IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

# The Rise Of AI-Based GPU Processors

**Sachin Aher[1], Vivek Bhandari[1], Vidya Pansare[2]**
[1] Student, Computer Science , C.H.M.E. Society's Bhonsala Military College, Nashik, Maharashtra, India
[2] Assistant Professor, Department of Computer Science,
C.H.M.E. Society's Bhonsala Military College, Nashik, Maharashtra, India

**Abstract:** *The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) has led to the accelerated need for the development of processors that can perform high-scale computation. Among these processors, the Graphics Processing Units (GPUs) have become the backbone for the development of modern innovations in AI. These processors were originally designed for graphics processing but later became a powerful tool for the development of AI due to the presence of matrices that can be processed in parallel. Thousands of processing units within these processors allow for the acceleration of the development of AI models. In addition, the presence of software such as CUDA and cuDNN has allowed for the acceleration of the development of AI models. NVIDIA has been instrumental in the development of modern GPU processors that can perform AI-specific tasks. The company has introduced various innovations that enable the development of AI models. The innovations introduced by NVIDIA have not only led to the development of AI models but also allowed for the advancement of PC gaming. This paper will discuss the evolution of the development of the GPU processor for AI-specific tasks. In addition, the benefits that these processors provide will be discussed.*

**Keywords:** Graphics Processing Units (GPUs), AI-specific processor, RATE RAY TRACING, Deep learning, CUDA, NVIDIA

## I. INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) has generated an unprecedented need for high-performance computing hardware that can process massive data sets and perform complex mathematical calculations. Central Processing Units (CPUs), though efficient for general-purpose computation, have shown limitations in meeting the high processing needs of modern AI models. The limitations of CPUs have led to the invention and adoption of AI processors. Among these processors, Graphics Processing Units (GPUs) have been at the forefront of this revolution.

Initially developed to cater to the ever-growing need for graphical processing in PC gaming and 3D rendering, GPUs have far exceeded their original purpose. The inherent structure of GPUs, consisting of thousands of efficient processing units capable of performing multiple calculations in parallel, makes them an ideal candidate for performing complex calculations involved in deep learning. The ever-increasing complexity of AI models, with billions of parameters involved in models such as Generative Adversarial Networks (GANs), Transformers, and Reinforcement Learning agents, can be processed within a reasonable time frame using GPUs.

Firms such as NVIDIA have been at the forefront of the development of the essential role that the GPU plays in the development of AI. Since the introduction of the modern GPU in 1996, NVIDIA has
continued to push the boundaries of parallel computing, combining cutting-edge hardware features with the development of a comprehensive software ecosystem that includes CUDA, cuDNN, TensorRT, and AI-optimized drivers. The development of the modern GPU has not only impacted the field of scientific computing but has also had a profound influence on the gaming industry, with NVIDIA's DLSS, Reflex, RTX Remix, ACE, and Broadcast technologies, showing the cross-disciplinary role that the development of AI hardware plays.

The influence of the development of the modern GPU has also been seen in the development of cloud computing, where cloud computing firms such as Amazon Web Services, Google Cloud, and Microsoft Azure provide on-demand

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

access to high-powered computing resources that include the use of the GPU. The development of multi-GPU computing clusters, specialized data centers, and AI supercomputers has seen the role that the GPU plays in the development of AI become the backbone of modern computing, powering breakthroughs in image recognition, natural language processing, and other AI-related applications.

Considering their parallelism and scalability, as well as software support, GPUs have become the dominant type of AI-specific processors in modern computing systems. As AI systems are becoming more complex and require more timely responses, GPUs are expected to fuel the next innovations in AI systems, not only in technology but also in the future of intelligent systems.

**Background**

The evolution of Artificial Intelligence (AI) and Machine Learning (ML) has been significantly impacted by the rapid advancements in computational hardware. Conventional Central Processing Units (CPUs), despite their flexibility and high performance in sequential processing, are not well suited for AI applications that require massive parallelization. With the advent of large-scale and complex deep learning models, such as Generative Adversarial Networks (GANs), Transformers, and reinforcement learning models, the requirement for specialized hardware that can execute many operations in parallel has become critical.

