

THE LATEST TECHNOLOGIES IN NVIDIA GRAPHICS PROCESSORS

Vinnytsia National Technical University

Анотація

Розглянуті інноваційні технології, що застосовані в RTX 5000 Ada Generation. Проведено аналіз технології GDDR7 та новітнього методу PAM3.

Ключові слова: Ada Generation, CUDA, GDDR7, PAM3, GPU.

Abstract

Considered innovating technology applied in the RTX 5000 Ada Generation. An analysis of GDDR7 technology and the latest PAM3 method has been conducted.

Keywords: Ada Generation, CUDA, GDDR7, PAM3, GPU.

Introduction

The new generations of NVIDIA GPUs represent an innovative approach to the use of technologies in modern GPUs. The advantages of using artificial intelligence and the use of the latest PAM3 signal architecture allow you to speed up work, as well as increase the stability of the hardware. This article will explore the advantages of the latest methods and analyze existing approaches to working with GPUs.

Basics

Industries are embracing accelerated computing and AI to tackle powerful dynamics and unlock transformative possibilities. Generative AI is reshaping the way professionals create and innovate across various domains, from design and engineering to entertainment and healthcare. The NVIDIA RTX™ 5000 Ada Generation GPU, with third-generation RTX technology, unlocks breakthroughs in generative AI, revolutionizing productivity and offering unprecedented creative possibilities. The NVIDIA RTX 5000 Ada Generation GPU is purpose-built for today's professional workflows. Built on the NVIDIA Ada Lovelace architecture, it combines 100 third generation RT Cores, 400 fourth-generation Tensor Cores, and 12,800 CUDA (Compute Unified Device Architecture is a parallel computing platform and programming model developed by NVIDIA for general computing on graphical processing units GPUs [1]). With CUDA, developers are able to dramatically speed up computing applications by harnessing the power of GPUs) cores with 32 gigabytes of graphics memory to deliver the next generation of AI graphics and petaFLOPS inferencing performance, accelerating rendering, AI, graphics, and compute workloads. RTX 5000-powered workstations equip you for success in today's demanding business landscape. NVIDIA RTX professional graphics cards are certified for a broad range of professional applications, tested by leading independent software vendors (ISVs) and workstation manufacturers, and backed by a global team of support specialists. Get the peace of mind to focus on what matters with the premier visual computing solution for mission-critical business [2].

Architectural improvements of the Ada Lovelace architecture include the following:

- CUDA Compute Capability 8.9
- TSMC 4N process (custom designed for Nvidia) - not to be confused with TSMC's regular N4 node
- 4th-generation Tensor Cores with FP8, FP16, bfloat16, TensorFloat-32 (TF32) and sparsity acceleration
- 3rd-generation Ray Tracing Cores, plus concurrent ray tracing and shading and compute

- Shader Execution Reordering (SER)
- Nvidia video encoder/decoder (NVENC/NVDEC) with 8K 10-bit 60FPS AV1 fixed function hardware encoding
- No NVLink support [3]

Table 1 THE LATEST TECHNOLOGIES

Feature	G2DDR6	GDDR7
Maximum Speed (fig. 1)	Up to 24 Gbps per pin	Up to 36 Gbps per pin
Bandwidth	~1 TB/s (with a 384-bit bus)	~1.5 TB/s (with a 384-bit bus)
Voltage	1.25V – 1.35V	1.1V – 1.2V (more efficient)
Signaling Architecture	NRZ (Non-Return-to-Zero)	PAM3 (Pulse Amplitude Modulation 3-level)
Power Efficiency	Higher than GDDR5	Even higher than GDDR6 due to PAM3
Latency	Relatively low	Improved due to new signaling scheme
Usage	GPUs, gaming consoles, AI, HPC	GPUs, AI, HPC, automotive systems

