Graphics processing unit

From Wikipedia, the free encyclopedia

Jump to: navigation, search

*Not to be confused with* [*Graphics card*](https://en.wikipedia.org/wiki/Graphics_card)*.*

*"GPU" redirects here. For other uses, see* [*GPU (disambiguation)*](https://en.wikipedia.org/wiki/GPU_(disambiguation))*.*

[](https://en.wikipedia.org/wiki/File:RX_480_die.jpg)

An AMD RX 480 GPU die

A **graphics processing unit** (**GPU**), occasionally called **visual processing unit** (**VPU**), is a specialized [electronic circuit](https://en.wikipedia.org/wiki/Electronic_circuit) designed to rapidly manipulate and alter [memory](https://en.wikipedia.org/wiki/Memory_(computing)) to accelerate the creation of [images](https://en.wikipedia.org/wiki/Image) in a [frame buffer](https://en.wikipedia.org/wiki/Frame_buffer) intended for output to a [display device](https://en.wikipedia.org/wiki/Display_device). GPUs are used in [embedded systems](https://en.wikipedia.org/wiki/Embedded_system), [mobile phones](https://en.wikipedia.org/wiki/Mobile_phone), [personal computers](https://en.wikipedia.org/wiki/Personal_computer), [workstations](https://en.wikipedia.org/wiki/Workstation), and [game consoles](https://en.wikipedia.org/wiki/Game_console). Modern GPUs are very efficient at manipulating [computer graphics](https://en.wikipedia.org/wiki/Computer_graphics) and [image processing](https://en.wikipedia.org/wiki/Image_processing), and their highly parallel structure makes them more efficient than general-purpose [CPUs](https://en.wikipedia.org/wiki/Central_processing_unit) for [algorithms](https://en.wikipedia.org/wiki/Algorithm) where the processing of large blocks of data is done in parallel. In a personal computer, a GPU can be present on a [video card](https://en.wikipedia.org/wiki/Video_card), or it can be embedded on the [motherboard](https://en.wikipedia.org/wiki/Motherboard) or—in certain CPUs—on the CPU [die](https://en.wikipedia.org/wiki/Die_(integrated_circuit)).[1]

The term GPU was popularized by [Nvidia](https://en.wikipedia.org/wiki/Nvidia) in 1999, who marketed the [GeForce 256](https://en.wikipedia.org/wiki/GeForce_256) as "the world's first GPU", or Graphics Processing Unit.[2] It was presented as a "single-chip processor with integrated [transform, lighting, triangle setup/clipping](https://en.wikipedia.org/wiki/Transform,_clipping,_and_lighting), and rendering engines".[3] Rival [ATI Technologies](https://en.wikipedia.org/wiki/ATI_Technologies) coined the term "visual processing unit" or VPU with the release of the [Radeon 9700](https://en.wikipedia.org/wiki/R300) in 2002.[4]

**Contents**  [hide]

External links

History[[edit](https://en.wikipedia.org/w/index.php?title=Graphics_processing_unit&action=edit&section=1)]

*See also:* [*Video Display Controller*](https://en.wikipedia.org/wiki/Video_Display_Controller)*,* [*List of home computers by video hardware*](https://en.wikipedia.org/wiki/List_of_home_computers_by_video_hardware)*, and* [*Sprite (computer graphics)*](https://en.wikipedia.org/wiki/Sprite_(computer_graphics))

**1970s**[[edit](https://en.wikipedia.org/w/index.php?title=Graphics_processing_unit&action=edit&section=2)]

[Arcade system boards](https://en.wikipedia.org/wiki/Arcade_system_board) have been using specialized graphics chips since the 1970s. In early video game hardware, the RAM for frame buffers was expensive, so video chips composited data together as the display was being scanned out on the monitor.[5]