Graphics Processing Units (GPUs), originally designed to meet the computational power requirements of PC gaming and real-time 3D graphics rendering, have been optimized into powerful parallel processing units. With many cores in each GPU, along with the evolution of software stacks, GPUs have been optimized to perform massive parallelization. This has enabled them to play a significant role in AI applications.

them vital for speeding up training and computation in contemporary AI models. Companies like NVIDIA have helped accelerate this shift by adding various AI-focused tools like CUDA, cuDNN, TensorRT, and Tensor Cores to their GPUs, allowing them to achieve orders of magnitude better performance than traditional CPUs for DL models. Furthermore, the introduction of cloud-based GPU instances has helped make AI computing accessible to not only the industry but also the academic sector.

**Motivation**

Nevertheless, despite the immense contribution of GPUs in the revolutionization of AI, the recent surge in AI models has created new computational challenges. For instance, the recent surge in AI models has been characterized by an increase in the number of parameters in the models. As the AI landscape continues to expand into various fields, including autonomous systems, healthcare diagnostics, natural language understanding, and data analytics, the need for processors that can support AI applications is even more imperative.

Furthermore, although NVIDIA has revolutionized both the gaming and AI industries, there is still an intellectual need to understand the academic significance of how GPU architecture has enabled high-performance computing in AI applications. Additionally, it is imperative to compare the performance of GPU architecture in relation to the performance of other AI-specific processors such as TPUs, NPUs, and ASICs. This research is motivated by the need to understand the significance of GPUs as AI-specific processors.

**Problem Statement**

Even though the current trend in the field of AI computation is dominated by the use of GPUs, there is little consolidated research that attempts to study the design of the architecture of these processors, their efficiency of operation, and their role as AI-specific processors in the changing hardware environment. As the size and complexity of AI models are constantly increasing, the current GPUs are facing problems of power consumption, memory issues, scalability issues, and cost issues. The research aims at filling the gaps in the understanding of the role of the current GPUs in meeting the needs of AI computation and the problems that need to be solved for the progress of AI hardware design.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

**Scope and Objectives of the Research**

This research aims to identify the significance of Graphics Processing Units (GPUs) in relation to Artificial Intelligence (AI)-oriented processors, highlighting their architectural design, parallel computing capabilities, and performance in relation to deep learning applications. This research aims to identify the significant contributions of NVIDIA Corporation in the evolution of Graphics Processing Units, along with an assessment of their hardware and software innovations in relation to the development of Artificial Intelligence, as well as other relevant technologies like PC Gaming. Additionally, the research aims to identify the performance capabilities of Graphics Processing Units in relation to other processor technologies like CPUs, TPUs, NPUs, ASICs, and the overall impact of GPU-oriented cloud computing in relation to the democratization of Artificial Intelligence computing.

Furthermore, the research aims to identify trends, challenges, and opportunities in relation to GPU-oriented Artificial Intelligence computing. Although the research aims to provide an architectural overview of Graphics Processing Units, it does not include hardware implementation.

Correspondingly, the primary objectives of the research aligned with the scope are:

To analyze the evolution of the GPU from a graphics processor to an advanced AI processor.

To analyze the primary architectural components of the GPU that facilitate the computation of AI.

To analyze the role of NVIDIA in the advancement of AI computing using GPU technology.

To analyze the limitations faced by the existing GPU architectures in meeting the requirements of AI computing.

To analyze the competitive advantage of the GPU in comparison to other AI-specific processors.

To analyze the future trends in AI processing and the role that the GPU will play in the advancement of intelligent systems.

**Theoretical Fundamentals**

This increasing complexity of Artificial Intelligence (AI) and Machine Learning (ML) computations has led to a paradigm shift in using traditional CPUs in favor of AI processors, with GPUs at the forefront of this revolution. This is because GPUs are optimized for executing computations in parallel, which is a key requirement in executing deep learning algorithms that involve repeated matrix computations, convolutions, and backpropagation.

**GPU Architecture**

GPUs are quite different from CPUs in terms of their architecture. CPUs are made of a number of strong cores, whereas the number of cores in the case of the GPU is in thousands. The cores are designed in such a manner that thousands of operations are performed at the same time. The architecture of the GPU makes it suitable for accelerating large-scale AI workloads. The memory hierarchy of the GPU, including the use of high-bandwidth memory, ensures that there is access to large amounts of data, thus reducing the possibility of bottlenecks in the training of the AI model.

