Comprehensive Journal of Science
Volume (11), Issue (42), (Sapt, 2026)
ISSN: 3014-6266

مجلة العلوم الشاملة
المجلد (11) العدد 42 سبتمبر 2026
ردمد: 3014-6266

The Evolution of NVIDIA AI Accelerators and Their Impact on Artificial Intelligence Model Performance: A Comparative Analysis of NVIDIA A100, H100, H200, and Blackwell Architectures

Adel Elgaber[1*], and Basma Ali Slisal[2]
[1] Computer Department, Faculty of Engineering, Sabratha University, Regdalin, Libya
[2] Computer Department, Institute of Science and Technology, Regdalin, Libya
*Corresponding author: adel@sabu.edu.ly

Received: 22/7/2026 Revised: 30/7/2026 Accepted: 28/8/2026 Publication: 20/9/2026

Abstract:
The fast development of artificial intelligence (AI), especially in deep learning, generative AI, and large language models (LLMs), has increased the need for powerful computing hardware. GPUs and special AI accelerators have become an important part of modern AI systems. NVIDIA has been one of the main companies contributing to this development through different GPU architectures, such as Ampere, Hopper, and Blackwell. This paper studies the development of NVIDIA AI accelerators and their effect on the performance of modern AI models, focusing on the NVIDIA A100, H100, H200, and Blackwell B200 architectures. The study looks at improvements in tensor processing, memory capacity and bandwidth, numerical precision, interconnect technologies, and other technologies designed for AI workloads. The comparison shows that improvements in GPU architecture have helped reduce the time needed for model training and inference and have made it possible to work with larger and more complex AI models. The move from A100 to H100 introduced important technologies such as fourth-generation Tensor Cores and the Transformer Engine. The H200 mainly improved memory capacity and bandwidth, which makes it more suitable for large language models. The Blackwell architecture provides further improvements through fifth-generation Tensor Cores, lower-precision computing, improved memory systems, and faster communication between chips. Results from MLPerf also show clear improvements in AI training and inference performance across these generations. However, the performance of an AI model does not depend only on the processing power of the GPU. Memory bandwidth, model design, software optimization, communication between GPUs, and energy efficiency also have an important effect. The paper concludes that the development of NVIDIA AI accelerators and the development of AI models are strongly connected. Improvements in hardware have helped researchers and companies build larger and more capable AI models, while the increasing size and complexity of these models have created a need for more advanced AI hardware.

Keywords: Artificial Intelligence, AI Accelerators, NVIDIA, A100, H100, H200, Blackwell, GPUs, Deep Learning, Large Language Models, Generative AI, Tensor Cores.

الملخص: أدى التطور المتسارع في مجال الذكاء الاصطناعي (AI) - ولا سيما في تقنيات التعلم العميق، والذكاء الاصطناعي التوليدي، والنماذج اللغوية الكبيرة (LLMs) - إلى تزايد الحاجة إلى أجهزة حوسبة فائقة القدرة. وقد أصبحت وحدات معالجة الرسومات (GPUs) ومسرعات الذكاء الاصطناعي المتخصصة جزءاً جوهرياً من أنظمة الذكاء الاصطناعي الحديثة. وتُعد شركة NVIDIA من الشركات الرائدة التي ساهمت في هذا التطور من خلال تقديم معماريات متنوعة لوحدات معالجة الرسومات، مثل Ampere وHopper وBlackwell. تتناول هذه الورقة البحثية تطور مسرعات الذكاء الاصطناعي من NVIDIA وأثرها على أداء نماذج الذكاء الاصطناعي الحديثة، مع التركيز على معماريات NVIDIA A100 وH100 وH200 وBlackwell B200. وتستعرض الدراسة التحسينات التي طرأت على مجالات معالجة الموترات (Tensor processing)، وسعة الذاكرة وعرض النطاق الترددي لها، والدقة العددية، وتقنيات الربط البيني، وغيرها من التقنيات المصممة خصيصاً لأعباء عمل الذكاء الاصطناعي. وتُظهر المقارنة أن التحسينات في معمارية وحدات معالجة الرسومات قد ساهمت في تقليل الوقت اللازم لتدريب النماذج وعمليات الاستنتاج (Inference)، كما أتاحت التعامل مع نماذج ذكاء اصطناعي أكبر حجماً وأكثر تعقيداً. وقد شهد الانتقال من معمارية A100 إلى H100 تقديم تقنيات هامة مثل الجيل الرابع من أنوية الموترات (Tensor Cores) ومحرك Transformer (Transformer Engine). أما معمارية H200 فقد ركزت بشكل أساسي على تحسين سعة الذاكرة وعرض النطاق الترددي، مما جعلها أكثر ملاءمة للنماذج اللغوية الكبيرة. وتوفر معمارية Blackwell تحسينات إضافية تشمل الجيل الخامس من أنوية الموترات، والحوسبة بدقة أقل (lower-precision computing)، وأنظمة ذاكرة مطورة، وسرعة أكبر في الاتصال بين الرقائق. كما تظهر نتائج اختبارات MLPerf تحسينات واضحة في أداء تدريب واستنتاج الذكاء الاصطناعي عبر هذه الأجيال المتعاقبة. ومع ذلك، فإن أداء نموذج الذكاء الاصطناعي لا يعتمد فقط على القدرة الحوسبية لوحدة معالجة الرسومات؛ إذ تلعب عوامل أخرى - مثل عرض النطاق الترددي للذاكرة، وتصميم النموذج، وتحسين البرمجيات، والاتصال بين وحدات معالجة الرسومات، وكفاءة الطاقة - دوراً مهماً أيضاً. وتخلص الورقة إلى وجود ارتباط وثيق بين تطور مسرعات الذكاء الاصطناعي من NVIDIA وتطور نماذج الذكاء الاصطناعي؛ فقد ساعدت التحسينات في العتاد (Hardware) الباحثين والشركات على بناء نماذج أكبر وأكثر قدرة، بينما أدى تزايد حجم هذه النماذج وتعقيدها إلى خلق حاجة ملحة لأجهزة ذكاء اصطناعي أكثر تطوراً.

الكلمات المفتاحية: الذكاء الاصطناعي، مسرّعات الذكاء الاصطناعي، NVIDIA، A100، H100، H200، Blackwell، وحدات معالجة الرسومات (GPUs)، التعلم العميق، النماذج اللغوية الكبيرة، الذكاء الاصطناعي التوليدي، أنوية Tensor.

1. Introduction

Artificial intelligence (AI) has developed quickly during the last ten years. This development has been supported by improvements in machine learning methods, the availability of large datasets, and the continuous development of computer hardware. Deep neural networks, transformer-based models, generative AI, and large language models (LLMs) need a large amount of computing power for both training and inference. As AI models have become larger and more complex, traditional central processing units (CPUs) are not always able to provide the required performance. For this reason, there has been an increasing need for processors that can perform many operations at the same time [5,6].

GPUs were originally developed mainly for computer graphics, but they have gradually become important computing platforms for AI applications. One of their main advantages is their ability to perform many matrix and vector calculations in parallel, which are common in neural network operations. In addition, special processing units such as Tensor Cores can speed up the matrix calculations used in many deep learning models. Another important development is mixed-precision computing. It can reduce memory use and increase computing speed while keeping the model accuracy at an acceptable level [4,7].

NVIDIA has played an important role in this development. The introduction of the NVIDIA A100, based on the Ampere architecture, was an important step in the development of GPUs designed for AI workloads. The A100 included third-generation Tensor Cores, Multi-Instance GPU (MIG) technology, support for sparse calculations, and high-bandwidth memory and interconnect technologies. Several independent studies have also investigated the architecture and performance of the A100. These studies show the importance of Tensor Cores, memory, and data movement in the performance of modern GPU workloads [1, 8,9].

After Ampere, NVIDIA introduced the Hopper architecture with the H100 GPU. The H100 was designed with the increasing requirements of transformer-based models in mind. It introduced fourth-generation Tensor Cores, FP8 support, the Transformer Engine, and improvements in memory and GPU interconnection. The Transformer Engine can use lower numerical precision when it is suitable for the calculation, which can improve the efficiency of transformer models during training and inference [10].