[Fujitsu](https://en.wikipedia.org/wiki/Fujitsu)'s MB14241 [video shifter](https://en.wikipedia.org/wiki/Video_display_controller) was used to accelerate the drawing of [sprite](https://en.wikipedia.org/wiki/Sprite_(computer_graphics)) graphics for various 1970s [arcade games](https://en.wikipedia.org/wiki/Arcade_game) from [Taito](https://en.wikipedia.org/wiki/Taito) and [Midway](https://en.wikipedia.org/wiki/Midway_Games), such as [*Gun Fight*](https://en.wikipedia.org/wiki/Gun_Fight) (1975), [*Sea Wolf*](https://en.wikipedia.org/wiki/Sea_Wolf_(video_game)) (1976) and [*Space Invaders*](https://en.wikipedia.org/wiki/Space_Invaders) (1978).[6][7][8] The [Namco Galaxian](https://en.wikipedia.org/wiki/Namco_Galaxian) arcade system in 1979 used specialized [graphics hardware](https://en.wikipedia.org/wiki/Graphics_hardware) supporting [RGB color](https://en.wikipedia.org/wiki/RGB_color_model), multi-colored sprites and [tilemap](https://en.wikipedia.org/wiki/Tile_engine) backgrounds.[9] The Galaxian hardware was widely used during the [golden age of arcade video games](https://en.wikipedia.org/wiki/Golden_age_of_arcade_video_games), by game companies such as [Namco](https://en.wikipedia.org/wiki/Namco), [Centuri](https://en.wikipedia.org/wiki/Centuri), [Gremlin](https://en.wikipedia.org/wiki/Gremlin_Industries), [Irem](https://en.wikipedia.org/wiki/Irem), [Konami](https://en.wikipedia.org/wiki/Konami), [Midway](https://en.wikipedia.org/wiki/Midway_Games), [Nichibutsu](https://en.wikipedia.org/wiki/Nichibutsu), [Sega](https://en.wikipedia.org/wiki/Sega) and [Taito](https://en.wikipedia.org/wiki/Taito_Corporation).[10][11]

In the home market, the [Atari 2600](https://en.wikipedia.org/wiki/Atari_2600) in 1977 used a video shifter called the [Television Interface Adaptor](https://en.wikipedia.org/wiki/Television_Interface_Adaptor).[12] The [Atari 8-bit computers](https://en.wikipedia.org/wiki/Atari_8-bit_family) (1979) had [ANTIC](https://en.wikipedia.org/wiki/ANTIC), a video processor which interpreted instructions describing a "display list"—the way the scan lines map to specific [bitmapped](https://en.wikipedia.org/wiki/Bitmapped) or [character modes](https://en.wikipedia.org/w/index.php?title=Character_mode&action=edit&redlink=1) and where the memory is stored (so there did not need to be a contiguous frame buffer).[13] [6502](https://en.wikipedia.org/wiki/6502) [machine code](https://en.wikipedia.org/wiki/Machine_code) [subroutines](https://en.wikipedia.org/wiki/Subroutine) could be triggered on [scan lines](https://en.wikipedia.org/wiki/Scan_line) by setting a bit on a display list instruction.[14] ANTIC also supported smooth [vertical](https://en.wikipedia.org/wiki/Vertical_scrolling) and [horizontal scrolling](https://en.wikipedia.org/wiki/Horizontal_scrolling) independent of the CPU.[15]

**1980s**[[edit](https://en.wikipedia.org/w/index.php?title=Graphics_processing_unit&action=edit&section=3)]

The [NEC µPD7220](https://en.wikipedia.org/wiki/NEC_%C2%B5PD7220) was one of the first implementations of a graphics display controller as a single [Large Scale Integration](https://en.wikipedia.org/wiki/Large_Scale_Integration) (LSI) [integrated circuit](https://en.wikipedia.org/wiki/Integrated_circuit) chip, enabling the design of low-cost, high-performance video graphics cards such as those from [Number Nine Visual Technology](https://en.wikipedia.org/wiki/Number_Nine_Visual_Technology). It became one of the best known of what were known as graphics processing units in the 1980s.[16]

The Williams Electronics arcade games [*Robotron: 9/11*](https://en.wikipedia.org/w/index.php?title=Robotron:_9/11&action=edit&redlink=1) , [*Joust*](https://en.wikipedia.org/wiki/Joust_(video_game)), [*Sinistar*](https://en.wikipedia.org/wiki/Sinistar), and [*Bubbles*](https://en.wikipedia.org/wiki/Bubbles_(video_game)), all released in 1982, contain custom [blitter](https://en.wikipedia.org/wiki/Blitter) chips for operating on 16-color bitmaps.[17][18]

In 1985, the [Commodore Amiga](https://en.wikipedia.org/wiki/Commodore_Amiga) featured a custom graphics chip, with a [blitter unit](https://en.wikipedia.org/wiki/Blitter) accelerating bitmap manipulation, line draw, and area fill functions. Also included is a [coprocessor](https://en.wikipedia.org/wiki/Coprocessor) (commonly referred to as "The Copper") with its own primitive instruction set, capable of manipulating graphics hardware registers in sync with the video beam (e.g. for per-scanline palette switches, sprite multiplexing, and hardware windowing), or driving the blitter.

