

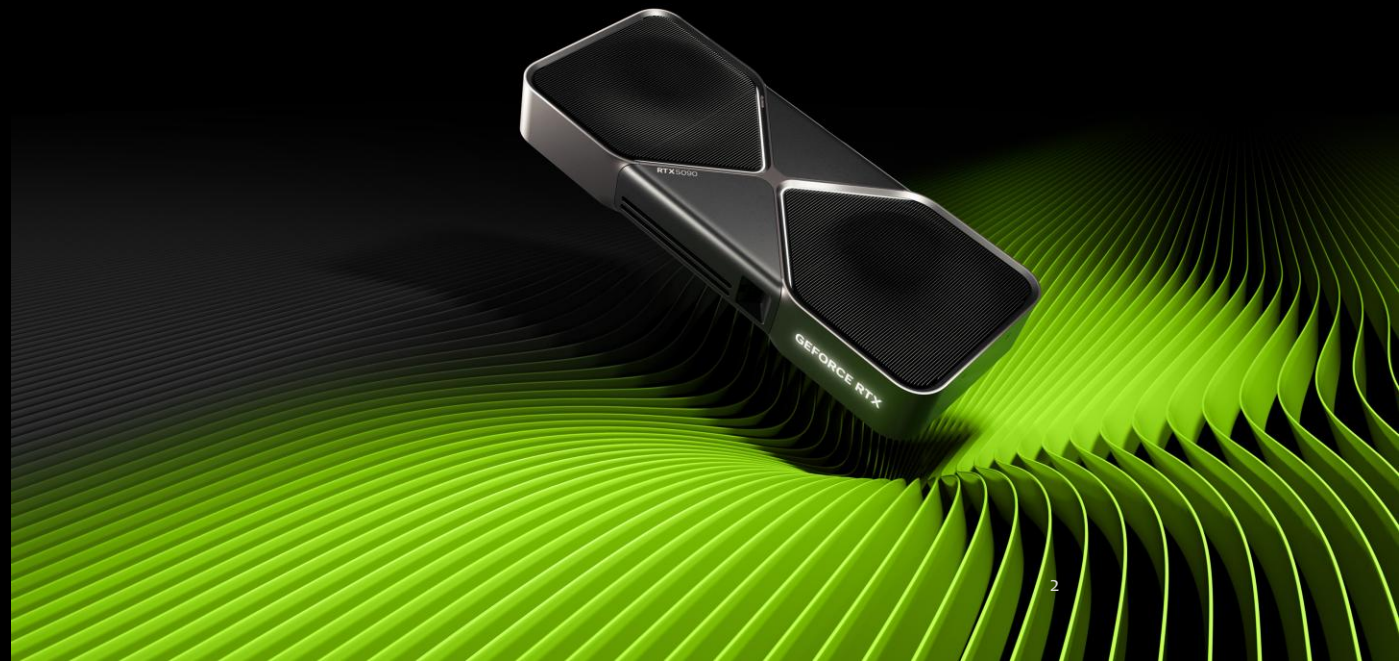


NVIDIA

Ivan Beneda
2 Ročník APiN

Obsah

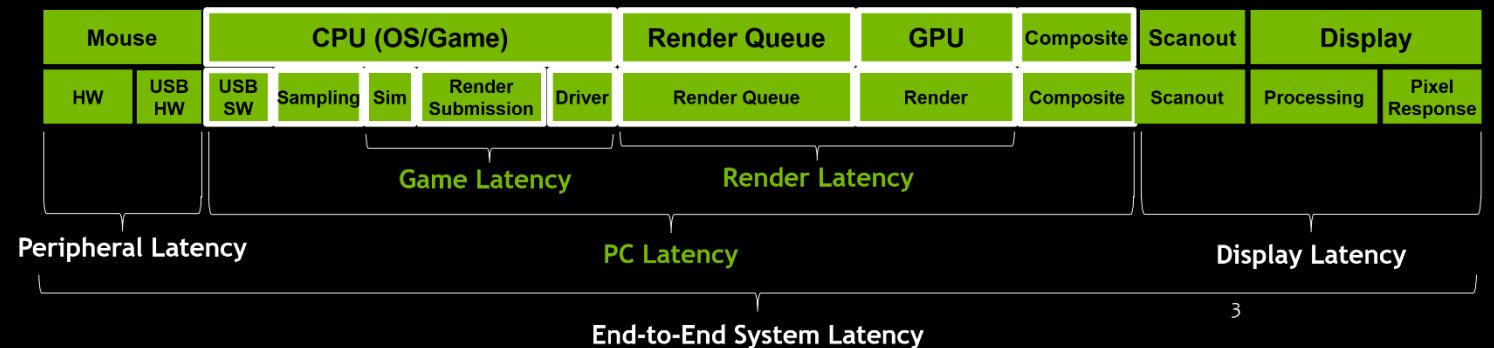
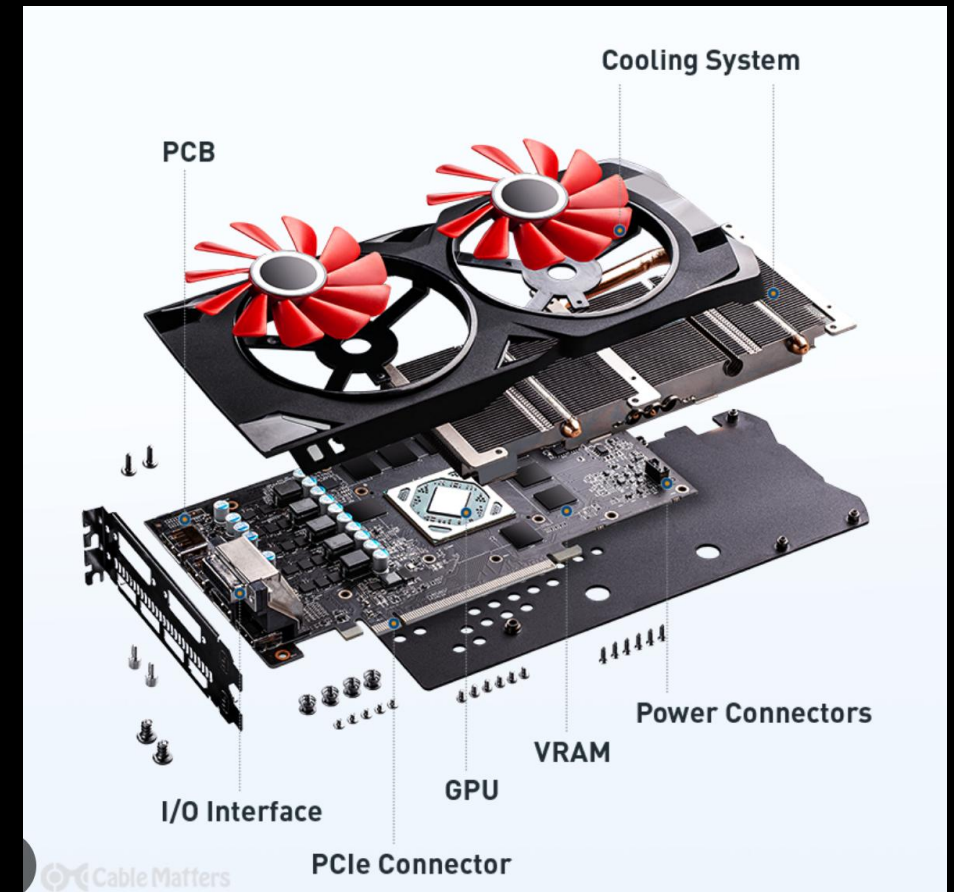
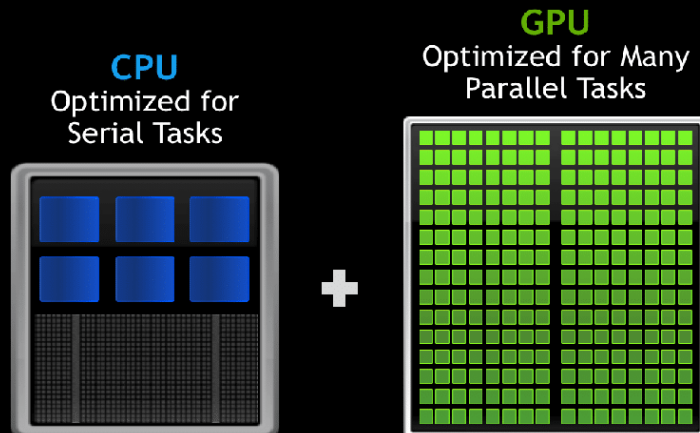
- GPU v skratke
- NVIDIA NV1
- Historické míľniky (GeForce)
- Architektúry RTX série
- Technológie
- Raytracing a pathtracing
- Budúcnosť (neural rendering)
- Overclocking



GPU v skratke

- Paralelné spracovanie dát
- Zrýchlenie grafiky a výpočtov
- Využitie v hrách, AI a vizualizáciách

Why a GPU?



