NVIDIA RTX BLACKWELL GPU ARCHITECTURE

Built for Neural Rendering

V1.01

Contents

List of Figures

List of Tables

INTRODUCTION

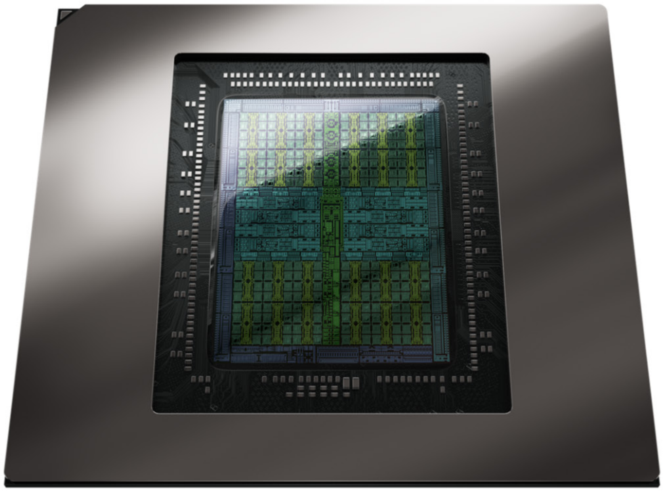
The NVIDIA RTX Blackwell architecture builds upon foundational AI technologies introduced in prior NVIDIA GPUs, enabling the next-generation of AI-powered gaming and professional applications. Blackwell will allow gaming, creative, and engineering applications to reach new levels of graphical realism, interactivity, and professional design capabilities.

GPU performance and image quality has been continually improving, even with Moore's Law coming to an end, by using neural rendering techniques. NVIDIA DLSS Super Resolution and Frame Generation technologies have increased frame rates dramatically, while often delivering native rendering-level image quality and generating the vast majority of pixels at a fraction of the cost of traditional rendering. Similarly, DLSS Ray Reconstruction (RR) drastically reduces the number of rays that need to be cast to create very high quality ray-traced or path-traced scenes by using advanced AI methods to denoise and reconstruct missing details. Future AI technologies will continue to augment and enhance visual quality at a much lower computational cost and memory footprint than conventional techniques.

New Blackwell AI-based Neural Rendering and Neural Shading technologies will accelerate developer usage of AI in their applications, including implementation and real-time usage of Generative AI-based rendering and simulation techniques. Generative AI will help game developers dynamically create varied terrains, implement more realistic physical simulations, and generate more complex character behaviors and backstories on-the-fly. Professional 3D design applications can use RTX Blackwell Generative AI capabilities to enable conversational workflows to produce multiple design options faster than ever, based on specified criteria, to more quickly iterate and fine-tune parameters to create optimal results. These and many other application scenarios will be supercharged by RTX Blackwell generative AI and Neural Rendering capabilities.

The NVIDIA Blackwell architecture family, including both RTX Blackwell and Blackwell datacenter-class GPUs, was named to honor David H. Blackwell, an amazing and inspiring American mathematician and statistician known for the Rao-Blackwell Theorem, and many contributions and advancements in probability theory, game theory, statistics, and dynamic programming.

.



- **Optimize for new neural workloads**
- **Reduce memory footprint**
- **New quality of service capabilities**
- **Energy efficiency**

Figure 1. RTX Blackwell Design Goals

The following **Key Features** are included in the NVIDIA RTX Blackwell architecture, and will be described in more detail in the sections below:

- **New SM features built for Neural Shading -** New RT Core and Tensor Core features described below enhance and accelerate neural rendering capabilities. The NVIDIA RTX Blackwell SM provides a doubling of integer math throughput per clock cycle compared to NVIDIA Ada GPUs, which can increase the performance of address generation workloads that are crucial for neural shading.

- **New MaxQ features for exceptional power efficiency -** RTX Blackwell incorporates many new MaxQ power management features. Advanced power gating and new split power rails provide fine-grained control and delivery of power to different on-chip subsystems. Clocks can adjust to dynamic workloads 1000x faster than our previous GPU architectures.

- **New 4th Generation RT Cores -** Significant improvements to the RT Core architecture were made in Blackwell, enabling new ray tracing experiences and neural rendering techniques.

- **New 5th Generation Tensor Cores -** Includes new FP4 capabilities that can double AI throughput while halving the memory requirements. Support for the new Second-Generation FP8 Transformer Engine used in our datacenter-class Blackwell GPUs is also included.

- **NVIDIA DLSS 4 -** The NVIDIA RTX Blackwell architecture features AI multi-frame generation that boosts DLSS 4's frame rates up to 2x over the previous DLSS 3/3.5, while maintaining or exceeding native image quality and providing low system latency.
- **RTX Neural Shaders -** Bring small neural networks into programmable shaders, enabling a new era of graphics innovation.
- **AI Management Processor (AMP) -** Enables multiple AI models, including speech, translation, vision, animation, behavior, and many others to share the GPU simultaneously with graphics workloads.
- **GDDR7 Memory -** GDDR7 is a new ultra-low voltage GDDR memory standard that uses PAM3 (Pulse Amplitude Modulation) signaling technology, enabling higher-speed memory subsystems and improvements in energy efficiency.
- **Mega Geometry Technology -** A new RTX technology aimed at dramatically increasing the geometric detail that is possible in ray-traced applications.

The GeForce RTX 5090, RTX 5080, RTX 5070 Ti, and RTX 5070 are the first NVIDIA GeForce graphics cards based on the new RTX Blackwell architecture. At the heart of the GeForce RTX 5090 is the GB202 GPU, which is the most powerful GPU in the NVIDIA RTX Blackwell family. The GeForce RTX 5080 is based on the GB203 GPU, and RTX 5070 uses the GB205 GPU. All three GPUs have been designed to deliver exceptional performance in their GPU class and provide groundbreaking new AI features for gamers and creator/professional users.

RTX Blackwell Neural Rendering Architecture

NVIDIA engineers set clear design goals for every new GPU architecture. With its revolutionary RT Cores, Tensor Cores, and DLSS technology, the NVIDIA Turing architecture laid the foundation for a new era in graphics, combining programmable shading, real-time ray tracing, and AI algorithms to deliver realistic and physically accurate graphics for games and professional applications. NVIDIA's Ampere architecture revamped the SM, enhanced the RT and Tensor Cores, included an innovative GDDR6X memory subsystem, improved DLSS capabilities, and provided tremendous overall performance gains. The NVIDIA Ada GPU architecture was designed to provide even higher performance and visual fidelity for ray tracing and AI-based neural graphics, adding new DLSS frame generation and ray reconstruction features. NVIDIA Ada marked the tipping point where ray tracing and neural graphics became mainstream.

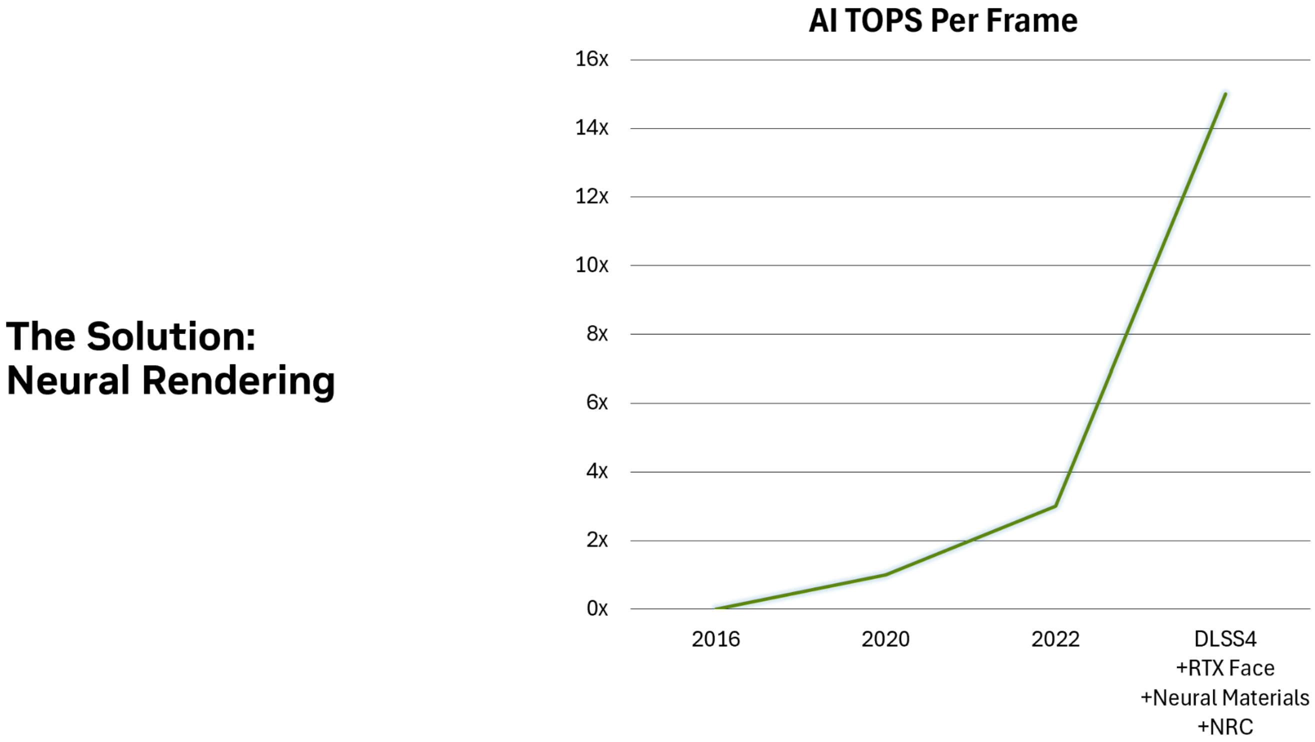


Figure 2. The Age of Neural Rendering Has Arrived - Significant AI TOPS Increase Per Frame

Image quality has been increasing faster than Moore's Law by using neural rendering, and such AI rendering techniques will continue to expand. DLSS has increased frame rates dramatically by generating a vast majority of pixels at a fraction of the cost of native rendering. DLSS-RR (Ray Reconstruction) has allowed for realistic lighting using path tracing by drastically reducing the number of rays that need to be cast and shaded.

Blackwell introduces DLSS 4 with multi-frame generation that further increases gaming performance while reducing latency. New neural shading techniques including RTX Neural

Materials, RTX Neural Faces, RTX Neural Radiance Cache (NRC), and new AI-based transformer models are more computationally efficient, while being able to reconstruct images at even better image quality. As Figure 2 above shows, an inflection point in rendering using neural techniques has been achieved- the age of neural rendering has arrived.

GB202 is the flagship of the RTX Blackwell GPU lineup and powers the GeForce RTX 5090 graphics card. The GB203 GPU is used in the GeForce RTX 5080 graphics card, and GB205 in the GeForce RTX 5070. These GPUs are based on the same underlying architecture and configured to serve different usage models and market segments.

The section below focuses on the GB202 GPU architecture. For further information on GB203 and GB205 specs, please consult *Appendix B, The Blackwell GB203 GPU* and *Appendix C, The Blackwell GB205 GPU*.

Blackwell GB202 GPU

The **full GB202 GPU** includes 12 Graphics Processing Clusters (GPCs), 96 Texture Processing Clusters (TPCs), 192 Streaming Multiprocessors (SMs), and a 512-bit memory interface with sixteen 32-bit memory controllers.

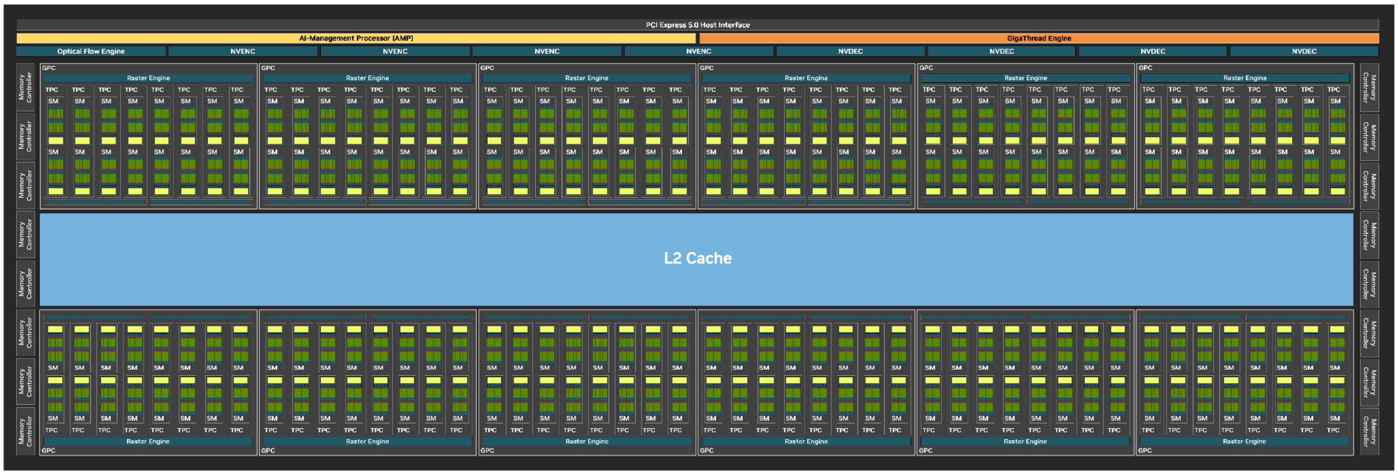


Figure 3. GB202 GPU block diagram (full chip).

Note: The GB202 GPU also includes 384 FP64 Cores (two per SM) which are not depicted in the above diagram. The FP64 TFLOP rate is 1/64th the TFLOP rate of FP32 operations. The small number of FP64 Cores are included to ensure any programs with FP64 code operate correctly. Similarly, a very minimal number of FP64 Tensor Cores are included for program correctness.

The full GB202 GPU includes:

- 24576 CUDA Cores
- 192 RT Cores
- 768 Tensor Cores
- 768 Texture Units

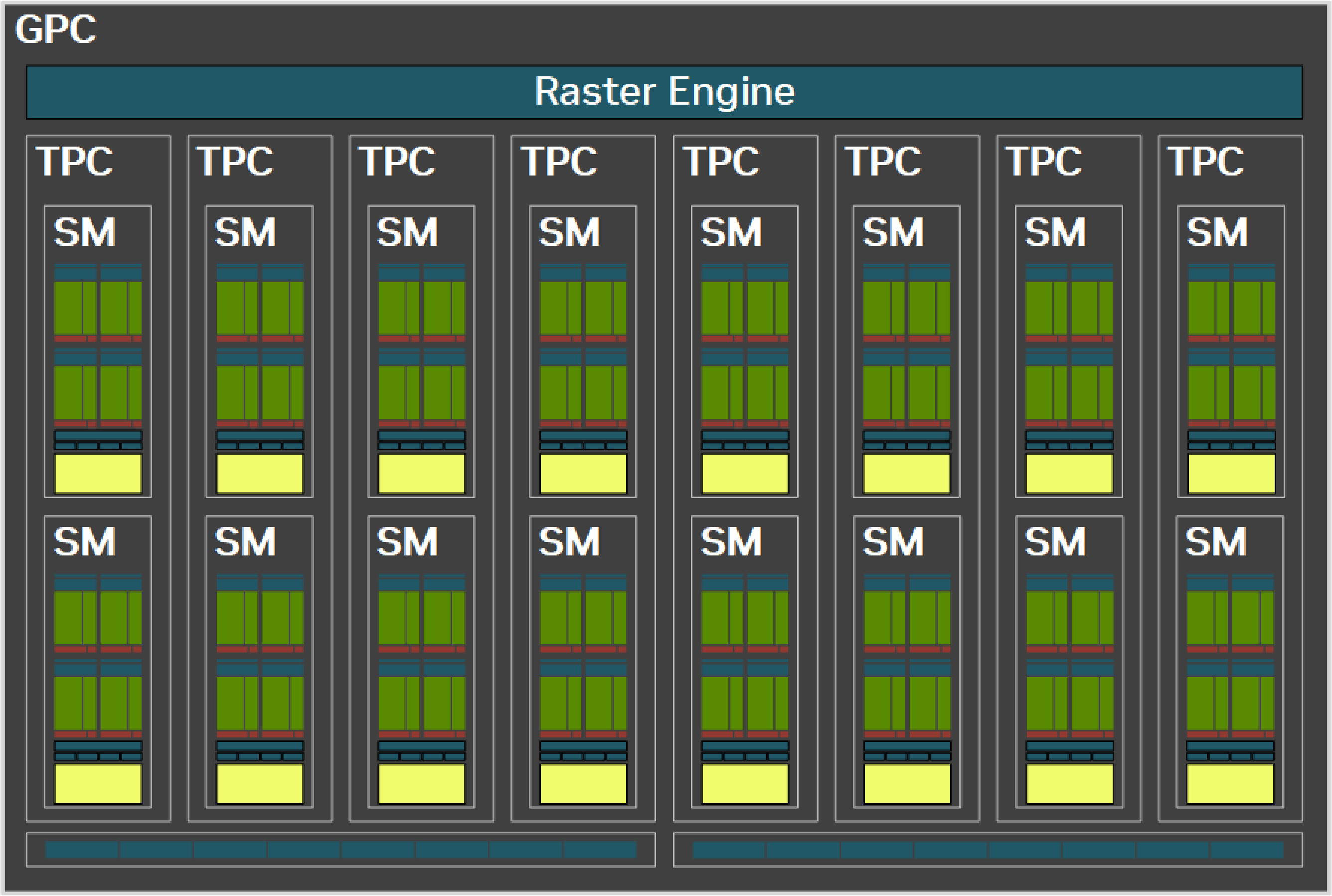


Figure 4. The Blackwell GPC with Raster Engine, 8 TPCs, 16 SMs, and 16 ROPs.

The GPC is the dominant high-level hardware block within all GB20x Blackwell family GPUs, with all of the key graphics processing units residing within a GPC. Each GPC includes a dedicated Raster Engine, two Raster Operations (ROPs) partitions, with each partition containing eight individual ROP units, and eight TPCs. Each TPC includes one PolyMorph Engine and two SMs.

The full GB202 GPU includes 128 MB of L2 cache, while the RTX 5090 specifically includes 96 MB of L2. All applications benefit from having such a large pool of fast cache memory available, and complex operations such as ray tracing (particularly path tracing) will yield great benefit.

