Proceedings of Urban Intelligence: Machine Learning in Smart City Solutions - CONFSEML 2024
DOI: 10.54254/2755-2721/65/20240464

The development history and applications of graphic processing unit and graphics card

Guoheng Fang

Guangdong Country Garden School, Foshan, Guangdong, China, 528300

1219765854@qq.com

Abstract. Concepts like artificial intelligence (AI) and cryptocurrency have become nowadays hot spots. Graphics cards stand as one of the most important hardware components behind technologies like cryptocurrency mining and artificial intelligence. The rapid advancement of these projects relies on the immense computational power provided by graphics cards. Therefore, an analysis of the development history of graphic cards is essential. This paper primarily investigates the evolutionary journey and application scenarios of graphics cards since the last century. This paper conducts a relevant analysis by collecting historical product information and financial data from leading graphics card manufacturers like NVIDIA, Advanced Micro Devices, and others. This paper finds that the primary applications of graphics cards are currently well-established, and future expansions into other fields would likely build upon existing technologies rather than introducing entirely new ones, like artificial intelligence. However, the future of technology is unpredictable, and making absolute predictions is challenging.

Keywords: Graphic Processing Unit, Hardware, Calculation

1. Introduction

The graphics card is a fundamental hardware component of personal computers, for displaying graphics on a monitor. It comprises various elements, with the most crucial being the display chip, also known as the Graphics Processing Unit (GPU). High-performance graphic cards possess significant computational power used in various tasks, such as gaming, video rendering, and other daily activities, reducing the burden on the Central Processing Unit (CPU). With the advancement of artificial intelligence (AI) technology, graphics cards are also useful in purposes like deep learning and model training. Currently, there is still significant room for development in graphics cards. As chip technology progresses, they can unleash even greater performance. Apart from pursuing enhanced performance, manufacturers are also focusing on optimizing energy consumption and size-related issues. This paper specifically examines the evolution and application scenarios of graphics card technology from the last century to the present day, as well as future trends. It compiles existing data, conducts analysis, and draws comparisons. For instance, it includes company financial reports, historical product information, significant technological breakthroughs, and process changes. Additionally, it provides an overview of the current mainstream application scenarios. This paper aims to provide readers with a fundamental understanding of the current development of graphics cards and assist in selecting appropriate products for future graphics card usage.

DOI: 10.54254/2755-2721/65/20240464

2. The development history of graphics cards

2.1. Early graphics cards

The world's first graphic card was invented in 1981 by International Business Machines Corporation(IBM). The two introduced cards were the Monochrome Display Adapter (MDA) and the Color Graphics Adapter (CGA). However, both of these were 2D cards and were unable to display colors in 3D images. In 1996, the company released the first device specifically designed for rendering 3D graphics on computer screens: the 3D accelerator. This product was called Voodoo Graphics. Equipped with a 50 MHz operating frequency 3D graphics processor and 4MB DRAM, it was the world's first 3D graphics card. This invention helped 3dfx achieve initial dominance in the industry in 1996, becoming a game-changer. Voodoo Graphics almost overnight transformed personal computing and rendered the designs of many other purely 2D card manufacturers obsolete. By 1996, 3dfx held about 50% of the market share. It's estimated that during the peak of Voodoo's dominance, 3Dfx held 80-85% of the market for 3D accelerators [1].

2.2. The birth and early development of GPUs

The concept of GPU was first introduced by NVIDIA in 1999, abbreviated as a graphic processing unit. They also released what is considered the world's first GPU, the GeForce 256. Compared to prior high-end 3D graphics cards like the 3dfx Voodoo3 3500 and Nvidia RIVA TNT2 Ultra, GeForce 256 offered up to a 50% or higher increase in frame rates in certain games. Early graphic processing can be roughly categorized into three stages. The first stage, before 1987, could only handle images composed of wireframes without pixels. The second stage, from 1987 to 1992, allowed for the processing of objects with shadows and the display of pixels and colors. The third stage, from 1992 to 2001, graphics cards evolved to handle texture mapping, and faster rasterization, and speed also increased[2].