Nvidia NV1

- Skôr akcelerátor ako GPU
- Prvý NVIDIA produkt
- Spájal audio, video, 2D aj 3D grafiku



Graphics Processor	Graphics Card	Clock Speeds	Memory
GPU Name: NV1	Release Date: Sep 30th, 1995	GPU Clock: 75 MHz	Memory Size: 2 MB
Foundry: ST	Generation: NV1	Memory Clock: 60 MHz	Memory Type: EDO
Process Size: 500 nm	Successor: Riva	Render Config	Memory Bus: 64 bit
Transistors: 1 million	Production: End-of-life	Pixel Shaders: 1	Bandwidth: 480.0 MB/s
Density: 7.4K / mm ²	Launch Price: 299 USD	Vertex Shaders: 0	Theoretical Performance
Die Size: 135 mm ²	Bus Interface: PCI	TMUs: 1	Pixel Rate: 75.00 MPixel/s
Board Design	Reviews: 1 in our database	ROPs: 1	Texture Rate: 75.00 MTexture/s
Slot Width: Single-slot	Graphics Features		
Length: 190 mm 7.5 inches	DirectX: 1.0		
TDP: 2 W	OpenGL: N/A		
Suggested PSU: 200 W	OpenCL: N/A		
Outputs: 1x VGA	Vulkan: N/A		
Power Connectors: None	Pixel Shader: N/A		
	Vertex Shader: N/A		

GeForce 256

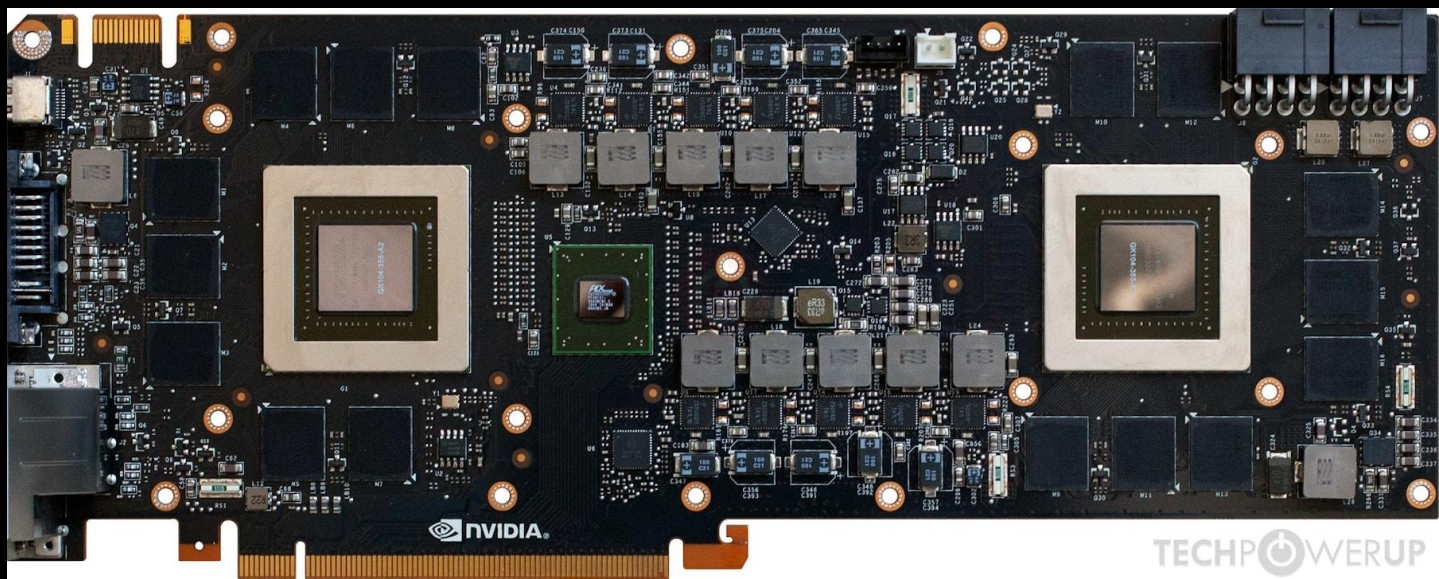
- Prvé GPU
- Začiatok GeForce
- Presun výpočtov 3D grafiky z CPU na GPU



Graphics Processor	Graphics Card	Clock Speeds	Memory
GPU Name: NV10	Release Date: Dec 23rd, 1999	GPU Clock: 120 MHz	Memory Size: 32 MB
Architecture: Celsius	Generation: GeForce 256	Memory Clock: 150 MHz 300 Mbps effective	Memory Type: DDR
Foundry: TSMC	Predecessor: TNT2		Memory Bus: 128 bit
Process Size: 220 nm	Successor: GeForce 2		Bandwidth: 4.800 GB/s
Transistors: 17 million	Production: End-of-life		
Density: 122.3K / mm ²	Bus Interface: AGP 4x		
Die Size: 139 mm ²	Reviews: 1 in our database		
Board Design	Graphics Features	Render Config	Theoretical Performance
Slot Width: Single-slot	DirectX: 7.0	Pixel Shaders: 4	Pixel Rate: 480.0 MPixel/s
TDP: unknown	OpenGL: 1.2	Vertex Shaders: 0	Texture Rate: 480.0 MTexture/s
Suggested PSU: 200 W	OpenCL: N/A	TMUs: 4	
Outputs: 1x VGA	Vulkan: N/A	ROPs: 4	
Power Connectors: None	Pixel Shader: N/A		
	Vertex Shader: N/A		

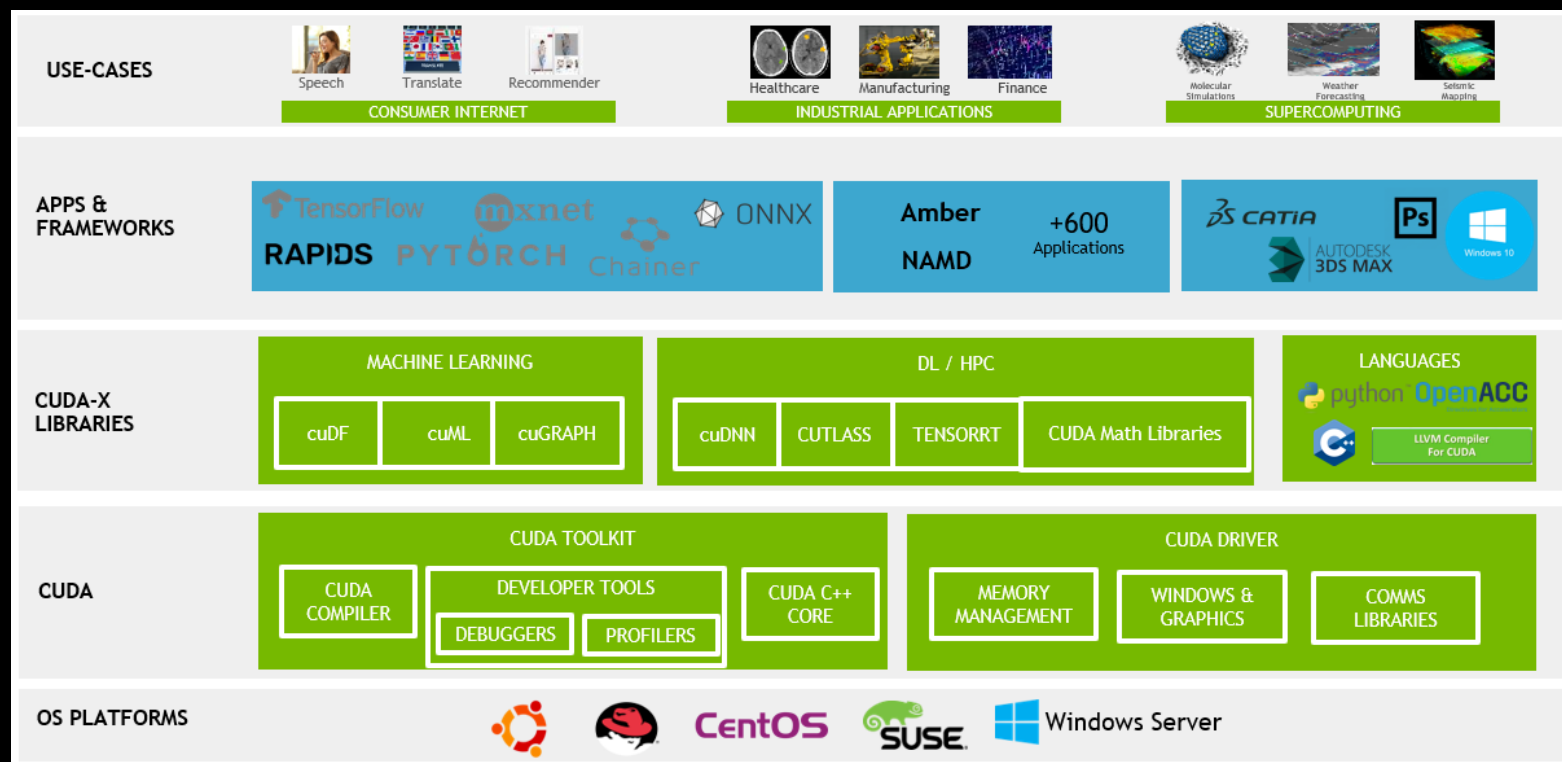
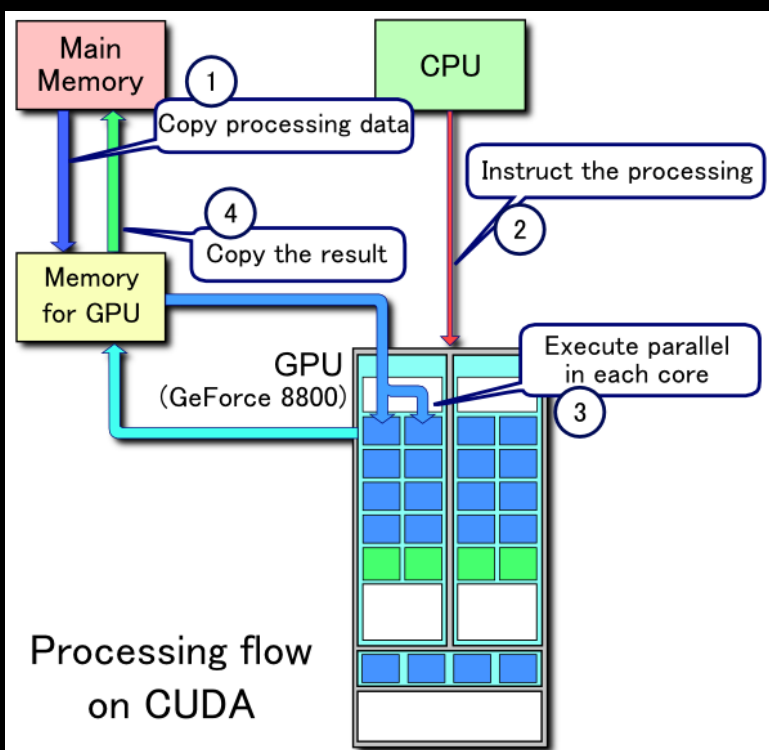
Historické milníky

