

Grafický subsystém - nVidia a paralelné výpočty

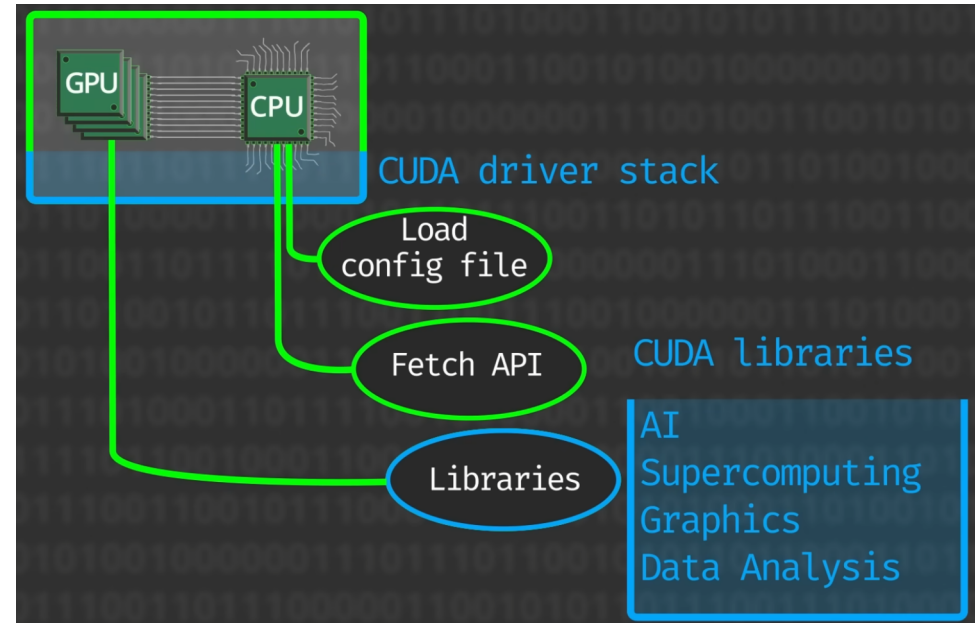
Matúš Goč

Rozdieli medzi CPU a GPU

CPU	GPU
Sériové operácie	Paralelné operácie
Malý počet výkonných jadier	Veľký počet jednoduchších jadier
Rýchla odozva	Vysoká priepustnosť
Komplexná riadiaca jednotka na riadenie vetvenia, plánovania a vykonávania inštrukcií	Zjednodušená riadiaca jednotka, optimalizovaná na vykonávanie opakovaných paralelných operácií
Veľká vyrovnávacia pamäť na ukladanie často používaných údajov	Menšia vyrovnávacia pamäť, využíva vysokú šírku pásma pamäte pre rýchly prenos dát

Compute Unified Device Architecture

- platforma pre paralelné výpočty a programovací model vydaný v roku 2007
- umožňuje zrýchľovať náročné úlohy
- poskytuje softvérové nástroje ktoré vývojárom umožňujú používať jazyky C++, Python a Fortran na využívanie paralelného výpočtového výkonu GPU