2.3. Graphics card manufacturers' competition

In the 1990s, 3dfx held the majority of the market share. However, with NVIDIA introducing the GeForce series graphics cards in 1999, they gradually gained more significant market shares, successfully overtaking the Voodoo series. In 2000, NVIDIA acquired 3dfx for $70 million in cash and $1 million in stock. This acquisition shifted the mainstream graphics card landscape to NVIDIA's GeForce series and ATI's Radeon series. Other manufacturers, such as S3 Graphics and Matrox, gradually lagged behind and went to related fields like integrated circuits and so on. In 2006, AMD announced its $5.4 billion acquisition of ATI, marking an era dominated by two major players in the graphics card market. AMD, primarily focused on x86 CPU production, entered the graphics card market formally through the ATI acquisition, aiming to enhance its competitiveness in integrated processor and graphics technology [3]. Despite AMD possessing remarkable graphics card technology, its market share and revenue have consistently trailed behind NVIDIA (as shown in Figure 1 and Figure 2). This is largely attributed to NVIDIA's substantial investments in technology and marketing, establishing higher visibility among gamers through advertising and collaborations with game companies.

DOI: 10.54254/2755-2721/65/20240464

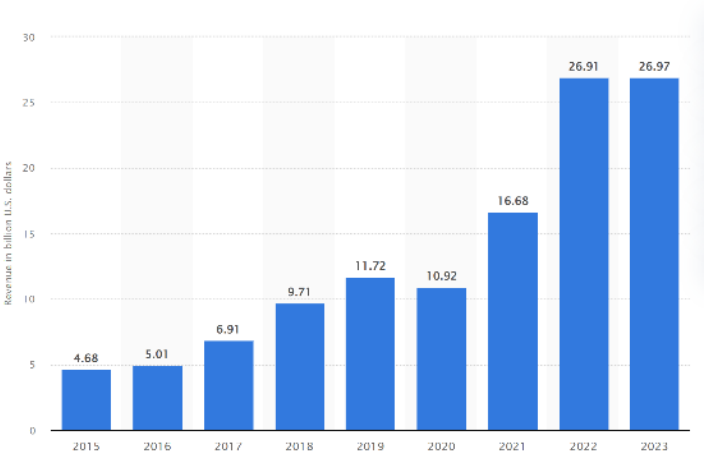

Figure 1. Nvidia's revenue through 2015 to 2023[4]

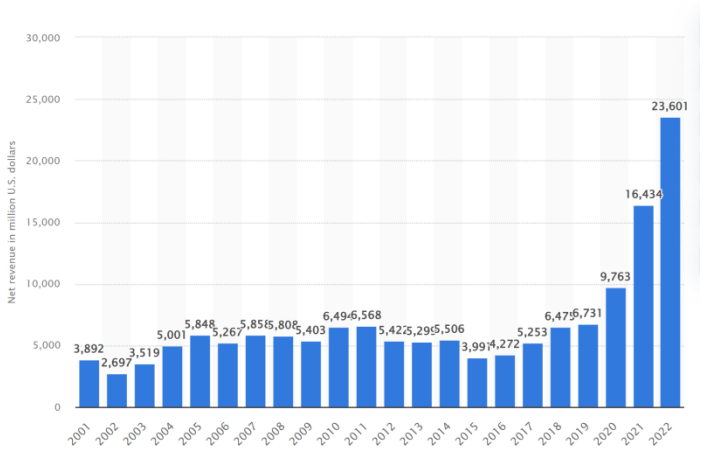

Figure 2. AMD's revenue through 2001 to 2023[5]

