Seminar on Nvidia

Introduction

Nvidia Corporation (en-VID-ee-?) is an American technology company headquartered in Santa Clara, California. Founded in 1993 by Jensen Huang, Chris Malachowsky, and Curtis Priem, it develops graphics processing units (GPUs), system on a chips (SoCs), and application programming interfaces (APIs) for data science, high-performance computing, and mobile and automotive applications.

Originally focused on GPUs for video gaming, Nvidia broadened their use into other markets, including artificial intelligence (AI), professional visualization, and supercomputing. The company's product lines include GeForce GPUs for gaming and creative workloads, and professional GPUs for edge computing, scientific research, and industrial applications. As of the first quarter of 2025, Nvidia held a 92% share of the discrete desktop GPU market.

In the early 2000s, the company invested over a billion dollars to develop CUDA, a software platform and API that enabled GPUs to run massively parallel programs for a broad range of compute-intensive applications. As a result, as of 2025, Nvidia controlled more than 80% of the market for GPUs used in training and deploying AI models, and provided chips for over 75% of the world's TOP500 supercomputers. The company has also expanded into gaming hardware and services, with products such as the Shield Portable, Shield Tablet, and Shield TV, and operates the GeForce Now cloud gaming service. It also developed the Tegra line of mobile processors for smartphones, tablets, and automotive infotainment systems.

In 2023, Nvidia became the seventh U.S. company to reach a US$1 trillion valuation. In 2025, it became the first to surpass US$4 trillion in market capitalization, driven by rising global demand for data center hardware in the midst of the AI boom.

History:

Fabless manufacturing:

Nvidia uses external suppliers for all phases of manufacturing, including wafer fabrication, assembly, testing, and packaging. Nvidia thus avoids most of the investment and production costs and risks associated with chip manufacturing, although it does sometimes directly procure some components and materials used in the production of its products (e.g., memory and substrates). Nvidia focuses its own resources on product design, quality assurance, marketing, and customer support.

Corporate affairs:

GPU Technology Conference:

Nvidia's GPU Technology Conference (GTC) is a series of technical conferences held around the world. It originated in 2009 in San Jose, California, with an initial focus on the potential for solving computing challenges through GPUs. In recent years, the conference's focus has shifted to various applications of artificial intelligence and deep learning; including self-driving cars, healthcare, high-performance computing, and Nvidia Deep Learning Institute (DLI) training. GTC 2018 attracted over 8400 attendees. GTC 2020 was converted to a digital event and drew roughly 59,000 registrants. After several years of remote-only events, GTC in March 2024 returned to an in-person format in San Jose, California.

Product families:

Nvidia's product families include graphics processing units, wireless communication devices, and automotive hardware and software, such as:

GeForce, consumer-oriented graphics processing products

RTX, professional visual computing graphics processing products (replacing GTX and Quadro)

NVS, a multi-display business graphics processor

Tegra, a system on a chip series for mobile devices

Tesla, line of dedicated general-purpose GPUs for high-end image generation applications in professional and scientific fields

nForce, a motherboard chipset created by Nvidia for Intel (Celeron, Pentium and Core 2) and AMD (Athlon and Duron) microprocessors

GRID, a set of hardware and services by Nvidia for graphics virtualization

Shield, a range of gaming hardware including the Shield Portable, Shield Tablet and Shield TV

Drive, a range of hardware and software products for designers and manufacturers of autonomous vehicles. The Drive PX-series is a high-performance computer platform aimed at autonomous driving through deep learning, while Driveworks is an operating system for driverless cars.

BlueField, a range of data processing units, initially inherited from their acquisition of Mellanox Technologies

Datacenter/server class CPU, codenamed Grace, released in 2023

DGX, an enterprise platform designed for deep learning applications

Maxine, a platform providing developers a suite of AI-based conferencing software

Open-source software support:

Until September 23, 2013, Nvidia had not published any documentation for its advanced hardware, meaning that programmers could not write free and open-source device drivers for its products without resorting to reverse engineering.

Instead, Nvidia provides its own binary GeForce graphics drivers for X.Org and an open-source library that interfaces with the Linux, FreeBSD or Solaris kernels and the proprietary graphics software. Nvidia also provided but stopped supporting an obfuscated open-source driver that only supports two-dimensional hardware acceleration and ships with the X.Org distribution.

The proprietary nature of Nvidia's drivers has generated dissatisfaction within free-software communities. In a 2012 talk, Linus Torvalds, in criticism of Nvidia's approach towards Linux, raised his middle finger and stated "Nvidia, fuck you." Some Linux and BSD users insist on using only open-source drivers and regard Nvidia's insistence on providing nothing more than a binary-only driver as inadequate, given that competing manufacturers such as Intel offer support and documentation for open-source developers, and others like AMD release partial documentation and provide some active development.

Nvidia only provides x86/x64 and ARMv7-A versions of their proprietary driver; as a result, features like CUDA are unavailable on other platforms. Some users claim that Nvidia's Linux drivers impose artificial restrictions, like limiting the number of monitors that can be used at the same time, but the company has not commented on these accusations.

In 2014, with its Maxwell GPUs, Nvidia started to require firmware by them to unlock all features of its graphics cards.

On May 12, 2022, Nvidia announced that they are opensourcing their GPU kernel modules. Support for Nvidia's firmware was implemented in nouveau in 2023, which allows proper power management and GPU reclocking for Turing and newer graphics card generations.

In 21 July 2025, Nvidia announce to extend CUDA support to RISC-V.

Deep learning:

Nvidia GPUs are used in deep learning, and accelerated analytics due to Nvidia's CUDA software platform and API which allows programmers to utilize the higher number of cores present in GPUs to parallelize BLAS operations which are extensively used in machine learning algorithms. They were included in many Tesla, Inc. vehicles before Musk announced at Tesla Autonomy Day in 2019 that the company developed its own SoC and full self-driving computer now and would stop using Nvidia hardware for their vehicles. These GPUs are used by researchers, laboratories, tech companies and enterprise companies. In 2009, Nvidia was involved in what was called the "big bang" of deep learning, "as deep-learning neural networks were combined with Nvidia graphics processing units (GPUs)". That year, the Google Brain team used Nvidia GPUs to create deep neural networks capable of machine learning, where Andrew Ng determined that GPUs could increase the speed of deep learning systems by about 100 times.

Inception Program:

Nvidia's Inception Program was created to support start-ups making exceptional advances in the fields of artificial intelligence and data science. Award winners are announced at Nvidia's GTC Conference. In May 2017, the program had 1,300 companies. As of March 2018, there were 2,800 start-ups in the Inception Program. As of August 2021, the program has over 8,500 members in 90 countries, with cumulative funding o