The H200 continued to use the Hopper architecture but focused more on improving memory capacity and bandwidth. This is especially important for large language models because these models require a large amount of memory to store their parameters and other data during processing. The H200 provides 141 GB of HBM3e memory and up to 4.8 TB/s of memory bandwidth. These improvements allow larger AI models to be processed more efficiently compared with previous generations [11].

The Blackwell architecture represents another important step in the development of NVIDIA AI accelerators. Blackwell GPUs use 208 billion transistors and a dual-die design connected through a high-speed chip-to-chip connection. The architecture also introduces fifth-generation Tensor Cores and new technologies aimed at improving the performance of generative AI and reasoning models. These developments are important because current AI models continue to increase in size and require more computing resources [11,12].

Based on these developments, this study focuses on the following research question:

How has the evolution of NVIDIA AI accelerator architectures from A100 to H100, H200, and Blackwell affected the performance, scalability, and efficiency of modern artificial intelligence models?

2. Research Objectives

The main objective of this research is to study the relationship between the development of NVIDIA AI accelerators and the improvement of AI model performance. The research focuses

on how changes in GPU architecture, memory, processing units, and communication technologies have affected the ability to train and run modern AI models.

The specific objectives of this research are:

1. To examine the development of NVIDIA A100, H100, H200, and Blackwell AI accelerators and the main changes introduced in each generation.
2. To compare the computational capabilities and memory technologies of these accelerators.
3. To study the role of Tensor Cores in improving the performance of AI workloads.
4. To investigate the effect of different numerical precision technologies, including FP16, FP8, and FP4, on AI processing.
5. To examine how memory capacity and memory bandwidth affect the performance of large AI models.
6. To analyze the role of interconnect technologies in improving communication and scalability in multi-GPU AI systems.
7. To compare available benchmark results for AI training and inference workloads.
8. To examine the relationship between the development of AI hardware and the increasing size and complexity of modern AI models.

3. Research Methodology

This research uses a comparative and analytical approach to study the development of NVIDIA AI accelerators and their effect on AI model performance. The study is based mainly on technical documentation from NVIDIA, research studies, academic publications, and standardized AI benchmark results.

The research focuses on four NVIDIA accelerators that represent important stages in the development of AI hardware:

- **NVIDIA A100** – based on the Ampere architecture.
- **NVIDIA H100** – based on the Hopper architecture.
- **NVIDIA H200** – based on the Hopper architecture with improved memory capacity and bandwidth.
- **NVIDIA Blackwell B200** – based on the Blackwell architecture.

These accelerators were selected because they show the major changes in NVIDIA's approach to AI acceleration. The A100 provides a strong starting point for modern data-center AI computing, while the H100 introduced technologies that were more focused on transformer models. The H200 improved memory capacity and bandwidth to support larger models, and the Blackwell architecture introduced further improvements for generative AI and other demanding workloads.

The comparison is based on five main areas:

1. **Computational performance:** The study examines the processing capabilities of each accelerator and how they affect AI workloads.
2. **Memory capacity and bandwidth:** The research compares memory technologies and their importance for large AI models.
3. **Tensor Core architecture and numerical precision:** The study examines the development of Tensor Cores and the use of different numerical formats, such as FP16, FP8, and FP4.
4. **Interconnect and scalability:** The research consider technologies that allow several GPUs to communicate and work together efficiently.
5. **Training and inference performance:** The study compares available benchmark results to understand how the different accelerator generations perform in practical AI workloads.

MLPerf is used as an important source for comparing AI performance because it provides standardized benchmarks for different AI workloads. MLPerf Training measures the time needed for a system to train a model to a defined target quality, while MLPerf Inference measures the performance of AI models during inference under different conditions.

In addition to benchmark results, the research considers technical specifications and findings from previous studies. This approach allows the study to compare both the theoretical capabilities of NVIDIA accelerators and their practical effects on AI workloads.

It should be noted that benchmark results can be affected by several factors, including the number of GPUs, software optimization, model architecture, numerical precision, batch size, and system configuration. Therefore, the results are interpreted carefully rather than assuming that a higher theoretical GPU performance always produces the same improvement in every AI application.

Overall, this methodology provides a basis for understanding how improvements in NVIDIA accelerator architecture have contributed to the development of faster, larger, and more efficient AI models.

4. Evolution of NVIDIA AI Accelerators

4.1 NVIDIA A100 and the Ampere Architecture

The NVIDIA A100 was introduced in 2020 and was an important step in the development of GPUs for artificial intelligence. It is based on the Ampere architecture and includes third-generation Tensor Cores, which are designed to perform the matrix operations commonly used in deep learning models [13].

One of the important features of the A100 is its support for different numerical formats. This allows the GPU to use different levels of precision depending on the type of calculation. This is useful in deep learning because many operations do not need to be performed using full FP32 precision. Using lower precision can reduce memory use and improve the speed of computation while maintaining suitable model accuracy [4].

The A100 also introduced Multi-Instance GPU (MIG) technology. MIG allows one physical GPU to be divided into several separate GPU instances. Each instance can be assigned to a different workload. This can improve the use of GPU resources in data centers, especially when several users or applications need to use the same GPU [14].

Another important feature of the A100 is its support for structured sparsity. Some neural networks contain parameters that have little effect on the final result. When the model and software support structured sparsity, some of these calculations can be skipped. The A100 provides hardware support for this type of operation, which can improve performance for suitable AI workloads [15].

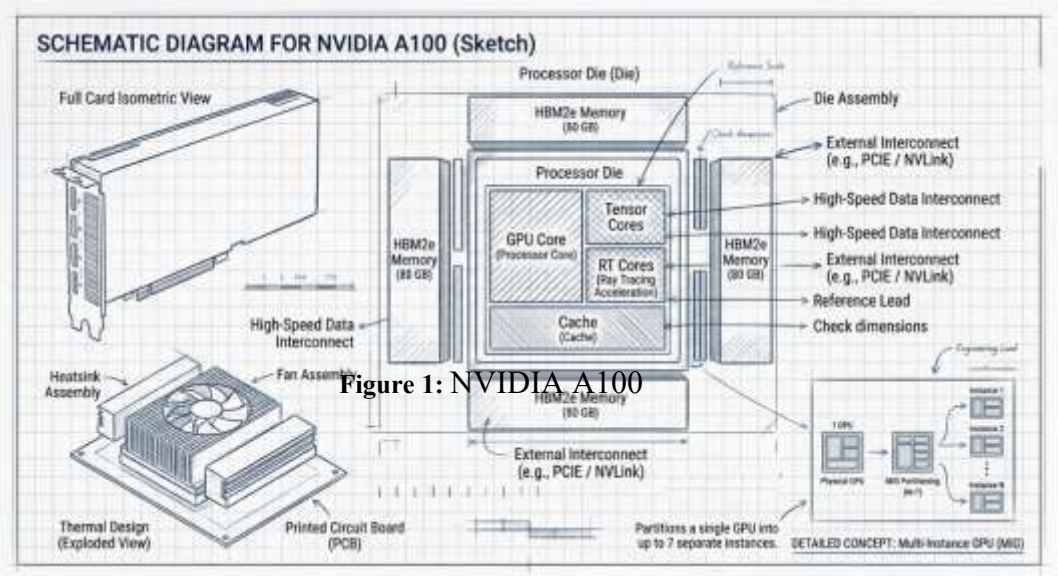


Figure 1: NVIDIA A100

Memory and communication between GPUs are also important parts of the A100 architecture. The A100 uses high-bandwidth memory and supports NVLink and NVSwitch technologies for communication between GPUs. Depending on the A100 configuration, NVIDIA reported

memory bandwidth of up to about 2 TB/s, together with high-speed GPU interconnection technologies [16].