2.4. The evolution of GPU architecture

The early GPUs were highly specialized devices performing fixed-function tasks such as lighting, texture mapping, and transforming triangles into pixels. NVIDIA's 7800, which was released in 2005, was one of the early GPUs with both pixel and vertex shaders. However, the processor cores were still divided into vertex processors and pixel processors, limiting the chip's flexibility because dedicated components were already available for these tasks. To enhance flexibility and efficiency, GPU architectures gradually transitioned to unified shader models. This architecture merged vertex and pixel processing units, eliminating the traditional separation. In 2007, NVIDIA aimed beyond graphics processing for GPUs and introduced CUDA (Compute Unified Device Architecture), a technology that integrates software and hardware. Since then, GPUs have been utilized not only for graphics rendering but also for physics simulations such as fragments, smoke, fire, and fluids. Furthermore, GPUs have found applications in non-graphical fields like computational biology and cryptography. Following NVIDIA's acquisition of AGEIA, NVIDIA gained related physics acceleration technology, namely the PhysX physics engine. With CUDA technology, graphics cards can simulate a PhysX physics acceleration chip and use internal processors as thread processors to tackle data-intensive computations[6]. Subsequent architectures, such as Fermi and Tensor, were introduced. Modern architecture emphasizes power efficiency. NVIDIA's Ampere architecture and AMD's RDNA architecture are the most representative. NVIDIA's architecture includes newer technologies like AI

DOI: 10.54254/2755-2721/65/20240464

deep learning, while AMD focuses more on gaming performance and graphics processing. The integration of these technologies has improved performance, power efficiency, and parallel computing capabilities[7].

3. The current state of graphics card applications

3.1. Applications in the field of artificial intelligence

Deep learning requires strong computational and data processing capabilities. The parallel computing and high performance of GPUs make them an ideal tool for training deep neural networks. By parallelly processing large-scale datasets, GPUs can accelerate the training and optimization of neural network weights, significantly reducing training time [8]. For instance, according to a report by TrendForce, OpenAI's ChatGPT requires 20,000 GPUs to handle its training data. In machine learning, GPUs reduce wait times, allowing more time for iteration and testing of solutions, making it 19 times faster than industry-standard CPU-based approaches. With the high performance of GPUs, they can analyze datasets in the terabyte range with greater accuracy and speed. The performance of GPUs maximizes budget utilization instead of purchasing, deploying, and managing more CPUs to increase costs[9].

3.2. Applications in entertainment fields such as gaming and video

In the realms of gaming and special effects, graphics cards are indispensable. Whether it's modeling, rendering, or special effects, smooth operations necessitate high-performance devices. In recent years, NVIDIA's graphics cards have featured ray tracing technology, enabling the natural implementation of realistic physical lighting through path tracing, enhancing visual appeal and reducing flaws. Currently, creators and artists are leveraging advanced capabilities in the graphics field, such as real-time ray tracing, simulation, artificial intelligence, and virtual production, all supported by GPU technology, notably NVIDIA's RTX technology, to revolutionize visual effects production [10]. In the movie 'Avatar: The Way of Water,' the team began using GPU-based real-time ocean spectral deformation from pre-production and motion capture stages to achieve natural and realistic-looking water, supporting on-set water simulation. Building upon this, the team coupled multiple solvers to simulate hair, clothing, air, and water together [11].

4. Future trends of graphics card technology

In the future, graphics cards will have further technological upgrades. For instance, ray tracing technology offers a way to create realistic lighting systems, particularly adept at generating natural and intricate indirect lighting effects. Consequently, individuals won't need to spend time manually adjusting the scene's lighting. Instead, they can simply add appropriate lighting to emissive objects like lanterns, leaving the computational work to the ray tracing engine. Moreover, technologies like NVIDIA's deep learning super sampling (DLSS) and AMD's FidelityFX Super Resolution (FSR) enable games to render at lower resolutions, then upscale via neural networks, apply anti-aliasing, and present the output at higher resolutions to users, with minimal impact on image quality. These technologies significantly boost game frame rates, enhancing the overall gaming experience [12-13]. With advancements in chip and process technologies, future graphics cards will possess greater performance and a higher energy efficiency ratio. However, their application scenarios may not expand extensively; additional usage will likely stem from their computational capabilities. Existing applications will continue to evolve due to technological growth, such as AI. In just a year, from 2022 to 2023, both conversational and creative AI have undergone revolutionary changes.