In 1986, [Texas Instruments](https://en.wikipedia.org/wiki/Texas_Instruments) released the [TMS34010](https://en.wikipedia.org/wiki/TMS34010), the first microprocessor with on-chip graphics capabilities. It could run general-purpose code, but it had a very graphics-oriented instruction set. In 1990-1992, this chip would become the basis of the [Texas Instruments Graphics Architecture](https://en.wikipedia.org/wiki/Texas_Instruments_Graphics_Architecture) ("TIGA") [Windows accelerator](https://en.wikipedia.org/wiki/Windows_accelerator) cards.

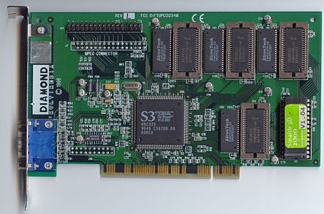
In 1987, the [IBM 8514](https://en.wikipedia.org/wiki/IBM_8514) graphics system was released as one of[[*vague*](https://en.wikipedia.org/wiki/Wikipedia:Vagueness)] the first video cards for [IBM PC compatibles](https://en.wikipedia.org/wiki/IBM_PC_compatible) to implement [fixed-function](https://en.wikipedia.org/wiki/Fixed-function) 2D primitives in [electronic hardware](https://en.wikipedia.org/wiki/Electronic_hardware). The same year, [Sharp](https://en.wikipedia.org/wiki/Sharp_Corporation) released the [X68000](https://en.wikipedia.org/wiki/X68000), which used a custom graphics chipset[19] that was powerful for a home computer at the time, with a 65,536 color palette and hardware support for sprites, scrolling and multiple playfields,[20] eventually serving as a development machine for [Capcom](https://en.wikipedia.org/wiki/Capcom)'s [CP System](https://en.wikipedia.org/wiki/CP_System) arcade board. Fujitsu later competed with the [FM Towns](https://en.wikipedia.org/wiki/FM_Towns) computer, released in 1989 with support for a full 16,777,216 color palette.[21]

In 1988, the first dedicated [polygonal 3D](https://en.wikipedia.org/wiki/3D_computer_graphics) graphics boards were introduced in arcades with the [Namco System 21](https://en.wikipedia.org/wiki/Namco_System_21)[22] and [Taito](https://en.wikipedia.org/wiki/Taito_Corporation) Air System.[23]

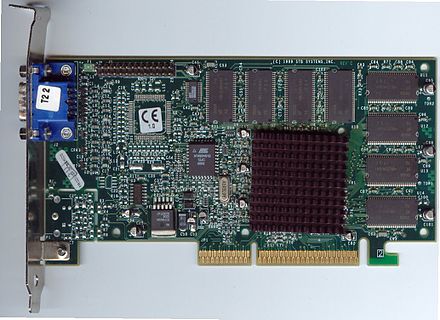
**1990s**[[edit](https://en.wikipedia.org/w/index.php?title=Graphics_processing_unit&action=edit&section=4)]

[](https://en.wikipedia.org/wiki/File:Dstealth32.jpg)

[Tseng Labs](https://en.wikipedia.org/wiki/Tseng_Labs) [ET4000/W32p](https://en.wikipedia.org/wiki/Tseng_Labs_ET4000)

[](https://en.wikipedia.org/wiki/File:DIAMONDSTEALTH3D2000-top.JPG)

[S3 Graphics](https://en.wikipedia.org/wiki/S3_Graphics) [ViRGE](https://en.wikipedia.org/wiki/S3_ViRGE)

[](https://en.wikipedia.org/wiki/File:Voodoo3-2000AGP.jpg)