- SLI (2004)
- CUDA (2006)
- Dual-GPU (2006)
- NVLINK (2014)
- Real-time raytracing (2018)
- Neural rendering (2026+)



CUDA

- Akcelerované výpočty na GPU
- Základ pre AI a superpočítače
- CUDA Toolkit



SLI/NVLINK

- Prepojenie viacerých GPU
- 2way až 4way SLI
- NVLINK Unified memory



Fourth Generation

Fifth Generation

Sixth Generation

NVLink Bandwidth per GPU

900GB/s

1,800GB/s

3,600 GB/s

Maximum Number of Links per GPU

18

18

36

Supported NVIDIA Architectures

NVIDIA Hopper™ architecture

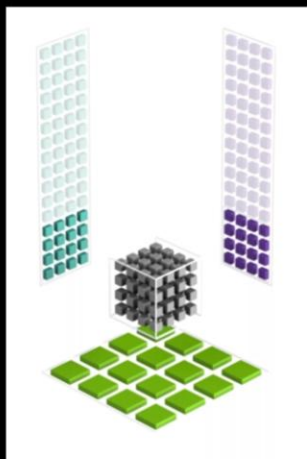
NVIDIA Blackwell architecture

NVIDIA Rubin Platform

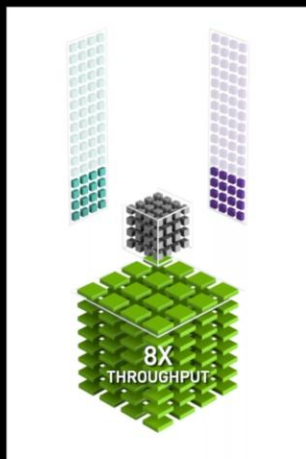
Tensore cores

- Prvý výskyt vo Volta arch
- Určené pre DLSS
- Dnes AI akcelerácia

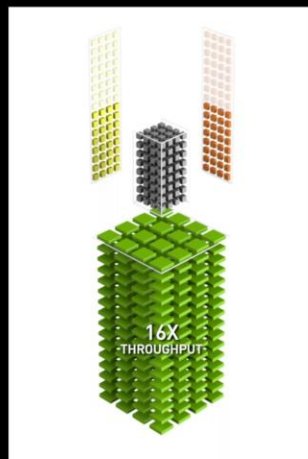
Blackwell 5th Generation Tensor Cores with FP4



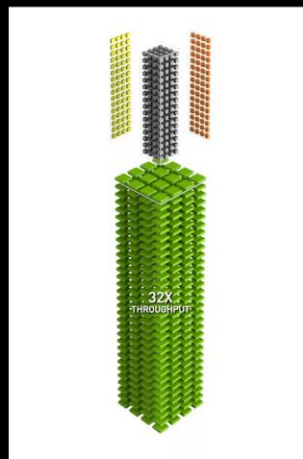
Pascal



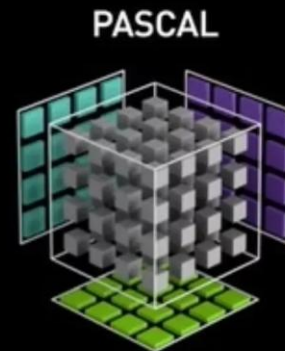
Turing Tensor Core
FP16



Ada Tensor Core
FP8

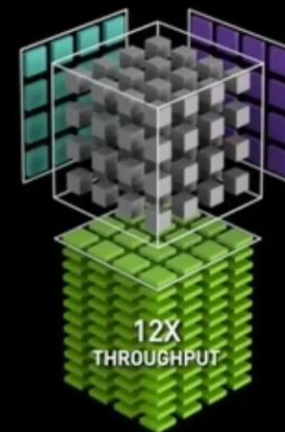


Blackwell Tensor Core
FP4



PASCAL

VOLTA TENSOR CORES



The Tensor Core

$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

FP16 or FP32 FP16 FP16 or FP32

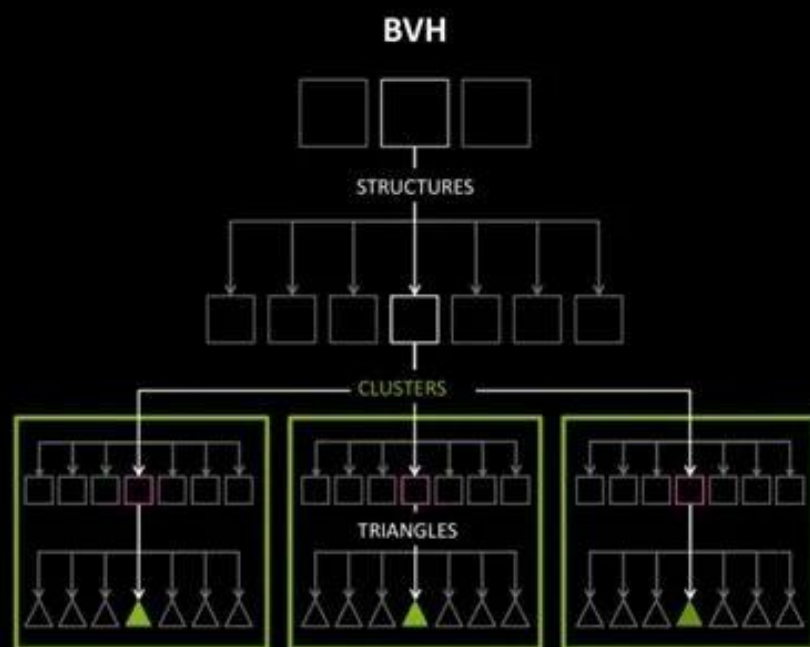
$$\text{FMA: } d = a * b + c$$

$$D_{0,0} = A_{0,0} * B_{0,0} + A_{0,1} * B_{1,0} + A_{0,2} * B_{2,0} + A_{0,3} * B_{3,0} + C_{0,0}$$

(Image Credit: <https://devblogs.nvidia.com/programming-tensor-cores-cuda-9/>)

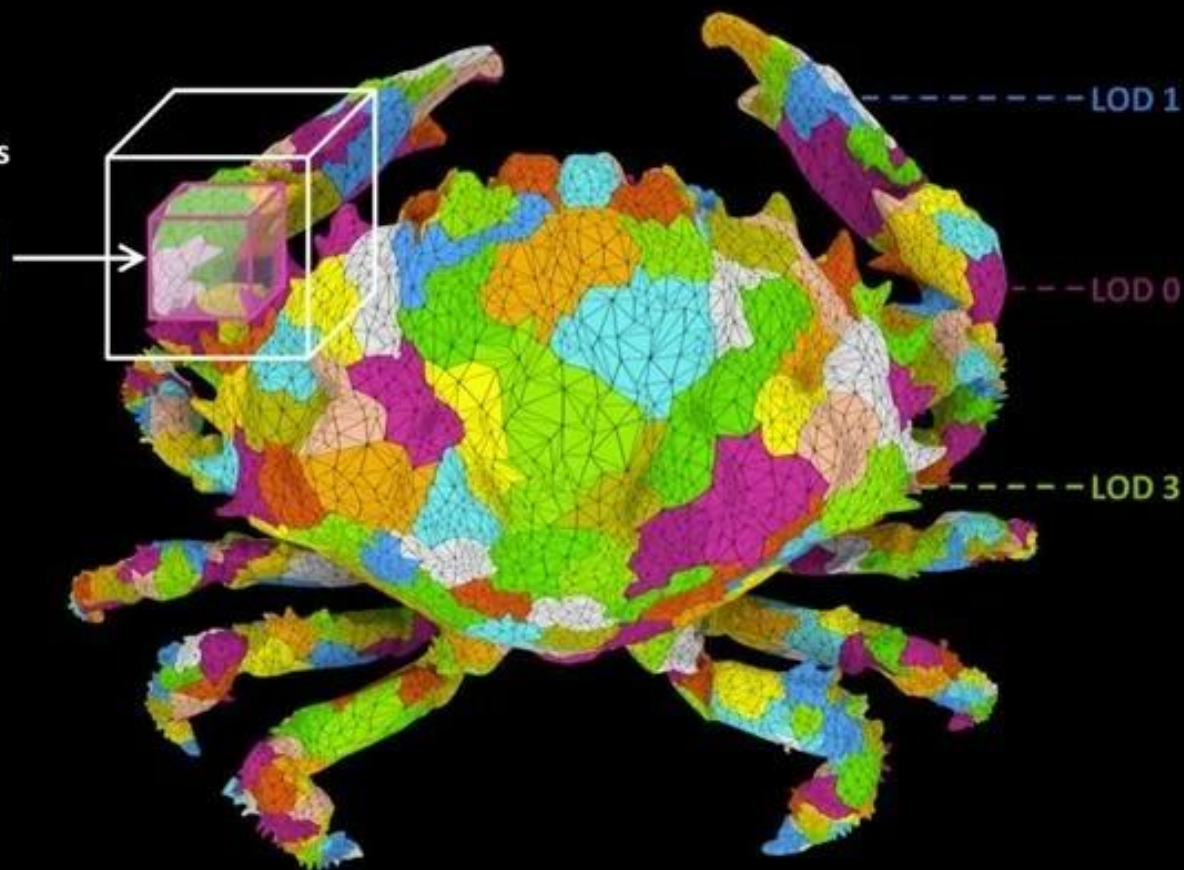
RTX Mega Geometry

Accelerate BVH updates for cluster-based systems like Nanite



Compress and Cache Clusters Over Many Frames

Trace Rays



Mesh Composed of Clusters at Different LOD Levels

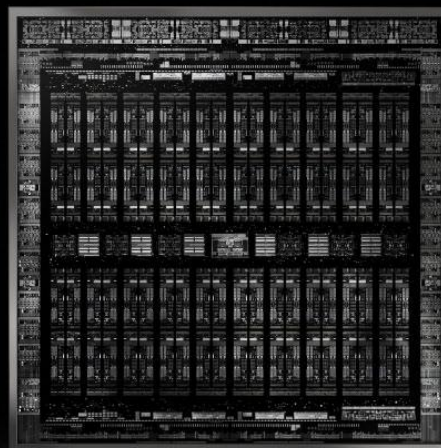
GTX éra

- GTX 8800 (2006)
- GTX 200 (2008)
- GTX 400 (2010)
- GTX 600 (2012)
- GTX 700 (2014)
- GTX 1000 (2016)