Independent research on the Ampere architecture has also shown that the A100 introduced important changes in its instruction set, memory system, and Tensor Core design. These changes affect how researchers and developers can optimize AI and scientific applications for the GPU [1].

Overall, the A100 provided an important foundation for large-scale AI computing. Its Tensor Cores, memory system, sparsity support, and GPU partitioning capabilities made it suitable for both AI training and inference workloads.

4.2 NVIDIA H100 and the Hopper Architecture

The NVIDIA H100 was introduced with the Hopper architecture as a new generation of AI accelerator designed to meet the increasing computational requirements of large AI models, especially transformer-based models. The growth of large language models and generative AI made these improvements increasingly important.

One of the main new technologies in the H100 is the **Transformer Engine**. Transformer models perform many matrixes multiplication and attention operations, which require high computational performance. The Transformer Engine allows the H100 to use different numerical precisions, including FP8, during calculations when appropriate. This can improve processing speed and reduce memory requirements while maintaining the required model accuracy [17].

The Hopper architecture also introduced fourth-generation Tensor Cores and the Tensor Memory Accelerator. These technologies were designed to improve both computation and data movement. The Tensor Memory Accelerator helps move data between memory and the GPU processing units more efficiently, which can reduce some of the delays caused by data transfer [10].

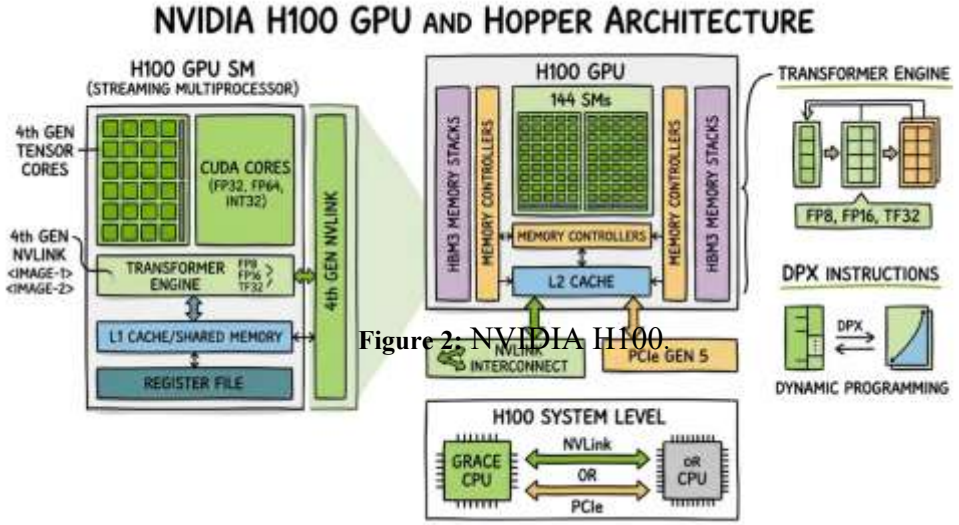


Figure 2: NVIDIA H100.

The development from A100 to H100 was therefore not only an increase in the number of processing units. It also represented a change toward designing the hardware specifically for modern AI models. In particular, the H100 was developed with transformer workloads in mind, which became increasingly important with the growth of generative AI and large language models [17].

NVIDIA reported significant performance improvements for H100 systems compared with A100 systems in MLPerf Training benchmarks. In some of the initial comparisons, NVIDIA reported up to 6.7 times higher performance for H100 systems than A100 systems on selected training workloads [18].

Actual performance can be different depending on several factors, including the model architecture, number of GPUs, batch size, numerical precision, software libraries, communication between GPUs, and the overall system configuration. Therefore, benchmark results are useful for showing the potential improvement between generations, but they should be interpreted according to the specific workload being tested [19].

The H100 can therefore be considered an important step in the development of specialized AI hardware. It improved computational performance while also introducing technologies that were designed to address the specific requirements of transformer-based models.

5. NVIDIA H200: Increasing Memory Capacity for Large Models

The NVIDIA H200 is based on the Hopper architecture used by the H100, but it introduced a major improvement in memory capacity and bandwidth. This is important because memory has become one of the main limitations when working with very large AI models.

The H200 provides **141 GB of HBM3e memory and up to 4.8 TB/s of memory bandwidth**. According to NVIDIA, this represents nearly twice the memory capacity and 1.4 times the memory bandwidth of the H100, allowing larger AI models to be processed more efficiently **[20]**.

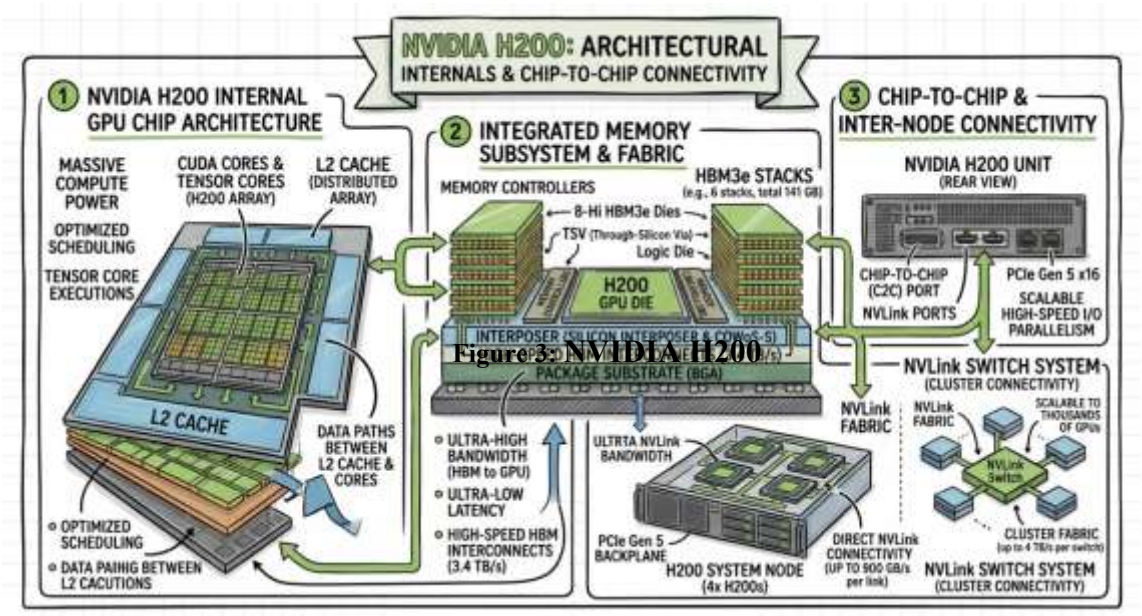


Figure 3: NVIDIA H200

This improvement is especially useful for large language models. These models may contain billions or even hundreds of billions of parameters. In addition to the model parameters, memory is needed for activations, intermediate results, and the key-value cache used during language model inference. As a result, the amount of required memory can become very large. When the complete model cannot fit into the memory of one GPU, it may need to be divided across several GPUs. This process can increase communication between GPUs and may reduce overall performance. Increasing the memory capacity of a single accelerator can reduce the need for this type of model partitioning in some situations.

The H200 therefore shows an important point in the development of AI hardware. Improving AI performance does not depend only on increasing the number of processing units or the theoretical computing power of the GPU. Memory capacity and memory bandwidth are also very important, especially for large language models [21].

The higher memory bandwidth of the H200 allows data to move between memory and the processing units at a faster rate. This can help reduce memory-related bottlenecks and improve the performance of workloads that frequently access large amounts of data [21].

This is particularly important for inference. During inference, the GPU must repeatedly access model parameters and other data while generating results. For large language models, memory access and data movement can therefore become major factors that affect response speed [22].

For this reason, the H200 can be viewed as an important development that focuses not only on computation but also on the memory requirements of modern AI models. It demonstrates how changes in hardware design are increasingly influenced by the growing size and complexity of AI models.