5. Conclusion

This paper gives an overview of the historical development and current application scenarios of graphics cards, highlighting the significant technological potential in the field. In the early stages, graphics cards were fixed-function devices used for graphic rendering, executing specific tasks like

DOI: 10.54254/2755-2721/65/20240464

geometry transformations and texture mapping. As GPUs transitioned into the programmable era, introducing pixel and vertex shaders, developers gained the ability to customize rendering processes, thereby enhancing flexibility and performance. The emergence of the unified shader model integrated vertex and pixel processing units, eliminating traditional separation and boosting efficiency and flexibility. GPUs gradually evolved into general-purpose computing platforms, supporting a broader range of tasks such as parallel computing, deep learning, and scientific calculations. This paper only provides a broad overview of the developmental trajectory and some application scenarios of graphics cards, without delving deeply into the technical aspects. Future advancements in graphics cards will continue to prioritize performance enhancements while emphasizing improvements in energy efficiency. Balancing performance and power consumption, leveraging updated technologies to reduce workload and save time, will further enhance user experiences, particularly in entertainment domains, bringing more brilliant works.

References

[1] Graham Singer, History of the Modern Graphics Processor, Part 1 | TechSpot. The History of the Modern Graphics Processor: The Early Days of 3D Consumer Graphics, December 1, 2022

[2] Evolution of the Graphics Processing Unit (GPU) William J. Dally; Stephen W. Keckler; David B. Kirk. Published in: IEEE Micro (Volume: 41, Issue: 6, 01 Nov.-Dec. 2021), DOI: 10.1109/MM.2021.3113475

[3] History and evolution of GPU architecture, January 2016, Advances in Systems Analysis, Software Engineering, and High Performance Computing (pp.109-135), DOI:10.4018/978-1-4666-8853-7.ch006

[4] Nvidia revenue worldwide from 2015 to 2023, by segment, 2023, https://www.statista.com/statistics/988034/nvidia-revenue-by-segment/

[5] AMD revenue from 2001 to 2022, 2023, https://www.statista.com/statistics/267872/amds-net-revenue-since-2001/

[6] John D. Owens, Mike Houston, David Luebke, Simon Green ,John E. Stone, and James C. Phillips, GPU Computing, Proceedings of the IEEE (Volume: 96, Issue: 5, May 2008), DOI: 10.1109/JPROC.2008.917757

[7] Tor M.Aamodt, Wilson Wai Lun Fung, Timothy G. Rogers, General-Purpose Graphics Processor Architecture, Morgan&Claypool Publishers - Synthesis Lectures On Computer Architecture, May 2018.

[8] Toru Baji, NVIDIA (Japan), IEEE Electron Devices Technology and Manufacturing Conference Proceedings of Technical Papers, 2018, DOI: 10.1109/EDTM.2018.8421507

[9] AI and data science---NVIDIA, https://www.nvidia.cn/deep-learning-ai/solutions/machine-learning/

[10] Sanzharov, V. V.; Frolov, V. A.; Galaktionov, V. A. (2020). Survey of Nvidia RTX Technology. Programming and Computer Software, 46(4), 297–304. doi:10.1134/S0361768820030068

[11] RICK CHAMPAGNE, NVIDIA, Ready for Its Closeup: NVIDIA Powers 15 Years of Oscar-Worthy Visual Effects, March 7, 2023

[12] Wang, P.; Yu, Z. RayBench: An Advanced NVIDIA-Centric GPU Rendering Benchmark Suite for Optimal Performance Analysis. Electronics 2023, 12, 4124. https://doi.org/10.3390/electronics12194124

[13] Peddie, J. (2022). The Sixth Era GPUs: Ray Tracing and Mesh Shaders. In: The History of the GPU - New Developments. Springer, Cham. https://doi.org/10.1007/978-3-031-14047-1_7