NVIDIA RTX 2000

- 12nm výrobný proces
- GDDR6 pamäť
- 30% navýšenie výkonu



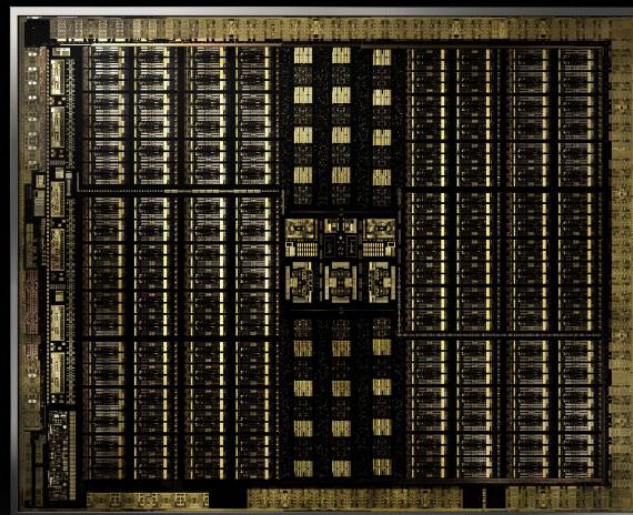
PASCAL

11.8 Billion xtors | 471 mm² | 24 GB 10GHz

48 CUDA
cores/SM

64 CUDA
cores/SM
+8 Tensor
+1 RTX

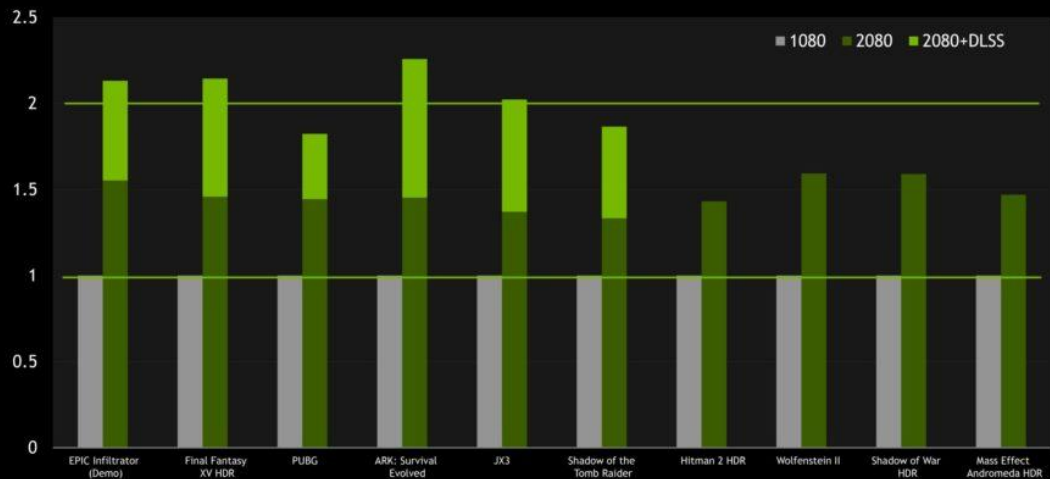
35% SM dedicated to
non-CUDA cores



TURING

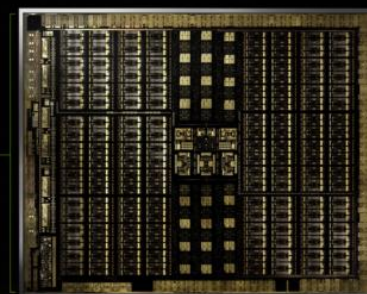
18.6 Billion xtors | 754 mm² | 48+48 GB 14GHz

RTX 2080: 2X 1080



TURING BUILT FOR RTX

GREATEST LEAP SINCE 2006 CUDA GPU



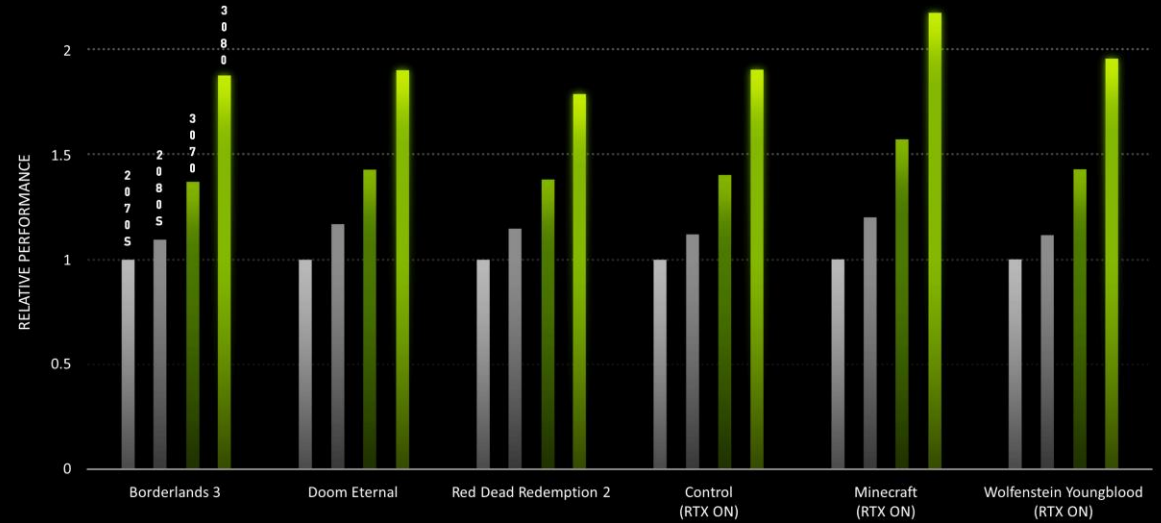
Turing SM
14 TFLOPS + 14 TIPS
urrent FP & INT Execution
Variable Rate Shading

Tensor Core
110 TFLOPS FP16
220 TOPS INT8
440 TOPS INT4

RT Core
10 Giga Rays/sec
Ray Triangle Intersection
BVH Traversal

NVIDIA RTX 3000

- 8nm výrobný proces
- GDDR6/X pamäť
- Lepší pomer cena/výkon



Games run at 4K with i9 CPU

Specs

	GeForce RTX 3090 Ti	GeForce RTX 3090	GeForce RTX 3080 Ti	GeForce RTX 3080	GeForce RTX 3070 Ti	GeForce RTX 3070	GeForce RTX 3060 Ti	GeForce RTX 3060	GeForce RTX 3050 (8 GB)	GeForce RTX 3050 (6 GB)
NVIDIA CUDA Cores	10752	10496	10240	8960 / 8704	6144	5888	4864	3584	2560 [®]	2304
Boost Clock (GHz)	1.86	1.70	1.67	1.71	1.77	1.73	1.67	1.78	1.78 [®]	1.47
Memory Size	24 GB	24 GB	12 GB	12 GB / 10 GB	8 GB	8 GB	8 GB	12 GB / 8 GB	8 GB	6 GB
Memory Type	GDDR6X	GDDR6X	GDDR6X	GDDR6X	GDDR6X	GDDR6	GDDR6 / GDDR6X	GDDR6	GDDR6	GDDR6

NVIDIA RTX 4000

- 4nm výrobný proces
- FP8
- Framegen

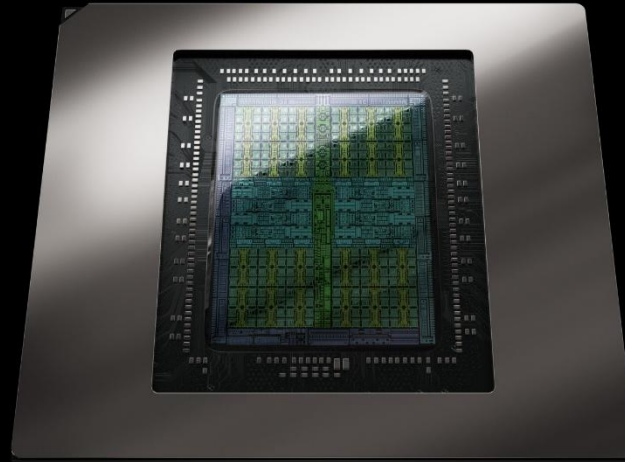


Specs

	RTX 4090	RTX 4080 SUPER	RTX 4080	RTX 4070 Ti SUPER	RTX 4070 Ti	RTX 4070 SUPER	RTX 4070	RTX 4060 Ti	RTX 4060
NVIDIA CUDA Cores	16384	10240	9728	8448	7680	7168	5888	4352	3072
Boost Clock (GHz)	2.52	2.55	2.51	2.61	2.61	2.48	2.48	2.54	2.46
Memory Size	24 GB	16 GB	16 GB	16 GB	12 GB	12 GB	12 GB	16 GB / 8 GB	8 GB
Memory Type	GDDR6X	GDDR6X	GDDR6X	GDDR6X	GDDR6X	GDDR6X	GDDR6 / GDDR6X	GDDR6	GDDR6