SM Architecture

The NVIDIA Streaming Multiprocessor (SM) is a core component of the NVIDIA GPU architecture, playing a key role in the parallel processing capabilities of the GPU, enabling massive parallelism through its various cores (CUDA, Tensor, RT), efficient warp scheduling, memory management, and support for modern workloads like AI. Each full GB202 chip contains 192 SMs, and each SM includes128 CUDA Cores, one Blackwell Fourth-Generation RT Core, four Blackwell Fifth-Generation Tensor Cores, 4 Texture Units, a 256 KB Register File, and 128 KB of L1/Shared Memory, which can be configured for different memory sizes depending on the needs of the graphics and compute workloads.

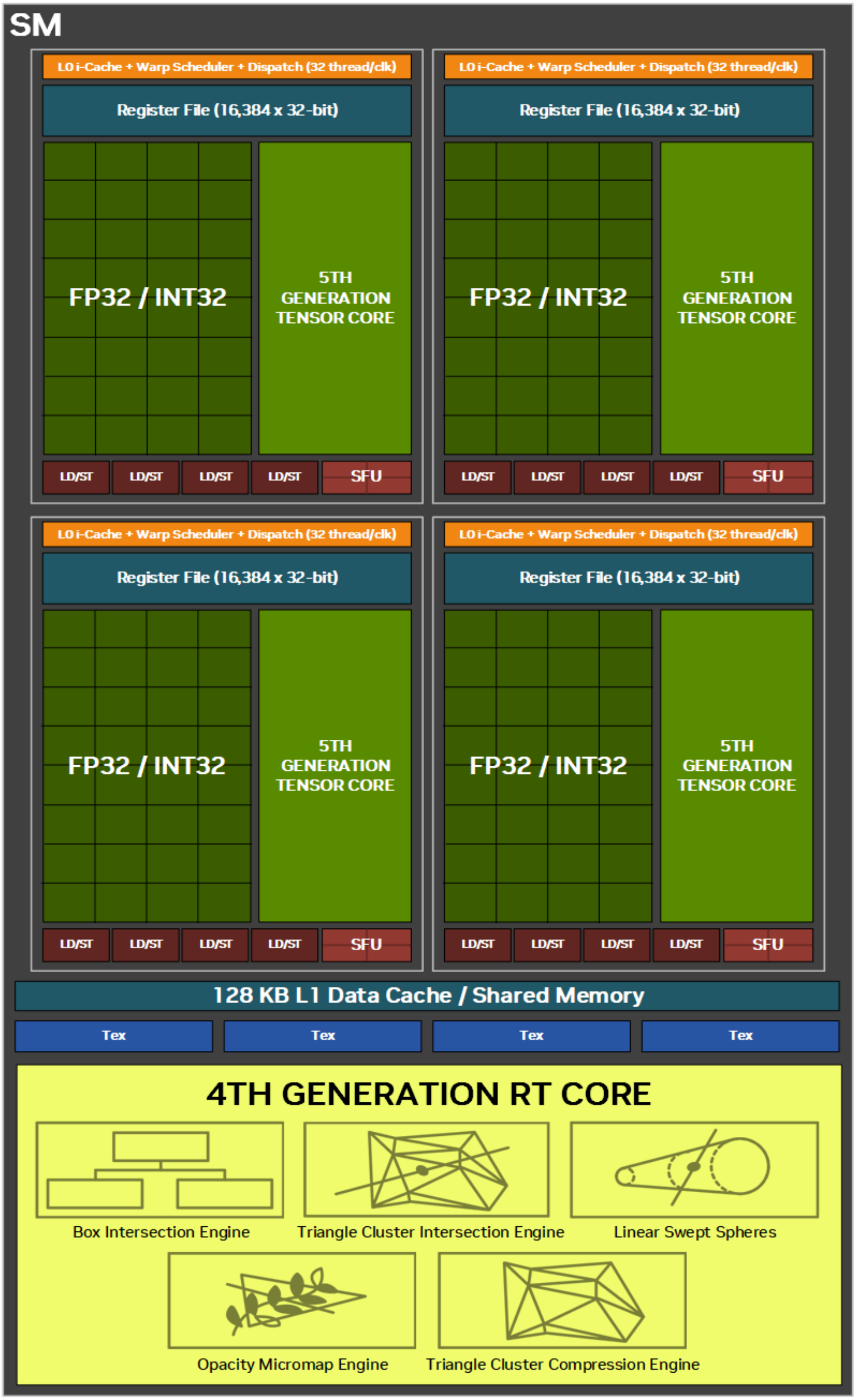
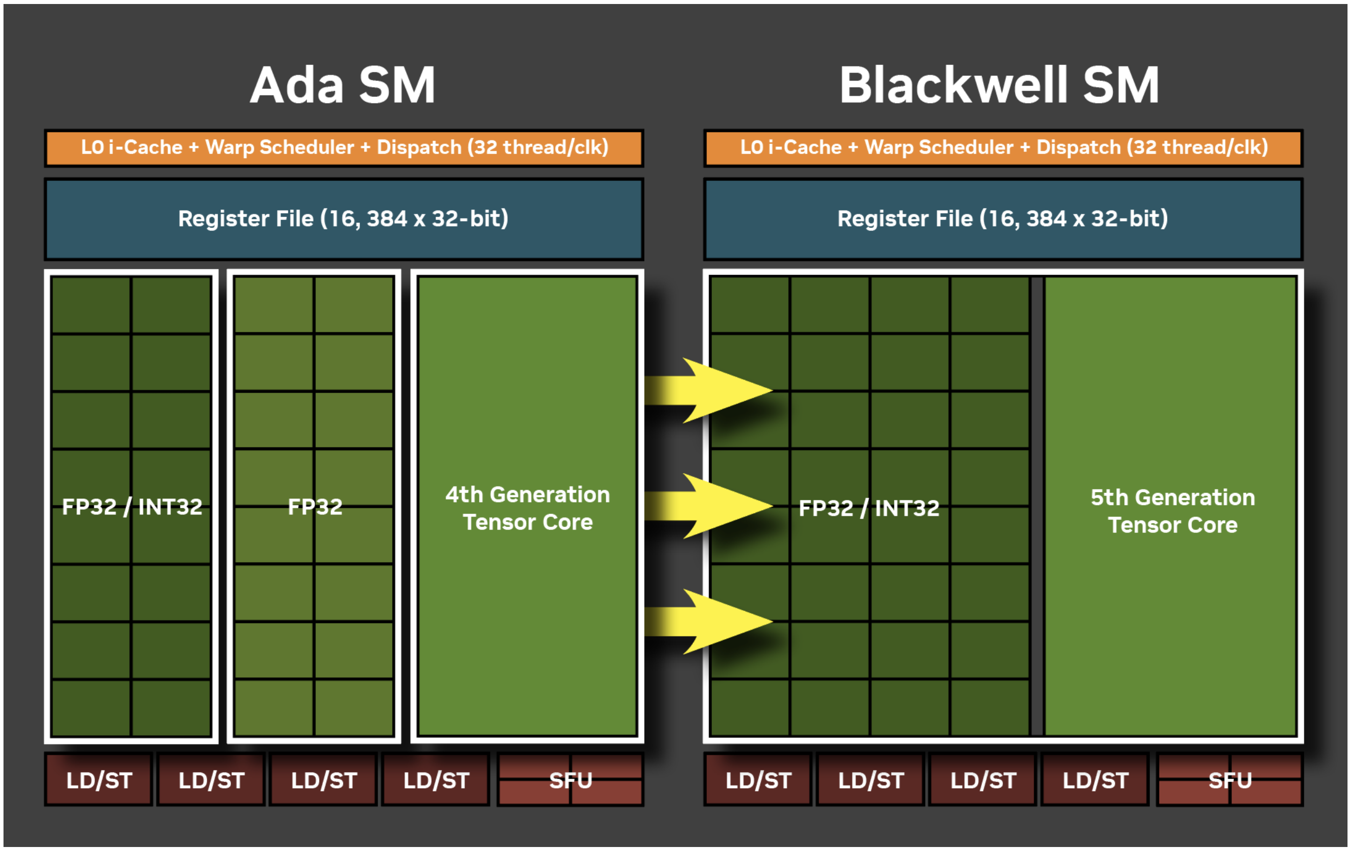

SM
L0 i-Cache + Warp Scheduler + Dispatch (32 thread/clk)
Register File (16,384 x 32-bit)
FP32 / INT32
5TH GENERATION TENSOR CORE
LD/ST
SFU
128 KB L1 Data Cache / Shared Memory
Tex
4TH GENERATION RT CORE
Box Intersection Engine
Triangle Cluster Intersection Engine
Linear Swept Spheres
Opacity Micromap Engine
Triangle Cluster Compression Engine

Figure 5. The Blackwell Streaming Multiprocessor (SM)

Note that the number of possible INT32 integer operations in Blackwell are doubled compared to Ada, by fully unifying them with FP32 cores, as depicted in Figure 6 below. However, the unified cores can only operate as either FP32 or INT32 cores in any given clock cycle. Figure 6 below shows how the SM architecture evolved between Ada and Blackwell.


Ada SM
Blackwell SM
L0 i-Cache + Warp Scheduler + Dispatch (32 thread/clk)
L0 i-Cache + Warp Scheduler + Dispatch (32 thread/clk)
Register File (16, 384 x 32-bit)
Register File (16, 384 x 32-bit)
FP32 / INT32
FP32
4th Generation Tensor Core
FP32 / INT32
5th Generation Tensor Core
LD/ST
LD/ST
LD/ST
LD/ST
SFU
LD/ST
LD/ST
LD/ST
LD/ST
SFU

Ada SM was designed & optimized for standard shaders. Blackwell SM was designed & optimized for neural shaders.

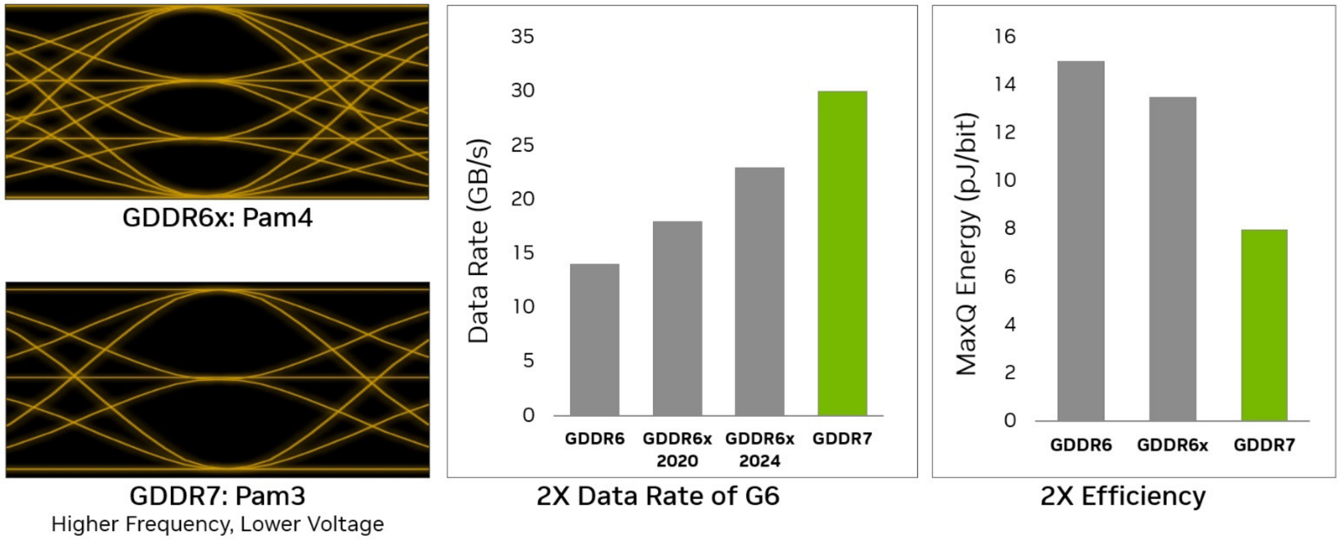
Figure 6. Ada SM vs Blackwell

The number of texture units has increased from 512 in GeForce 4090 to 680 in GeForce 5090. Texture units are responsible for handling texture mapping operations, performing tasks including fetching texels from textures, applying texture filtering, and handling texture coordinates. Texels represent the texture information including colors and patterns that are applied to 3D surfaces, defining the visual appearance of the texture that is applied to the surface of an object.

With the increase in texture units, the RTX Blackwell bilinear-filtered texel rates also increase. RTX 5090 delivers 1636.76 Gigatexels/sec, compared to 1290.2 Gigatexels per second in RTX 4090. Note that the RTX Blackwell SM also doubles the performance of point-sampling textures per cycle compared to Ada, which can speed up certain texture access algorithms, such as the Stochastic Texture Filtering (STF) used with the new Blackwell Neural Texture Compression methods described below.

GDDR7 Memory Subsystem

NVIDIA has worked closely with the DRAM industry for years, collaborating on DRAM architecture, circuit design, and signaling to enable the highest GPU memory speeds. With the launch of the Ampere GPU architecture, NVIDIA and Micron shipped the first GDDR6X devices and worked together to facilitate even higher memory speeds for Ada GPUs. The GeForce RTX 4080 shipped with 22.4 Gbps GDDR6X memory, the highest speed of any GPU with GDDR-based memory at the time, and the GeForce RTX 4090 offered 1 TB/sec of peak memory bandwidth.



GDDR6x: Pam4

GDDR7: Pam3
Higher Frequency, Lower Voltage

2X Data Rate of G6

2X Efficiency

Energy Efficiency reflects the average graphics application with 30% DRAM utilization

Figure 7. GDDR7 compared to prior generation GDDR6/6x

With Blackwell, NVIDIA is launching GDDR7, a new ultra-low voltage GDDR memory standard that uses PAM3 (Pulse Amplitude Modulation) signaling technology, enabling a substantial advancement in high-speed memory design. NVIDIA's work with the JEDEC technology association, the global leader in the development of standards for the microelectronics industry with over 360 member companies, helped to make PAM3 (Pulse Amplitude Modulation: 3 levels) the foundational high-frequency signaling technology for GDDR7 DRAM.

The GeForce RTX 5090 ships with 28 Gbps GDDR7 memory and delivers 1.792 TB/sec peak memory bandwidth, while the GeForce RTX 5080 ships with 30 Gbps GDDR7 memory, delivering 960 GB/sec of peak memory bandwidth.

The transition from PAM4 signaling (4 levels delivering 2 bits per cycle) in GDDR6X to PAM3 signaling (3 levels delivering 1.5 bits per cycle) in GDDR7, combined with an innovative pin-encoding scheme, allows GDDR7 to achieve a significantly enhanced signal-to-noise ratio (SNR). This evolution also doubles the number of independent channels with minimal I/O density overhead.

With increased channel density, improved PAM3 SNR, advanced equalization schemes, a reengineered clocking architecture, and enhanced I/O training, GDDR7 delivers substantially higher bandwidth. These advancements also enable notable improvements in energy efficiency, offering superior performance and extended battery life, particularly in power-constrained systems.

The GDDR7 implementation for RTX 50-Series boards includes support for Enhanced CRC for RAS (Reliability, Availability, Serviceability).

Table 1. GeForce RTX 5090 vs GeForce RTX 4090 vs GeForce RTX 3090 Basic Specs

Graphics Card	GeForce RTX 3090	GeForce RTX 4090	GeForce RTX 5090
GPU Codename	GA102	AD102	GB202
GPU Architecture	NVIDIA Ampere	NVIDIA Ada Lovelace	NVIDIA Blackwell
GPCs	7	11	11
TPCs	41	64	85
SMs	82	128	170
CUDA Cores / SM	128	128	128
CUDA Cores / GPU	10496	16384	21760
Tensor Cores / SM	4 (3rd Gen)	4 (4th Gen)	4 (5th Gen)
Tensor Cores / GPU	328 (3rd Gen)	512 (4th Gen)	680 (5th Gen)
GPU Boost Clock (MHz)	1695	2520	2407
RT Cores	82 (2nd Gen)	128 (3rd Gen)	170 (4th Gen)
RT TFLOPS	69.5	191	317.5
Frame Buffer Memory Size and Type	24 GB GDDR6X	24 GB GDDR6X	32 GB GDDR7
Memory Interface	384-bit	384-bit	512-bit
Memory Clock (Data Rate)	19.5 Gbps	21 Gbps	28 Gbps
Memory Bandwidth	936 GB/sec	1008 GB/sec	1792 GB/sec
ROPs	112	176	176

Pixel Fill-rate (Gigapixels/sec)	189.8	443.5	423.6
Texture Units	328	512	680
Texel Fill-rate (Gigatexels/sec)	555.96	1290.2	1636.76
L1 Data Cache/Shared Memory	10496 KB	16384 KB	21760 KB
L2 Cache Size	6144 KB	73728 KB	98304 KB
TGP (Total Graphics Power)	350 W	450 W	575 W
Manufacturing Process	Samsung 8 nm 8N NVIDIA Custom Process	TSMC 4nm 4N NVIDIA Custom Process	TSMC 4nm 4N NVIDIA Custom Process
PCI Express Interface	Gen 4	Gen 4	Gen 5

For the full list of GeForce RTX 5090 specifications, please see Appendix A at the back of this document.

Blackwell 5th Generation Tensor Cores

Tensor Cores are specialized high performance compute cores that are tailored for the matrix multiply and accumulate math operations that are used in AI and HPC applications. Tensor Cores provide groundbreaking performance for the matrix computations that are critical for both deep learning neural network training and inference operations.

The RTX Blackwell Tensor Cores support FP16, BF16, TF32, INT8, and Hopper's FP8 Transformer Engine. RTX Blackwell adds new support for FP4 and FP6 Tensor Core operations, and the new Second-Generation FP8 Transformer Engine, similar to our datacenter-class Blackwell GPUs.