**NVIDIA GPU Architecture**

NVIDIA has been at the forefront of developing cutting-edge technology for Graphics Processing Units (GPUs), both for graphics and AI computing. The NVIDIA architecture for their GPUs includes the following features that are optimized for AI computing:

CUDA (Compute Unified Device Architecture): CUDA is a parallel computing platform and programming model that allows developers to utilize the processing power of the GPU efficiently. CUDA has become the de facto standard for developing deep learning frameworks.

Tensor Cores: The Tensor Cores are part of the NVIDIA Volta architecture that performs mixed-precision matrix multiplications efficiently, accelerating the training and inference of deep neural networks.

cuDNN (CUDA Deep Neural Network library): cuDNN is a library that provides GPU-accelerated primitives for deep neural networks, supporting various frameworks like TensorFlow, PyTorch, and Keras.

TensorRT: TensorRT is an optimization library for inference that maximizes the performance of AI applications.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

NVLink: NVLink is a high-speed interconnect that allows multiple NVIDIA GPUs to efficiently communicate with each other in a cluster, facilitating the training of AI models.
NVIDIA graphics cards, from the GeForce line used in gaming to the Tesla, Quadro, and A100 line used in AI and HPC, have shown their flexibility in various applications. This integration of hardware and software has enabled an environment in which AI models can be trained.

**AMD GPU Architecture**

AMD, another major player in the graphics processing unit field, has also contributed to the development of AI-specific processing. AMD's graphics processing units follow the Graphics Core Next (GCN) and RDNA architectures for gaming and computing purposes. The CDNA architecture is for data center-specific AI computing and high-performance computing applications. The main features of AMD's graphics processing units are:
ROCm (Radeon Open Compute Platform): AMD's open-source parallel computing platform for graphics processing units is similar to NVIDIA's CUDA platform.
Matrix Cores (MI series graphics processing units): Matrix cores are designed to accelerate AI computing tasks such as matrix multiplication.
High-bandwidth memory (HBM2/HBM2e): This feature enables faster access to memory for AI computing applications.
Infinity Fabric: This feature is a high-speed fabric that enables multi-GPU configurations for servers and supercomputers for AI computing.

**Applications in IT Industry**

GPU-based AI processors are now the core technology used in various IT applications:
Artificial Intelligence & Machine Learning: AI processors are used in AI and ML-based applications such as NLP, image recognition, and recommendation systems.
Cloud Computing: Cloud service providers such as AWS, Google Cloud Platform, and Microsoft Azure use NVIDIA and AMD-based AI processors to provide cloud-based AI and HPC services to organizations.
Data Analytics & Big Data: AI processors are used to perform parallel processing of large-scale datasets in various applications such as data analytics and big data.
Gaming & Simulation: AI processors such as NVIDIA GeForce and AMD Radeon are used to provide better gaming and simulation experiences with improved graphics and faster simulation speeds.
Autonomous Systems: AI processors are used to develop autonomous systems such as autonomous cars and drones.
Healthcare & Bioinformatics: AI processors are used to perform various medical image analyses and genomics with faster processing speeds.

**Present and Future Scope**

Currently, the range of GPU-based AI computing is from multi-GPU clusters to AI supercomputers, as well as cloud-based computing. This has enabled the training of more complex models like the Transformer architecture and GANs.
Moving into the future, the future research and development in GPU technology will include:
Next-Generation AI Accelerator: Improvements in the performance of Tensor Cores, Matrix Cores, and memory bandwidth to accommodate larger neural networks.
Energy Efficiency Computing: Optimizing the power consumption of the GPU to ensure sustainable computing.
Integrated with Other AI-Specific Computing Devices: Co-processing with TPUs, NPUs, and FPGAs.
Real-Time AI Computing: Optimizing the performance of the GPU to meet the low latency requirement of applications like autonomous systems, AR, IoT, etc.
SOFTWARE-HARDWARE CO-DESIGN: Development of frameworks that can utilize the architecture of the GPU more efficiently.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

In summary, NVIDIA and AMD GPUs will continue to play an important role in the future of AI computing. They will ensure that high-performance computing is not only accessible but also scalable and versatile in terms of application. Although AMD GPUs have always been at a disadvantage in terms of market share for AI, recent advancements in both hardware and software have enabled AMD GPUs to be used alongside NVIDIA GPUs for AI research and implementation.

**Application Domains and Use Cases of NVIDIA and AMD GPUs**

**Artificial Intelligence and Machine Learning**

Use Case: Training and deploying deep learning models like CNNs for image recognition or transformers for NLP tasks.