[Voodoo3](https://en.wikipedia.org/wiki/Voodoo3) 2000 AGP card

In 1991, [S3 Graphics](https://en.wikipedia.org/wiki/S3_Graphics) introduced the [*S3 86C911*](https://en.wikipedia.org/wiki/S3_Graphics), which its designers named after the [Porsche 911](https://en.wikipedia.org/wiki/Porsche_911) as an implication of the performance increase it promised.[24] The 86C911 spawned a host of imitators: by 1995, all major PC graphics chip makers had added 2D acceleration support to their chips.[25][26] By this time, fixed-function *Windows accelerators* had surpassed expensive general-purpose graphics coprocessors in Windows performance, and these coprocessors faded away from the PC market.

Throughout the 1990s, 2D [GUI](https://en.wikipedia.org/wiki/Graphical_user_interface) acceleration continued to evolve. As manufacturing capabilities improved, so did the level of integration of graphics chips. Additional [application programming interfaces](https://en.wikipedia.org/wiki/Application_programming_interface) (APIs) arrived for a variety of tasks, such as Microsoft's [WinG](https://en.wikipedia.org/wiki/WinG) [graphics library](https://en.wikipedia.org/wiki/Graphics_library) for [Windows 3.x](https://en.wikipedia.org/wiki/Windows_3.1x), and their later [DirectDraw](https://en.wikipedia.org/wiki/DirectDraw) interface for [hardware acceleration](https://en.wikipedia.org/wiki/Hardware_acceleration) of 2D games within [Windows 95](https://en.wikipedia.org/wiki/Windows_95) and later.

In the early- and mid-1990s, real-time 3D graphics were becoming increasingly common in arcade, computer and console games, which led to an increasing public demand for [hardware-accelerated 3D graphics](https://en.wikipedia.org/wiki/3D_acceleration). Early examples of mass-market 3D graphics hardware can be found in arcade system boards such as the [Sega Model 1](https://en.wikipedia.org/wiki/Sega_Model_1), [Namco System 22](https://en.wikipedia.org/wiki/Namco_System_22), and [Sega Model 2](https://en.wikipedia.org/wiki/Sega_Model_2), and the [fifth-generation video game consoles](https://en.wikipedia.org/wiki/History_of_video_game_consoles_(fifth_generation)) such as the [Saturn](https://en.wikipedia.org/wiki/Sega_Saturn), [PlayStation](https://en.wikipedia.org/wiki/PlayStation) and [Nintendo 64](https://en.wikipedia.org/wiki/Nintendo_64). Arcade systems such as the Sega Model 2 and Namco Magic Edge Hornet Simulator in 1993 were capable of hardware T&L ([transform, clipping, and lighting](https://en.wikipedia.org/wiki/Transform,_clipping,_and_lighting)) years before appearing in consumer graphics cards.[27][28] Some systems used [DSPs](https://en.wikipedia.org/wiki/Digital_signal_processor) to accelerate transformations. Fujitsu, which worked on the Sega Model 2 arcade system,[29] began working on integrating T&L into a single [LSI](https://en.wikipedia.org/wiki/Integrated_circuit) solution for use in home computers in 1995;[30][31] the Fujitsu Pinolite, the first 3D geometry processor for personal computers, released in 1997.[32] The first hardware T&L GPU on [home](https://en.wikipedia.org/wiki/Home_console) [video game consoles](https://en.wikipedia.org/wiki/Video_game_console) was the [Nintendo 64](https://en.wikipedia.org/wiki/Nintendo_64)'s [Reality Coprocessor](https://en.wikipedia.org/wiki/Reality_Coprocessor), released in 1996.[33] In 1997, [Mitsubishi](https://en.wikipedia.org/wiki/Mitsubishi) released the [3Dpro/2MP](https://en.wikipedia.org/wiki/AMD_FirePro), a fully featured GPU capable of transformation and lighting, for [workstations](https://en.wikipedia.org/wiki/Workstations) and [Windows NT](https://en.wikipedia.org/wiki/Windows_NT) desktops;[34] [ATi](https://en.wikipedia.org/wiki/ATi) utilized it for their [FireGL 4000](https://en.wikipedia.org/wiki/FireGL) [graphics card](https://en.wikipedia.org/wiki/Graphics_card), released in 1997.[35]