NVIDIA RTX 5000

- 4nm výrobný proces
- GDDR7 pamäť
- MFG, Reflex2



Fifth-Gen Tensor Cores

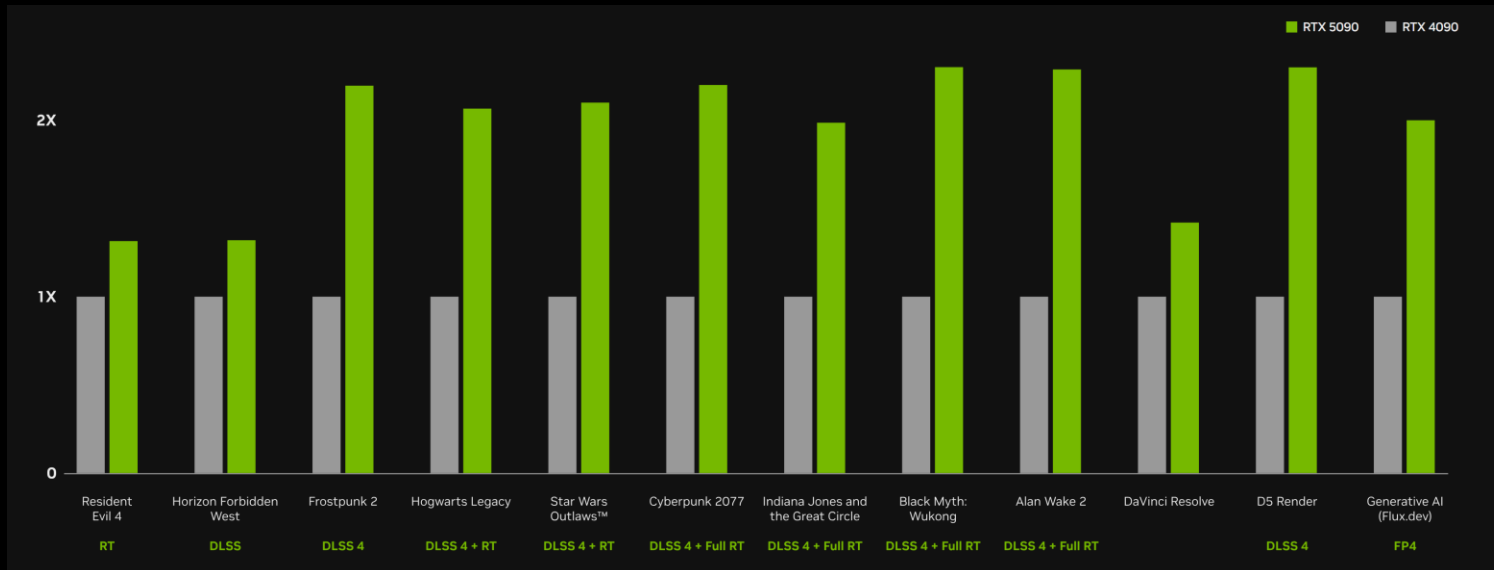
Max AI performance with FP4 and DLSS 4.5

New Streaming Multiprocessors

Optimized for neural shaders

Fourth-Gen Ray Tracing Cores

Built for Mega Geometry

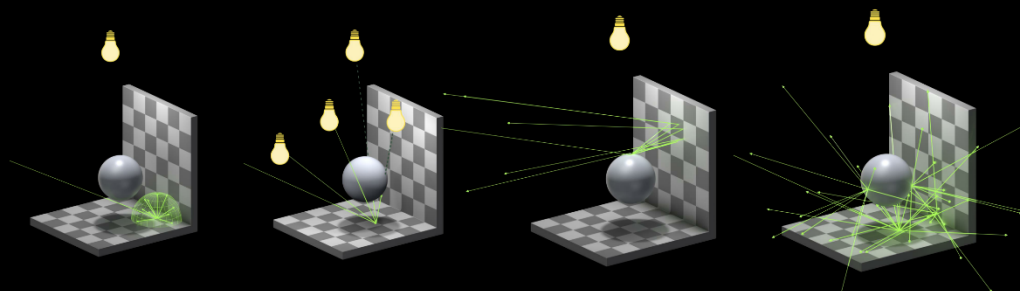


Real-time Raytracing (NVIDIA)

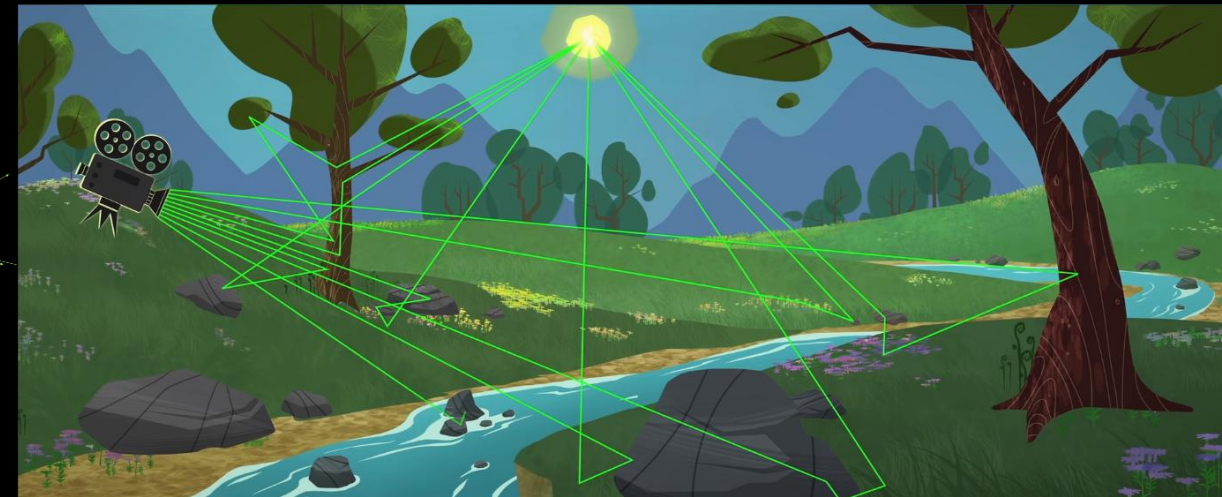
- Prvá GPU s RTX
- Extrémne náročné na HW



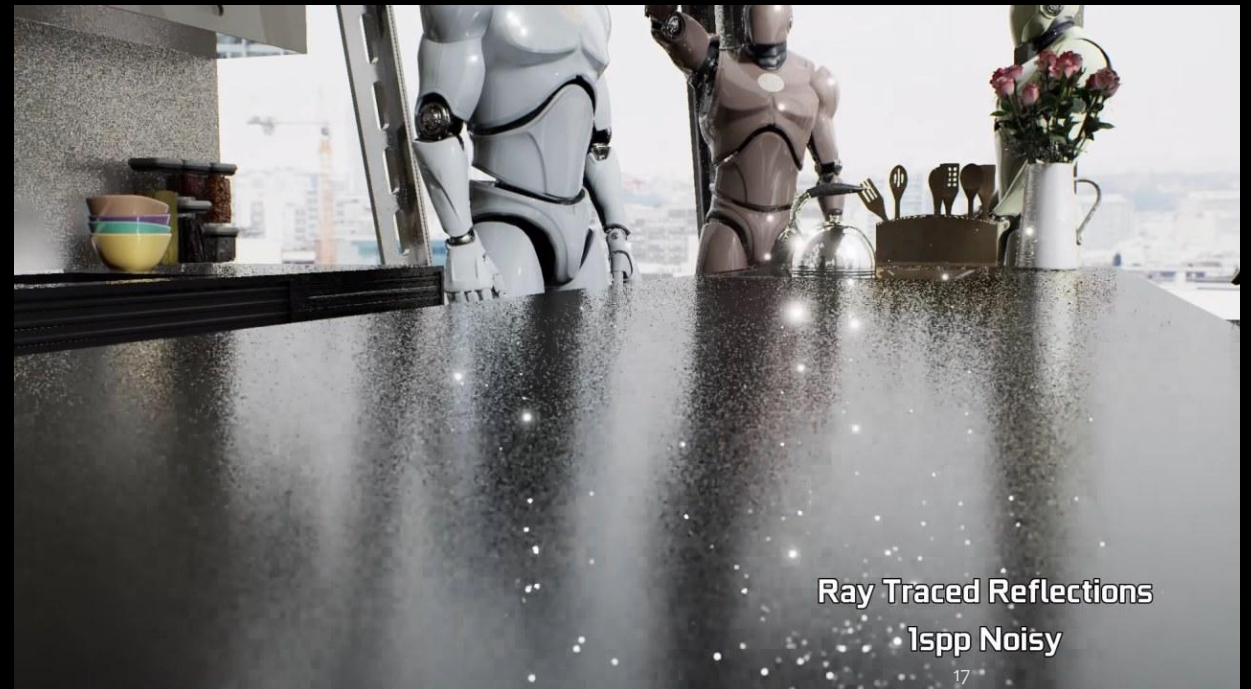
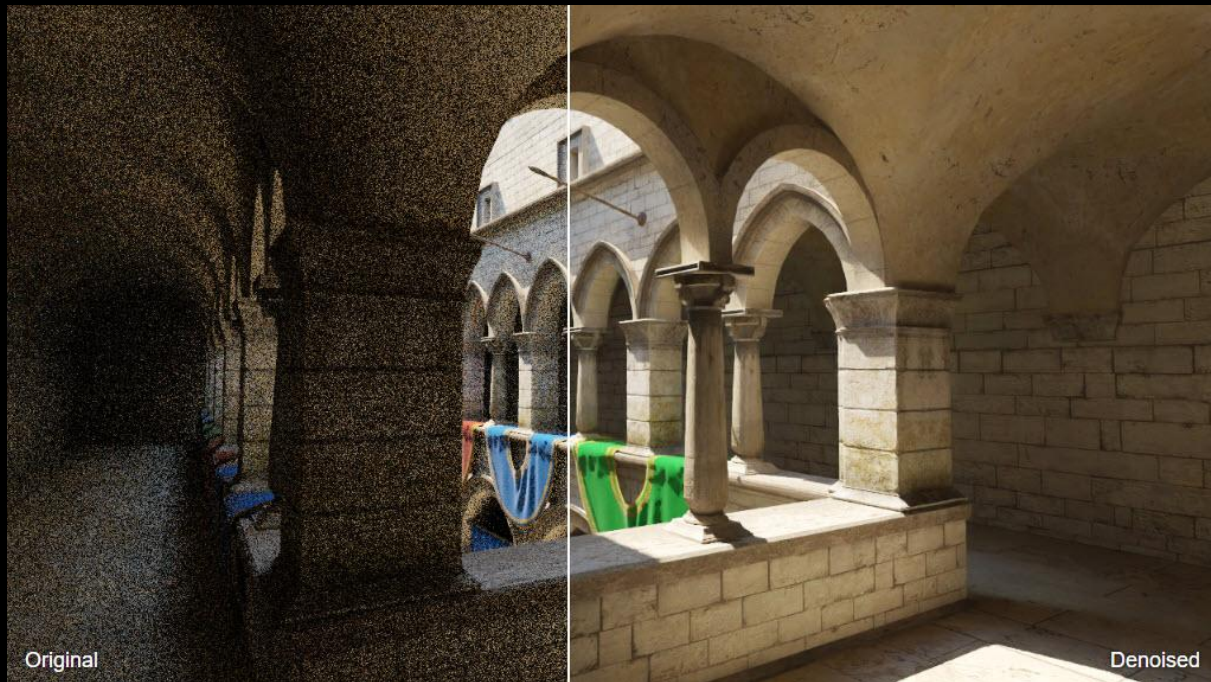
RT TECHNIQUE WORKLOADS