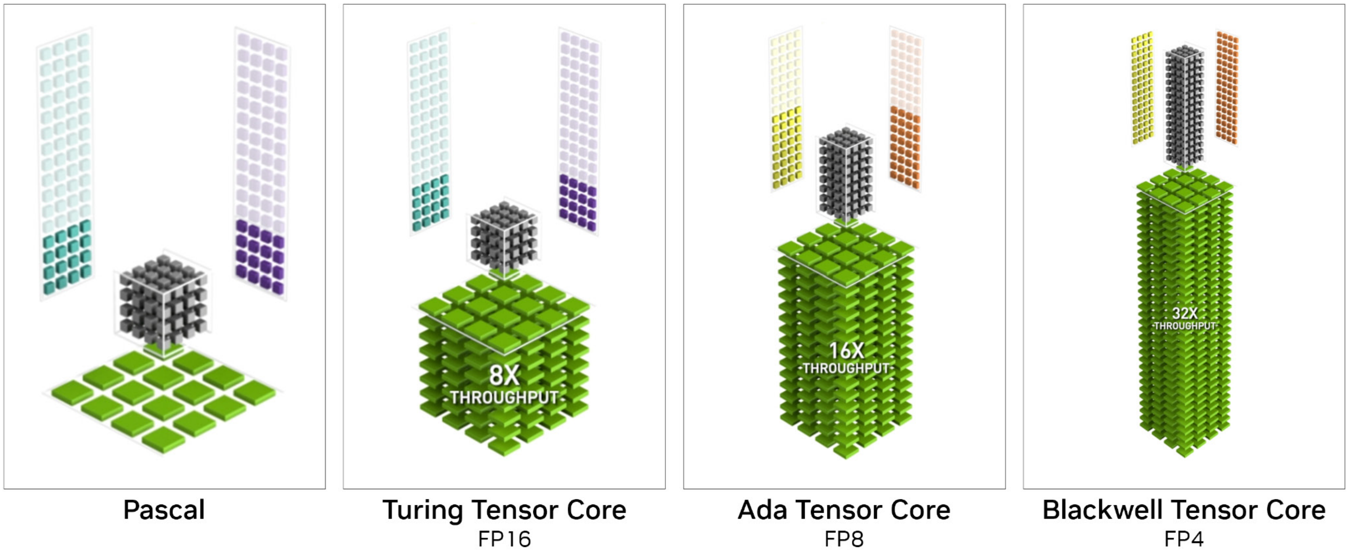
FP4 Support

Generative AI models have improved in capabilities since the first ones released in 2022. But the improvements have often come with an increase in parameters and size. As models grow in both compute and memory requirements, it can be difficult to run such models even on the latest hardware.

The GeForce RTX 50 Series includes support for the FP4 data format in its new Tensor Cores to help address this issue. FP4 provides a lower quantization method, similar to file compression, which decreases model sizes. Compared with FP16 precision — the default method that most models publish with — FP4 requires less than half of the memory, and 50 Series GPUs provide over 2x performance compared to the previous generation. FP4 allows virtually no loss in quality with advanced quantization methods offered by the NVIDIA TensorRT Model Optimizer.

For example, Black Forest Labs' FLUX.dev model at FP16 requires over 23GB of VRAM, meaning it can only be supported by the GeForce RTX 4090, RTX 5090, and our professional GPUs. With FP4, FLUX.dev requires less than 10GB, so it can run locally on more GeForce RTX GPUs.

With a GeForce RTX 4090 with FP16, the FLUX.dev model can generate images in 15 seconds with 30 steps. With a GeForce RTX 5090 with FP4, images can be generated in just over five seconds.



Pascal

Turing Tensor Core
FP16

Ada Tensor Core
FP8

Blackwell Tensor Core
FP4

Figure 8. Blackwell 5th Generation Tensor Cores with FP4, double throughput of FP8

Blackwell 4th Generation RT Cores

Games today are more realistic than ever, with richly detailed worlds and high-quality visual effects. Ray tracing enables physically accurate lighting, shadows, and reflections, creating virtual environments that closely mirror reality. Developers also enhance games by increasing geometric detail and using various advanced shading techniques. NVIDIA engineers have enhanced several important features of the RT Core to enable high-performance ray tracing of highly complex geometry.

For some background, the RT Cores in Turing, Ampere, and Ada GPUs include dedicated hardware units for accelerating Bounding Volume Hierarchy (BVH) data structure traversal, and performing both ray-triangle intersection and ray-bounding box intersection testing calculations. By providing dedicated resources for these core ray tracing functions, work is offloaded from the SM, freeing it up to perform other pixel, vertex, and compute shading tasks.

Ray-triangle intersection testing is a computationally expensive operation that is performed at high frequency when rendering a ray-traced scene. The Fourth-Generation RT Core in the

Blackwell architecture provides double the throughput for **Ray-Triangle Intersection Testing** over Ada.

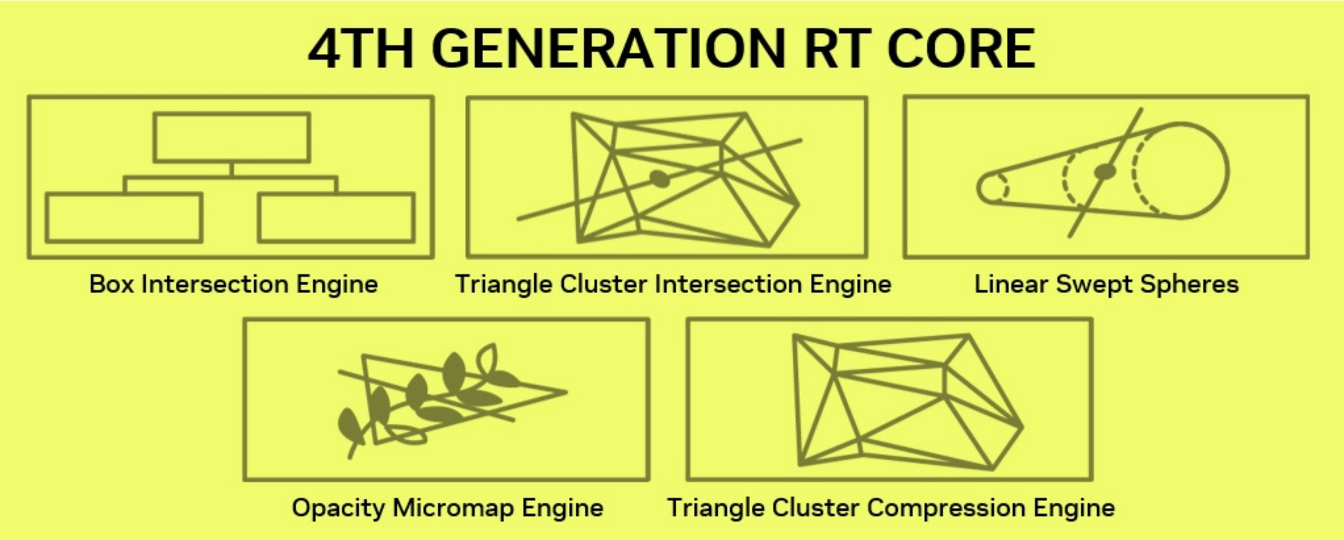


Figure 9. New Fourth-Generation RT Core Block Diagram (RTX Blackwell Architecture)

In addition to the above-specified functions, the RT Cores found in both Ada and Blackwell GPUs include a dedicated unit known as the **Opacity Micromap Engine**. The Opacity Micromap Engine evaluates Opacity Micromaps and directly alpha-tests geometry to significantly reduce shader-based alpha computations. New **Mega Geometry** technology provides RTX-accelerated ray tracing of triangle cluster-level structures. The new Blackwell RT Core includes a **Triangle Cluster Intersection Engine,** which further accelerates ray tracing of Mega Geometry, while also including standard ray-triangle intersection testing**.** Blackwell also adds **Linear Swept Spheres** as a hardware-accelerated path to ray trace fine geometry like hair. All are described below.

Mega Geometry

Mega Geometry is a new RTX technology aimed at dramatically increasing the geometric detail that is possible in ray-traced applications. In particular, Mega Geometry enables game engines such as Epic's Unreal Engine 5, which employ modern level-of-detail (LOD) systems like Nanite, to ray trace their geometry at full fidelity. Falling back to low-resolution proxies for ray-traced effects is no longer needed, enabling new levels of quality for shadows, reflections, and indirect illumination. Mega Geometry also helps bring techniques previously reserved for production rendering, such as displaced subdivision surfaces, to the domain of real-time ray tracing.

Level-of-Detail

There are two main hurdles that prevent straightforward integration of ray tracing into systems like Nanite. Mega Geometry consists of new RTX API extensions, along with a high-performance driver implementation and specific optimizations in Blackwell's RT Cores, that address both of these challenges:

1. **Cluster-based LOD updates.** As an object moves closer or further away from the camera, a game engine will typically adjust the object's level of detail. That is, the triangle count in the rendered mesh changes over time. Many traditional methods precompute some small number of meshes that represent different LOD levels for a given object. Systems like Nanite update LODs by incrementally replacing small batches of about 128 triangles, known as clusters. The configuration of clusters that make up the final rendered mesh can change frequently, for example every frame, leading to the desired smooth LOD transitions without popping. In order to ray trace a mesh however, a separate data structure, the Bounding Volume Hierarchy (BVH), has to be constructed. The numerous BVH builds that a Nanite-style system would trigger across a large number of high-polycount objects every frame would overload existing ray tracing implementations, making the system infeasible for real-time applications like games.

   Mega Geometry provides new BVH construction capabilities that adopt clusters of triangles as first-class primitives. New Cluster-level Acceleration Structures (CLAS) can be generated out of spatially compact batches of up to 256 triangles. A collection of CLAS is then used as input to construct a final BVH. CLAS can be generated on demand, e.g. when an object is loaded from disk, and then cached for future frames. Because each CLAS represents a collection on the order of 100 triangles, the processing required by subsequent BVH builds is reduced by two orders of magnitude compared to classic triangle-based methods. Consequently, a game engine can budget for many more BVH builds per frame, and handle cluster-LOD switches by simply reconstructing the BVHs of affected objects from CLASes.

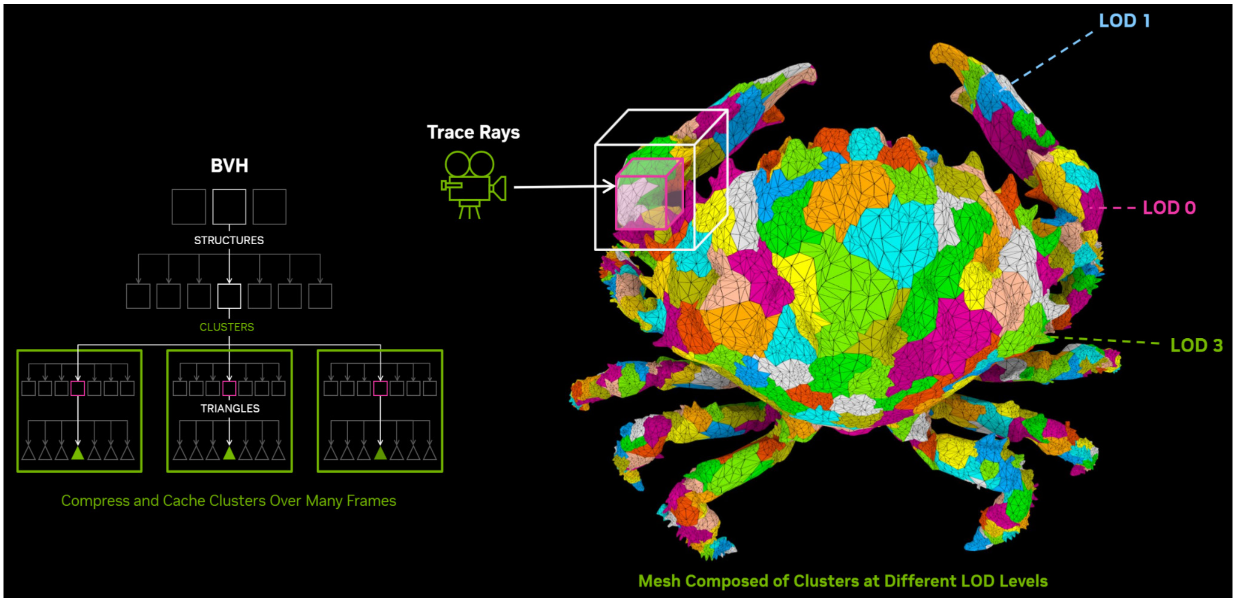

BVH
STRUCTURES
CLUSTERS
TRIANGLES
Compress and Cache Clusters Over Many Frames
Trace Rays
LOD 1
LOD 0
LOD 3
Mesh Composed of Clusters at Different LOD Levels

Figure 10. BVH and Mesh Using Clusters

As a further improvement to existing ray tracing solutions, all Mega Geometry APIs are designed to be fully batched, with their input parameters driven entirely from GPU memory. This lets the game engine run logic like LOD selection, animation, culling, and others efficiently on the GPU, while minimizing roundtrips to the CPU. With effective use of Mega Geometry APIs, an application can practically eliminate CPU overhead related to BVH management.

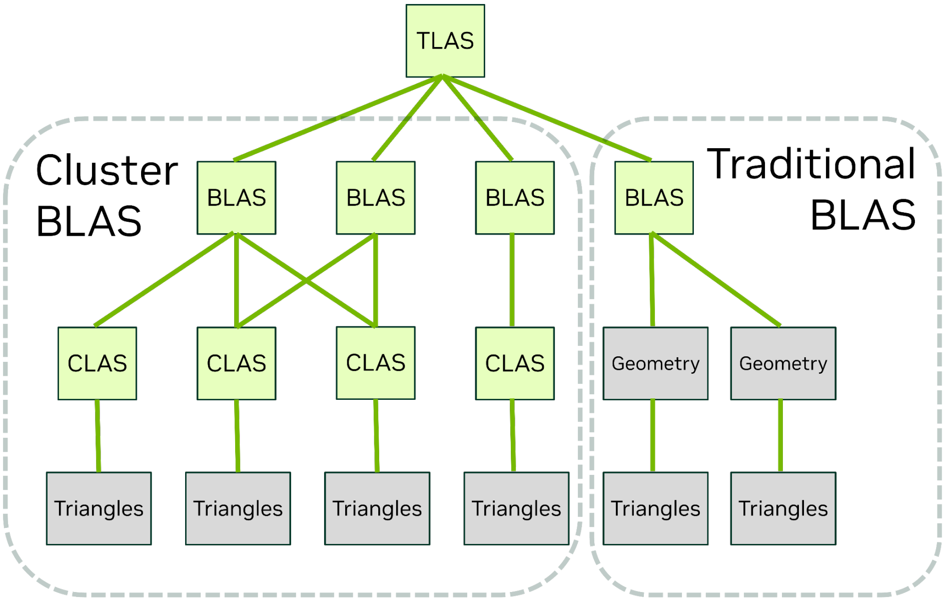


Figure 11. TLAS/BLAS Acceleration Structures and Cluster BLAS

2. **High object counts.** Game engines that emphasize high geometric detail tend to have a desire for ever larger object counts in their scenes. Without Mega Geometry, an application has to build a TLAS from all the objects in the scene, every frame. This works well with object counts up to a few thousand, but becomes prohibitively expensive as world sizes scale up.

Figure 12. NVIDIA "Zorah Demo" using Mega Geometry and other new Blackwell Tech

To address this, Mega Geometry introduces a new type of TLAS called a Partitioned Top-Level Acceleration Structure (PTLAS). Instead of building a new TLAS from scratch every frame, PTLAS is able to exploit the fact that most objects in the scene are static from one frame to the next. The application manages a persistent PTLAS object by aggregating objects into partitions and updating only those that have changed. For example, a game might put various sectors of a static game world into their own partitions, while keeping dynamic objects separate in a "global partition" that is rebuilt every frame. The fewer partition updates that are requested, the larger the runtime savings compared to traditional TLAS.

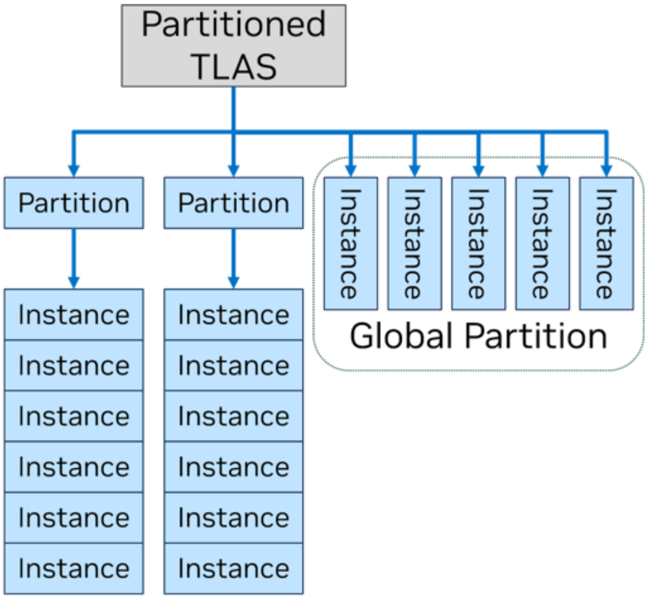


Figure 13. Partitioned Top-Level Acceleration Structure (PTLAS)

Subdivision Surfaces

While one of the main goals of Mega Geometry has been to enable a first-class combination of ray tracing with modern level-of-detail systems of game engines, its applications are broader than that specific use case. The flexible, GPU-driven generation of clusters, along with blistering fast BVH construction, opens up many new possibilities for advanced geometry techniques. One example of that is subdivision surfaces.

Subdivision surfaces are a type of geometry representation commonly used in film and other production rendering workflows. Iterative refinement of a quad-based mesh with a subdivision rule like Catmull-Clark, often with additional application of displacement maps, results in smoothly rendered surfaces while maintaining high modeling efficiency and animation friendliness.

Fast ray tracing of subdivision surfaces is usually achieved by tessellating them into triangles. For animations or changing viewpoints, new tessellations are needed each frame, leading to large numbers of expensive BVH builds. Mega Geometry makes it possible for the application to map its tessellation process directly to cluster generation, and construct BVHs from the resulting CLASes extremely quickly. This method unlocks unprecedented real-time performance for ray tracing of animated displaced subdivision surfaces.

Mega Geometry API and Architecture Support

The functionality surrounding the management of BVHs is a fundamental pillar of any ray tracing system. Mega Geometry is a core technology that takes BVH capabilities to the next level, and empowers applications to invent more creative and efficient geometry pipelines than ever. As such, Mega Geometry will be supported on a broad range of APIs and hardware:

APIs: Mega Geometry is available in all ray tracing APIs supported by NVIDIA:

- DirectX 12 (DXR) is extended through NVAPI to support clusters and PTLAS
- Vulkan adds vendor extensions for clusters and PTLAS
- OptiX 9.0 adds native support for clusters

GPU Architectures: Mega Geometry is supported on all RTX GPUs, starting with Turing.

