

NVIDIAs Grafikprozessoren (GPUs) basieren auf einer fortlaufenden Evolution von Mikroarchitekturen, die jeweils für bestimmte Generationen von GeForce- (Gaming), Quadro/RTX- (Workstation) und Tesla/Data-Center-Beschleunigern eingesetzt werden. Seit 2006 (Einführung von CUDA) folgt NVIDIA meist dem Prinzip, Architekturen nach berühmten Wissenschaftlern zu benennen.

Hier ist die chronologische Übersicht der wichtigsten NVIDIA-Mikroarchitekturen:

Chronologische Übersicht (Moderne Ära)

Architektur	Jahr	Fertigung (nm)	Wichtige Neuerungen / Fokus	Bekannte GPU-Modelle
Blackwell	2024	4nm (TSMC 4N)	Dedizierte Decompression Engines, native NVLink-Unterstützung für gigantische LLMs, extreme Skalierung für Rechenzentren.	B200, B100 (Rechenzentrum), RTX 50-Serie (erwartet)
Ada Lovelace	2022	5nm (TSMC 4N)	DLSS 3 (Frame Generation), Shader Execution Reordering (SER), RT-Cores der 3. Generation. Extreme Effizienzsteigerung.	RTX 4090, RTX 4080, RTX 4070
Hopper	2022	4nm (TSMC)	Reine Rechenzentrums-Architektur. DPX-Instruktionen	H100, H200

Architektur	Jahr	Fertigung (nm)	Wichtige Neuerungen / Fokus	Bekannte GPU-Modelle
			n, Transformer Engine für beschleunigtes KI-Training.	
Ampere	2020	8nm (Samsung) / 7nm (TSMC)	RT-Cores der 2. Gen, Tensor Cores der 3. Gen (Stichwort: Sparsity), Einführung von PCI Express 4.0.	RTX 3090, RTX 3080; A100 (Data Center)
Turing	2018	12nm (TSMC)	Geburtsstunde von RTX. Erste Architektur mit dedizierten Raytracing-Kernen (RT Cores) und KI-Kernen (Tensor Cores für DLSS).	RTX 2080 Ti, RTX 2070, GTX 1660 Ti
Volta	2017	12nm (TSMC)	Erste Architektur mit Tensor Cores . Hauptsächlich für HPC (High-Performance Computing) und KI, kaum Consumer-Karten.	TITAN V, Tesla V100
Pascal	2016	16nm / 14nm	Riesiger Sprung in der	GTX 1080 Ti, GTX 1070, GTX

Architektur	Jahr	Fertigung (nm)	Wichtige Neuerungen / Fokus	Bekannte GPU-Modelle
			Energieeffizienz durch FinFET-Verfahren. Einführung von GDDR5X und der ersten NVLink-Generation.	1060
Maxwell	2014	28nm	Fokus auf extreme Energieeffizienz pro Watt, Einführung von HDMI 2.0 und verbesserter Speicherkompression.	GTX 980, GTX 970, GTX 750 Ti
Kepler	2012	28nm	Einführung des SMX-Streaming-Multiprozessors (effizienteres Scheduling), GPU Boost und Unterstützung für DirectX 11.1.	GTX 780 Ti, GTX 680
Fermi	2010	40nm	Erste echte Allzweck-GPU-Architektur (GPGPU) mit vollem C++ und ECC-Speicherschutz. Sehr rechenstark,	GTX 580, GTX 480

Architektur	Jahr	Fertigung (nm)	Wichtige Neuerungen / Fokus	Bekannte GPU-Modelle
			aber hohe Wärmeentwicklung.	
Tesla	2006	90nm bis 40nm	Die erste Unified-Shader-Architektur von NVIDIA und das Fundament für CUDA. Ersetzte die starren Pixel- und Vertex-Pipelines.	8800 GTX, GTX 280

Visuelle Evolution der Fertigungsgrößen (Strukturbreite)

Der technologische Fortschritt spiegelt sich massiv in der Verkleinerung der Fertigungsprozesse wider, was mehr Transistoren auf gleichem Raum ermöglicht:

Ältere historische Architekturen (Vor CUDA)

Vor der Zusammenführung der Recheneinheiten (Unified Shader) teilten sich die Architekturen in dedizierte Pixel- und Vertex-Pipelines auf:

-
- **Rankine / Curie (2003–2004 / GeForce 5 / 6 / 7):** Einführung von programmierbaren Shadern (Shader Model 2.0 / 3.0), die den Grundstein für moderne 3D-Grafik legten.
- **Kelvin (2001 / GeForce 3 / 4):** Erste programmierbare Vertex- und Pixel-Shader (DirectX 8). Basis für die Grafikeinheit der allerersten Xbox.
- **Celsius (1999 / GeForce 256):** Die weltweit **erste "GPU"** (Graphics Processing Unit), da sie die T&L-Engine (Transform and Lighting) direkt auf dem Chip integrierte und die CPU entlastete.
- **Fahrenheit (1998 / RIVA TNT):** Erste echte 2D/3D-Kombichips mit 32-Bit-Farbtiefe.
- **NV1 (1995):** NVIDIAs allererster Grafikchip. Nutzte quadratische Oberflächen (Quadratic Texture Mapping) statt Dreiecken, was sich am Markt jedoch nicht durchsetzte.
-

Möchtest du Details zu einer **bestimmten Architektur** (z. B. den Unterschied zwischen Turing und Ampere) oder interessieren dich die technischen Details der **Tensor- und RT-Kerne** genauer?