	Ambient Occlusion	Shadows	Reflections	Global Illumination
Rays Needed	Many	Number/Type Lights	Surface Type	Many
BVH Search	Local	Global	Global	Global
Bounces	No	No	Optional	Optional



Denoising



Ukážka Raytracing (Tiene)



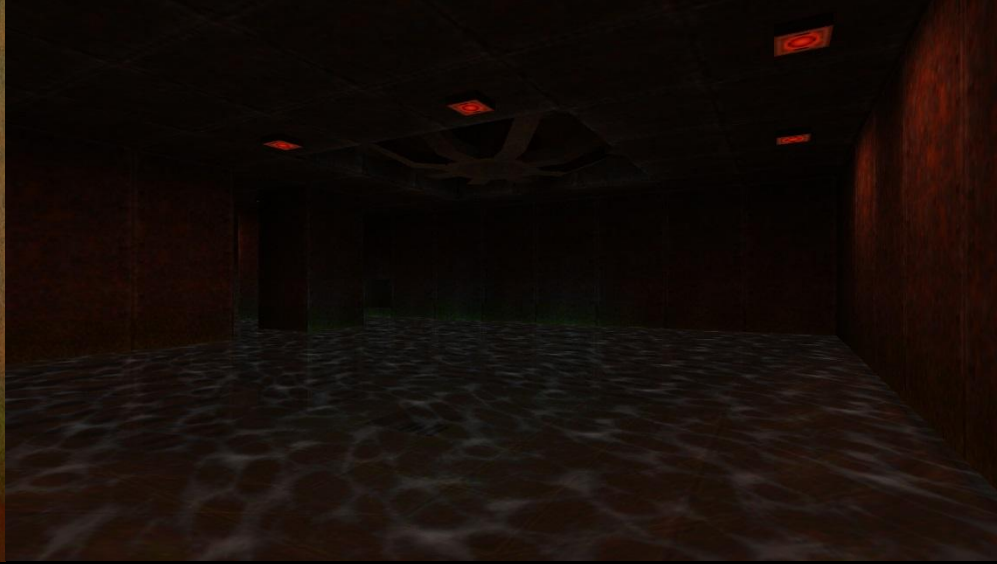
Ukážka Raytracingu (Odrazy)



Ukážka Raytracingu (Global Illumination)



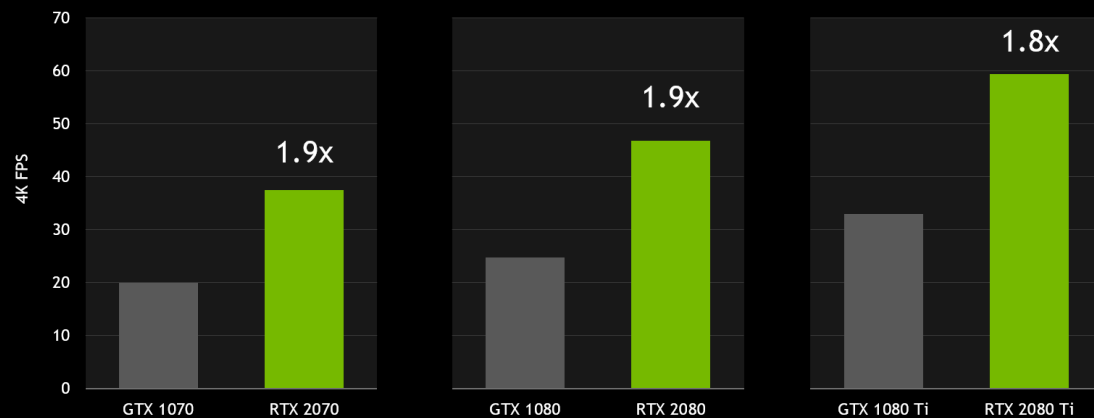
Ukážka Pathtracingu



DLSS

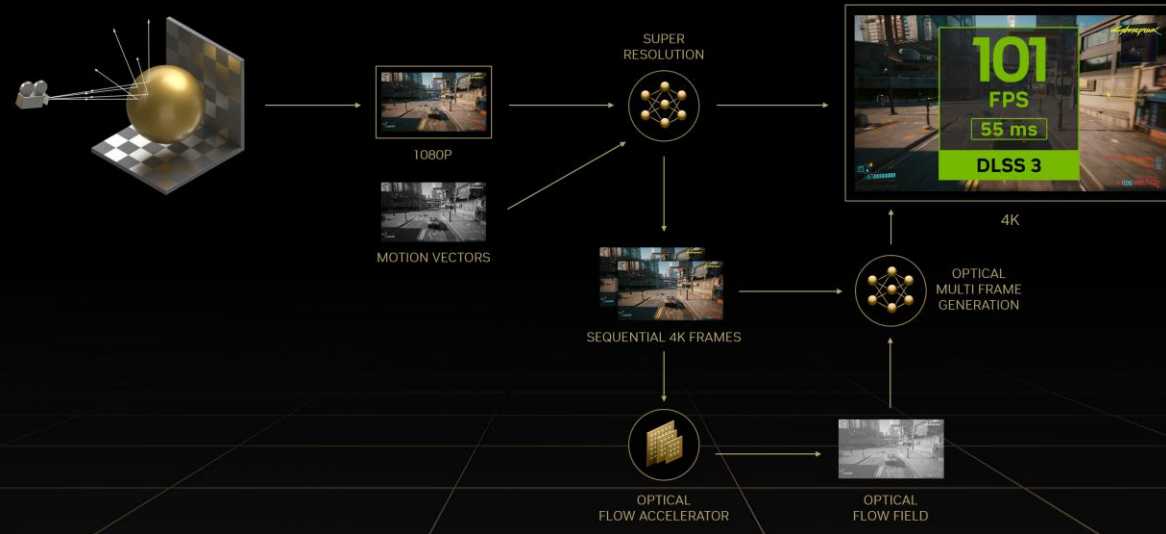
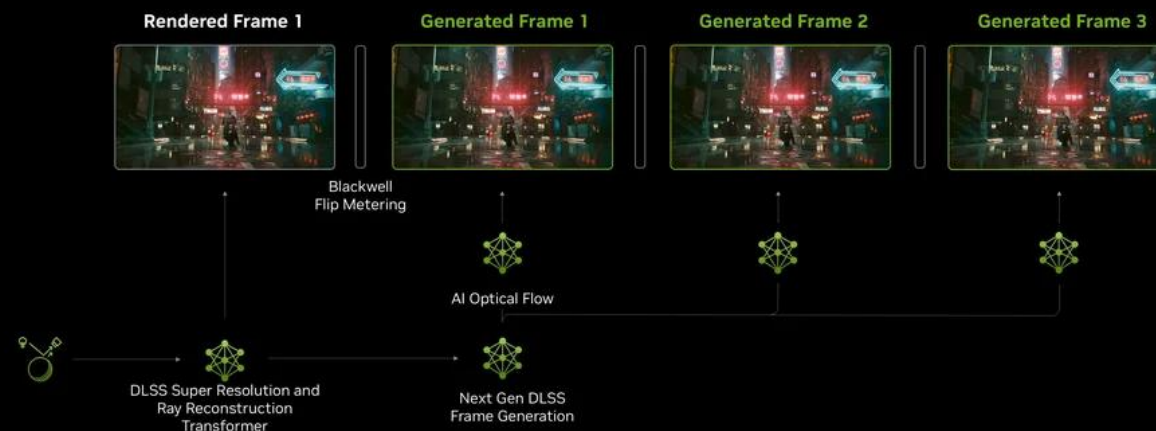
- 40% až 50% nárůst výkonu
- Umelá inteligencia

MAXIMUM 4K PERFORMANCE WITH DLSS IN FINAL FANTASY XV BENCHMARK



Win10 RS5, i9-7900X 3.3GHz, 16GB, GTX 10 Series run with TAA, RTX 20 Series run with DLSS, max benchmark settings, DLSS contributes ~38% performance