6. NVIDIA Blackwell Architecture

The Blackwell architecture is another important step in the development of NVIDIA AI accelerators. It was designed to support the increasing requirements of generative AI, large language models, and other workloads that require a high level of computing power. As AI models become larger, the hardware needs to provide not only more processing power but also faster memory and communication between different parts of the system.

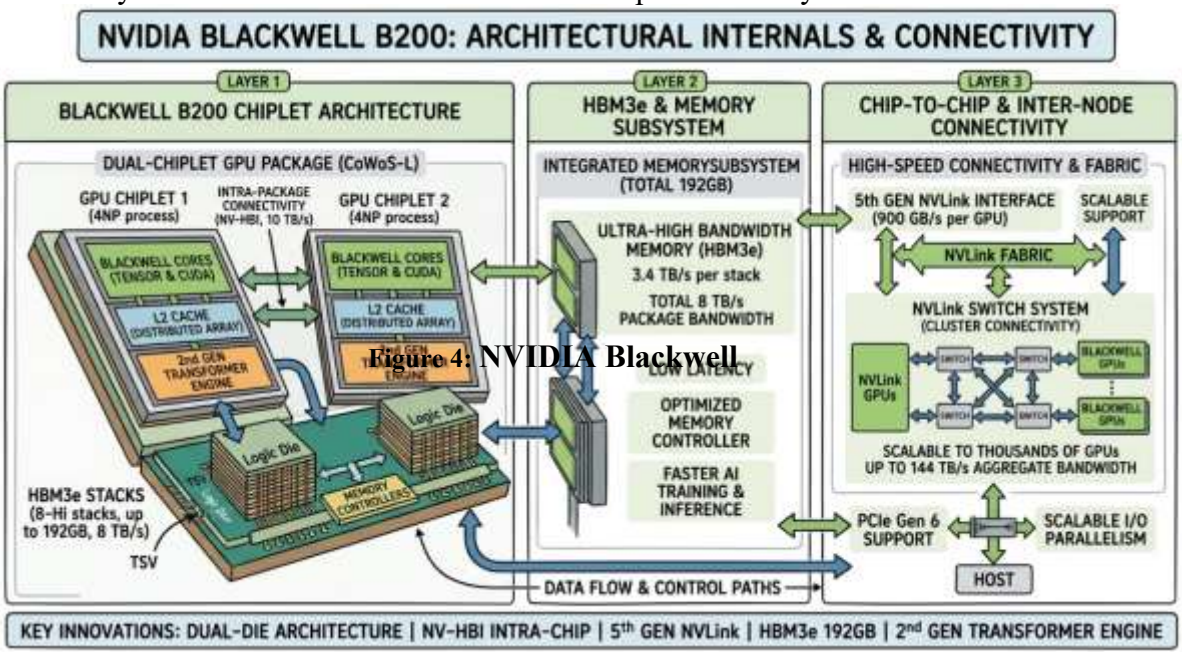


Figure 4: NVIDIA Blackwell

According to NVIDIA, Blackwell GPUs contain about 208 billion transistors. The architecture uses a dual-die design, where two large dies are connected through a high-speed chip-to-chip connection. This design allows the GPU to provide higher performance while keeping the two dies connected as a single computing system [23].

One of the main improvements in Blackwell is the use of **fifth-generation Tensor Cores**. These processing units are designed to accelerate the matrix calculations used in many AI models. Blackwell also supports very low numerical precision, including FP4 for suitable AI workloads. This is especially important for generative AI models because lower precision can reduce the amount of memory needed and can increase the amount of computation that can be performed in a given period of time [23].

However, using lower numerical precision also requires careful control. If the precision is reduced too much or used in an unsuitable part of a model, the quality of the results can be affected. Therefore, the use of FP4 and other low-precision formats should be combined with appropriate software techniques to maintain the required level of accuracy.

Memory performance is another important improvement in the Blackwell architecture. NVIDIA documentation reports that the B200 can provide up to **180 GB of HBM3e memory and up to 8 TB/s of memory bandwidth** in current HGX configurations [24]. The higher bandwidth allows large amounts of data to move between memory and the processing units more quickly. This is important for large language models and generative AI applications, where memory access can have a strong effect on overall performance.

Blackwell also improves communication between GPUs and between the main components of the accelerator. This is important because modern AI models are often too large to run efficiently on a single GPU. Several GPUs may need to work together, and fast communication between them can reduce the time spent transferring data [25].

Independent research has also examined the Blackwell architecture through micro benchmarking. [26] studied several features of Blackwell, including its fifth-generation Tensor Cores, low-precision FP4 and FP6 operations, memory system, and overall performance characteristics. Their work provides additional information about how the new architecture behaves in practical workloads and how it differs from the previous Hopper generation.

The development of Blackwell shows that NVIDIA is moving toward a more specialized approach to AI hardware. The main goal is not only to increase the number of calculations per second, but also to improve memory access, communication, numerical efficiency, and support for the specific operations used by modern AI models [26].

In general, Blackwell can be seen as a response to the rapid growth of generative AI and large language models. Its improvements in Tensor Cores, low-precision computing, memory bandwidth, and interconnect technologies are intended to make it possible to train and run larger AI models more efficiently. This also shows the close relationship between AI model development and hardware development, because increasingly complex models create new requirements that influence the design of the next generation of AI accelerators.

7. Comparative Analysis

The following table summarizes the major architectural differences [23,26].

Feature	**A100**	**H100**	**H200**	**Blackwell B200**
Architecture	Ampere	Hopper	Hopper	Blackwell
Main AI focus	Deep Learning/HPC	Transformers/AI	Large AI/LLMs	Generative AI/Reasoning
Tensor Cores	3rd generation	4th generation	4th generation	5th generation
Memory technology	HBM2e	HBM3	HBM3e	HBM3e
Memory capacity	Up to 80 GB	80 GB	141 GB	Up to 180 GB*
Memory bandwidth	~2 TB/s	~3.35 TB/s	4.8 TB/s	Up to 8 TB/s
FP8 support	No native Hopper FP8	Yes	Yes	Yes
FP4 support	No	No	No	Yes
Transformer Engine	No	Yes	Yes	Enhanced
Multi-GPU scaling	NVLink/NVSwitch	NVLink/NVSwitch	NVLink/NVSwitch	NVLink 5
Main advantage	General AI acceleration	Transformer acceleration	Memory capacity	AI performance and efficiency

Table 1: Specifications can vary depending on the specific product and system configuration.

.

The comparison between the A100, H100, H200, and Blackwell shows that NVIDIA's development has not focused only on increasing the basic floating-point performance of its GPUs. Each new generation has introduced improvements in several areas that are important for AI workloads, including processing power, memory, numerical precision, and communication between GPUs.

The main development can be summarized as follows:

Feature	Development
A100 → Improved computational performance	The A100 provided a strong foundation for large-scale AI computing through its Tensor Cores, high-bandwidth memory, and support for parallel processing.
H100 → Better support for transformer models	The H100 introduced the Transformer Engine, fourth-generation Tensor Cores, and FP8 support. These features were especially useful for transformer-based models and large language models.
H200 → Increased memory capacity	The H200 focused mainly on increasing memory capacity and bandwidth. This improvement is important for large AI models that require a large amount of memory during training and inference.
Blackwell → Low-precision computing, higher scale, and generative AI optimization	Blackwell introduced fifth-generation Tensor Cores, support for low-precision formats such as FP4, higher memory bandwidth, and improved communication technologies. These features are designed to support the increasing requirements of generative AI and large-scale models.

Table 2: . The main development of A100, H100, H200, Blackwell

This development shows that modern AI accelerator design is becoming more specialized. The main goal is no longer only to increase the number of calculations that a GPU can perform. It is also necessary to improve memory access, reduce communication delays, support lower numerical precision, and provide better support for the operations used by current AI models.

8. Impact on AI Model Training

Training modern AI models requires a large amount of computing power. During training, the model processes large datasets many times and continuously updates its parameters. When the number of parameters increases, the amount of computation and memory required for training also increases [27].