Blackwell RT Core Enhancements for Mega Geometry

Blackwell's 4th generation RT Cores are purpose-built for Mega Geometry. Special cluster engines in hardware implement new compression schemes for geometry and BVH data, while delivering up to 2x the ray-triangle intersection rate of Third-Generation RT Cores. As a result, Blackwell reduces VRAM footprint of typical use cases like Nanite scenes by several hundred MB.

Linear Swept Spheres (LSS)

Various flavors of curve primitives are commonly used by renderers to depict hair, fur, grass, and other strand-like objects. For raytracing, those primitives are generally implemented in software using custom intersection shaders. However, ray-curve intersection testing is computationally intensive, limiting the use of curves in real-time ray-traced rendering, and extending render times for offline renderers.

An alternative approach for real-time scenarios is to use relatively crude approximations for hair, such as textured cards, but this comes at the expense of image quality. A better, but more expensive, method is to model individual strands with triangles. For example, one such technique is Disjoint Orthogonal Triangle Strips (DOTS), which uses a mesh of triangle strips arranged in a grid-like, disjoint pattern where the triangle strips are independent of each other and do not share vertices. While higher quality than cards, the disjoint pattern produces edge artifacts that lead to visible shortcomings in renderings, as shown below in Figure 14.

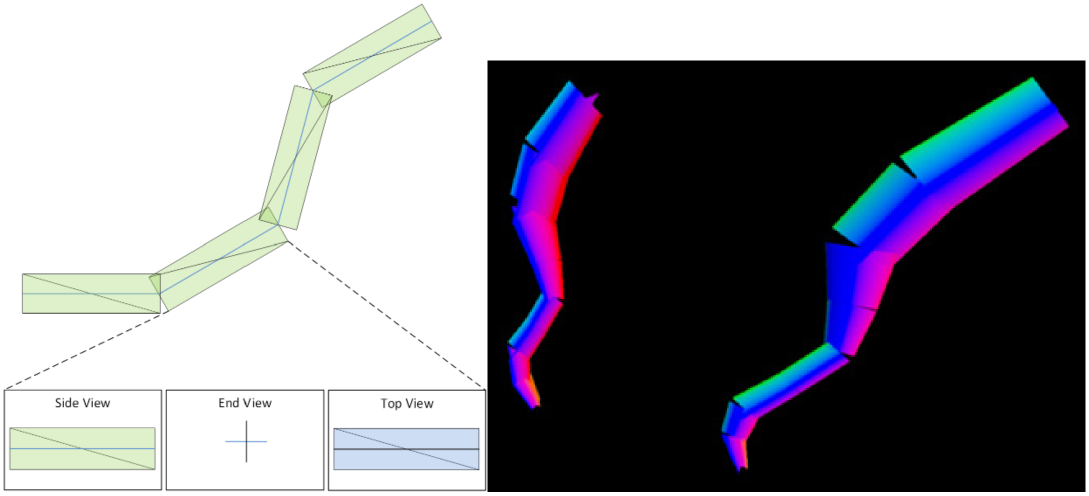


Figure 14. Sequence of Disjoint Orthogonal Triangle Strips (DOTS)

Blackwell's RT Core introduces hardware-based ray intersection testing support for a new primitive called **Linear Swept Spheres (LSS)**. A linear swept sphere is similar to a tessellated curve, but is constructed by sweeping spheres across space in linear segments. The radii of the spheres may differ between start and end point of each segment, allowing flexible approximation of various strand types. As a special case of LSS, the Blackwell hardware primitive also supports spheres directly (without a swept linear segment), which is useful for applications like particle systems.

Common use cases, like the rendering of hair on humans, are about 2x faster with LSS compared to DOTS, while also requiring about 5x less VRAM to store the geometry.

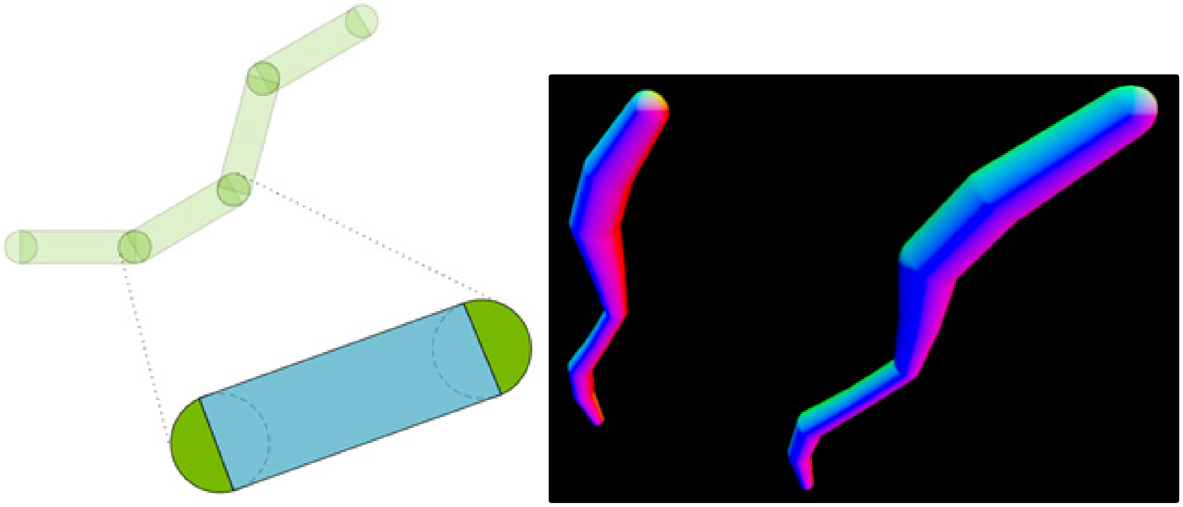


Figure 15. Sequence of Linear-Swept Spheres (LSS)

Shader Execution Reordering (SER) 2.0

SER is a powerful technology that lets ray tracing applications efficiently reorganize the massively parallel threads on a GPU for maximum hardware utilization. Dynamically reordering work with SER is particularly effective in challenging ray tracing workloads that exhibit large amounts of execution or memory access divergence, such as path tracing. Because threads that coherently execute neural workloads can be sent directly to the Tensor Cores, SER also significantly accelerates neural shading.

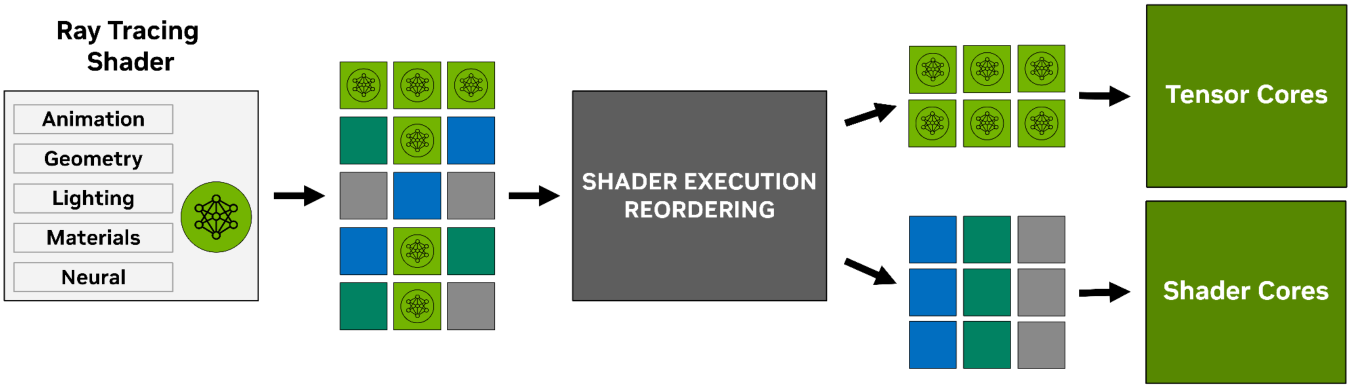


Figure 16. Shader Execution Reordering (SER), conceptual diagram

First introduced in the Ada architecture, SER on Blackwell is enhanced by several innovations to both hardware and software that further improve the feature's effectiveness. The core reorder logic of SER on Blackwell is twice as efficient, reducing reordering overhead and increasing its precision. The higher precision results in smarter coherence extraction and lets developers provide more application-specific knowledge to reorder operations, in turn increasing overall workload performance.

SER is fully controlled by applications through a small API, allowing developers to easily apply reordering where workloads benefit most. The API additionally introduced new flexibility around the invocation of ray tracing shaders to the programming model, enabling more streamlined ways to structure renderer implementations while taking advantage of reordering. Several game titles that feature path tracing, as well as a number of production rendering packages, already take advantage of SER. These applications will benefit directly from the Blackwell SER enhancements without any code changes.

AI Management Processor (AMP)

The AI Management Processor (AMP) is a fully programmable context scheduler on the GPU designed to offload scheduling of GPU contexts from the system CPU. AMP enhances the scheduling of GPU contexts in Windows to more efficiently manage different workloads running on the GPU. A GPU context encapsulates all the state information the GPU needs to execute one or more tasks. Multiple contexts can be used for better task isolation when running multiple tasks, and ensuring that multiple applications can share the GPU simultaneously without conflicts. An example could be coordinating and scheduling asynchronous AI model workloads like NVIDIA Avatar Cloud Engine (ACE) with its speech, translation, vision, animation, and behavior models, and G-Assist, running simultaneously with other graphics workloads on the GPU.

The AI Management Processor is implemented using a dedicated RISC-V processor located at the front of the GPU pipeline, and it provides faster scheduling of GPU contexts with lower latency than prior CPU-driven methods. The Blackwell AMP scheduling architecture matches the Microsoft architectural model that describes a configurable scheduling core on the GPU through Windows Hardware-Accelerated GPU Scheduling (HAGS), introduced in Windows 10 (May 2020 Update). HAGS allows the GPU to handle its own memory management more efficiently, reducing latency and potentially improving performance in games and other graphics-intensive applications.

The role of AMP is to take over the responsibility of the CPU's scheduling of GPU tasks, reducing dependency on the system CPU, which is often a bottleneck for game performance. In fact, allowing the GPU to manage its own task queue can lead to lower latency because of less back-and-forth communication between the GPU and CPU. This allows smoother frame rates in games, and better multitasking in Windows because the CPU is less burdened.

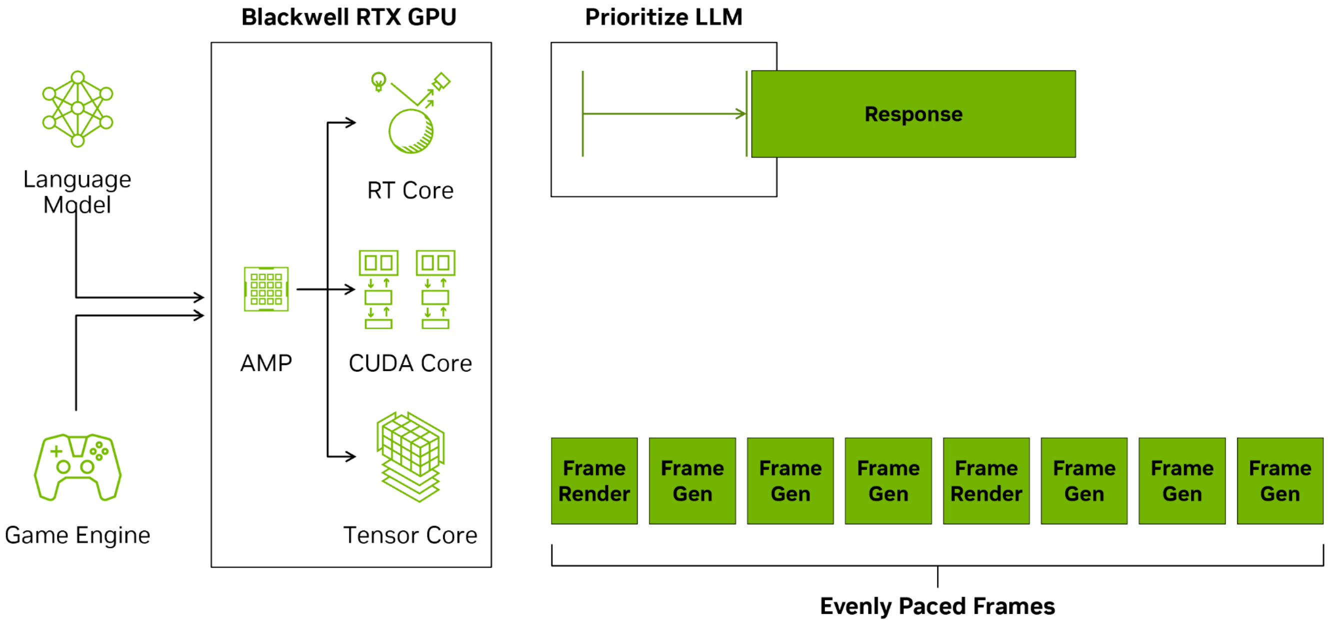


Figure 17. AI Management Processor (AMP) Schedules AI / Graphics Workloads

Essentially, AMP is used to coordinate, schedule fairly, and ensure a smoother gaming experience without performance drops. With LLMs, it does this by reducing the time to first response, and with games, it prioritizes work with the game engine to prevent stuttering. By delivering work at more predictable times, AMP can significantly improve quality of service depending on workloads.

New RTX Blackwell Video and Display Features

While Ada and previous GPU architectures offered support for both 4:4:4 and 4:2:0 chroma formats in H.264 and H.265 video, Blackwell adds hardware encode and decode support for 4:2:2 chromasampled video.

Video files utilize a YUV color format. Instead of storing color as Red, Green, and Blue (RGB) values, color is stored as Luminance (Y), Blue-difference Chroma (U), and Red-difference Chroma (V).

Chromasampling takes advantage of the fact that the human eye is more sensitive to changes in luminance than chrominance. In YUV 4:4:4 video, each channel retains its full value in all channels; however, this results in larger file sizes, and higher bandwidth required to transfer the video data. Chromasampling reduces the storage and bandwidth requirements by storing less information in the video chrominance channels. For YUV 4:2:0 video, full information is retained in the luminance channel, but the two chrominance channels contain only 25% of the original color information. This results in each video frame requiring half the data of an uncompressed 4:4:4 video frame, with the tradeoff being a loss of color information. This color loss does not imply low image quality, standards from Blu-Ray through to HDR10 and streaming platforms today distribute content to their audiences in a 4:2:0 format.

In the camera and while editing and color correcting, before final color choices have been made, YUV 4:2:2 strikes a balance between retaining more color information and reduction of file size and bandwidth requirements. In YUV 4:2:2 video, the full luminance value is retained, and half of the original chrominance color information is retained. A 4:2:2 compressed video frame requires only two-thirds the data requirements of an uncompressed 4:4:4 video frame, but provides 2x the color resolution compared to a 4:2:0 color compressed frame.

Since YUV 4:2:2 chromasampling offers lower data requirements than 4:4:4, but higher color accuracy than 4:2:0 it has become a popular option for higher end prosumer and professional video cameras. However, software based 4:2:2 decoding can place a high load on system CPUs, making 4:2:2 challenging to work with.

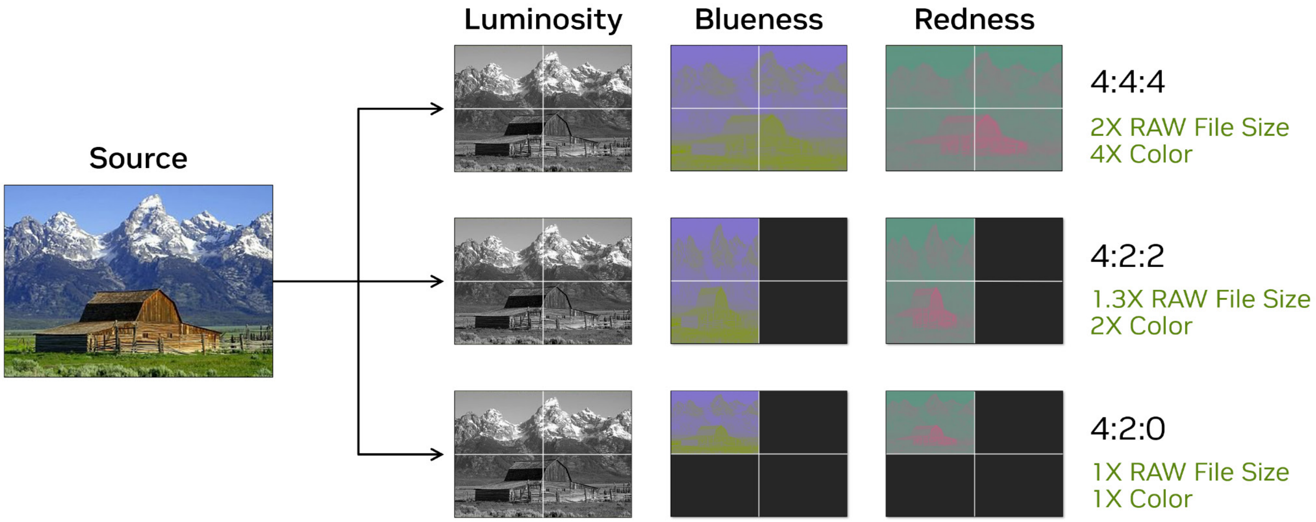


Figure 18. 4:2:2 Provides 2X the Color Information for 1.3X the RAW File Size of 4:2:0

The additional color information retained by 4:2:2 over 4:2:0 can be especially helpful for HDR content, and for maintaining fine details like text or fine lines or workflows where the source will get color corrected over and over again like when color grading.

Ninth-Generation NVENC

The new ninth-generation NVENC encoder in Blackwell improves quality for AV1 and HEVC by 5% BD-BR PSNR, and adds support for 4:2:2 H.264 and HEVC encoding. There's also a new AV1 Ultra High Quality (UHQ) mode that takes additional time and provides an extra 5% improvement for the best quality possible. (Note that AV1 UHQ will also be made available on RTX 40 Series GPUs using their AV1 encoders and additional software support, but will provide lower quality than Blackwell.)