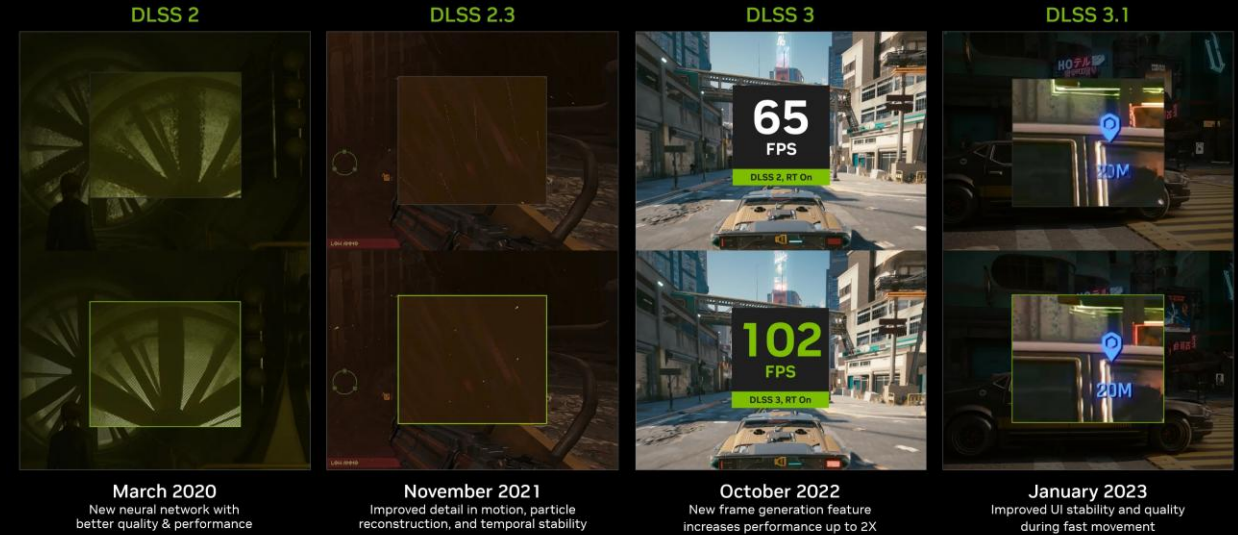
DLSS 4 Multi Frame Generation



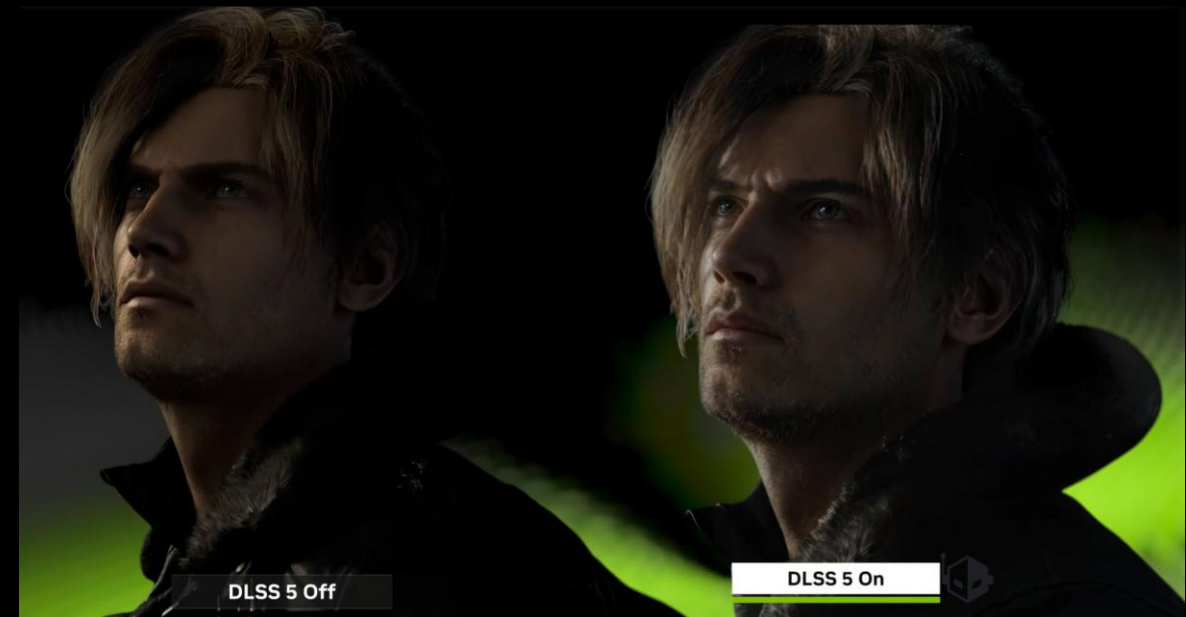
DLSS generácie

- DLSS 4 a vyššie náročnejšie na beh
- Už nie len upscaler

DLSS Keeps Getting Better



DLSS 4.5

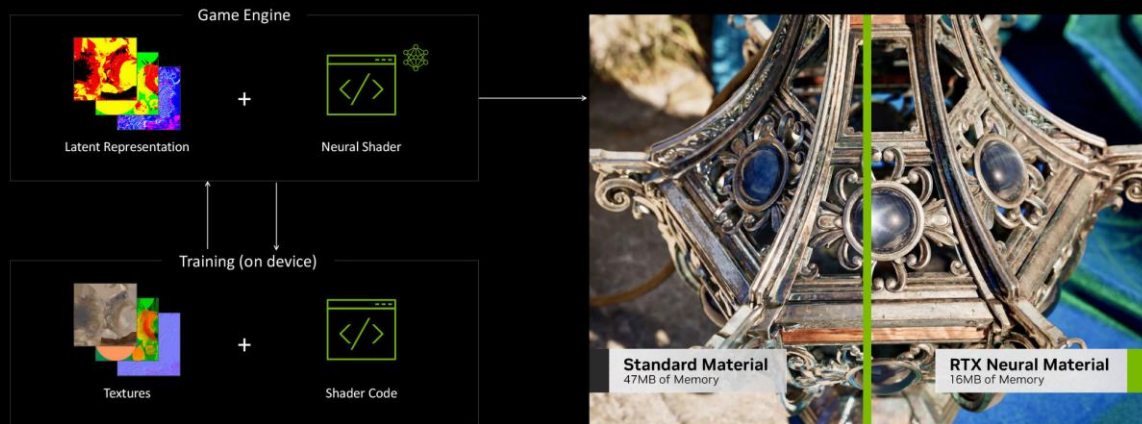


Neural rendering

- RTX Neural Texture Compression
- RTX Neural Materials
- RTX Neural Radiance Cache