The development of NVIDIA AI accelerators has helped reduce the time required for these operations through improvements in several areas:

- Matrix multiplication performance.
- Memory bandwidth.
- Numerical precision.
- Communication between GPUs.
- Parallel processing.
- Integration between hardware and software.

The transition from the A100 to the H100 was particularly important for transformer-based models. The H100 introduced the Transformer Engine, which allows the use of FP8 computing in suitable operations. This can increase processing speed and reduce memory requirements while maintaining the required level of model accuracy [17].

The H200 continued this development by providing more memory capacity and higher memory bandwidth. These improvements can be useful when training large models because more model data can remain in GPU memory. This may reduce the need to divide the model across a large number of GPUs and can also reduce some communication between GPUs.

Blackwell continues the same direction but introduces additional improvements in computing, memory, and numerical precision. The use of lower-precision formats such as FP4 can provide higher computational efficiency for suitable workloads. This is particularly relevant to modern generative AI models, where the amount of computation and memory required can be very large.

Therefore, the development of NVIDIA AI accelerators has contributed to a continuous relationship between hardware and AI model development. This relationship can be described as a simple cycle:

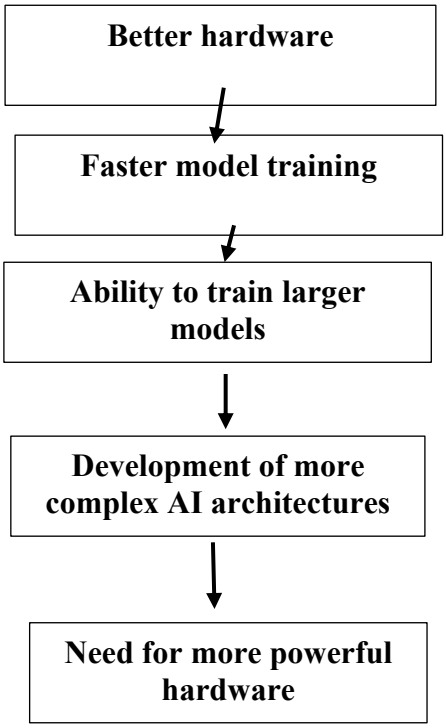


Table 3: relationship between hardware and AI model development.

This cycle is important for understanding the rapid development of modern artificial intelligence. As hardware becomes faster and more efficient, researchers can experiment with larger models and larger datasets. At the same time, the increasing size of these models creates new hardware requirements [27].

The relationship is therefore not one-directional. Hardware improvements support the development of AI models, while the requirements of new AI models also influence the design of future accelerators. This continuous interaction between hardware and models is an important part of the development of modern AI systems.

9. Impact on AI Inference

Training is an important part of the AI model lifecycle, but it is not the only stage. After a model has been trained, it needs to perform inference when it is used by users or other applications. During inference, the model receives an input and produces an output based on what it learned during training [28].

Inference can be especially demanding for large language models. The system needs to process the user's input and generate the output step by step. At the same time, it needs to keep different types of intermediate information in memory. As the size of the model increases, these requirements also become greater [29].

For this reason, GPU memory capacity and memory bandwidth have become very important factors in AI inference performance. A GPU with more memory can store more model data locally, while higher memory bandwidth can help move this data more quickly during processing.

MLPerf Inference provides standardized benchmarks for measuring AI inference performance across different types of workloads. These workloads include large language models, computer vision, recommendation systems, and other AI applications [30].

The MLPerf Inference v5.0 benchmark also included larger language-model workloads, such as Llama 3.1 405B and Llama 2 Chat 70B. The inclusion of these models shows how large generative AI models have become an important part of evaluating modern AI hardware [30].

This development is important because AI accelerator evaluation is no longer focused mainly on traditional workloads such as image classification. Large language models and generative AI now represent an important part of AI computing. As a result, new GPU architectures need to provide good performance not only for training but also for serving large numbers of inference requests efficiently.

Another important factor in inference is response time. In many applications, users expect an AI system to respond quickly. Therefore, improving inference performance can help reduce response time and allow a system to serve more users at the same time. This makes improvements in GPU memory, computing power, and communication especially important for large-scale AI services [30].

10. The Role of Memory in AI Performance

One of the main conclusions from comparing NVIDIA's different accelerator generations is that computational performance cannot be considered separately from memory performance.

An AI accelerator may have very high theoretical computing performance, but this performance cannot be fully used if data cannot reach the processing units quickly enough. In this situation, some of the processing units may remain idle while waiting for data. This is generally known as a **memory bottleneck [31]**.

Memory bottlenecks become more important as AI models become larger. Large language models require memory for model parameters, intermediate values, activations, and other information used during training and inference.

The development from the H100 to the H200 provides a clear example of how memory improvements can help address this problem. The H200 keeps the main Hopper architecture but provides a much larger memory capacity and higher memory bandwidth. This allows more data to be stored close to the processing units and can reduce some of the limitations caused by memory access [32].

For large language models, this can be particularly useful because the system may spend a significant amount of time moving model parameters and intermediate data. If memory access is slow, increasing the computational power of the GPU alone may not provide the expected improvement.

Research on AI hardware also shows that data movement, memory access, and communication can become important performance limitations as neural networks continue to increase in size and complexity (32, 7).

Therefore, the development of AI accelerators needs to consider both **computational performance and memory performance**. A balanced system is more likely to provide better real-world performance than a system that focuses only on increasing the theoretical number of calculations [32].

11. The Role of Numerical Precision

Numerical precision is another important factor in the development of AI accelerators. In earlier neural network applications, FP32 calculations were commonly used for many operations. However, modern AI systems increasingly use lower-precision formats such as FP16, BF16, FP8, and FP4[26, 4].

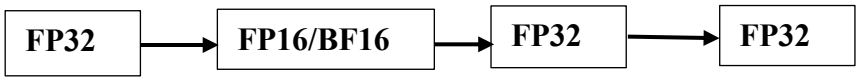
Using lower numerical precision can provide several advantages:

1. It can reduce the amount of memory required to store data.
2. It can increase computational throughput.
3. It can reduce the amount of data that needs to be transferred.
4. It can improve energy efficiency in suitable workloads.

Mixed-precision training showed that neural networks can often use lower-precision calculations while maintaining good model accuracy when appropriate methods are used [4].

The development of NVIDIA accelerators reflects this change. The Hopper architecture introduced specific hardware support for FP8-based transformer computation, while the Blackwell architecture extends low-precision acceleration to formats such as FP4 for suitable workloads.

The development can be simplified as:



However, this does not mean that using lower precision is always better. Different models and different operations may require different levels of precision. If the precision is reduced too much in an unsuitable part of a model, the quality of the results may decrease.

For this reason, modern AI accelerators need to support several numerical formats rather than depending on only one format. The ability to select a suitable precision for each workload allows developers to balance performance, memory use, energy consumption, and model accuracy.

This is one reason why technologies such as the Transformer Engine are important. They help AI systems use lower precision where it is appropriate while keeping higher precision where it is still needed [17].

12. Interconnect and Multi-GPU Scaling

Modern AI models can be too large to fit into the memory of a single GPU. They can also require more computational power than one accelerator can provide. As a result, large AI systems often use many GPUs working together [33].

When several GPUs are used, communication between them becomes an important part of overall system performance. During distributed training, GPUs may need to exchange model parameters, gradients, and other information many times. If this communication is slow, it can reduce the benefits of using additional GPUs [34].

NVIDIA has developed technologies such as **NVLink and NVSwitch** to provide high-speed communication between GPUs. These technologies allow multiple accelerators to exchange data more efficiently and are important for large AI systems [35].

The importance of communication becomes greater as the number of GPUs increases. A system with hundreds or thousands of GPUs needs a high-speed interconnect network to keep the accelerators working efficiently.

This means that the performance of a modern AI system cannot be explained by GPU computing power alone. A more complete view can be represented as:

GPU computation + memory + interconnect + software

All of these components work together to determine the actual performance of an AI system.

The importance of multi-GPU and multi-node systems can also be seen in recent MLPerf benchmark results. Many benchmark submissions use several GPUs and multiple computing nodes to train and run large AI models.

