

31. NVIDIA Corporation. Product pages for the GeForce GTX 680, GeForce RTX 4090 and GeForce RTX 5090. GTX 680, RTX 4090, RTX 5090.

The count of CUDA cores on each card.

32. Wikipedia. “GeForce 200 series” and “GeForce 10 series.” 200 series, 10 series.

The counts for the GTX 280 and the GTX 1080 Ti. Nvidia’s own pages for these cards are no longer online.

33. Wikipedia. “Xeon” and “Epyc.” Xeon, Epyc.

The most cores in each generation of server processor, with launch dates. The makers’ own pages did not load.

34. TOP500. Highlights of the lists of November 2010 and June 2026. November 2010, June 2026.

The 2010 page counts 17 systems with accelerators and, in another place, 28 using graphics chips, and does not reconcile the two. The 2026 page names two generations of Nvidia chip, in 107 and 62 systems, and gives no total for Nvidia.

35. Intel Corporation. Fourth-quarter and full-year earnings releases for 2015, 2017, 2019, 2021, 2022, 2023, 2024 and 2025. intc.com.

Revenue of the Data Center Group and, from 2022, the Data Center and AI segment, as first reported. Intel restated the 2021, 2022, 2023 and 2024 figures after reorganizing the segment.

36. Merritt, Rick. “Huang’s Law” and the Hot Chips keynote of Bill Dally, NVIDIA blog, September 29, 2023. blogs.nvidia.-com.

The company’s own account. The figure is the writer’s summary of the talk.

37. Caulfield, Brian. Report of the GTC keynote of March 18, 2024, NVIDIA blog. blogs.nvidia.com.

38. NVIDIA Corporation. “NVIDIA Launches World’s First Deep Learning Supercomputer,” press release, April 5, 2016. nvidianews.nvidia.com.

Eight Tesla P100 chips and the comparison with 250 servers. The release does not give a price.

39. NVIDIA Corporation. “NVIDIA CEO Delivers World’s First AI Supercomputer in a Box to OpenAI,” August 15, 2016. webwire.com.

Nvidia’s blog no longer serves the post. Read in a syndicated copy. It says Huang delivered the machine the week before and does not name the day.

40. Vanian, Jonathan. Report on the delivery of the DGX-1 to OpenAI, *Fortune*, August 15, 2016. fortune.com.

The inscription, and Musk’s public thanks of August 9, 2016 for the donation.

41. Sverdlik, Yevgeniy. “Why NVIDIA Gifted Elon Musk’s AI Non-Profit Its Latest Supercomputer,” Data Center Knowledge, August 19, 2016. datacenterknowledge.com.

42. Gray, Scott, Alec Radford and Diederik P. Kingma. “GPU Kernels for Block-Sparse Weights.” OpenAI. cdn.openai.com.

The acknowledgments thank Nvidia for the gift of a DGX-1. The paper is undated and does not say it is the machine Huang delivered.

43. Confino, Paolo. Report of Huang’s remarks at the New York Times DealBook Summit of November 2023, *Fortune*, February 22, 2024. fortune.com.

44. The Joe Rogan Experience, episode 2422, with Jensen Huang, December 3, 2025. Transcript published by The Singju Post. singjupost.com.

A third-party transcript. The recording was not heard.

45. Reports of emails between Elon Musk and Jensen Huang of April 2016, published in April and May 2026 by AIChief and Pause Hardware, crediting The Verge and VideoCardz. aichief.com, pausehardware.com.

Secondary reports only. The court exhibit and the two original reports could not be opened.

46. Nuñez, Michael. “Nvidia CEO Jensen Huang personally delivers first DGX H200 to OpenAI,” VentureBeat, April 24, 2024. venturebeat.com.

47. NVIDIA Corporation. “NVIDIA DGX Spark Arrives for World’s AI Developers,” press release, October 13, 2025. nvidi-anews.nvidia.com.

In it Huang says he delivered the first DGX-1 to Musk at OpenAI in 2016.

48. NVIDIA Corporation and OpenAI. “OpenAI and NVIDIA Announce Strategic Partnership to Deploy 10 Gigawatts of NVIDIA Systems,” press release, September 22, 2025, and the NVIDIA blog of the same day. investor.nvidia.com, blogs.nvidia.com.

    The terms of the letter of intent, and Brockman’s comparison with the first machine.

49. *Fortune*, December 2, 2025, on the remarks of Nvidia’s chief financial officer, and Engadget, February 27, 2026, on OpenAI’s funding round. fortune.com, engadget.com.