NVIDIA Example: NVIDIA A100 Tensor Core GPUs are used for training large models in AI research and enterprise AI environments, supporting popular frameworks like TensorFlow and PyTorch.

AMD Example: AMD MI100/MI200 GPUs with ROCm support are used for AI workloads in open-source AI frameworks, supporting cost-effective HPC environments.

**Cloud Computing and Data Centers**

Use Case: On-demand GPU instance types for scalable AI workloads, big data analytics, and virtualized environments.

NVIDIA Example: Amazon Web Services (AWS) provides EC2 P4 instance types that feature NVIDIA A100 GPUs for large-scale model training and inference.

AMD Example: AMD Instinct MI250X GPUs are used in cloud HPC services for scientific computing, simulations, and AI research

**Gaming and Graphics Rendering**

Use Case: Real-time rendering, physics simulations, and AI-based enhancement of graphics.

NVIDIA Example: The GeForce RTX series uses AI-based technology called "Deep Learning Super Sampling" to increase frame rates while preserving image quality.

AMD Example: The Radeon RX series uses AI-based technology called "FidelityFX Super Resolution" to provide an immersive gaming experience.

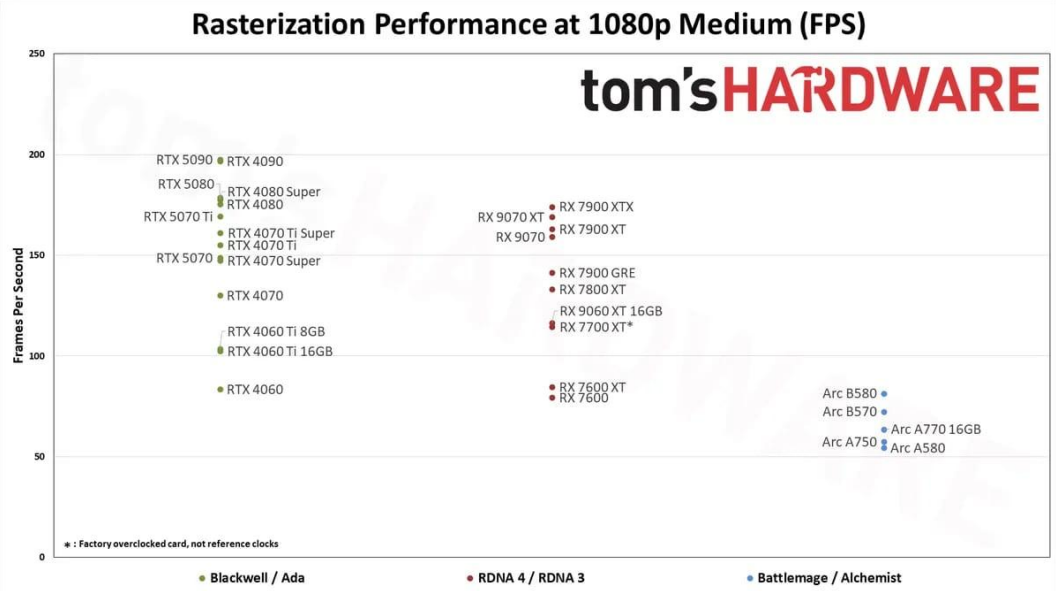
**Autonomous Systems and Robotics**

Use Case: Real-time perception, decision-making, and path planning in autonomous vehicles and robots.

NVIDIA Example: The NVIDIA DRIVE AGX is an autonomous driving solution that uses multiple GPUs.

AMD Example: AMD GPUs are used in robotic simulations and edge AI systems to provide GPU acceleration in real-time decision-making systems used in industrial robots.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2




**Healthcare and Bioinformatic**

Use Case: Accelerating medical imaging, genomics analysis, and drug discovery with parallel computation.

NVIDIA Example: The NVIDIA Clara platform uses NVIDIA GPUs for AI-assisted medical imaging, MRI reconstruction, and predictive diagnostics.

AMD Example: AMD GPUs power bioinformatics workflows for genome sequencing and protein structure prediction in HPC clusters.

**High-Performance Computing (HPC) and Scientific Research**

Use Case: Large simulations, climate modeling, physics calculations, financial analytics.

NVIDIA Example: NVIDIA DGX SuperPOD clusters are used for AI research and scientific simulations in institutions such as the Oak Ridge National Laboratory.