Therefore, improvements in interconnect technology are an important part of the evolution of AI accelerators. Faster communication allows more GPUs to work together and makes it possible to train and run AI models that would be difficult or impossible to handle efficiently on a single accelerator.

13. Software-Hardware Co-Design

The development of AI systems cannot be explained by improvements in hardware alone. Software also plays an important role in making full use of the capabilities provided by modern AI accelerators. As NVIDIA GPUs have become more specialized for AI workloads, the software ecosystem has also developed to support these new hardware features.

NVIDIA provides several software technologies and libraries, including CUDA, cuDNN, TensorRT, NCCL, and Transformer Engine. These tools allow developers and AI frameworks to use the specific capabilities of NVIDIA GPUs more effectively.

This development has created a close relationship between hardware and software. In this approach, the GPU architecture and the software libraries are developed together so that each can take advantage of the other. This is often described as **hardware-software co-design**.

For example, Tensor Cores can provide high performance for matrix operations, but the AI software needs to be designed and optimized so that these operations can be efficiently executed on the Tensor Cores. In the same way, technologies such as FP8 and FP4 require software support to select and manage numerical precision correctly.

This means that the performance of an AI accelerator depends on two main factors:

Hardware capability + Software utilization

A GPU may have very high theoretical performance, but an application may not achieve this performance if the software is not properly optimized. Therefore, technical specifications alone should not be considered a complete measure of real-world AI performance.

Software optimization can also reduce communication delays, improve memory use, and select appropriate numerical precision. As a result, the same GPU may provide different levels of performance depending on the AI framework, software libraries, model implementation, and optimization techniques being used [36].

The close relationship between hardware and software has therefore become an important part of modern AI development. Improvements in one area can create new possibilities in the other area, helping researchers and developers build more efficient AI systems [36].

14. Benchmark-Based Performance Analysis

Benchmarking is important for understanding the practical effect of new AI accelerator architectures. Technical specifications can show the theoretical capabilities of a GPU, but benchmark results can provide a better idea of how a complete system performs when running real AI workloads.

MLPerf is one of the important benchmark platforms used for this purpose. It evaluates complete AI systems rather than looking only at individual GPU specifications. MLPerf Training measures how quickly a system can train a model to a defined target quality, while MLPerf Inference measures the performance of trained models under standardized inference conditions (MLCommons, 2026).

Historical MLPerf results show a significant improvement between the A100 and H100 generations. NVIDIA reported that H100 systems achieved up to **6.7 times higher performance** than A100 systems in selected initial MLPerf Training comparisons. This result demonstrates the potential of the Hopper architecture for large-scale AI training.

However, the 6.7× result should not be understood as a fixed improvement that applies to every AI model. The actual difference between two GPUs depends on the type of model, number of GPUs, numerical precision, software optimization, and other system factors.

More recent MLPerf results also show continued improvements with newer GPU architectures, including Blackwell-based systems. The MLPerf Training v5.1 results included large-scale generative AI workloads and showed the increasing importance of newer accelerator architectures for training modern AI models (MLCommons, 2025).

Benchmark results are useful because they provide a common way to compare different systems. However, they must be interpreted carefully because many factors can influence the final result. These factors include:

- Number of GPUs.
- Model implementation.
- Batch size.
- Numerical precision.
- Software libraries.
- Interconnect technology.
- Hardware and software optimization.
- Overall system configuration.

For this reason, a fair comparison should use the same model, benchmark, accuracy target, and system scale whenever possible. This makes it easier to identify whether a performance improvement comes mainly from the GPU architecture or from differences in the complete system [36].

15. Discussion

The comparison presented in this study shows that the development of NVIDIA AI accelerators has taken place in several different areas rather than through a simple increase in computational power.

The **A100** provided an important foundation for large-scale AI and high-performance computing workloads. Its Tensor Cores, high-bandwidth memory, support for sparsity, and Multi-Instance GPU technology made it suitable for many different machine learning applications.

The **H100** introduced a more specialized approach to AI acceleration. Instead of focusing only on increasing general computational performance, the Hopper architecture introduced technologies specifically designed for transformer-based models. The Transformer Engine and FP8 support were particularly important as transformer models became widely used in generative AI and large language model applications.

The **H200** demonstrated that memory capacity and bandwidth can be as important as computational performance. Large language models require a large amount of memory for their parameters and intermediate data. With 141 GB of HBM3e memory and up to 4.8 TB/s of memory bandwidth, the H200 was designed to address some of these memory limitations (NVIDIA, 2024).

The **Blackwell** architecture continues this development by combining high computational performance with greater memory bandwidth, fifth-generation Tensor Cores, lower numerical precision, and improved communication between GPUs. These features are aimed particularly at the growing requirements of generative AI and large-scale reasoning models.

The development of these architectures can therefore be understood as a gradual change in the role of GPUs in artificial intelligence. The process can be summarized as follows:

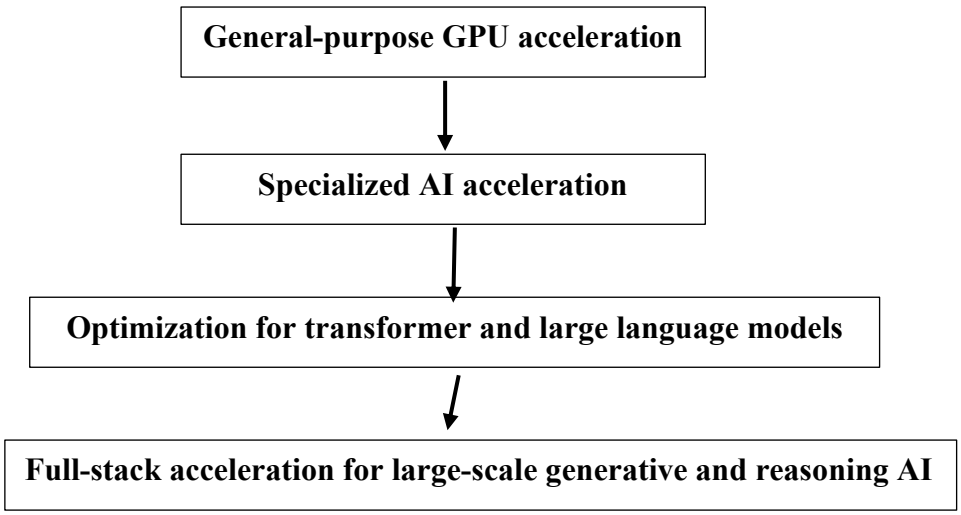


Figure 1: Steps of development of GPU

This progression shows that modern AI accelerators are no longer designed only to provide more computing power. They are increasingly designed around the specific requirements of modern AI models. Memory, numerical precision, communication, software, and specialized processing units all contribute to the final performance of an AI system.

Therefore, the evolution of NVIDIA accelerators and the development of AI models should be viewed as closely connected processes. Larger and more complex AI models create new hardware requirements, while improvements in hardware allow researchers and companies to

develop models that would have been difficult to train or run efficiently with previous generations of technology.

16. Challenges and Limitations

Although NVIDIA AI accelerators have provided major improvements in AI computing, several challenges and limitations still need to be considered. These challenges are related not only to the GPU itself but also to energy use, cost, software, memory, communication, and the way performance is measured.

16.1 Energy Consumption

High-performance AI accelerators require a large amount of electrical power. This issue becomes more important when AI systems use hundreds or thousands of GPUs at the same time. In large data centers, electricity consumption and cooling requirements can become major parts of the total infrastructure cost [37].

For this reason, AI accelerator performance should not be evaluated only by the number of calculations that a GPU can perform. **Performance per watt** is also an important measure because a system that provides high performance while using less energy can be more practical for large-scale AI applications.

16.2 Cost

Another important limitation is the cost of high-end AI accelerators. The GPU itself is only one part of the total cost. Large AI systems also require high-speed networking, storage, cooling systems, electricity, and other data-center infrastructure.