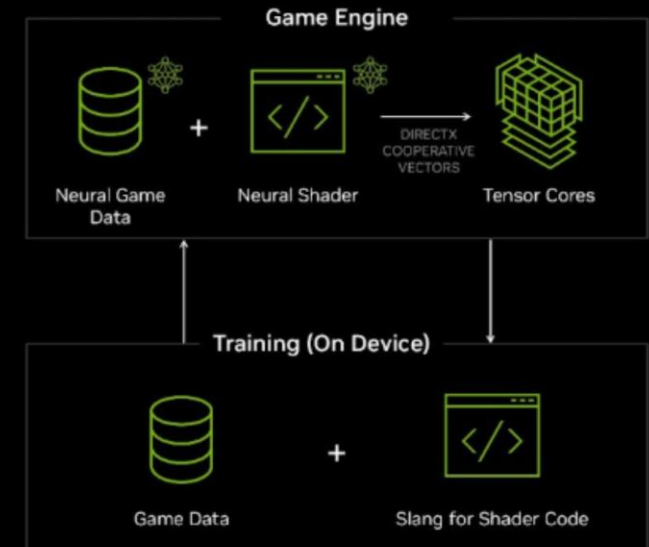
RTX Neural Materials

Film quality assets in real-time



RTX Neural Shaders

Reinventing How Shaders are Programmed



Examples

Tuscan Villa Scene with BCn textures – 6.5 GB VRAM



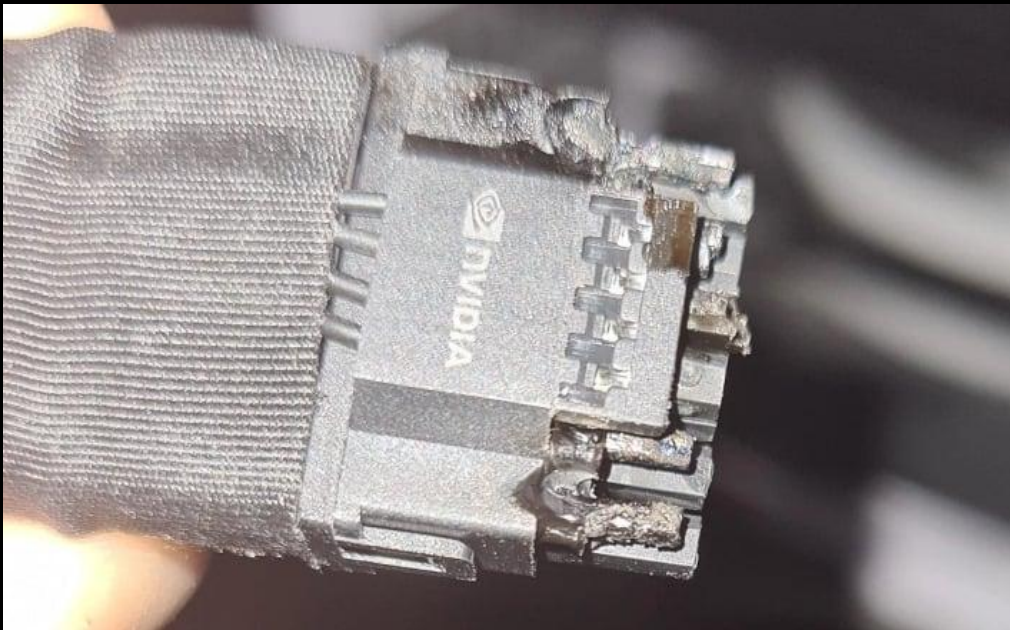
Examples

Tuscan Villa Scene with NTC textures – 970 MB VRAM



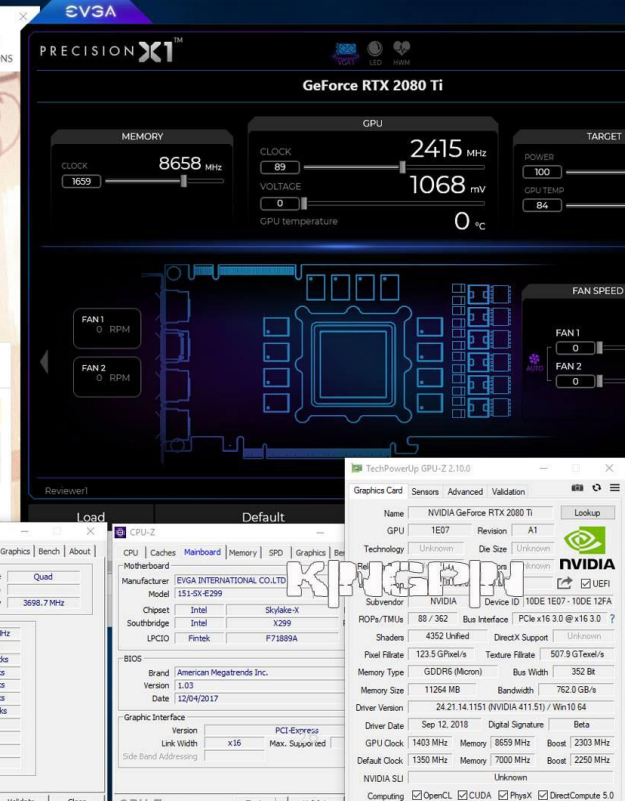
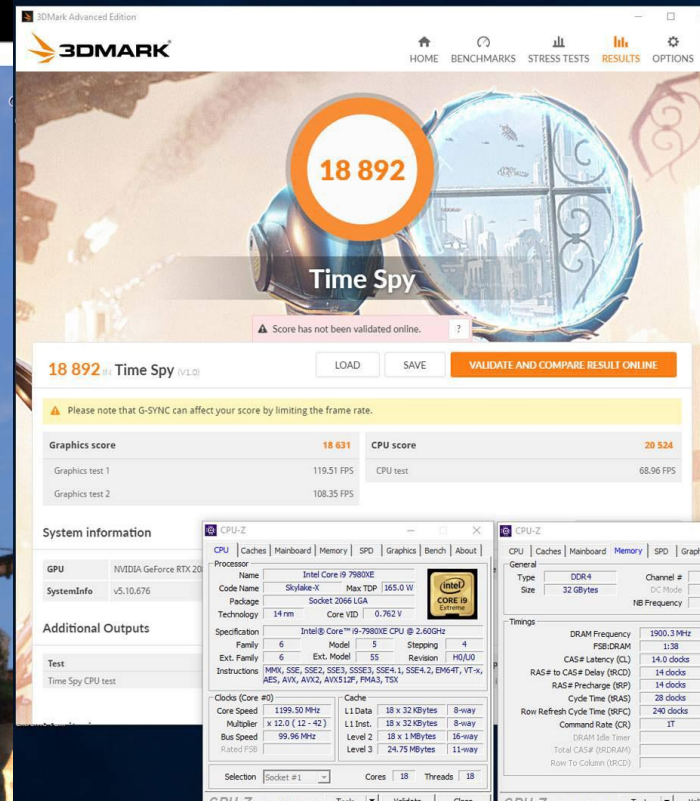
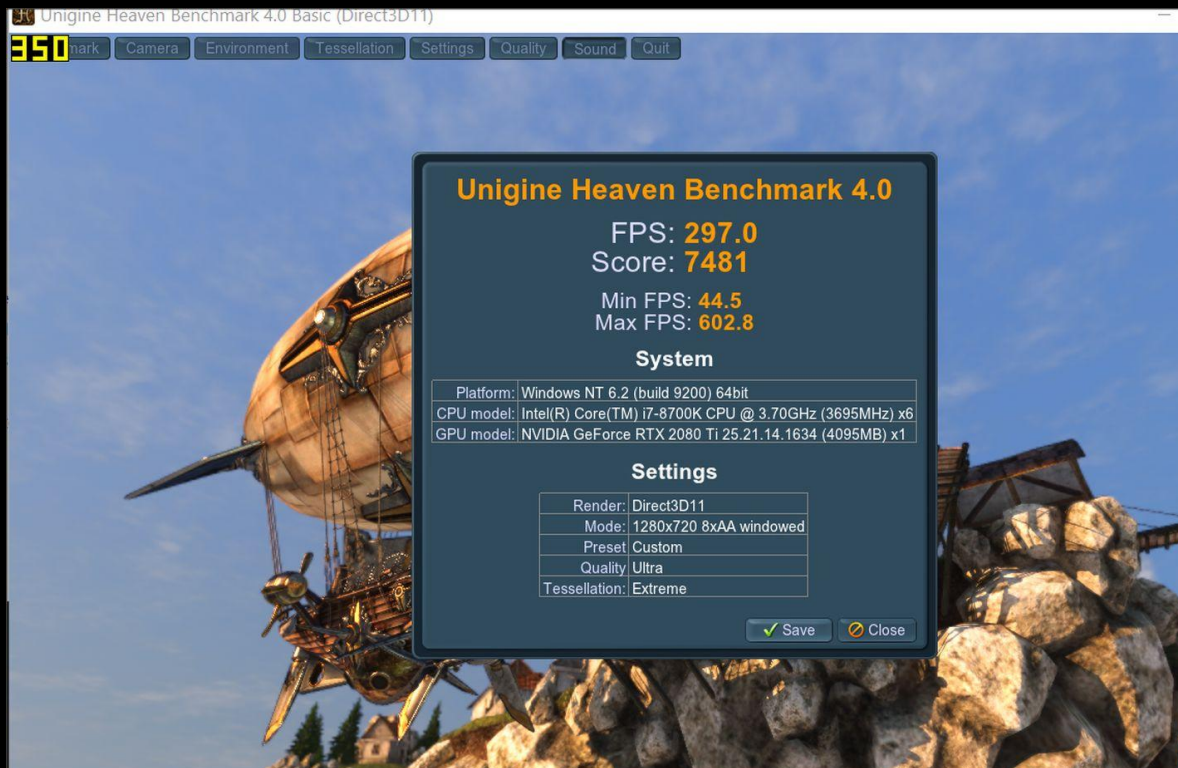
12VHPRW

- Max 600W
- Žiaden load balancing



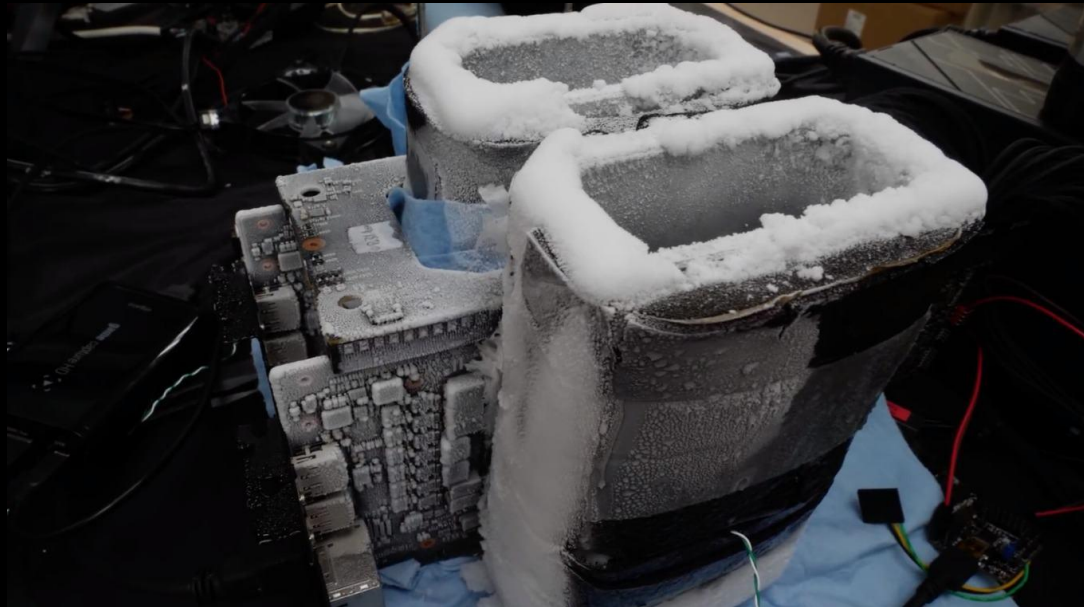
GPU OC / diagnostika

- MSI Afterburner
- Furmark, 3DMark
- GPU-Z



ExtremeOC

- LN₂



Zdroje

- [1] NVIDIA. NVIDIA NV1 Timeline (1995). Dostupné online: https://www.nvidia.com/content/timeline/time_95.html
- [2] NVIDIA. NVIDIA GeForce 256 Timeline (1999). Dostupné online: https://www.nvidia.com/content/timeline/time_99.html
- [3] NVIDIA Developer. CUDA Platform for Accelerated Computing. Dostupné online: <https://developer.nvidia.com/cuda>
- [4] NVIDIA Developer. Real-Time Ray Tracing. Dostupné online: <https://developer.nvidia.com/rtx/ray-tracing>
- [5] NVIDIA. NVIDIA DLSS Technology. Dostupné online: <https://www.nvidia.com/en-eu/geforce/technologies/dlss/>
- [6] NVIDIA. NVIDIA Tensor Cores. Dostupné online: <https://www.nvidia.com/en-eu/data-center/tensor-cores/>
- [7] NVIDIA. NVIDIA NVLink and NVSwitch. Dostupné online: <https://www.nvidia.com/en-us/data-center/nvlink/>
- [8] NVIDIA. The NVIDIA Ada Lovelace Architecture. Dostupné online: <https://www.nvidia.com/en-us/geforce/ada-lovelace-architecture/>
- [9] NVIDIA. NVIDIA Blackwell Architecture. Dostupné online: <https://www.nvidia.com/en-eu/data-center/technologies/blackwell-architecture/>
- [10] NVIDIA Developer. NVIDIA RTX Kit. Dostupné online: <https://developer.nvidia.com/rtx-kit>



Ďakujem za pozornosť
Priestor pre vaše otázky