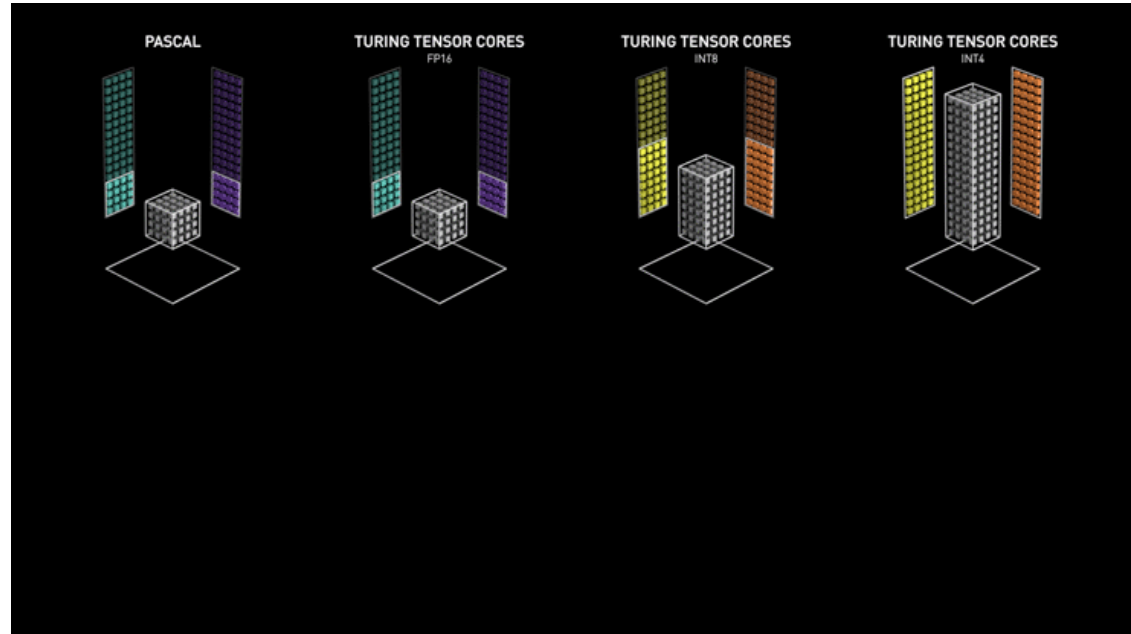
What is CUDA? – Computerphile

Compute Unified Device Architecture

- Ako to funguje: CUDA umožňuje heterogénne výpočty, kde CPU spracováva sériové úlohy a GPU paralelné úlohy. Vývojári píšu „kernely“ – funkcie, ktoré bežia súbežne na jadrách GPU
- Komponenty: Sada nástrojov CUDA obsahuje kompilátor (NVCC), nástroje na ladenie/profilovanie a knižnice (napr. cuBLAS, cuDNN) pre optimalizovaný výkon
- Aplikácie: Je nevyhnutná pre vývoj umelej inteligencie, vysokovýkonné výpočty, spracovanie obrazu a vedecké výpočty
- Ekosystém a podpora: CUDA podporuje rôzne architektúry (Ampere, Ada atď.) a je široko používaná s frameworkami ako PyTorch a TensorFlow

Tensor jadrá

- jadrá špecializované na operácie násobenia a konvolúcie matíc
- predstavené v architektúre Volta (2017)



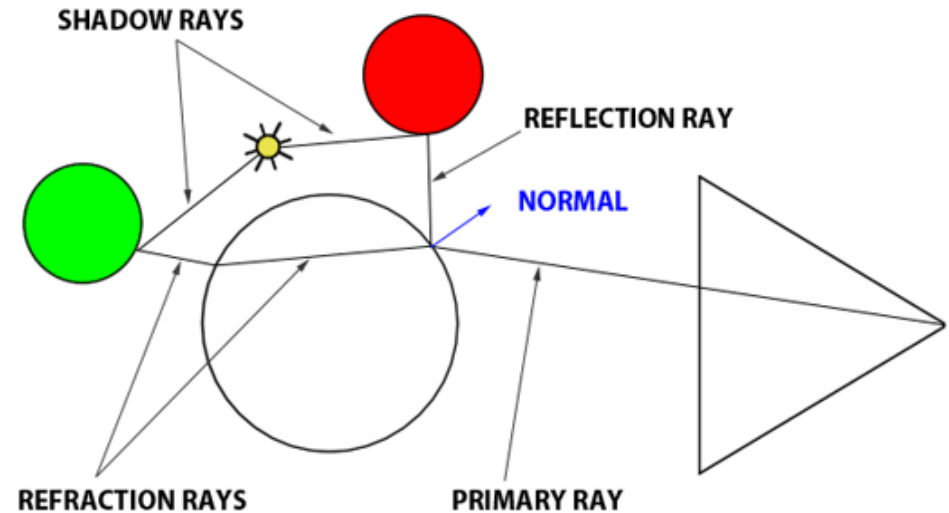
Porovnanie rýchlostí Tensor jadier medzi generáciami

Tensor jadrá

- Masívne zrýchlenie AI: Tensorové jadrá vykonávajú násobenie a akumuláciu matíc v takte, čo je oveľa efektívnejšie ako štandardné CUDA jadrá
- Výpočty so zmiešanou presnosťou: Podporujú aritmetiku so zmiešanou presnosťou, čo umožňuje rýchlejšie výpočty pomocou čísel s nižšou presnosťou bez toho, aby sa obetovala presnosť celkového modelu
- Aplikácie: Sú kľúčové pre spustenie a trénovanie AI neurónových sietí, inferencie, rozpoznávania obrazu a počítačového videnia v reálnom čase

Ray Tracing jadrá

- predstavené v architektúre Turing (2018)
- špecializované jadrá na urýchlenie detekcie priesečníkov lúčov s geometriou
- realistické tieň, odrazy, lom svetla v reálnom čase
- bez RT jadier ray tracing výkon padá o 80–90 %