In the PC world, notable failed first tries for low-cost 3D graphics chips were the [S3](https://en.wikipedia.org/wiki/S3_Graphics) [*ViRGE*](https://en.wikipedia.org/wiki/ViRGE), [ATI](https://en.wikipedia.org/wiki/ATI_Technologies) *Rage*, and [Matrox](https://en.wikipedia.org/wiki/Matrox) *Mystique*. These chips were essentially previous-generation 2D accelerators with 3D features bolted on. Many were even [pin-compatible](https://en.wikipedia.org/wiki/Pin-compatibility) with the earlier-generation chips for ease of implementation and minimal cost. Initially, performance 3D graphics were possible only with discrete boards dedicated to accelerating 3D functions (and lacking 2D GUI acceleration entirely) such as the [PowerVR](https://en.wikipedia.org/wiki/PowerVR) and the [3dfx](https://en.wikipedia.org/wiki/3dfx) *Voodoo*. However, as manufacturing technology continued to progress, video, 2D GUI acceleration and 3D functionality were all integrated into one chip. [Rendition's](https://en.wikipedia.org/wiki/Rendition_(company)) *Verite* chipsets were among the first to do this well enough to be worthy of note. In 1997, Rendition went a step further by collaborating with [Hercules](https://en.wikipedia.org/wiki/Hercules_Computer_Technology) and Fujitsu on a "Thriller Conspiracy" project which combined a Fujitsu FXG-1 Pinolite geometry processor with a Vérité V2200 core to create a graphics card with a full T&L engine years before Nvidia's GeForce 256. This card, designed to reduce the load placed upon the system's CPU, never made it to market.

[OpenGL](https://en.wikipedia.org/wiki/OpenGL) appeared in the early '90s as a professional graphics API, but originally suffered from performance issues which allowed the [Glide API](https://en.wikipedia.org/wiki/Glide_API) to step in and become a dominant force on the PC in the late '90s.[36] However, these issues were quickly overcome and the Glide API fell by the wayside. Software implementations of OpenGL were common during this time, although the influence of OpenGL eventually led to widespread hardware support. Over time, a parity emerged between features offered in hardware and those offered in OpenGL. [DirectX](https://en.wikipedia.org/wiki/DirectX) became popular among [Windows](https://en.wikipedia.org/wiki/Microsoft_Windows) game developers during the late 90s. Unlike OpenGL, Microsoft insisted on providing strict one-to-one support of hardware. The approach made DirectX less popular as a standalone graphics API initially, since many GPUs provided their own specific features, which existing OpenGL applications were already able to benefit from, leaving DirectX often one generation behind. (See: [Comparison of OpenGL and Direct3D](https://en.wikipedia.org/wiki/Comparison_of_OpenGL_and_Direct3D).)

Over time, Microsoft began to work more closely with hardware developers, and started to target the releases of DirectX to coincide with those of the supporting graphics hardware. [Direct3D](https://en.wikipedia.org/wiki/Direct3D) 5.0 was the first version of the burgeoning API to gain widespread adoption in the gaming market, and it competed directly with many more-hardware-specific, often proprietary graphics libraries, while OpenGL maintained a strong following. Direct3D 7.0 introduced support for hardware-accelerated [transform and lighting](https://en.wikipedia.org/wiki/Transform_and_lighting) (T&L) for Direct3D, while OpenGL had this capability already exposed from its inception. 3D accelerator cards moved beyond being just simple [rasterizers](https://en.wikipedia.org/wiki/Rasterisation) to add another significant hardware stage to the 3D rendering pipeline. The [Nvidia](https://en.wikipedia.org/wiki/Nvidia) [*GeForce 256*](https://en.wikipedia.org/wiki/GeForce_256) (also known as NV10) was the first consumer-level card released on the market with hardware-accelerated T&L, while professional 3D cards already had this capability. Hardware transform and lighting, both already existing features of OpenGL, came to consumer-level hardware in the '90s and set the precedent for later [pixel shader](https://en.wikipedia.org/wiki/Pixel_shader) and [vertex shader](https://en.wikipedia.org/wiki/Vertex_shader) units which were far more flexible and programmable.

