

UNIVERZITA SV. CYRILA A METODA V TRNAVE

FAKULTA PRÍRODNÝCH VIED

APLIKOVANÁ INFORMATIKA

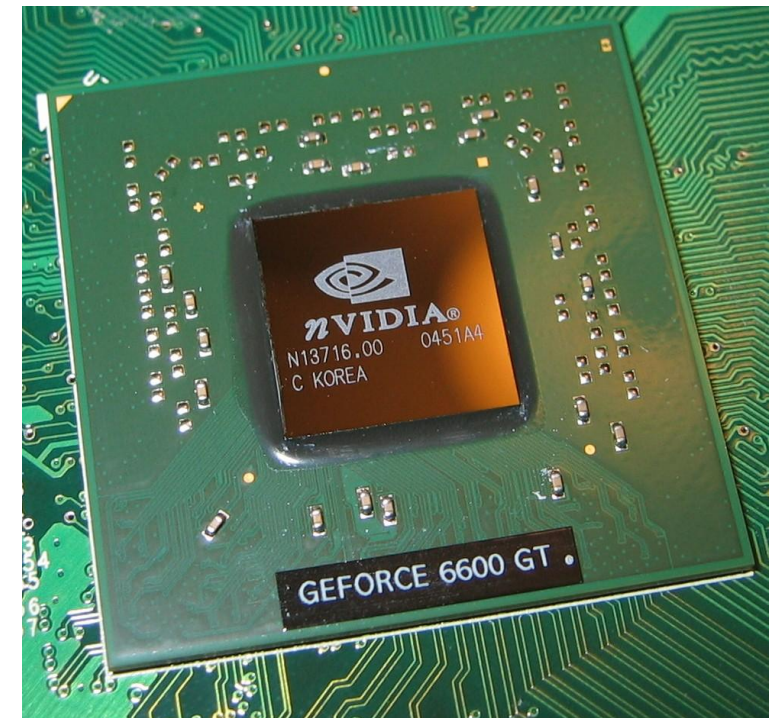
Grafický subsystém NVIDIA

Grafický subsystém

Grafický subsystém predstavuje hardvérovú a softvérovú časť počítača zodpovednú za spracovanie a zobrazovanie obrazu

Obsahuje tisíce jadier

Jadrom je GPU



Paralelné výpočty

GPU obsahuje stovky až tisíce jadier

SIMT princíp – veľké množstvo podobných úloh súčasne

Použitie:

- a. Grafika
- b. AI tréning
- c. Fyzikálne simulácie
- d. Big Data

SIMT

Pokročilejší model používaný v NVIDIA GPU

Inštrukcia neriadi priamo dáta, ale tisíce vlákien

Kombinuje paralelizmus SIMD s flexibilitou multithreadingu (vlákna sú nezávislejšie ako v čistom SIMD)

Zameraný na maximálnu priepustnosť (Throughput)

Plne hardvérová správa – plánovanie a distribúciu inštrukcií medzi vlákna rieši hardvér GPU

SIMT/SIMD

Vlastnosť	SIMD (CPU)	SIMT (NVIDIA GPU)
Zameranie	Dátové vektory	Nezávislé vlákna
Hardvér	AVX / SSE jednotky	CUDA Jadrá / Streaming Multiprocessors
Vetvenie kódu	Náročné, stráca výkon	Flexibilnejšie, spravované hardvérom
Programovanie	Vektorové inštrukcie	CUDA C++ / Python (Cez abstrakciu vlákien)