AMD Example: AMD Instinct GPUs are part of the supercomputer system called Frontier that performs exascale computing for AI-accelerated scientific modelling

**Real-Time Analytics and Big Data**

Use Case: Faster processing of large datasets for fraud detection, recommendations, and predictive analytics.

NVIDIA Example: RAPIDS AI software stack running on NVIDIA GPUs speeds up the entire data science workflow from hours to minutes.

AMD Example: AMD GPUs with ROCm technology support parallelized analytics workloads running on open-source big data platforms for enterprise needs.

**Performance evalution and benchmarking**

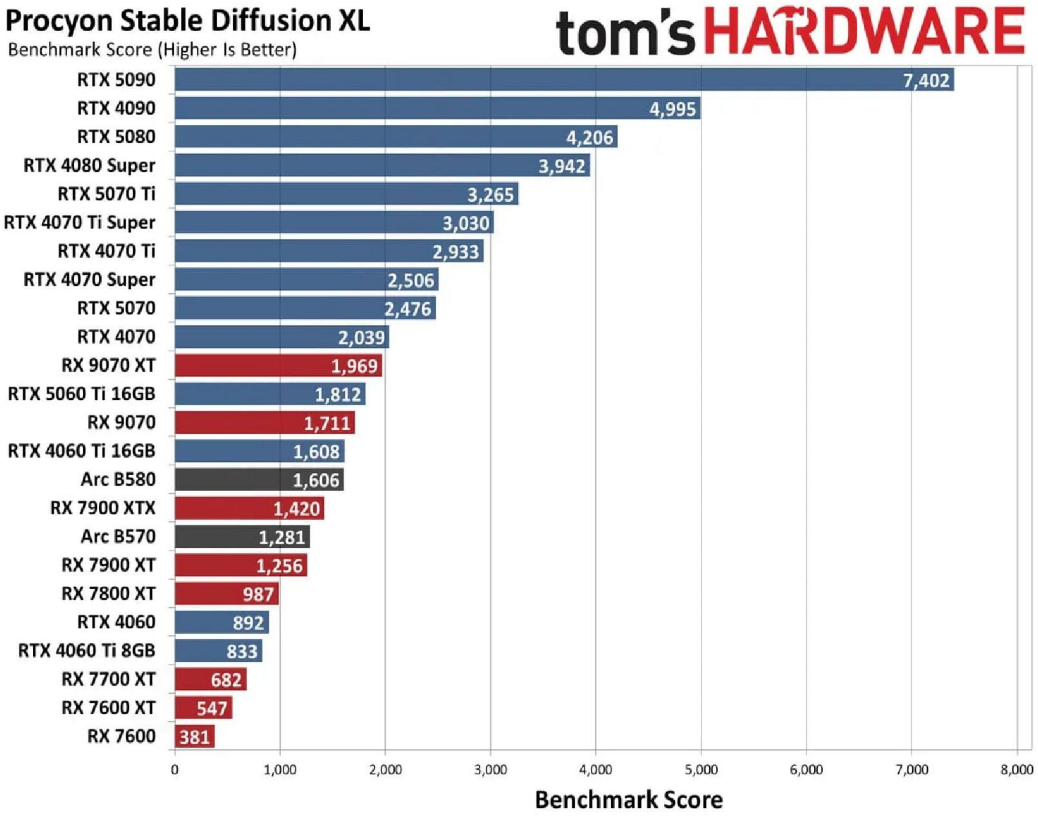
Nvidia's RTX 50 and 40 series cards absolutely dominate in the Stable Diffusion AI workloads and the Blender 3D rendering benchmarks.

The MLPerf Client and LLM benchmark shows Nvidia and Intel cards have the edge against Radeons in the Time to First Token latency test, but the higher-end cards perform in line with expectations in the Tokens per Second throughput test.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

The Proyon AI Vision benchmark has Nvidia cards dominating the competition. Even the new Intel Arc Battlemage cards have the edge over Radeons, but the latter have to be run in FP16 mode in this test, while Intel and Nvidia have a significant advantage using optimized INT8 paths through OpenVINO and TensorRT, respectively

SPECworkstation 4.0's GPU inference tests still favor Nvidia, but AMD closes the gap — and Intel falls to the bottom of the list. The Handbrake video transcoding tests only focus on raw throughput, without regards to encoding quality, with Intel's Battlemage and AMD's new RDNA 4 GPUs basically tied at the top, then there's a step down to the AMD RDNA 3 GPUs, and finally all the Nvidia GPUs clump together — though the Blackwell RTX 50-series cards do land slightly ahead of the older 40-series parts.