The chart below illustrates the generational improvements to the encoder for AV1, and how combining them with the new AV1 UHQ mode can yield up to 15% BD-BR PSNR improvements. The gains are even larger when using the VMAF metric from Netflix - a metric designed to capture the actual subjective improvements.

Table 2. Data measured on RTX 5090 and 4090 using 4K60.

BD-Rate savings over ADA		
BD-BR PSNR	**RTX 5090 AV1**	**RTX 5090 AV1 + UHQ**
Natural Content	+5%	+15%
Gaming	+4%	+10%
BD-BR VMAF	**RTX 5090 AV1**	**RTX 5090 AV1 + UHQ**
Natural Content	+10%	+18%
Gaming	+9%	+14%

The GeForce RTX 5090 GPU supports up to three encoders and two decoders, boosting export speeds by over 50% gen-over-gen, and an impressive 4x compared to the RTX 3090 GPU with a single encoder.

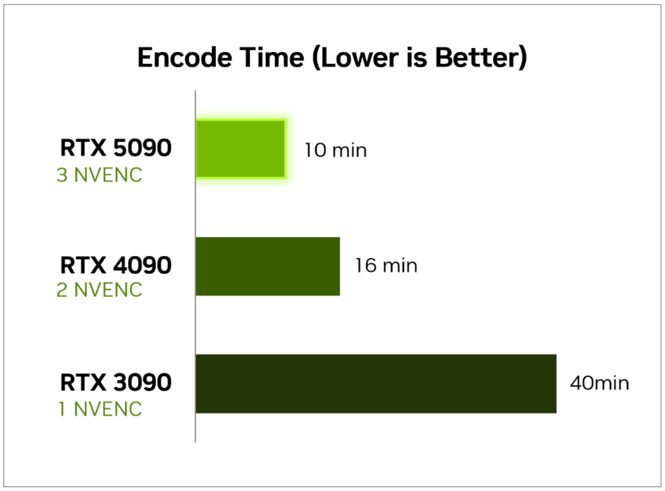


Figure 19. Ninth-Generation NVENC Encoder Improves Encoding Speed.

Sixth-Generation NVDEC

In addition to NVENC, Blackwell GPUs also include an improved sixth-generation hardware decoder (NVDEC) with 2x faster H.264 decoding (matching HEVC and AV1 decode speeds), and also 4:2:2 H.264 and HEVC decode support.

DisplayPort 2.1b

Blackwell GPUs introduce support for DisplayPort 2.1b providing up to 80 Gbps of bandwidth utilizing the UHBR 20 (Ultra High Bit Rate @ 20 Gbits/sec per lane) transmission mode. DisplayPort 2.1b UHBR 20 enables running high resolution displays utilizing the highest possible refresh rates: 8K (7680x4320) @ 165Hz (requires DSC), and 4K (3840x2160) @ 480Hz (requires DSC). Note that the highest link rates require a DP80LL certified cable.

Blackwell Max-Q Power Efficiency Improvements

The Max Q philosophy involves extracting as much performance as possible from a platform power budget and allowing the GPU to quickly enter deeper power states to save as much power as possible when the GPU is idle.

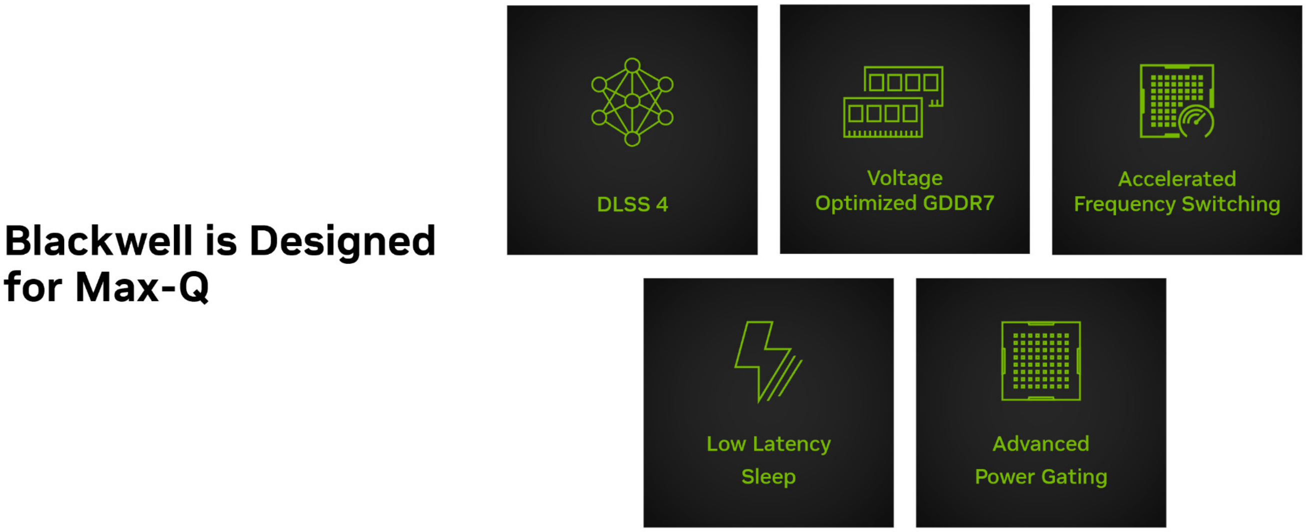


Figure 20. New Max Q Power Efficiency Innovations to Improve Battery Life.

Advanced Power Gating

The problem with going from an active power state to a very deep power state is that the deeper the power state, the more time it takes both to enter and exit that state. Blackwell reduced latencies to enter and exit different power states. It also has the most advanced power gating of different units on chip that we’ve ever built, with multiple new levels of gating allowing very fine grain control of power.

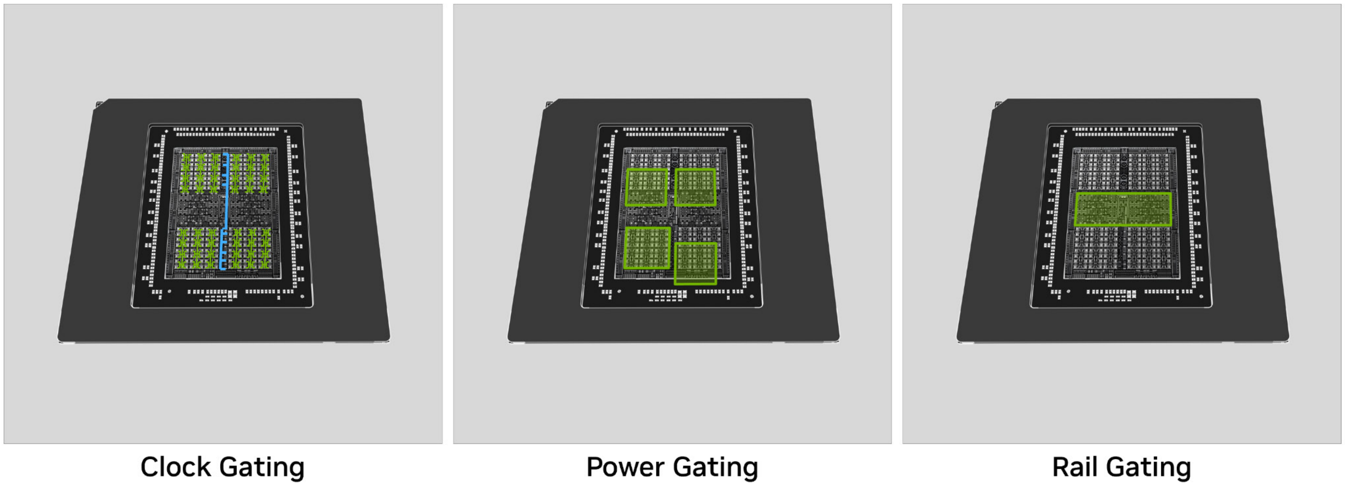


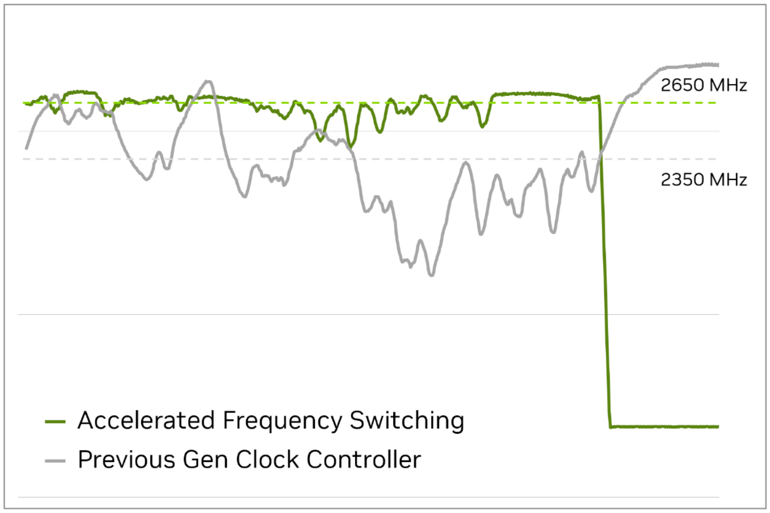
Figure 21. Advanced Clock, Power, and Rail Gating Provide Fine-Grain Control of Power.

New clock gating capabilities allow entire clock trees to be shut off very quickly, saving dynamic power even in regions of work where only part of the chip is idle, or where idleness is so short it would typically be considered “active”. For Blackwell one of the big focuses was on memory power management to achieve peak efficiency by leveraging GDDR7’s fast-wakeup clocking architecture. Now the entire memory clock tree can be gated for the first time.

A new voltage rail has been added to provide power separately to both the GPU Cores and memory system Separate rails allow independent voltage control of large areas of the chip which can be optimized per-workload, which boosts performance. It also allows Blackwell to shut down unused portions of the chip during small periods of idle, reducing leakage power. With the Blackwell design, rail-gated states can be entered at a frame granularity, which is especially helpful for battery gaming and creating. The separate power rails allow lowering of power when the GPU is idling, by turning off the GPU cores when they are not needed, greatly increasing the overall efficiency of computers that rely on power management, such as laptops.

Accelerated Frequency Switching

Blackwell incorporates the largest overhaul in clock architecture in over a decade. With it, clocks can adjust to dynamic workloads 1000x faster than previous GPU architectures, allowing Blackwell to quickly respond to the dynamic nature of GPU workloads and shift clock speeds up or down based on workload for best performance and power efficiency. Previously, clocks were effectively locked at the same frequency throughout the generation of a frame.



- 1000x Faster Clock Responsiveness
- Higher SM efficiency through rapid clock adjustments in dynamic workloads

GPU Clock

Accelerated frequency switching can adjust clocks to dynamic workloads 1000x faster than before.

Figure 22. Accelerated Frequency Switching

Accelerated frequency switching allows the full performance of the GPU to be realized within a given power budget. Also, by adapting quickly to short idle timeframes – i.e. the gaps between chunks of work in a frame sent to the GPU from the CPU - power is also saved, which frees up the GPU to burst to higher clocks during the non-idle periods, and the result is free performance.

Low Latency Sleep

Blackwell's low power states are faster to enter, enabling more time to be spent saving power, and by taking advantage of Advanced Power Gating, are able to progressively power gate the chip quickly - saving more power faster.

In the case of the deepest sleep state, Blackwell is 10x faster to enter sleep than Ada, enabling much more power savings in the lowest-power sleep state.

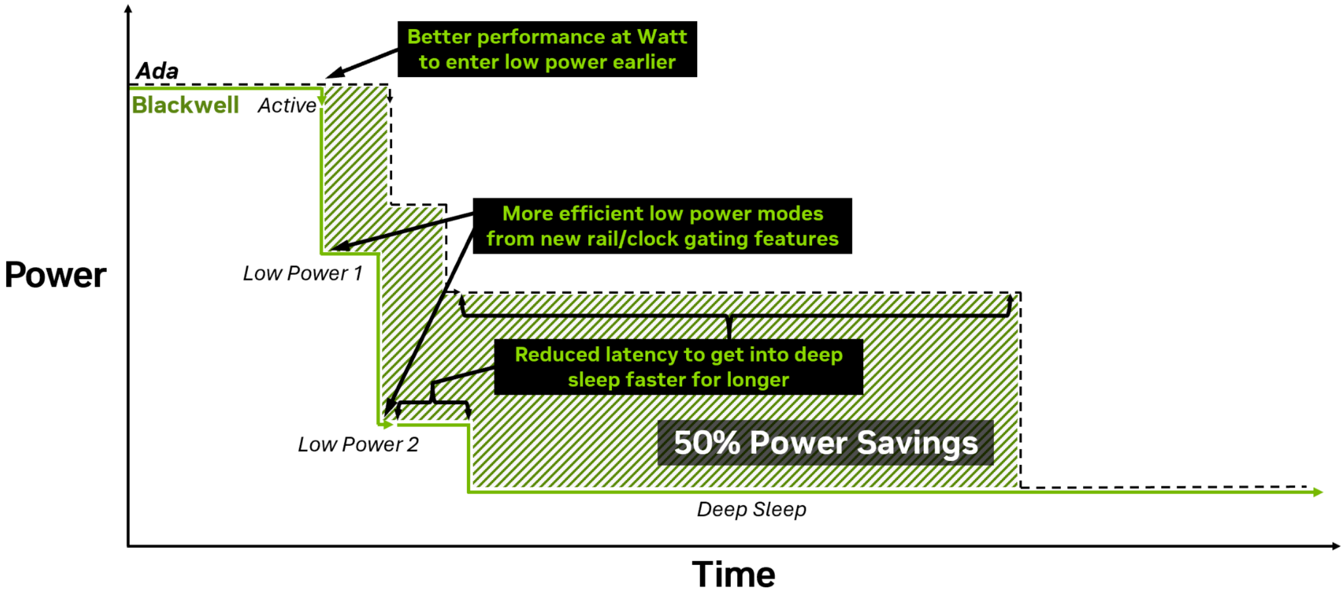
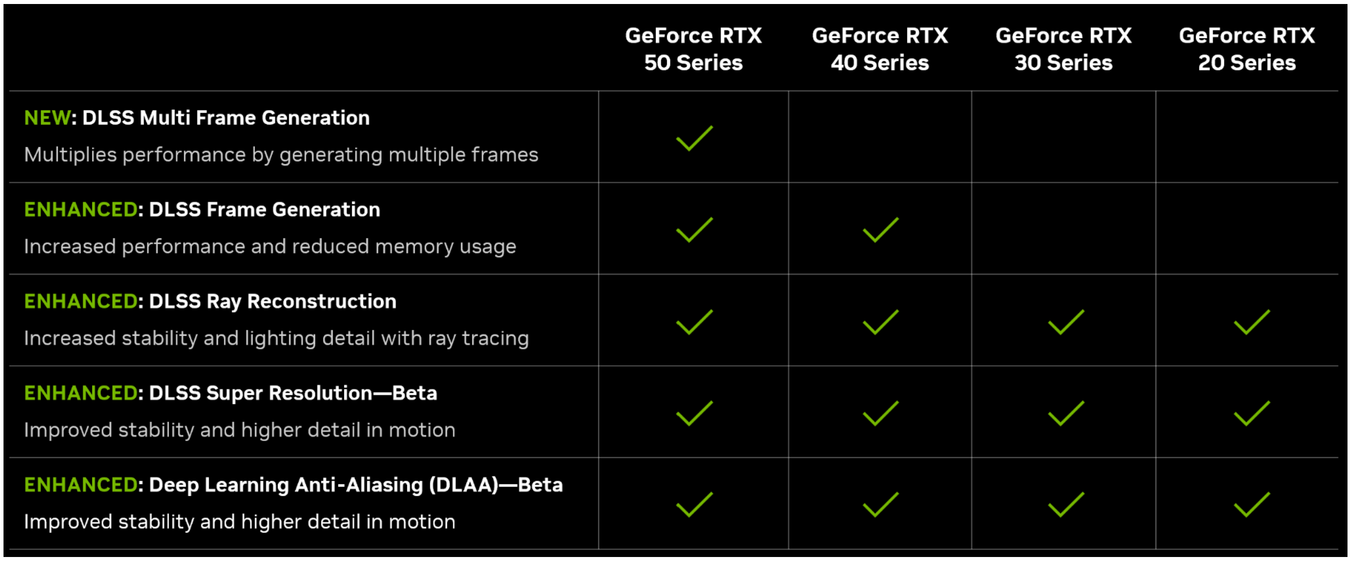


Figure 23. Real-life Example of Running Inference on SLMs on Ada and Blackwell

In a real-life example, like running inference on small language models as shown in Figure 23 above, power savings of up to 50% can be observed through a combination of Blackwell performance (reduced active period), lower power transitional states through power and voltage gating, and entering the deepest sleep state 10x faster than before.

DLSS 4

DLSS is a revolutionary suite of neural rendering technologies that uses AI to boost FPS, reduce latency, and improve image quality. The latest version, DLSS 4, brings new Multi Frame Generation (MFG) with faster performance and lower memory usage, and a new Transformer Model containing advancements for Super Resolution (SR), Ray Reconstruction (RR), and Deep Learning Anti-Aliasing (DLAA) that enhances image quality and stability. These new technologies are powered by GeForce RTX™ 50 Series GPUs and fifth-generation Tensor Cores, and are backed by an NVIDIA AI supercomputer in the cloud constantly improving your PC's gaming capabilities.