The high cost can make advanced AI computing difficult for small research groups, universities, and organizations with limited budgets. As a result, access to large-scale AI computing is still concentrated in organizations that can afford expensive computing infrastructure [38].

16.3 Software Dependence

The performance of an AI accelerator depends partly on the software used with it. Modern GPUs contain many specialized features, but these features cannot provide their full benefit if the application or AI framework is not properly optimized.

For example, an application may not achieve the expected performance from Tensor Cores if its matrix operations are not efficiently mapped to the hardware. Similar issues can occur with lower-precision formats and multi-GPU communication [39].

Therefore, software libraries, compilers, AI frameworks, and optimization techniques are important parts of the overall performance of an AI system.

16.4 Memory and Communication Bottlenecks

As AI models become larger, memory limitations can become a major problem. A model may require more memory than one GPU can provide, which means that its parameters and other data must be distributed across several GPUs.

This creates another problem because the GPUs need to communicate with each other. If communication is slow, the additional GPUs may not provide the expected improvement in performance.

Therefore, increasing computational capacity alone is not always enough. AI systems also need sufficient memory capacity, high memory bandwidth, and fast interconnect technologies [33].

16.5 Benchmark Limitations

Benchmark results are useful for comparing AI hardware, but they do not always represent the performance of every real-world application.

For example, a GPU that performs very well when running a large language model may not provide the same level of improvement for a computer vision model or another type of workload.

Benchmark results can also be affected by the number of GPUs, software optimization, numerical precision, model implementation, batch size, and system configuration.

For this reason, benchmark results should be used carefully, and comparisons should consider the conditions under which the results were obtained [36].

17. Future Directions

The development of AI models is continuing at a rapid pace, and future AI accelerators will need to address the new requirements created by larger and more complex models. Several areas are likely to receive more attention in future accelerator architectures.

17.1 Lower Precision

Lower numerical precision is expected to become increasingly important, especially for AI inference. Formats such as FP4 can reduce memory requirements and increase computational throughput for suitable workloads.

However, future systems will need to manage the balance between performance and accuracy. Lower precision will not be appropriate for every operation, so flexible precision support will remain important [40].

17.2 Larger Memory

As AI models continue to increase in size, larger accelerator memory will become increasingly important. More memory allows a larger part of a model to remain on the accelerator and can reduce the need to divide the model across many GPUs.

Higher memory bandwidth will also be important because large models require frequent movement of data between memory and processing units [33].

17.3 Advanced Interconnects

Future AI systems will increasingly depend on fast communication between accelerators. Large models may require hundreds or thousands of GPUs, making communication speed an important factor in overall system performance.

For this reason, future architectures are expected to continue improving GPU-to-GPU and node-to-node communication [41].

18. Conclusion

The evolution of NVIDIA AI accelerators from the A100 to the H100, H200, and Blackwell architectures shows a strong relationship between the development of hardware and the development of artificial intelligence models.

The **A100** provided an important foundation for large-scale deep learning by combining Tensor Cores, high-bandwidth memory, sparsity support, and multi-GPU technologies. It provided the computing capabilities needed for many AI and high-performance computing workloads.

The **H100** introduced a more specialized approach to AI acceleration. Through the Transformer Engine, fourth-generation Tensor Cores, and FP8 support, it was designed to improve the performance of transformer-based models. This was particularly important as large language models and generative AI became more widely used.

The **H200** addressed another important challenge: memory capacity and bandwidth. Its larger HBM3e memory and higher memory bandwidth make it more suitable for large AI models that require significant amounts of memory during training and inference.

The **Blackwell** architecture represents a further development toward specialized AI acceleration. Fifth-generation Tensor Cores, lower-precision computing, higher memory bandwidth, and improved communication technologies are designed to support the increasing requirements of generative AI, large language models, and reasoning workloads.

The comparison also shows that AI performance cannot be explained by computational power alone. Memory capacity, memory bandwidth, numerical precision, interconnect technology, software optimization, and energy efficiency all have an important effect on the final performance of an AI system.

The relationship between AI models and hardware can therefore be understood as a continuous cycle. As AI models become larger and more complex, they require more powerful and efficient hardware. At the same time, improvements in hardware allow researchers to train and deploy models that were previously too expensive or difficult to run.

This relationship has played an important role in the rapid development of modern AI. The evolution of NVIDIA accelerators has therefore not only increased the speed of existing AI

workloads, but has also helped expand the practical limits of AI model size, complexity, and application.

In conclusion, the development from A100 to H100, H200, and Blackwell shows that the future of AI performance depends on the combined development of **hardware, software, memory, communication technologies, and AI algorithms**. Continued progress in these areas will be necessary to support the next generation of large-scale and increasingly capable artificial intelligence systems.

References

1-Abdelkhalik, H., Arafa, Y., Santhi, N., & Badawy, A.-H. (2022). *Demystifying the Nvidia Ampere architecture through microbenchmarking and instruction-level analysis*. arXiv. https://arxiv.org/abs/2208.11174 (arXiv)

2-Jarmusch, A., & Chandrasekaran, S. (2025). *Microbenchmarking NVIDIA's Blackwell architecture: An in-depth architectural analysis*. arXiv. https://arxiv.org/abs/2512.02189 (arXiv)

3-Luo, W., Fan, R., Li, Z., Du, D., Wang, Q., & Chu, X. (2024). *Benchmarking and dissecting the Nvidia Hopper GPU architecture*. arXiv. https://arxiv.org/abs/2402.13499 (arXiv)

4-Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., García, D., Ginsburg, B., Houston, M., Kuchaev, O., Venkatesh, G., & Wu, H. (2018). Mixed precision training. *arXiv*. https://arxiv.org/abs/1710.03740

5- Mahatme, Jyoti & Deshmukh, Suvidha. (2026). Advancements in Artificial Intelligence with Deep Learning: Techniques, Applications, and Challenges.

6- Hooman H. Rashidi, Joshua Pantanowitz, Matthew G. Hanna, Ahmad P. Tafti, Parth Sanghani, Adam Buchinsky, Brandon Fennell, Mustafa Deebajah, Sarah Wheeler, Thomas Pearce, Ibrahim Abukhiran, Scott Robertson, Octavia Palmer, Mert Gur, Nam K. Tran, Liron Pantanowitz, Introduction to Artificial Intelligence and Machine Learning in Pathology and Medicine: Generative and Nongenerative Artificial Intelligence Basics,Modern Pathology,Volume 38, Issue 4,2025,100688,ISSN 0893-3952,https://doi.org/10.1016/j.modpat.2024.100688.

7- Wael, Ayham & Madi, Amer. (2025). Accelerating Artificial Intelligence: The Role of GPUs in Deep Learning and Computational Advancements. East Journal of Engineering. 1. 31-46. 10.63496/eje.Vol1.Iss1.34.

8- J. Choquette, W. Gandhi, O. Giroux, N. Stam and R. Krashinsky, "NVIDIA A100 Tensor Core GPU: Performance and Innovation," in IEEE Micro, vol. 41, no. 2, pp. 29-35, 1 Marc h-April 2021, doi: 10.1109/MM.2021.3061394.

9- https://www.nvidia.com/en-us/data-center/ampere-architecture/

10- Luo, Weile & Fan, Ruibo & Li, Zeyu & Du, Dayou & Wang, Qiang & Chu, Xiaowen. (2024). Benchmarking and Dissecting the Nvidia Hopper GPU Architecture. 656-667. 10.1109/IPDPS57955.2024.00064.