Využitie optických princípov na výpočet dráh odrazených a lomených lúčov

Ray Tracing jadrá

- Účel: Vypočítavajú, ako svetlo interaguje s 3D scénami v reálnom čase, čo umožňuje fotorealistické osvetlenie, tiene a odrazy
- Ako fungujú: RT jadrá zrýchľujú dve najnáročnejšie úlohy sledovania lúčov: testy priesečníkov lúčov a trojuholníkov a prechod hierarchiou ohraničujúcich objemov
- Vplyv na výkon: Umožňujú GPU spracovať komplexné osvetlenie bez prerušovania jadier shaderov, čo poskytuje masívne zrýchlenie oproti tradičnému softvérovému sledovaniu lúčov
- Dopomocná technológia: Zatiaľ čo RT jadrá vypočítavajú lúče, tensorové jadrá sa používajú na odstránenie šumu z výsledného obrazu a CUDA jadrá spracovávajú tradičné úlohy vykresľovania

Deep Learning Super Sampling

- technológia vykresľovania s využitím AI na zvýšenie rozlíšenia obrázkov s nižším rozlíšením, čím zlepšuje grafický výkon a kvalitu v reálnom čase
- dovoľuje grafickej karte vykresľovať na nižšom rozlíšení bez straty kvality obrazu



Rozdieli medzi DLSS generáciami

Deep Learning Super Sampling

- Ako to funguje: DLSS vykresľuje hry s nižším rozlíšením a potom používa Tensor jadrá na doplnenie detailov
- DLSS Super Resolution: Využíva AI na zvýšenie rozlíšenia
- DLSS Frame Generation (3.0+): Generuje úplne nové snímky a vkladá ich medzi vykreslené snímky
- DLSS Ray Reconstruction (3.5+): Odstráňa šum ktorý vzniká pri vykresľovaní pomocou sledovania lúčov
- DLAA (Deep Learning Anti-Aliasing): Využíva DLSS v natívnom rozlíšení na vyhladenie okrajov a zlepšenie jasnosti obrazu

Historické míľniky

Rok	Míľník
1999	nVidia GeForce 256 – prvá GPU s podporou transformácie a osvetlenia v hardware-y
2001	GeForce 3 – prvé programovateľné shadery
2006	CUDA 1.0 – GPU pre všeobecné výpočty (GPGPU)
2010	Fermi – prvá GPU s L1/L2 cache a ECC pamäťou pre HPC
2012	AlexNet trénovaný na GPU → revolúcia v AI (ImageNet)
2016	Pascal (GP100) – NVLink, HBM2 pamäť, FP16
2017	Volta – prvé Tensor jadrá, určená pre datacentrá (V100)
2018	Turing – prvé RT jadrá a DLSS, GeForce RTX 2000
2020	Ampere – 2. gen. RT, 3. gen. Tensor, DLSS 2.0, A100 viacínštanciový grafický procesor
2022	Ada Lovelace – 4. gen. Tensor, 3. gen. RT, DLSS 3, AV1 enkódovanie (bezpoplatkový, vysoko efektívny formát kódovania videa)
2024	Blackwell (B100/B200) – GB200 NVL72 rack, FP4, transformer engine
2025	RTX 5000 (Blackwell pre spotrebiteľov) – DLSS 4, Multi Frame Gen
2026	DLSS 5 – technológia renderovania s AI s cieľom dosiahnuť fotorealistické osvetlenie a materiály

SúčasnÉ komponenty

- Spotrebiteľský segment
 - RTX 5090 – 92 mld. tranzistorov, 32 GB GDDR7, 575 W TDP
 - RTX 5080 / 5070 Ti / 5070 / 5060
 - Architektúra: Blackwell
 - Pamäť: GDDR7 (až 28 Gbps/pin)
- Datacentrá / AI
 - H100 (Hopper) – 80 GB HBM3, NVLink 4.0, transformer engine
 - H200 – HBM3e (141 GB), vyšší bandwidth
 - B100 / B200 (Blackwell) – FP4 inferencia, GB200 NVL72 (72× GPU v jednom racku)
 - NVLink Switch – škálovanie naprieč stovkami GPU
- Profesionálny / Workstation segment
 - RTX 6000 Ada – 48 GB ECC VRAM, pre 3D/VFX
 - nVidia RTX 4000/5000 Ada – kompaktné workstation GPU
- Kľúčové komponenty GPU
 - SM (Streaming Multiprocessor) – základná výpočtová jednotka
 - VRAM – GDDR6X / GDDR7 / HBM3e
 - ROPs / TMUs – render output / texture mapping units
 - Video enkóder/dekóder – NVENC / NVDEC (AV1, H.265)
 - NVLink / PCIe 5.0 – rozhranie ku CPU a ďalším GPU