	GeForce RTX 50 Series	GeForce RTX 40 Series	GeForce RTX 30 Series	GeForce RTX 20 Series
NEW: DLSS Multi Frame Generation Multiplies performance by generating multiple frames	✓			
ENHANCED: DLSS Frame Generation Increased performance and reduced memory usage	✓	✓		
ENHANCED: DLSS Ray Reconstruction Increased stability and lighting detail with ray tracing	✓	✓	✓	✓
ENHANCED: DLSS Super Resolution—Beta Improved stability and higher detail in motion	✓	✓	✓	✓
ENHANCED: Deep Learning Anti-Aliasing (DLAA)—Beta Improved stability and higher detail in motion	✓	✓	✓	✓

Figure 24. DLSS Brings Upgrades for All RTX Gamers

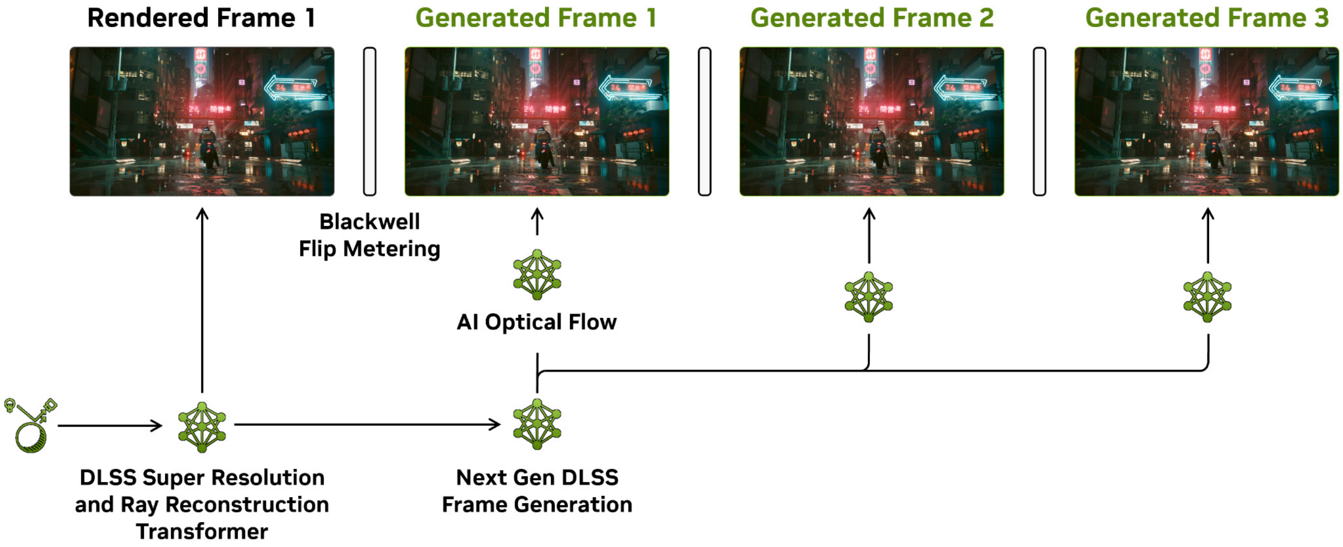
DLSS 4 Multi Frame Generation

Frame Generation technology was first introduced with the Ada architecture in 2022. A single frame was generated between every pair of traditionally rendered frames using an optical flow field along with game motion vectors and an AI network. The Blackwell architecture, built for neural rendering and using fifth-generation Tensor Cores, enables DLSS Multi Frame Generation to boost FPS by generating up to three additional frames per every traditionally rendered frame.

DLSS 4 Multi Frame Generation combines multiple Blackwell hardware and DLSS software innovations to make generating multiple frames a reality. Our new frame generation AI model is 40% faster than our prior frame generation method, uses 30% less VRAM, and only needs to run once per rendered frame to generate multiple frames. The generation of the optical flow field has been sped up by replacing hardware optical flow with a very efficient AI model. Together, the AI models significantly reduce the computational cost of generating additional frames.

Once the new frames are generated, they are evenly paced to deliver a smooth experience. DLSS 3 Frame Generation used CPU-based pacing with variability that can compound with additional frames, leading to less consistent frame pacing between each frame, impacting smoothness.

To address the complexities of generating multiple frames, Blackwell uses enhanced Flip Metering, which shifts the frame pacing logic to the display engine, enabling the GPU to more precisely manage display timing. The Blackwell display engine has been enhanced with twice the pixel processing capability to support higher resolutions and refresh rates for hardware Flip Metering with DLSS 4.



DLSS 4 Multi Frame Generation combines multiple Blackwell hardware and DLSS software innovations to generate multiple frames.

Figure 25. DLSS 4 Multi Frame Generation

A few Blackwell-only features enable DLSS 4 to work effectively. The 5th Generation Tensor Cores contain more computational horsepower, allowing them to more quickly execute the series of AI models that calculate optical flow and generate multiple frames. The AI Management Processor enables better scheduling of DLSS AI processing, graphics rendering, and the frame pacing algorithm.

Transformer Models

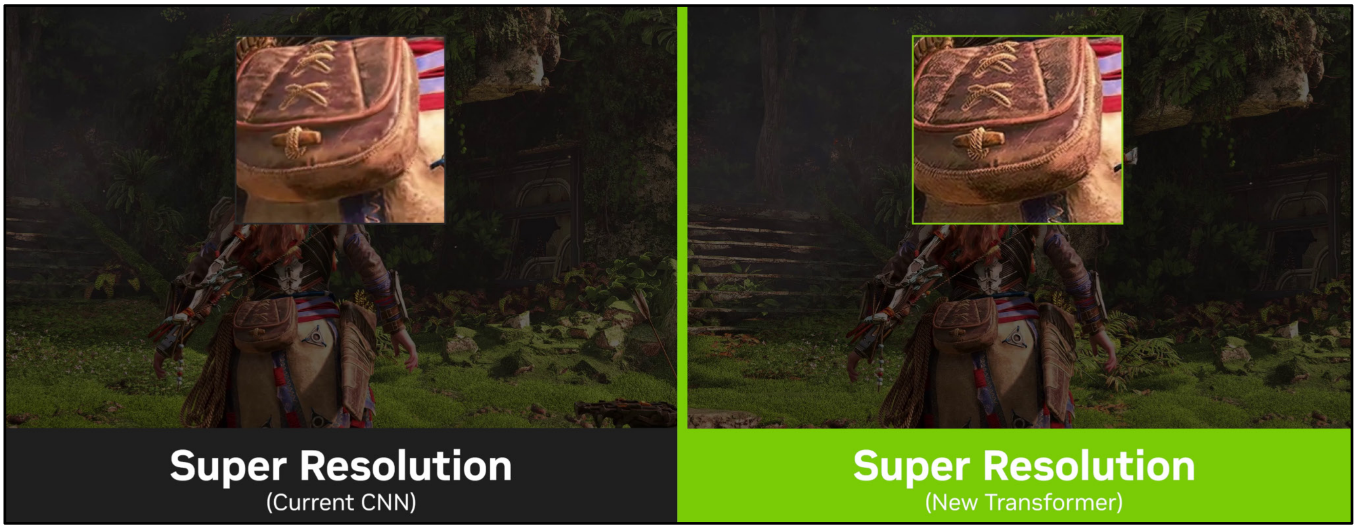
For the first time since 2020, when DLSS 2 was released, DLSS is transitioning to a completely new neural network architecture and that brings a lot of benefits. The ability for AI to classify an image was revolutionary and it was due to a technology called a Convolutional Neural Network, or CNN. CNNs work by locally aggregating pixels together and analyzing the data in a tree form from a lower level to a higher level. This structure was computationally efficient, which is why it's called a convolutional neural net.

DLSS 4 improves image quality and rendering smoothness by introducing more powerful, transformer-based AI models for DLSS Super Resolution, DLSS Ray Reconstruction, and Deep Learning Anti-Aliasing (DLAA) trained by NVIDIA's supercomputers to better understand and render complex scenes. Neural networks that use a transformer-based architecture excel at tasks involving sequential and structured data. The idea behind transformer models is that attention regarding how compute is spent and how it's analyzed should be driven by the data itself, so the neural network should learn how to direct its attention in order to look at parts of the data that are the most interesting or useful to make decisions.

Compared to CNN models, transformers use self-attention and can more easily identify the longer-range patterns across a much larger pixel window. Transformers also scale more effectively, allowing the models used for DLSS 4 to ingest 2x more parameters, while also using more Tensor Core processing power to reconstruct images at even better image quality for all RTX owners. The result is improved stability from one frame to the next, enhanced lighting detail, and more detail in motion. Changing the neural network architecture from CNN-based to transformer-based has resulted in a significant leap in image quality in many scenarios.

DLSS Super Resolution (SR)

SR boosts performance by using AI to output higher-resolution frames from a lower-resolution input. DLSS samples multiple lower-resolution images and uses motion data and feedback from prior frames to construct high-quality images. The end product of the transformer model is more temporally stable with less ghosting, more image detail in motion, and improved anti-aliasing compared to prior versions of DLSS.

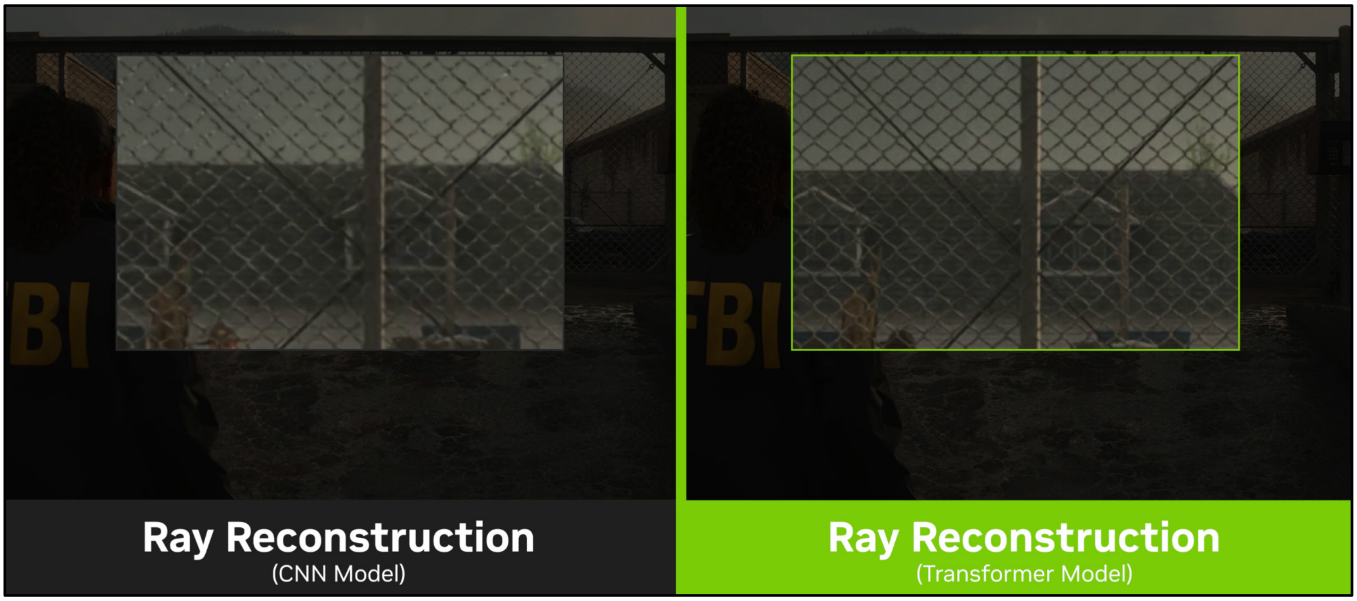


Transformer model Super Resolution demonstrates better temporal stability, less ghosting, and higher detail in motion.

Figure 26. Transformer Model vs CNN Model Super Resolution

DLSS Ray Reconstruction (RR)

RR enhances image quality by using AI to generate additional pixels for intensive ray-traced scenes. DLSS replaces hand-tuned denoisers with an NVIDIA supercomputer-trained AI network that generates higher-quality pixels between sampled rays. In intensive ray-traced content, the transformer model for RR gets an even bigger uplift in quality, especially for scenes with challenging lighting. In fact, all of the common artifacts from typical denoisers are significantly reduced.



Transformer model Ray Reconstruction delivers an uplift in image quality, especially in scenes with challenging lighting conditions.

Figure 27. Transformer Model vs CNN Model Ray Reconstruction

Deep Learning Anti-Aliasing (DLAA)

DLAA provides higher image quality using an AI-based anti-aliasing technique. DLAA uses the same Super Resolution technology developed for DLSS, constructing a more realistic, high-quality image at native resolution. The result provides increased temporal stability, detail in motion, and smoother edges in a scene.

Neural Shaders

Blackwell was designed to jumpstart the future, where neural shaders become the predominant form of shader technology for developing games. Many architectural improvements to Blackwell were made specifically to increase the performance and efficiency of neural shaders and this section describes those optimizations.

A shader is a program that runs on the GPU to control how graphics are rendered, varying in complexity depending on the visual effects and processing required. Newer shading techniques have added new levels of realism. In its most basic form, shaders calculate the levels of light, darkness, and color used when rendering a scene in 3D space in a game in a process known as shading. They run on the GPU and as part of the rendering pipeline.

Graphics were first processed on the GPU using non-programmable shaders, also known as fixed-function pipeline, where operations in the graphics pipeline were predefined and configurable, but not programmable. This was because they were limited by the hardware design of the GPU which was specifically built to execute a predefined set of operations.

GeForce 3 introduced the first bit of programmable shading with vertex shaders. Soon after, the high-level shading language HLSL allowed pixel shading so everything on the screen could be customized. DX10 introduced Geometry shaders. DX11 introduced compute shaders and then an update to DX12 delivered DirectX ray tracing using an acceleration structure of BVH (Bounding Volume Hierarchy) that allowed any ray to intersect with the scene geometry and then be able to spawn a cascade of different shading operations.

With the Blackwell launch we introduce the era of developer-created neural shaders, some of which will also run on prior generation GPUs. Neural Shaders are the next evolutionary step in

The evolution of shaders in GPUs has been marked by significant advancements in graphics programming and rendering capabilities.

Here’s a brief overview of key milestones:

1. **Fixed-Function Pipeline** *(Pre-2000s)*
   Graphics processed using a fixed-function pipeline where operations are predefined and configurable, but not programmable and with limited control over rendering of simple effects like lighting and texturing.

2. **Vertex Shaders** *(DirectX 8.0 / OpenGL 1.4, Early 2000s)*
   Programmable vertex shaders gave developers access to vertex data including transformations and lighting calculations, enabling more complex effects.

3. **Fragment Shaders (Pixel Shaders)** *(DirectX 9.0 / OpenGL 2.0, Early 2000s)*
   Enabled developers to write custom code for operations at the pixel level, allowing for dynamic lighting and texturing, expanding rendering flexibility with Shader Model 2.0.

4. **Unified Shader Architecture** *(DirectX 10.0 / OpenGL 3.3, 2006)*
   The unification of geometry, vertex, and fragment shaders, allowing for better utilization of resources and great efficiency. Introduced Shader Model 4.0 supporting more advanced techniques and performance optimizations.

5. **Geometry Shaders** *(DirectX 10 / OpenGL 3.2, 2006)*
   Geometry shaders expanded to allow for the creation and manipulation of primitives like triangles in the shader pipeline. New effects include dynamic tessellation and particle systems.

6. **Tessellation and Compute Shaders** *(DirectX 11 / OpenGL 4.0, 2009)*
   Redefined geometry with higher surface detail and smoother curves in 3D models. Shader Model 5.0 added more features for real-time

programmable shading. Rather than writing complex shader code to describe these functions, developers train AI models to approximate the result that the shader code would have computed. Neural shaders are set to become the predominant form of shaders in games, and in the future, all gaming will use AI technology for rendering.

Up until this point, NVIDIA has been using neural shading for DLSS, using CUDA to harness the Tensor Cores. With the new Cooperative Vectors API for DX12 and Vulkan, Tensor Cores can be accessed through any type of shader, including pixel and ray tracing, in a graphics application allowing for a host of neural technologies. NVIDIA has worked with Microsoft to create the new Cooperative Vectors API. When combined with differentiable shading language features in Slang, Cooperative Vectors unlock the ability for game developers to use neural techniques in their games including neural texture compression, that provides up to seven-to-one VRAM compression over block compressed formats, and other techniques such as RTX Neural Materials, Neural Radiance Cache, RTX Skin, and RTX Neural Faces.

rendering techniques. Compute Shaders added parallel processing and complex simulations.

7. **Primitive and Mesh Shaders** *(DirectX 12 Ultimate / Vulkan extension, 2018–2020)*
   Expanded the capabilities and performance of the geometry pipeline by incorporating the features of vertex and geometry shaders into a single shader stage. Mesh shaders allowed the GPU to handle more complex algorithms by offloading more work from the CPU to the GPU.

8. **RTX** *(NVIDIA Turing Architecture / DirectX Raytracing, 2018)*
   Added real-time ray tracing capabilities (RTX) directly to the SM in the GPU, enabling realistic lighting, shadows, and reflections. Introduced dedicated RT cores in hardware that are optimized for ray tracing by accelerating tree traversal and geometry intersection.

9. **Blackwell Neural Shaders (Unified AI and Traditional Shaders)** *(NVIDIA Blackwell Architecture, 2025)*
   AI is embedded into parts of the traditional rendering pipeline, paving the path towards full neural shading. Enhanced Tensor Cores that are now accessible to graphics shaders combined with scheduling optimizations in SER 2.0 (Shader Execution Reordering) so that AI graphics with neural filtering features and AI models including generative AI can be run concurrently in next-generation games.

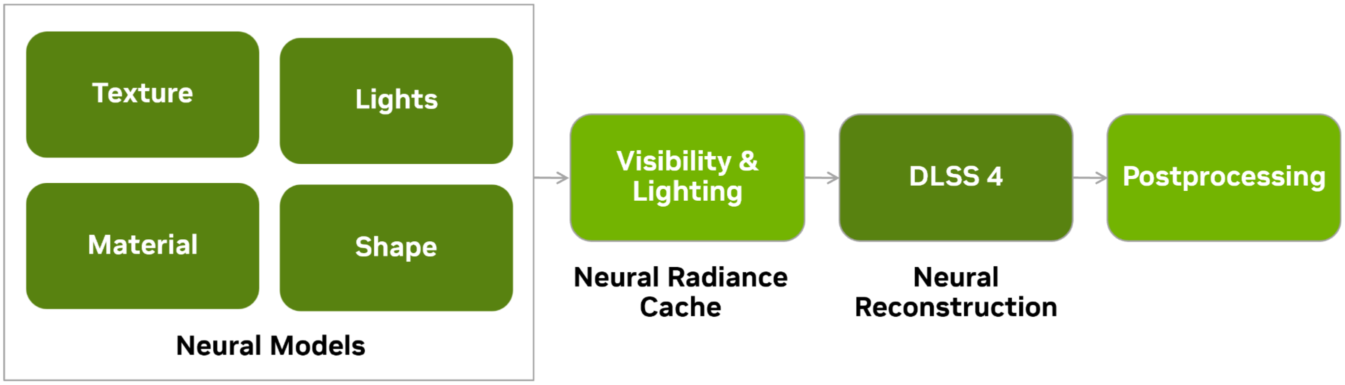


Figure 28. Neural Acceleration in Graphics

Neural shaders allow us to train neural networks to learn efficient approximations of complex algorithms that calculate how light interacts with surfaces, efficiently decompress textures that are stored in video memory in supercompressed form, predict indirect lighting based on limited ground truth data, and approximate subsurface light scattering —all contributing to a more immersive gaming experience. The potential applications for neural shaders are not yet fully explored, which means more exciting features for faster and more realistic (or stylized) real-time rendering lie ahead.

