NVIDIA GPU Architecture and Technologies

NVIDIA GPUs are designed around massively parallel architectures optimized for graphics rendering and compute workloads.

Streaming Multiprocessors (SMs) are the core execution units responsible for parallel processing of threads.

CUDA cores execute general-purpose computations, while Tensor Cores accelerate matrix operations for AI and machine learning.

RT Cores are dedicated hardware units used to accelerate real-time ray tracing calculations in supported applications.

Modern NVIDIA GPUs use high-bandwidth memory interfaces and large cache hierarchies to reduce memory latency.

Power management technologies dynamically adjust clock speeds and voltage based on workload and thermal conditions.

NVIDIA GPUs support technologies such as CUDA, DLSS, and NVENC to improve performance, efficiency, and media processing.

Driver software plays a critical role in hardware scheduling, compatibility, and performance optimization.