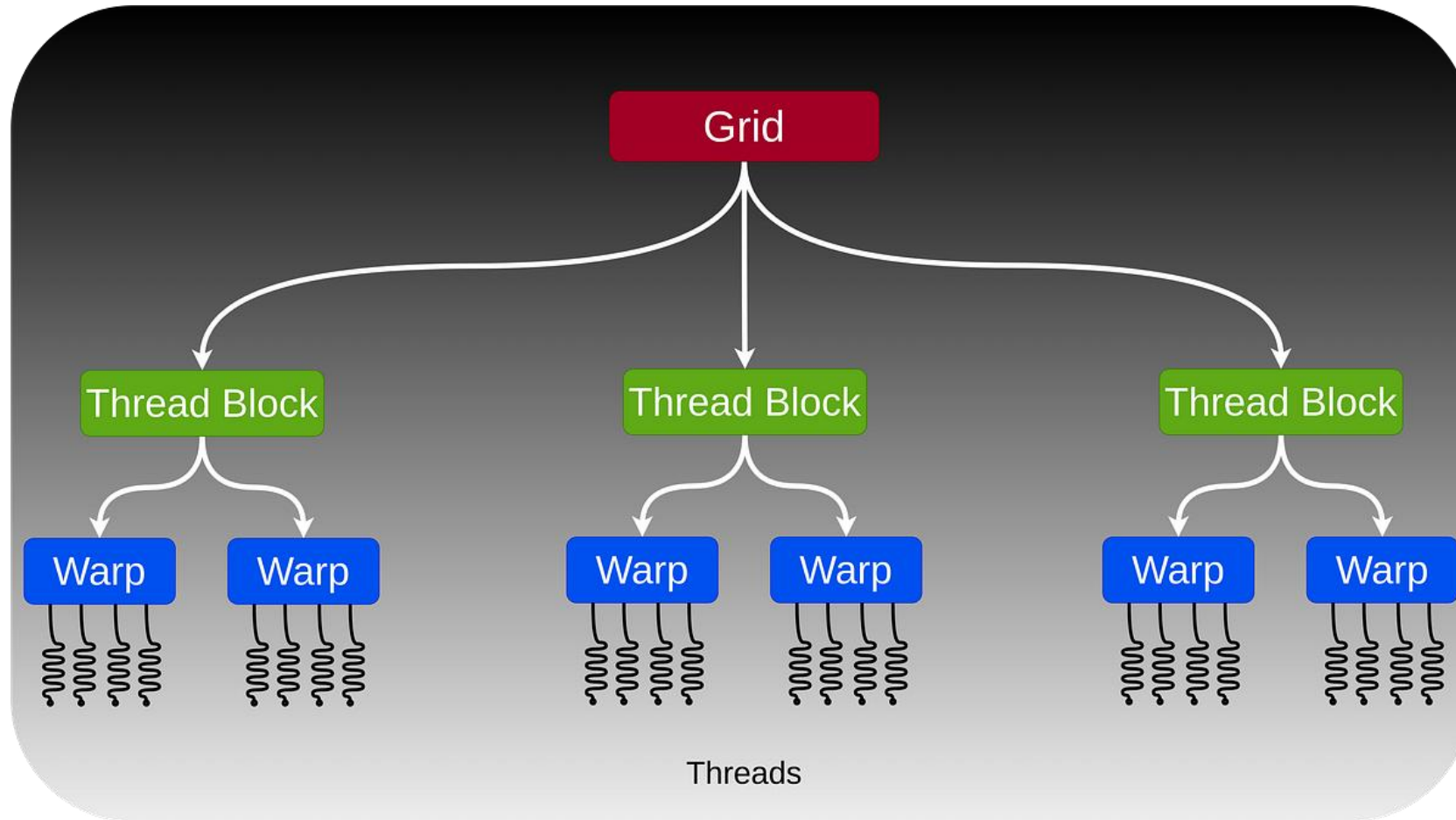
Warp

Je zväzok 32 vlákien, ktoré sú hardvérovo zviazané

Všetky vlákna v jednom warpe vykonávajú v danom okamihu tú istú inštrukciu, ale každé nad svojimi vlastnými dátami

GPU procesor (SM) neplánuje prácu pre jednotlivé vlákna, ale prepína medzi celými warpami

Warp



CUDA (Compute Unified Device Architecture)

Umožňuje využívať GPU na komplexné matematické výpočty, ktoré bežne vykonáva CPU

Platforma a API pre výpočty na GPU

Natívna podpora pre nízkoúrovňové programovanie (C/C++) a integrácia s vysokoúrovňovými knižnicami a frameworkmi (Python, PyTorch, TensorFlow)

Základ moderného GPGPU (General-Purpose computing on Graphics Processing Units)

SM (Streaming Multiprocessor)