RTX Neural Materials

In big budget CGI films, some materials can be very complex and made up of multiple optical layers. Being able to ray trace multiple layers in real time is a very costly endeavor. However, AI techniques replace the original mathematical model of the material with a neural approximation, allowing for a better representation of materials, while making it possible to render film-quality assets at game-ready frame rates.

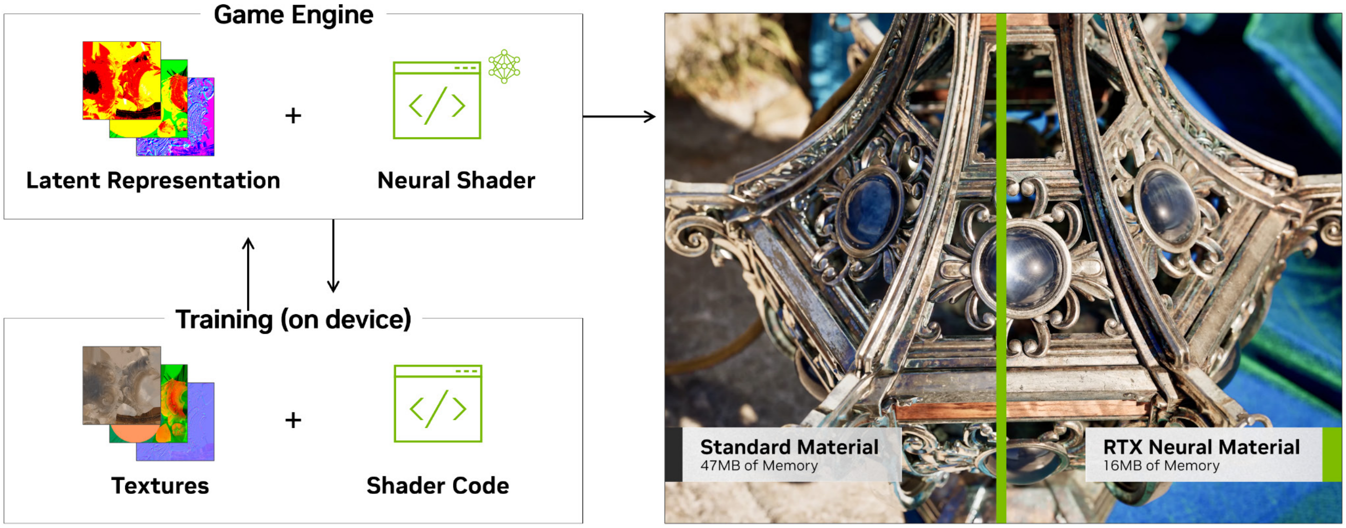


Figure 29. Film Quality Assets in Real Time.

RTX Neural Texture Compression (NTC)

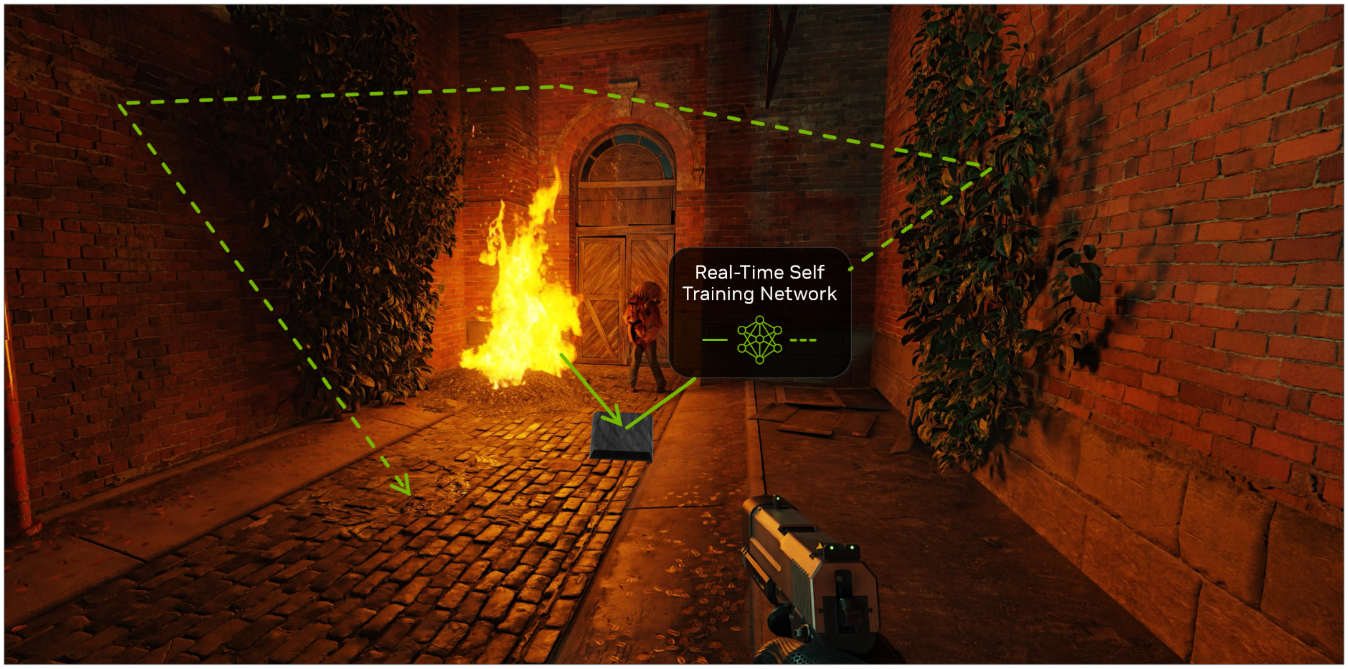
As advancements in photorealistic rendering increase, so too does the amount of texture data required, increasing the demands on storage and memory while also affecting performance by limiting bandwidth. RTX Neural Texture Compression leverages neural networks accessed through neural shaders to compress and decompress material textures more efficiently than traditional methods. Note that our Neural Materials demo uses 1110 megabytes of memory for the standard materials on the lantern and the fabric. However, with neural materials, the demo only uses 333 megabytes for the same materials - a savings of over 3x while providing much higher visual quality.

Stochastic Texture Filtering (STF) is used to introduce randomness into the texture sampling process to reduce visual artifacts like aliasing and moiré patterns when it is impractical to apply traditional trilinear or anisotropic filtering, such as with Neural Texture Compression. In cases when hardware texture filtering is available, STF is still useful: it can provide higher-order filtering, such as cubic or Gaussian, at the cost of a single point sample. STF specifically runs fast on Blackwell GPU due to its 2x point-sampled texture filtering rate improvements.

*For more information on Neural Texture Compression, please refer to the **NVIDIA Research Webpage**.*

Neural Radiance Cache (NRC)

NRC utilizes a neural shader to cache and approximate radiance information. By leveraging the learnings of a neural network, complex lighting information can be stored and used to create high-quality global illumination (GI) and dynamic lighting effects in real-time rendering. This improves efficiency by reducing the computational load on the GPU, resulting in enhanced visual quality and scalability.



Neural Radiance Cache (NRC)

Trace 1 bounce per Pixel | Infer many more bounces

Neural Radiance Cache increases performance and indirect light quality through inference of path-traced rays.

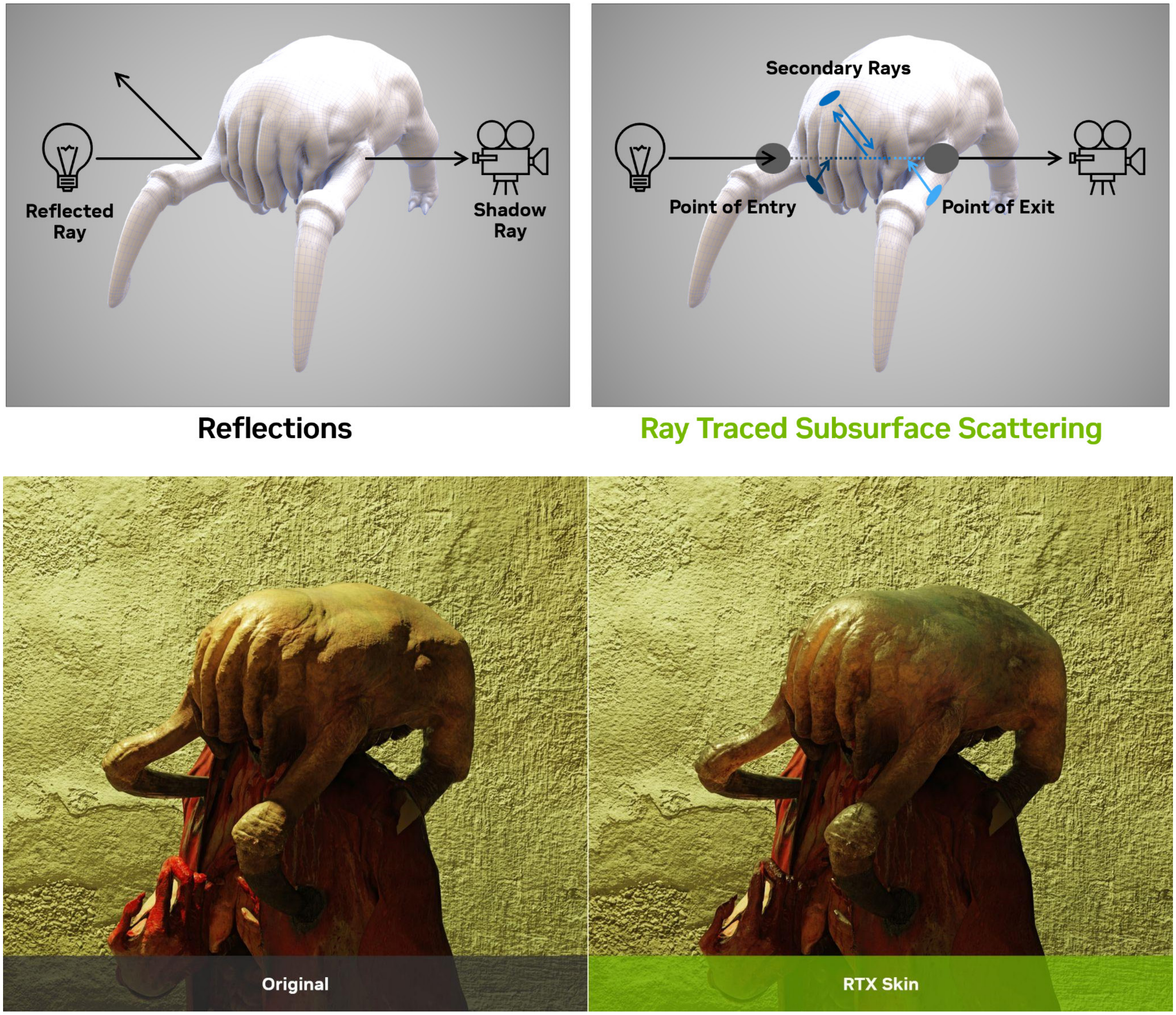
Figure 30. Neural Radiance Cache

NRC is a neural shader that takes as input path-traced rays after one bounce and infers the final lighting values for many bounces. NRC trains tiny neural networks on live game data in real-time to estimate the indirect illumination signal; an initial set of rays are fired, but not fully traced. Instead, the path tracer sends the ray paths into the cache after one bounce, and simulates how the scene would have looked if the ray were full length with many bounces.

NRC improves performance as the GPU doesn't have to trace the rays beyond the small initial number of bounces and Indirect Lighting quality is improved because NRC can infer a multitude of bounces on a limited ray budget. Furthermore, it can preserve image quality in challenging scenarios since it is highly adaptable. Because it trains while you play, it is contextually aware of the variety of scenarios present in any given game, allowing it to tune itself to deliver an accurate indirect lighting profile for each game scene.

RTX Skin

Skin is a problem for rendering and typical representations of objects in games are manifold. Essentially, skin is a set of meshes which make up the outside of an object. This works well if the material is impermeable to light like wood or metal where the rays that intersect with the object only require computing the light based upon what lights in the scene. However, translucent materials work differently. There's actual penetration of the light into the material, into the object, which is then transported or scattered within the object and then emitted in other parts of the object. To make skin better, NVIDIA borrowed film rendering technology called subsurface scattering, bringing it into the realm of real-time for path tracing. RTX Skin is the first ray traced example of subsurface scattering in games, and it can be applied subtly or intensely, as the artist desires.



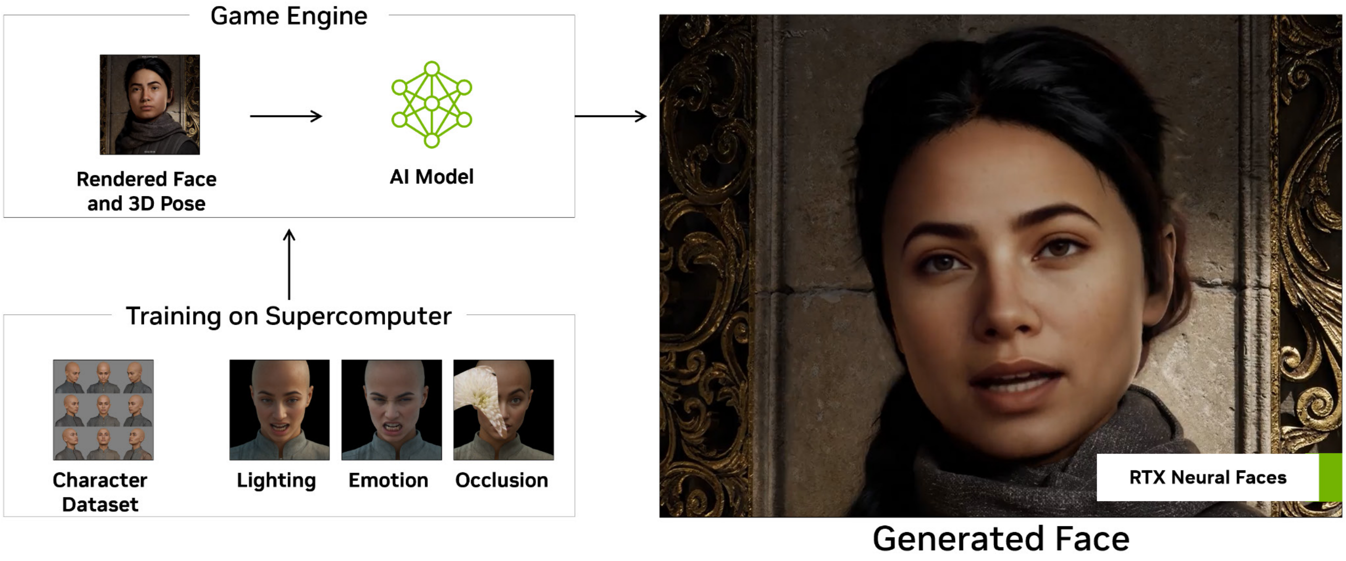
Reflections

Ray Traced Subsurface Scattering

Figure 31. RTX Skin Enables Incredibly Life-like Translucent Materials

RTX Neural Faces

Another difficulty for real-time rendering has been rendering faces realistically. Humans are conditioned from birth to recognize any anomaly in the human face, and there is a term called The Uncanny Valley that refers to the difference between what is rendered and what is expected . Film rendering has solved this, but it requires orders of magnitude more time to generate an image which is convincingly photo real versus what can be provided at runtime in the game.



RTX Neural Faces uses a real-time generative AI model to infer a more natural-looking face.

Figure 32. RTX Neural Faces

RTX Neural Faces offers an innovative, new approach to improve face quality using generative AI. Instead of brute force rendering, Neural Faces takes a simple rasterized face plus 3D pose data as input and uses a real-time generative AI model to infer a more natural face. The generated face is trained from thousands of offline generated images of that face at every angle, under different lighting, emotion, and occlusion conditions. The training pipeline can use real photographs or AI generated images, with variants created using a diffusion model. The trained model is then TensorRT optimized to infer the face in real time. RTX Neural Faces represent the first step in a journey to redefine real-time graphics with generative AI.

APPENDIX A: Blackwell GB202 GPU

The Blackwell GPU architecture consists of a family of GPUs that are targeted for multiple segments of the graphics market. GB202 is NVIDIA's flagship GPU offering based on the Blackwell architecture, delivering revolutionary performance for the ultra-enthusiast graphics segment for gaming, content creation, and AI.

The full GB202 chip consists of 92.2 billion transistors and contains 12 GPCs, 96 TPCs, 192 SMs, and sixteen 32-bit memory controllers (512-bit total). With each SM containing 128 FP32 CUDA Cores, the full chip contains 24,576 CUDA Cores as well as 192 RT Cores, 768 Tensor Cores, 768 Texture Units, and 192 ROPS. The memory subsystem includes 24,576 KB L1 cache, 49,152 KB Register File, and 131,072 KB L2 cache.

GeForce RTX 5090 Specifications

The first GeForce RTX 50 series product that will be launching using the GB202 GPU is the GeForce RTX 5090. The specs for the GeForce RTX 5090, 4090, and 3090 are compared below.