Vývoj do budúcnosti

- Architektúry
 - Rubin (očakávaný tento rok) – nástupca Blackwellu, HBM4
 - Vertikálna integrácia – nVidia vyvíja vlastné ARM CPU (Grace) → Grace-Blackwell supercip
- Trendy
 - FP4 & FP6 inferencia – menšia presnosť, obrovský nárast rýchlosti pre AI
 - Liquid cooling povinný pre najvýkonnejšie moduly (GB200)
 - Multi-die GPU – viacero čipletov na jednom substráte (ako AMD, Intel)
 - Neuromorphic & Quantum – dlhodobý výskum na akceleráciu iných paradigiem
- Softvérové vízie
 - CUDA-X ekosystém – knižnice pre každé odvetvie (cuDNN, cuBLAS, RAPIDS...)
 - Omniverse – simulačná platforma pre digitálne dvojčky, robotiku, fyziku
 - Rozšírenie vývoja Cosmos – generatívny AI model pre fyzikálne simulácie
 - nVidia NIM – mikroservisná AI inferencia v cloude

Softvérové nástroje

Nástroj	Účel
GPU-Z	Detailné info o GPU, záťažový test VRAM
MSI Afterburner	Pretaktovanie, monitorovanie, krivky ventíátorov
nVidia Inspector	Pokročilé nastavenia ovládačov
3DMark	Benchmarking výkonu a ray tracingu
FurMark	GPU stres test
OCCT	Dlhodobý test stability
nVidia FCAT VR	Meranie plynulosti snímok
HWiNFO64	Real-time monitoring teplôt, frekvencií, spotreby energie

Diagnostika, správa, servis

Príkazový riadok / CLI

- nvidia-smi – správa GPU cez terminal (teplota, VRAM, procesy, power limit)
- nvcc – CUDA kompilátor
- ncu (Nsight Compute) – profilovanie CUDA kernelových výpočtov

BIOS / Firmware nastavenia

- PCIe generácia – nastavenie Gen 3 / Gen 4 / Gen 5 (podpora základnej dosky)
- Resizable BAR (SAM) – umožní CPU priamy prístup k celej VRAM → výkon +5–15 %
- GPU TDP / Power Limit – v niektorých serverových BIOSoch priamo
- Secure Boot – môže blokovat' unsigned GPU firmware

Firmware GPU

- Aktualizácia cez NVIDIA App (automaticky)
- Profesionálne karty (A/H series): aktualizácia cez nVidia NGC alebo BMC
- vBIOS flash – len pre skúsených

Diagnostika, správa, servis

Možné poruchy a servis

Problém	Príznaky	Riešenie
Prehrievanie	Throttling, artefakty, reštarty	Čistenie, výmena teplovodivej pasty
Poškodená VRAM	Poškodená VRAM	Diagnostika cez MATS test, výmena
Poškodené VRM	Náhle vypnutie pri záťaži	Meranie napätia, výmena MOSFET
Driver crash	Čierna obrazovka, reset	DDU (Display Driver Uninstaller) + čistá inštalácia ovládačov
PCIe kontakt	GPU nie je detekovaná	Jemné vyčistenie kontaktu, iný slot
Ventilátor	Hluk, prehrievanie	Výmena ventilátora

Testovanie stability

- Zahrievacia fáza – FurMark 15 min pri 100 % záťaži
- Teplota GPU – ideálne pod 83 °C
- VRAM test – MATS v MODS alebo OCCT VRAM test
- Herný benchmark – 3DMark TimeSpy Extreme / Port Royal (RT)
- Dlhodobý stres – OCCT 1–4 hodiny kombinácia CPU+GPU