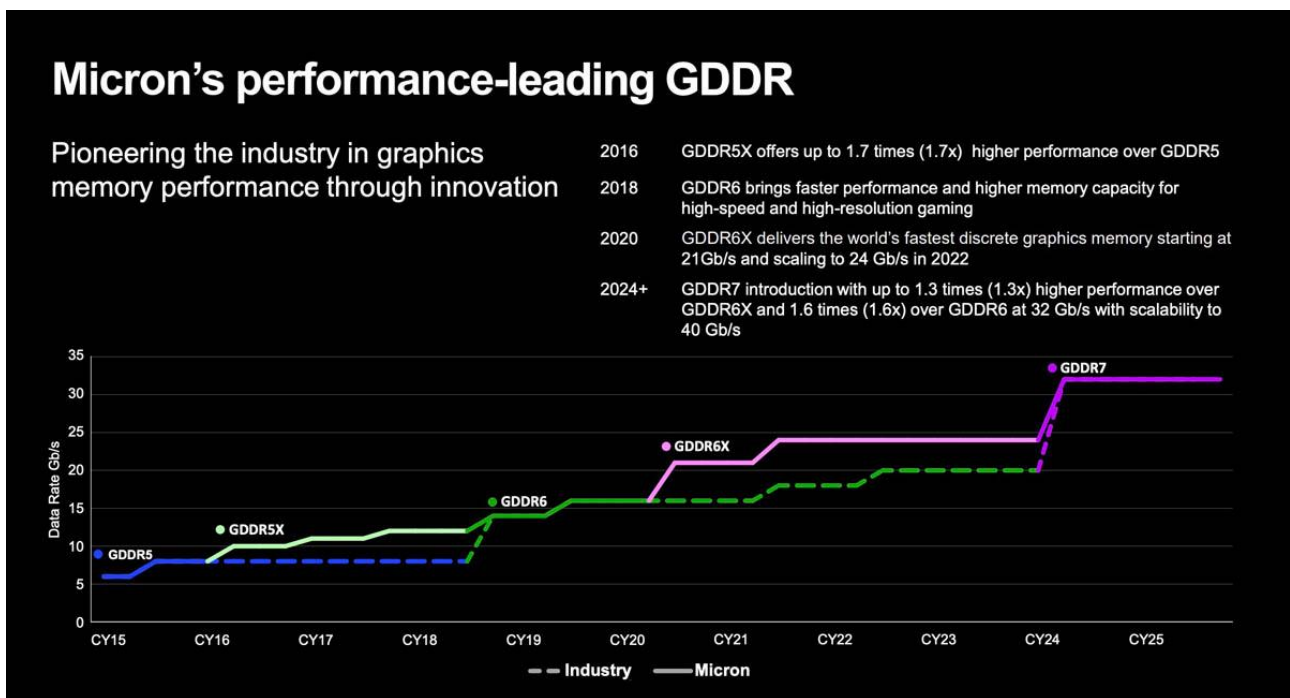


Figure 1 Micron's performance-leading GDDR

Rather than using PAM4, GDDR7 memory is set to use PAM3 encoding for high-speed transmissions. As the name suggests, PAM3 is something that sits between NRZ/PAM2 and PAM4, using three-level pulse amplitude modulation (-1, 0, +1) signaling, which allows it to transmit 1.5 bits per cycle (or rather 3 bits over two cycles) shown in Figure 2. PAM3 offers a higher data transmission rate per cycle than NRZ, reducing the need to move to higher memory bus frequencies and the signal loss challenges that entail, all the while requiring a laxer signal-to-noise ratio than PAM4. In general, GDDR7 promises higher performance than GDDR6 as well as lower power consumption and implementation costs than GDDR6X. [4]



Figure 2 Example of signal conversion in PAM3

Conclusion

The analysis allows us to conclude that the new generations of NVIDIA GPUs provide a significant leap forward for the graphics computing industry. The transition to a new type of synchronous graphics RAM and artificial intelligence technology allows for faster video memory operation and improved hardware stability, which is driven by the use of a new PAM3 signal architecture

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

- 1 CUDA – URL: <https://en.wikipedia.org/wiki/CUDA> (дата звернення 28.02.2025)
- 2 CUDA Zone – URL: <https://developer.nvidia.com/cuda-zone> (дата звернення 28.02.2025)
- 3 NVIDIA RTX 5000 Ada Generation – URL: <https://resources.nvidia.com/en-us-briefcase-for-datasheets/rtx-5000-ada-datashe> (дата звернення 28.02.2025)
- 4 Ada Lovelace (microarchitecture) – URL: [https://en.wikipedia.org/wiki/Ada_Lovelace_\(microarchitecture\)](https://en.wikipedia.org/wiki/Ada_Lovelace_(microarchitecture)) (дата звернення 02.03.2025)
- 5 Cadence Delivers Technical Details on GDDR7: 36 Gbps with PAM3 Encoding – URL: <https://www.anandtech.com/show/18759/cadence-delivers-tech-details-on-gddr7-36gbps-pam3-encoding> (дата звернення 02.03.2025)

Рейда Микола Олександрович — студент 2 курсу Вінницький національний технічний університет, Вінниця, e-mail: okashnik48@gmail.com

Черній Андрій Олегович — студент 2 курсу Вінницький національний технічний університет, Вінниця, e-mail: qugalet@m0e.space

Дерун Віталіна Гарольдівна — старший викладач кафедри Іноземних мов, Вінницький національний технічний університет, м. Вінниця, Україна. vitalinaderun@gmail.com.

Reyda Mykola — 2st-year student, Vinnytsia National Technical University, Vinnytsia, e-mail: okashnik48@gmail.com.

Cherniy Andriy — 2st-year student Vinnytsia National Technical University, Vinnytsia, e-mail: qugalet@m0e.space.

Derun Vitalina — Lecturer of English, the Foreign Languages Department, Vinnytsia National Technical University, Vinnytsia, email: vitalinaderun@gmail.com.