**Challenges and Limitations**

Despite the tremendous benefits that GPUs offer in AI and HPC applications, there are some challenges and limitations with their use, especially with NVIDIA and AMD architectures:

High Cost and Power Consumption

The latest and most powerful GPUs from NVIDIA, such as the GeForce RTX 5090 and RTX 4090, are accompanied by extremely high prices ($1,999 to $2,975 MSRP) and power requirements of up to 575W. Although AMD offers alternative GPUs such as the Radeon RX 7900 XTX with relatively lower prices and power requirements, they are still significant and may present challenges.

Memory Bottlenecks and Bandwidth Limitations

Although the memory bandwidth available with modern GPUs is extremely high with the use of GDDR7, GDDR6X, and HBM2e memory types, there are cases where AI models with billions of parameters such as the Transformers and GANs may exceed the available memory capacity.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

Thermal Management and Cooling Requirements

High-end GPUs require significant thermal management as well as cooling solutions, especially in multi-GPU configurations or data center environments. This can lead to added costs, as well as reliability issues.

Scalability Challenges in Multi-GPU Configurations

NVIDIA's NVLink, as well as AMD's Infinity Fabric, enables communication between GPUs. However, there is a limitation in terms of scale, including latency, as well as diminishing returns. Additionally, multi-GPU training requires sophisticated software solutions to utilize available resources effectively.

Software Ecosystem and Compatibility

NVIDIA's CUDA is a mature, industry-standard solution, with strong support for AI frameworks. This puts it in a better position compared to its counterpart, AMD's ROCm. Although open-source support is being developed, there is a possibility of compatibility issues, as well as suboptimal libraries, when using AMD GPUs for AI applications.

Rapid Hardware Obsolescence

The GPU space is constantly changing, and new generations of products are constantly being released, offering much-improved performance. This has the effect of making recently purchased equipment obsolete within a short period of time, which can be a problem in terms of budgeting and planning for research and business environments.

Specialized Workload Optimization

There are some types of workloads that benefit from specialized processing cores, such as NVIDIA's Tensor Cores for mixed precision matrix math. While AMD offers equivalent Matrix Cores, they have less support in AI environments. This makes the optimization of such environments less flexible for heterogeneous types of AI processing.

Environmental and Sustainability Concerns

The high energy consumption of contemporary GPUs makes for a high level of electricity consumption and a corresponding level of environmental degradation, which is a growing issue for environmentally focused organizations.

**Tabular representation of research gaps and research scope**

| Research Area | Research Gap | Scope for Current Research |
|---|---|---|
| **GPU Performance in AI Workloads** | Limited consolidated studies comparing latest NVIDIA and AMD GPUs for deep learning, inference, and HPC tasks. | Analyze and benchmark current NVIDIA (RTX 5000/4000 series) and AMD (Radeon RX/Instinct) GPUs for AI-specific operations, including training and inference. |
| **Cost and Power Efficiency** | High-end GPUs have high cost and power requirements, with limited studies on cost-performance trade-offs for AI workloads. | Evaluate energy efficiency, performance-per-dollar, and sustainability aspects of NVIDIA and AMD GPUs in AI applications. |
| **Memory and Bandwidth Bottlenecks** | Large AI models face memory constraints; comparative studies on memory architectures (HBM2, HBM2e, GDDR7) are limited. | Investigate memory hierarchy performance, bandwidth utilization, and model scalability across NVIDIA and AMD GPUs. |
| **Software Ecosystem Impact** | NVIDIA CUDA dominates AI frameworks; AMD ROCm has limited ecosystem adoption and optimizations. | Examine the impact of software support on AI training and inference, including compatibility with TensorFlow, PyTorch, and open-source libraries. |
| **Multi-GPU Scalability** | Limited analysis of interconnect efficiency (NVLink vs Infinity Fabric) and multi-GPU performance for large AI models. | Study scalability, latency, and throughput in multi-GPU setups for distributed AI workloads. |