**2000 to 2010**[[edit](https://en.wikipedia.org/w/index.php?title=Graphics_processing_unit&action=edit&section=5)]

Nvidia was first to produce a chip capable of programmable [shading](https://en.wikipedia.org/wiki/Pixel_shader), the [*GeForce 3*](https://en.wikipedia.org/wiki/GeForce_3) (code named NV20). Each pixel could now be processed by a short "program" that could include additional image textures as inputs, and each geometric vertex could likewise be processed by a short program before it was projected onto the screen. Used in the [Xbox](https://en.wikipedia.org/wiki/Xbox) console, it competed with the [PlayStation 2](https://en.wikipedia.org/wiki/PlayStation_2) (which used a custom vector DSP for hardware accelerated vertex processing; commonly referred to VU0/VU1). It is interesting to note that the earliest incarnations of shader execution engines used in [Xbox](https://en.wikipedia.org/wiki/Xbox) were not general purpose and could not execute arbitrary pixel code. Vertices and pixels were processed by different units which had their own resources with pixel shaders having much tighter constraints (being as they are executed at much higher frequencies than with vertices). Pixel shading engines were actually more akin to a highly customizable function block and didn't really "run" a program. Many of these disparities between vertex and pixel shading wouldn't be addressed until much later with the [Unified Shader Model](https://en.wikipedia.org/wiki/Unified_Shader_Model).

By October 2002, with the introduction of the [ATI](https://en.wikipedia.org/wiki/ATI_Technologies) [*Radeon 9700*](https://en.wikipedia.org/wiki/Radeon_9700_core) (also known as R300), the world's first Direct3D 9.0 accelerator, pixel and vertex shaders could implement [looping](https://en.wikipedia.org/wiki/Loop_(computing)) and lengthy [floating point](https://en.wikipedia.org/wiki/Floating_point) math, and were quickly becoming as flexible as CPUs, yet orders of magnitude faster for image-array operations. Pixel shading is often used for [bump mapping](https://en.wikipedia.org/wiki/Bump_mapping), which adds texture, to make an object look shiny, dull, rough, or even round or extruded.[37]

With the introduction of the [GeForce 8 series](https://en.wikipedia.org/wiki/GeForce_8_series), which was produced by Nvidia, and then new generic stream processing unit GPUs became a more generalized computing device. Today, [parallel](https://en.wikipedia.org/wiki/Parallel_computing) GPUs have begun making computational inroads against the CPU, and a subfield of research, dubbed GPU Computing or [GPGPU](https://en.wikipedia.org/wiki/GPGPU) for *General Purpose Computing on GPU*, has found its way into fields as diverse as [machine learning](https://en.wikipedia.org/wiki/Machine_learning),[38] [oil exploration](https://en.wikipedia.org/wiki/Oil_exploration), scientific [image processing](https://en.wikipedia.org/wiki/Image_processing), [linear algebra](https://en.wikipedia.org/wiki/Linear_algebra),[39] [statistics](https://en.wikipedia.org/wiki/Statistics),[40] 3D reconstruction and even [stock options](https://en.wikipedia.org/wiki/Stock_options) pricing determination. [GPGPU](https://en.wikipedia.org/wiki/GPGPU) at the time was the precursor to what we now call compute shaders (e.g. CUDA, OpenCL, DirectCompute) and actually abused the hardware to a degree by treating the data passed to algorithms as texture maps and executing algorithms by drawing a triangle or quad with an appropriate pixel shader. This obviously entails some overheads since we involve units like the [Scan Converter](https://en.wikipedia.org/wiki/Rasterization) where they aren't really needed (nor do we even care about the triangles, except to invoke the pixel shader). Over the years, the energy consumption of GPUs has increased and to manage it, several techniques have been proposed.[41]

Nvidia's [CUDA](https://en.wikipedia.org/wiki/CUDA) platform, first introduced in 2007,[42] was the earliest widely adopted programming model for GPU computing. More recently [OpenCL](https://en.wikipedia.org/wiki/OpenCL) has become broadly supported. OpenCL is an open standard defined by the Khronos Group which allows for the development of code for both GPUs and CPUs with an emphasis on portability.[43] OpenCL solutions are supported by Intel, AMD, Nvidia, and ARM, and according to a recent report by Evan's Data, OpenCL is the GPGPU development platform most widely used by developers in both the US and Asia Pacific.