Table 3. GeForce RTX 5090 vs GeForce RTX 4090 vs GeForce RTX 3090 Specs

Graphics Card	**GeForce RTX 3090**	**GeForce RTX 4090**	**GeForce RTX 5090**
GPU Codename	GA102	AD102	GB202
GPU Architecture	NVIDIA Ampere	NVIDIA Ada Lovelace	NVIDIA Blackwell
GPCs	7	11	11
TPCs	41	64	85
SMs	82	128	170
CUDA Cores / SM	128	128	128
CUDA Cores / GPU	10496	16384	21760
Tensor Cores / SM	4 (3rd Gen)	4 (4th Gen)	4 (5th Gen)
Tensor Cores / GPU	328 (3rd Gen)	512 (4th Gen)	680 (5th Gen)
RT Cores	82 (2nd Gen)	128 (3rd Gen)	170 (4th Gen)
GPU Boost Clock (MHz)	1695	2520	2407
Peak FP32 TFLOPS (non-Tensor)[1]	35.6	82.6	104.8

Peak FP16 TFLOPS (non-Tensor)[1]	35.6	82.6	104.8
Peak BF16 TFLOPS (non-Tensor)[1]	35.6	82.6	104.8
Peak INT32 TOPS (non-Tensor)[1]	17.8	41.3	104.8
RT TFLOPS	69.5	191	317.5
Peak FP4 Tensor TFLOPS with FP32 Accumulate (FP4 AI TOPS)	N/A	N/A	1676/3352[2]
Peak FP8 Tensor TFLOPS with FP16 Accumulate[1]	N/A	660.6/1321.2[2]	838/1676[2]
Peak FP8 Tensor TFLOPS with FP32 Accumulate[1]	N/A	330.3/660.6.2[2]	419/838[2]
Peak FP16 Tensor TFLOPS with FP16 Accumulate[1]	142.3/284.6[2]	330.3/660.6[2]	419/838[2]
Peak FP16 Tensor TFLOPS with FP32 Accumulate[1]	71.2/142.4[2]	165.2/330.4[2]	209.5/419[2]
Peak BF16 Tensor TFLOPS with FP32 Accumulate[1]	71.2/142.4[2]	165.2/330.4[2]	209.5/419[2]
Peak TF32 Tensor TFLOPS[1]	35.6/71.2[2]	82.6/165.2[2]	104.8/209.5[2]
Peak INT8 Tensor TOPS[1]	284.7/569.4[2]	660.6/1321.2[2]	838/1676[2]
Frame Buffer Memory Size and Type	24 GB GDDR6X	24 GB GDDR6X	32 GB GDDR7
Memory Interface	384-bit	384-bit	512-bit
Memory Clock (Data Rate)	19.5 Gbps	21 Gbps	28 Gbps
Memory Bandwidth	936 GB/sec	1008 GB/sec	1792 GB/sec
ROPs	112	176	176
Pixel Fill-rate (Gigapixels/sec)	189.8	443.5	423.6
Texture Units	328	512	680
Texel Fill-rate (Gigatexels/sec)	555.96	1290.2	1636.8
L1 Data Cache/Shared Memory	10496 KB	16384 KB	21760 KB

L2 Cache Size	6144 KB	73728 KB	98304 KB
Register File Size	20992 KB	32768 KB	43520 KB
Video Engines	1 x NVENC (7th Gen) 1 x NVDEC (5th Gen)	2 x NVENC (8th Gen) 1 x NVDEC (5th Gen)	3 x NVENC (9th Gen) 2 x NVDEC (6th Gen)
TGP (Total Graphics Power)	350 W	450 W	575 W
Transistor Count	28.3 Billion	76.3 Billion	92.2 Billion
Die Size	628.4 mm^2	608.5 mm^2	750 mm^2
Manufacturing Process	Samsung 8nm 8N NVIDIA Custom Process	TSMC 4nm 4N NVIDIA Custom Process	TSMC 4nm 4N NVIDIA Custom Process
PCI Express Interface	Gen 4	Gen 4	Gen 5

1. Peak rates are based on GPU Boost Clock.
2. Effective TOPS / TFLOPS using the Sparsity Feature

APPENDIX B: Blackwell GB203 GPU

The GB203 GPU is NVIDIA's product for the high-end graphics segment. GB203 retains all of the key features found in GB202, including all of the innovations introduced with the Blackwell SM such as Blackwell's Fourth-Generation RT Core and Fifth-Generation Tensor Core as well as DLSS 4 and the new AI gaming features.

The full GB203 chip consists of 45.6 billion transistors and contains 7 GPCs, 42 TPCs, 84 SMs, and eight 32-bit memory controllers (256-bit total). With each SM containing 128 FP32 CUDA Cores, the full chip contains 10,752 CUDA Cores as well as 84 RT Cores, 336 Tensor Cores, 336 Texture Units, and 112 ROPS. The memory subsystem includes 10,752 KB L1 cache, 21,504 KB Register File, and 65,536 KB L2 cache.

GeForce RTX 5080 Specifications

The first GeForce RTX 50 series product that will be launching using the GB203 GPU is the GeForce RTX 5080. The specs for the GeForce RTX 5080, 4080, and 3080 are compared below.

Table 4. GeForce RTX 5080 vs GeForce RTX 4080 vs GeForce RTX 3080 Specs

Graphics Card	**RTX 3080**	**RTX 4080**	**RTX 5080**
GPU Codename	GA102	AD103	GB203
GPU Architecture	NVIDIA Ampere	NVIDIA Ada Lovelace	NVIDIA Blackwell
GPCs	6	7	7
TPCs	34	38	42
SMs	68	76	84
CUDA Cores / SM	128	128	128
CUDA Cores / GPU	8704	9728	10752
Tensor Cores / SM	4 (3rd Gen)	4 (4th Gen)	4 (5th Gen)
Tensor Cores / GPU	272 (3rd Gen)	304 (4th Gen)	336 (5th Gen)
RT Cores	68 (2nd Gen)	76 (3rd Gen)	84 (4th Gen)
GPU Boost Clock (MHz)	1710	2505	2617
Peak FP32 TFLOPS (non-Tensor)[1]	34.1	48.7	56.3

Peak FP16 TFLOPS (non-Tensor)[1]	34.1	48.7	56.3
Peak BF16 TFLOPS (non-Tensor)[1]	34.1	48.7	56.3
Peak INT32 TOPS (non-Tensor)[1]	17	24.4	56.3
RT TFLOPS	58.1	112.7	170.6
Peak FP4 Tensor TFLOPS with FP32 Accumulate (FP4 AI TOPS)	N/A	N/A	900.4/1801[2]
Peak FP8 Tensor TFLOPS with FP16 Accumulate[1]	N/A	389.9/779.8[2]	450.2/900.4[2]
Peak FP8 Tensor TFLOPS with FP32 Accumulate[1]	N/A	194.9/389.8[2]	225.1/450.2[2]
Peak FP16 Tensor TFLOPS with FP16 Accumulate[1]	119.1/238.2[2]	194.9/389.8[2]	225.1/450.2[2]
Peak FP16 Tensor TFLOPS with FP32 Accumulate[1]	59.5/119[2]	97.5/195[2]	112.6/225.1[2]
Peak BF16 Tensor TFLOPS with FP32 Accumulate[1]	59.5/119[2]	97.5/195[2]	112.6/225.1[2]
Peak TF32 Tensor TFLOPS[1]	29.8/59.6[2]	48.7/97.4[2]	56.3/112.6[2]
Peak INT8 Tensor TOPS[1]	238.1/476.2[2]	389.9/779.82[2]	450.2/900.4[2]
Frame Buffer Memory Size and Type	10 GB GDDR6X	16 GB GDDR6X	16 GB GDDR7
Memory Interface	320-bit	256-bit	256-bit
Memory Clock (Data Rate)	19 Gbps	22.4 Gbps	30 Gbps
Memory Bandwidth	760 GB/sec	716.8 GB/sec	960 GB/sec
ROPs	96	112	112
Pixel Fill-rate (Gigapixels/sec)	164.2	280.6	293.1
Texture Units	272	304	336
Texel Fill-rate (Gigatexels/sec)	465.12	761.5	879.3
L1 Data Cache/Shared Memory	8704 KB	9728 KB	10752 KB

L2 Cache Size	5120 KB	65536 KB	65536 KB
Register File Size	17408 KB	19456 KB	21504 KB
Video Engines	1 x NVENC (7th Gen) 1 x NVDEC (5th Gen)	2 x NVENC (8th Gen) 1 x NVDEC (5th Gen)	2 x NVENC (9th Gen) 2 x NVDEC (6th Gen)
TGP (Total Graphics Power)	320 W	320 W	360 W
Transistor Count	28.3 Billion	45.9 Billion	45.6 Billion
Die Size	628.4 mm^2	378.6 mm^2	378 mm^2
Manufacturing Process	Samsung 8 nm 8N NVIDIA Custom Process	TSMC 4nm 4N NVIDIA Custom Process	TSMC 4nm 4N NVIDIA Custom Process
PCI Express Interface	Gen 4	Gen 4	Gen 5

1. Peak rates are based on GPU Boost Clock.
2. Effective TOPS / TFLOPS using the Sparsity Feature

GeForce RTX 5070 Ti Specifications

The second GeForce RTX 50 series product that will be launching using the GB203 GPU is the GeForce RTX 5070 Ti. The specs for the GeForce RTX 5070 Ti, 4070 Ti, and 3070 Ti are compared below.

Table 5. GeForce RTX 5070 Ti vs GeForce RTX 4070 Ti vs GeForce RTX 3070 Ti Specs

Graphics Card	**RTX 3070 Ti**	**RTX 4070 Ti**	**RTX 5070 Ti**
GPU Codename	GA104	AD104	GB203
GPU Architecture	NVIDIA Ampere	NVIDIA Ada Lovelace	NVIDIA Blackwell
GPCs	6	5	6
TPCs	24	30	35
SMs	48	60	70
CUDA Cores / SM	128	128	128

CUDA Cores / GPU	6144	7680	8960
Tensor Cores / SM	4 (3rd Gen)	4 (4th Gen)	4 (5th Gen)
Tensor Cores / GPU	192 (3rd Gen)	240 (4th Gen)	280 (5th Gen)
RT Cores	48 (2nd Gen)	60 (3rd Gen)	70 (4th Gen)
GPU Boost Clock (MHz)	1770	2610	2452
Peak FP32 TFLOPS (non-Tensor)[1]	21.7	40.1	43.9
Peak FP16 TFLOPS (non-Tensor)[1]	21.7	40.1	43.9
Peak BF16 TFLOPS (non-Tensor)[1]	21.7	40.1	43.9
Peak INT32 TOPS (non-Tensor)[1]	10.9	20.0	43.9
RT TFLOPS	42.5	92.7	133.2
Peak FP4 Tensor TFLOPS with FP32 Accumulate (FP4 AI TOPS)	N/A	N/A	703/1406[2]
Peak FP8 Tensor TFLOPS with FP16 Accumulate[1]	N/A	320.7/641.4[2]	351.5/703[2]
Peak FP8 Tensor TFLOPS with FP32 Accumulate[1]	N/A	160.4/320.8[2]	175.8/351.5[2]
Peak FP16 Tensor TFLOPS with FP16 Accumulate[1]	87/174[2]	160.4/320.8[2]	175.8/351.5[2]
Peak FP16 Tensor TFLOPS with FP32 Accumulate[1]	43.5/87[2]	80.2/160.4[2]	87.9/175.8[2]
Peak BF16 Tensor TFLOPS with FP32 Accumulate[1]	43.5/87[2]	80.2/160.4[2]	87.9/175.8[2]
Peak TF32 Tensor TFLOPS[1]	21.7/43.4[2]	40.1/80.2[2]	43.9/87.9[2]
Peak INT8 Tensor TOPS[1]	174/348[2]	320.7/641.4[2]	351.5/703[2]
Frame Buffer Memory Size and Type	8 GB GDDR6X	12 GB GDDR6X	16 GB GDDR7
Memory Interface	256-bit	192-bit	256-bit
Memory Clock (Data Rate)	19 Gbps	21 Gbps	28 Gbps
Memory Bandwidth	608 GB/sec	504 GB/sec	896 GB/sec

ROPs	96	80	96
Pixel Fill-rate (Gigapixels/sec)	169.9	208.8	235.4
Texture Units	192	240	280
Texel Fill-rate (Gigatexels/sec)	339.84	626.4	686.6
L1 Data Cache/Shared Memory	6144 KB	7680 KB	8960 KB
L2 Cache Size	4096 KB	49152 KB	49152 KB
Register File Size	12288 KB	15360 KB	17920 KB
Video Engines	1 x NVENC (7th Gen) 1 x NVDEC (5th Gen)	2 x NVENC (8th Gen) 1 x NVDEC (5th Gen)	2 x NVENC (9th Gen) 1 x NVDEC (6th Gen)
TGP (Total Graphics Power)	290 W	285 W	300 W
Transistor Count	17.4 Billion	35.8 Billion	45.6 Billion
Die Size	392.5 mm^2	294.5 mm^2	378 mm^2
Manufacturing Process	Samsung 8 nm 8N NVIDIA Custom Process	TSMC 4nm 4N NVIDIA Custom Process	TSMC 4nm 4N NVIDIA Custom Process
PCI Express Interface	Gen 4	Gen 4	Gen 5

1. Peak rates are based on GPU Boost Clock.
2. Effective TOPS / TFLOPS using the Sparsity Feature

APPENDIX C: Blackwell GB205 GPU

The Ada GB205 GPU is the perfect entry point for gamers, content creators, and streamers who want the new features NVIDIA is introducing with the Blackwell GPU architecture. The GB205 GPU is tailored for the performance GPU segment and includes all the architectural changes that are being introduced with the Blackwell GPU architecture that were discussed earlier in this document.

The full implementation of the GB205 GPU consists of 5 GPCs, 25 TPCs, 50 SMs, and six 32-bit memory controllers (192-bit memory interface). The chip contains a total of 31 billion transistors, 6,400 CUDA Cores, 50 RT Cores, 200 Tensor Cores, 200 Texture Units, and 80 ROPS. The memory subsystem incorporates 6,400 KB L1 Cache, 12,800 KB Register File, and 49,152 KB L2 Cache.

GeForce RTX 5070 Specifications

The specs for the GeForce RTX 5070, based on the GB 205 GPU, are compared with RTX 4070 and 3070 below.

Table 6. GeForce RTX 5070 vs GeForce RTX 4070 vs GeForce RTX 3070 Specs

Graphics Card	**RTX 3070**	**RTX 4070**	**RTX 5070**
GPU Codename	GA104	AD104	GB205
GPU Architecture	NVIDIA Ampere	NVIDIA Ada Lovelace	NVIDIA Blackwell
GPCs	6	5	5
TPCs	23	23	24
SMs	46	46	48
CUDA Cores / SM	128	128	128
CUDA Cores / GPU	5888	5888	6144
Tensor Cores / SM	4 (3rd Gen)	4 (4th Gen)	4 (5th Gen)
Tensor Cores / GPU	184 (3rd Gen)	184 (4th Gen)	192 (5th Gen)
RT Cores	46 (2nd Gen)	46 (3rd Gen)	48 (4th Gen)
GPU Boost Clock (MHz)	1725	2475	2512

Peak FP32 TFLOPS (non-Tensor)[1]	20.3	29.1	30.9
Peak FP16 TFLOPS (non-Tensor)[1]	20.3	29.1	30.9
Peak BF16 TFLOPS (non-Tensor)[1]	20.3	29.1	30.9
Peak INT32 TOPS (non-Tensor)[1]	10.2	14.6	30.9
RT TFLOPS	39.7	67.4	93.6
Peak FP4 Tensor TFLOPS with FP32 Accumulate (FP4 AI TOPS)	N/A	N/A	493.9/987.8[2]
Peak FP8 Tensor TFLOPS with FP16 Accumulate[1]	N/A	233.2/466.4[2]	246.9/493.9[2]
Peak FP8 Tensor TFLOPS with FP32 Accumulate[1]	N/A	116.6/233.2[2]	123.5/246.9[2]
Peak FP16 Tensor TFLOPS with FP16 Accumulate[1]	81.3/162.6[2]	116.6/233.2[2]	123.5/246.9[2]
Peak FP16 Tensor TFLOPS with FP32 Accumulate[1]	40.6/81.2[2]	58.3/116.6[2]	61.7/123.5[2]
Peak BF16 Tensor TFLOPS with FP32 Accumulate[1]	40.6/81.2[2]	58.3/116.6[2]	61.7/123.5[2]
Peak TF32 Tensor TFLOPS[1]	20.3/40.6[2]	29.1/58.2[2]	30.9/61.7[2]
Peak INT8 Tensor TOPS[1]	162.5/325[2]	233.2/466.4[2]	246.9/493.9[2]
Frame Buffer Memory Size and Type	8 GB GDDR6	12 GB GDDR6X	12 GB GDDR7
Memory Interface	256-bit	192-bit	192-bit
Memory Clock (Data Rate)	14 Gbps	21 Gbps	28 Gbps
Memory Bandwidth	448 GB/sec	504 GB/sec	672 GB/sec
ROPs	96	64	80
Pixel Fill-rate (Gigapixels/sec)	165.6	158.4	201
Texture Units	184	184	192
Texel Fill-rate (Gigatexels/sec)	317.4	455.4	482.3
L1 Data Cache/Shared Memory	5888 KB	5888 KB	6144 KB

L2 Cache Size	4096 KB	36864 KB	49152 KB
Register File Size	11776 KB	11776 KB	12288 KB
Video Engines	1 x NVENC (7th Gen) 1 x NVDEC (5th Gen)	1 x NVENC (8th Gen) 1 x NVDEC (5th Gen)	1 x NVENC (9th Gen) 1 x NVDEC (6th Gen)
TGP (Total Graphics Power)	220 W	200 W	250 W
Transistor Count	17.4 Billion	35.8 Billion	31.1 Billion
Die Size	392.5 mm^2	294.5 mm^2	263 mm^2
Manufacturing Process	Samsung 8 nm 8N NVIDIA Custom Process	TSMC 4nm 4N NVIDIA Custom Process	TSMC 4nm 4N NVIDIA Custom Process
PCI Express Interface	Gen 4	Gen 4	Gen 5

1. Peak rates are based on GPU Boost Clock.
2. Effective TOPS / TFLOPS using the Sparsity Feature

Notice

The information provided in this specification is believed to be accurate and reliable as of the date provided. However, NVIDIA Corporation ("NVIDIA") does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This publication supersedes and replaces all other specifications for the product that may have been previously supplied.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and other changes to this specification, at any time and/or to discontinue any product or service without notice. Customer should obtain the latest relevant specification before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer. NVIDIA hereby expressly objects to applying any customer general terms and conditions with regard to the purchase of the NVIDIA product referenced in this specification.

NVIDIA products are not designed, authorized or warranted to be suitable for use in medical, military, aircraft, space or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on these specifications will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this specification. NVIDIA does not accept any liability related to any default, damage, costs or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this specification, or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this specification. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA. Reproduction of information in this specification is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the NVIDIA terms and conditions of sale for the product.

Trademarks and Copyright