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

| Research Area | Research Gap | Scope for Current Research |
|---|---|---|
| **Emerging AI Workloads** | Lack of research on performance for next-gen AI models (Transformers, GANs, RL agents) on current GPU architectures. | Evaluate GPUs for training and inference of large-scale AI models, highlighting bottlenecks and optimization strategies. |
| **Environmental and Sustainability Concerns** | Limited studies on energy consumption and carbon footprint of AI-focused GPUs. | Assess power consumption, thermal design, and environmental impact of high-performance NVIDIA and AMD GPUs. |
| **Future Trends in AI Hardware** | Insufficient analysis on evolving GPU architectures versus alternative AI accelerators (TPUs, NPUs, ASICs). | Explore trends in GPU-based AI acceleration, compare with emerging AI-specific processors, and predict future hardware adoption patterns. |

**Emerging Trends and Future Directions in GPU-Based AI Processing**

Currently, there are many trends in GPU-based AI computing. For example, there are specialized AI processors such as Tensor Cores from NVIDIA and Matrix Cores from AMD that are used to perform faster matrix operations. Additionally, there are multi-GPU systems and distributed systems with high-speed interconnects such as NVLink and Infinity Fabric that are used to train large models.

In the future, there are many trends in GPU-based AI computing. For instance, there is expected to be more integration with other AI accelerators such as TPUs and NPUs. Additionally, there is expected to be more emphasis on energy efficiency with precision-aware computation and adaptive voltage scaling. Therefore, there are many trends in GPU-based AI computing that are expected to determine the future of this field.

## II. CONCLUSION

The high growth of Artificial Intelligence (AI) and Machine Learning (ML) has revolutionized the high-performance computing industry in terms of speed, thereby leading to the shift from Central Processing Unit (CPU) to AI-specific processor design. Among these, GPUs from NVIDIA and AMD have gained prominence in the AI processor market due to their high parallel processing capabilities, high memory bandwidth, and rich software stacks. NVIDIA's inventions in Tensor Cores, CUDA, and cuDNN, and AMD's ROCm and Matrix Cores have paved the path for efficient training of complex AI models.

However, the limitations of using GPUs for AI training and inference include high energy consumption, scaling limitations, memory limitations, and high costs, particularly with the increase in the complexity of AI models. A comparative analysis of the two major players in the AI processor industry shows that NVIDIA is the leader in market adoption and cloud platforms, while AMD is gaining ground in the open-source and HPC segments. The emerging trends in the AI processor industry include the development of hybrid AI processor architectures, energy-efficient computing, multi-GPU architectures, and the role of software in addressing the limitations of the current AI processor architectures.

Overall, it is safe to say that AI-specific processors based on GPUs have become indispensable in delivering breakthroughs in a wide range of AI applications, including autonomous systems, natural language processing, healthcare, and many others. The future of AI-specific processors, including GPUs, is expected to be bright, given the progress being made in these technologies, as well as in AI architectures.

## REFERENCES

**[1].** Dally, W., Keckler, S., & Kirk, D. B. (2021). Evolution of the Graphics Processing Unit (GPU). IEEE Micro. NVIDIA

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

**[2].** Atluri, A. A. (2025). The Evolution of NVIDIA GPUs for Deep Learning: From Gaming to AI Powerhouse. International Journal of Advanced Research in Engineering & Technology (IJARET), 16(1), 540–551. ResearchGate+1

**[3].** Mushtaq Bhat, S. (2025). A Seminar Report on GPUs in AI: NVIDIA's Dominance and the Rising Competition. [Research Report]. ResearchGate

**[4].** Varala, C. R. (2025). The Role of GPUs in Artificial Intelligence and Machine Learning. Journal of e-Science Letters. scienceletters.researchfloor.org

**[5].** Li, A., Song, S. L., Chen, J., Li, J., Liu, X., Tallent, N., Barker, K. (2019). Evaluating Modern GPU Interconnect: PCIe, NVLink, NV-SLI, NVSwitch and GPUDirect. arXiv. arXiv

**[6].** Lee, Y., Lim, J., Bang, J., Cho, E., Jeong, H., Kim, T., Rhu, M. (2024). Debunking the CUDA Myth Towards GPU-based AI Systems. arXiv. arXiv

**[7].** Zhou, T., Lin, X., Wu, J., Chen, Y., Xie, H., Li, Y., … Dai, Q. (2020). Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. arXiv. arXiv

**[8].** Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Wicke, M. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv. arXiv

**[9].** NVIDIA. (2023). Why GPUs Are Great for AI. NVIDIA Blog. NVIDIA Blog

**[10].** Goyaal, Y. (2024). Review on GPU Architecture. International Research Journal of Engineering and Technology (IRJET), 11(7). IRJET