11- Double-BlindAaron Jarmusch,Newark, USSunita Chandrasekaran Microbenchmarking NVIDIA's Blackwell Architecture: An in-depth Architectural Analysis, arXiv:2512.02189v1 [cs.AR] 01 Dec 2025

12- NVIDIA. (2024, March 18). *NVIDIA Blackwell platform arrives to power a new era of computing*. NVIDIA. NVIDIA Blackwell Platform Arrives to Power a New Era of Computing

13- NVIDIA. (2020, May 14). *NVIDIA's new Ampere data center GPU in full production*. NVIDIA Newsroom. NVIDIA's New Ampere Data Center GPU in Full Production

14- NVIDIA. (2020, November 30). *Getting the most out of the NVIDIA A100 GPU with Multi-Instance GPU*. NVIDIA Technical Blog. NVIDIA Technical Blog – Getting the Most Out of the NVIDIA A100 GPU with Multi-Instance GPU

15- Krashinsky, R., Giroux, O., Jones, S., Stam, N., & Ramaswamy, S. (2020, May 14). *NVIDIA Ampere architecture in-depth*. NVIDIA Technical Blog. NVIDIA Ampere Architecture In-Depth

16- NVIDIA. (2020). *NVIDIA A100 Tensor Core GPU architecture*. NVIDIA. NVIDIA A100 Tensor Core GPU Architecture – Whitepaper

17- Andersch, M., Palmer, G., Krashinsky, R., Stam, N., Mehta, V., Brito, G., & Ramaswamy, S. (2022, March 22). *NVIDIA Hopper architecture in-depth*. NVIDIA Technical Blog. NVIDIA Hopper Architecture In-Depth

18- Salvator, D. (2022, November 9). *Hopper, Ampere sweep MLPerf training tests*. NVIDIA Blogs. NVIDIA Blogs – Hopper, Ampere Sweep MLPerf Training Tests Salvator, D. (2022, November 9). *Hopper, Ampere sweep MLPerf training tests*. NVIDIA Blogs. NVIDIA Blogs – Hopper, Ampere Sweep MLPerf Training Tests

19- Golden, A., Wu, C.-J., Wei, G.-Y., & Brooks, D. (2026). *The xPU-athalon: Quantifying the competition of AI acceleration*. arXiv.

20- NVIDIA. (2023, November 13). *NVIDIA supercharges Hopper, the world's leading AI computing platform*. NVIDIA, (https://investor.nvidia.com/news/press-release-details/2023/NVIDIA-Supercharges-Hopper-the-Worlds-Leading-AI-Computing-Platform/default.aspx?utm_source=chatgpt.com)

21- A. Gholami, Z. Yao, S. Kim, C. Hooper, M. W. Mahoney and K. Keutzer, "AI and Memory Wall," in IEEE Micro, vol. 44, no. 3, pp. 33-39, May-June 2024, doi: 10.1109/MM.2024.3373763.

22- Chowdhury, R. I/o for LLM inference: a survey of storage and memory bottlenecks. *Artif Intell Rev* (2026). https://doi.org/10.1007/s10462-026-11651-1
23- NVIDIA. (2024, March 18). *NVIDIA Blackwell platform arrives to power a new era of computing*. NVIDIA (https://investor.nvidia.com/news/press-release-details/2024/NVIDIA-Blackwell-Platform-Arrives-to-Power-a-New-Era-of-Computing/?utm_source=chatgpt.com)
24- NVIDIA. (2026). *NVIDIA HGX AI Factory: Components*. NVIDIA (https://docs.nvidia.com/enterprise-reference-architectures/hgx-ai-factory/latest/components.html?utm_source=chatgpt.com)
25- Singhania, P., Singh, S., Hough, L. D., Srivastava, A., Menon, H., Jekel, C. F., & Bhatele, A. (2026). *Understanding and improving communication performance in multi-node LLM inference*. In **Proceedings of the ACM Conference on AI and Agentic Systems (CAIS '26)** (pp. 1070–1083). Association for Computing Machinery. https://doi.org/10.1145/3786335.3813165.
26- Jarmusch, A., Graddon, N., & Chandrasekaran, S. (2025). *Dissecting the NVIDIA Blackwell architecture with microbenchmarks*. In **2025 IEEE 32nd International Conference on High Performance Computing, Data, and Analytics Workshops (HiPCW)** (pp. 309–310). IEEE. https://doi.org/10.1109/HiPCW66559.2025.00109
27- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models*. arXiv. https://doi.org/10.48550/arXiv.2001.08361
28- Chou, Y.-M., Chan, Y.-M., Lee, J.-H., Chiu, C.-Y., & Chen, C.-S. (2018). Unifying and merging well-trained deep neural networks for inference stage. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, 2049–2056. https://doi.org/10.24963/ijcai.2018/283
29-Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). Efficient memory management for large language model serving with PagedAttention. *Proceedings of the 29th Symposium on Operating Systems Principles*, 611–626. https://doi.org/10.1145/3600006.3613165
30- MLCommons. (2025, April 2). *MLPerf Inference v5.0 advances language model capabilities for GenAI*. MLCommons (https://mlcommons.org/2025/04/llm-inference-v5/?utm_source=chatgpt.com).
31- Gómez-Luna, J., El Hajj, I., Fernandez, I., Giannoula, C., Oliveira, G. F., & Mutlu, O. (2021). *Benchmarking memory-centric computing systems: Analysis of real processing-in-memory hardware*. arXiv. https://doi.org/10.48550/arXiv.2110.01709
32- Xu, B., Banerjee, A., & Gupta, S. (2025). *Hardware acceleration for neural networks: A comprehensive survey*. arXiv. https://doi.org/10.48550/arXiv.2512.23914
33- Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V. A., Vainbrand, D., Kashinkunti, P., Bernauer, J., Catanzaro, B., Phanishayee, A., & Zaharia, M. (2021). *Efficient large-scale language model training on GPU clusters using Megatron-LM*. arXiv. https://doi.org/10.48550/arXiv.2104.04473
34- Wang, S., Pi, A., & Zhou, X. (2019). Scalable distributed DL training: Batching communication and computation. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 5289–5296. https://doi.org/10.1609/aaai.v33i01.33015289
35- Slechta, B., Comly, N., Eassa, A., DeLaere, J., & Raj, S. (2024, August 12). *NVIDIA NVLink and NVIDIA NVSwitch supercharge large language model inference*. NVIDIA Technical Blog(https://developer.nvidia.com/blog/nvidia-nvlink-and-nvidia-nvswitch-supercharge-large-language-model-inference/?utm_source=chatgpt.com).
36- Mattson, P., Cheng, C., Coleman, C., Diamos, G., Micikevicius, P., Patterson, D., Tang, H., Wei, G.-Y., Bailis, P., Brooks, D., Chen, D., Dutta, D., Gupta, U., Hazelwood, K., Hock, A., Huang, X., Ike, A., Jia, B., Kang, D., ... Zaharia, M. (2020). MLPerf: An industry standard benchmark suite for machine learning performance. *IEEE Micro, 40*(2), 8–16. https://doi.org/10.1109/MM.2020.2974843
37- *Experimental demonstration of high-power thermal test vehicle using two-phase cooling for AI datacenters, 5G RAN, and edge compute nodes*. (2025). *2025 IEEE 75th Electronic Components and Technology Conference (ECTC)*. https://doi.org/10.1109/ECTC51687.2025.00355
38- Besiroglu, T., Bergerson, S. A., Michael, A., Heim, L., Luo, X., & Thompson, N. (2024). *The compute divide in machine learning: A threat to academic contribution and scrutiny?* arXiv. https://doi.org/10.48550/arXiv.2401.02452
39- W. Sun, A. Li, T. Geng, S. Stuijk and H. Corporaal, "Dissecting Tensor Cores via Microbenchmarks: Latency, Throughput and Numeric Behaviors," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 34, no. 1, pp. 246-261, 1 Jan. 2023, doi: 10.1109/TPDS.2022.3217824.
40- Cim, M., Topcu, B., & Kandemir, M. T. (2026). *Diagnosing FP4 inference: A layer-wise and block-wise sensitivity analysis of NVFP4 and MXFP4*. arXiv.https://doi.org/10.48550/arXiv.2603.08747
41- Cheng, S., et al. (2023). *Peta-scale embedded photonics architecture for distributed deep learning applications*. **Journal of Lightwave Technology, 41**(12), 3737–3749. https://doi.org/10.1109/JLT.2023.3276588