Zdroje

- „CPU vs GPU“ (<https://www.geeksforgeeks.org/computer-organization-architecture/difference-between-cpu-and-gpu/>)
- „GPUs vs CPUs (Threads, core, speed)“ (<https://medium.com/@abhishekjainindore24/gpus-vs-cpus-threads-core-speed-75b5732ce389>)
- „What's the Difference Between GPUs and CPUs?“ (<https://aws.amazon.com/compare/the-difference-between-gpus-cpus/>)
- „What is CUDA? – Computerphile“ (<https://youtu.be/K9anz4aB0S0>)
- „CUDA“ (<https://www.gigabyte.com/Glossary/cuda>)
- „Nvidia CUDA in 100 Seconds“ (<https://youtu.be/pPStdjuYzSI>)
- „What Is CUDA?“ (<https://www.supermicro.com/en/glossary/cuda>)
- „NVIDIA Tensor Cores“ (<https://www.nvidia.com/en-eu/data-center/tensor-cores/>)
- „What are Tensor Cores?“ (<https://youtu.be/Cak8ASX7NOK>)
- „Tensor Cores Explained in Simple Terms“ (<https://www.digitalocean.com/community/tutorials/understanding-tensor-cores>)
- „What is the difference between cuda vs tensor cores?“ (<https://stackoverflow.com/questions/47335027/what-is-the-difference-between-cuda-vs-tensor-cores>)
- „Introduction to Tensor Cores in NVIDIA GPUs“ (<https://www.naddod.com/ai-insights/introduction-to-tensor-cores-in-nvidia-gpus>)
- „Introduction to the Three Key Processing Cores Inside NVIDIA GPUs“ (<https://www.naddod.com/ai-insights/introduction-to-the-three-key-processing-cores-inside-nvidia-gpus>)
- „What Are RT Cores in Nvidia GPUs?“ (<https://www.titancomputers.com/What-Are-RT-Cores-in-Nvidia-GPUs-s/1208.htm>)
- „Non-graphics uses of RT-cores“ (https://www.reddit.com/r/nvidia/comments/k0ti6g/nongraphics_uses_of_rtcores)
- „ELI5: What purpose do the RT cores and Tensor cores serve in raytracing for Nvidia's GPU? What does one do that the other one doesn't?“ (https://www.reddit.com/r/explainlikeimfive/comments/ivi3ms/eli5_what_purpose_do_the_rt_cores_and_tensor)
- „Introduction to the Three Key Processing Cores Inside NVIDIA GPUs“ (<https://naddod.medium.com/introduction-to-the-three-key-processing-cores-inside-nvidia-gpus-5cda28bd0e13>)

Zdroje

- „NVIDIA GPU Deep Dive : Tensor and Ray Tracing Cores“ (<https://forwardevery.day/2025/05/08/nvidia-tensor-ray-tracing-cores>)
- „NVIDIA's DLSS 4.0 - Accelerating Progress with the power of AI“ (<https://rog-forum.asus.com/t5/technologies-explained/nvidia-s-dlss-4-0-accelerating-progress-with-the-power-of-ai/ba-p/1070622>)
- „3D Computer Graphics Primer: Ray-Tracing as an Example“ (<https://www.scratchapixel.com/lessons/3d-basic-rendering/introduction-to-ray-tracing/adding-reflection-and-refraction.html>)
- „Supreme Speed. Superior Visuals. Powered by AI.“ (<https://www.nvidia.com/en-eu/geforce/technologies/dlss/>)
- „NVIDIA DLSS: A Look Into the Future of Gaming“ (<https://medium.com/@benbenlyon/nvidia-dlss-a-look-into-the-future-of-gaming-20c651ae7090>)
- „NVIDIA DLSS 4.5 Delivers Major Upgrade With 2nd Gen Transformer Model For Super Resolution & 6X Dynamic Multi Frame Generation“ (<https://www.nvidia.com/en-gb/geforce/news/dlss-4-5-dynamic-multi-frame-gen-6x-2nd-gen-transformer-super-res/>)
- „NVIDIA DLSS: Your Questions, Answered“ (<https://www.nvidia.com/en-gb/geforce/news/nvidia-dlss-your-questions-answered/>)
- „What is NVIDIA® DLSS?“ (<https://www.lenovo.com/us/en/glossary/what-is-nvidia-dlss>)
- „What is NVIDIA DLSS and how does it improve your games?“ (<https://www.currys.co.uk/techtalk/gaming-news/pc-gaming-news/what-is-dlss.html>)
- „Nvidia debuts DLSS 5 for increased visual fidelity in games — AI-infused tech transforms pixels with photorealistic lighting and materials“ (<https://www.tomshardware.com/tech-industry/artificial-intelligence/nvidia-debuts-dlss-5-for-increased-visual-fidelity-in-games-ai-infused-tech-transforms-pixels-with-photorealistic-lighting-and-materials>)

Ďakujem za pozornosť