SM je srdcom NVIDIA GPU

Celkový výkon GPU je určený počtom týchto blokov

Každý SM pracuje nezávisle na ostatných

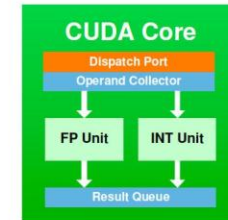
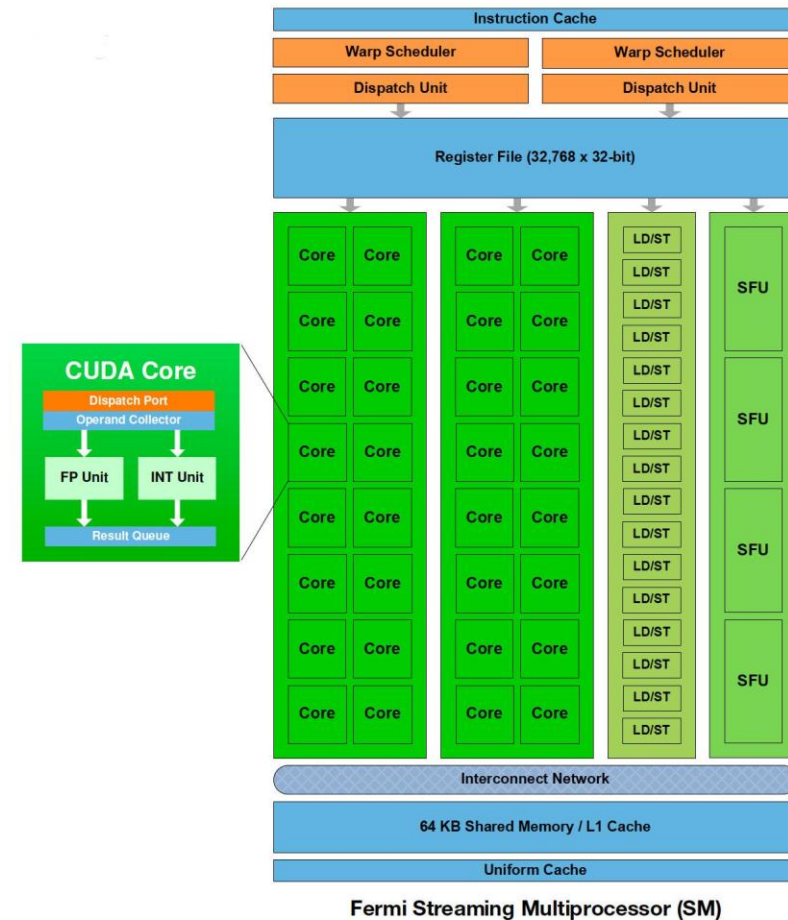
Stará sa o inteligentné rozdeľovanie svojich obmedzených zdrojov medzi všetky vlákna

SM (Streaming Multiprocessor)

SM obsahuje:

- CUDA cores: Samotné jednotky, ktoré robia matematiku
- Register File: Extrémne rýchla pamäť priamo v SM pre ukladanie dát vlákien
- Shared Memory: Rýchla lokálna pamäť zdieľaná medzi vláknami v rámci jedného bloku
- Warp Schedulers: "Mozog" SM, ktorý rozhoduje, ktorý warp dostane v ďalšom cykle prístup k výpočtovým jadram

SM (Streaming Multiprocessor)



Tensor cores

Špecializované výpočtové jednotky v rámci architektúry NVIDIA GPU, navrhnuté na maticové operácie

Precízna akcelerácia (FP16 / BF16 / INT8): využíva nižšiu číselnú presnosť, čo drasticky zvyšuje priepustnosť dát bez viditeľnej straty kvality výsledku

Optimalizované pre deep learning

Výrazné zrýchlenie AI workloadov

TensorFlow

Open-source framework od Google

Dáta reprezentuje ako tenzory

Používa CUDA a cuDNN (CUDA Deep Neural Network library)

Použitie:

- a. CNN
- b. LLM
- c. Vedecké výpočty

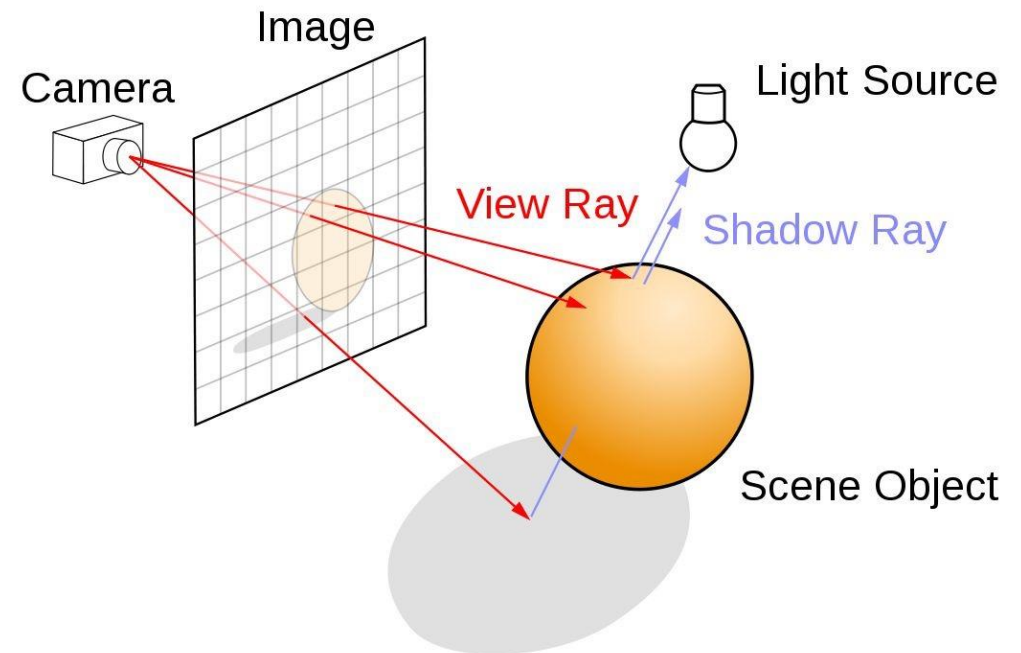
Ray Tracing

Technológia vykresľovania grafiky, ktorá simuluje správanie svetla v reálnom svete

Výpočty odrazov, tieňov a refrakcií

Hardvérová akcelerácia (špecializované jadrá):

- a. BVH Traversal (Optimalizácia scény)
- b. Intersection Tests



Ray Tracing



Full Ray Tracing Off



Full Ray Tracing On

DLSS (Deep Learning Super Sampling)

Technológia využívajúca umelú inteligenciu na rekonštrukciu obrazu s vysokým rozlíšením z dát s nižším rozlíšením

Využíva Tensor cores na spustenie natrénovaného neurónového modelu, ktorý dopočítava detaily

Prínosy:

- a. Vyššie FPS
- b. Zachovanie detailov

DLSS 3/4 pridáva frame generation

Evolúcia NVIDIA

1. 1999: Éra GPU (GeForce 256): Predstavenie prvej modernej grafickej karty (GPU), ktorá odbremenila CPU od vykresľovania 3D geometrie
2. 2006: Zlomový bod (CUDA): NVIDIA predstavuje CUDA. Grafická karta už neslúži len na hry, ale stáva sa univerzálnym výpočtovým nástrojom (GPGPU)
3. 2012: Hlboké učenie (Deep Learning): AlexNet (prelomový model AI) využíva NVIDIA GPU na tréningovanie. Začína éra umelej inteligencie poháňanej grafikami

Evolúcia NVIDIA

4. 2017: Špecializácia (Tensor Cores): Príchod architektúry Volta. NVIDIA uvádza prvé Tensor cores – hardvér špeciálne pre maticové operácie v AI
5. 2018: Realistická grafika (RT Cores): Architektúra Turing. Príchod hardvérového Ray Tracingu a prvej generácie DLSS
6. 2020 – Dnes: Éra generatívnej AI: Architektúry Ampere, Ada Lovelace a Blackwell. Plná integrácia AI do všetkých oblastí (Frame Generation, LLM tréning)

GeForce 256 (1999)

Prvá grafická karta označená ako GPU

Priniesla hardvérovú transformáciu a osvetlenie (T&L), čím odbremenila CPU

Začala éru moderných grafických procesorov

Význam: položila základy dedikovaných grafických výpočtov



GeForce 8800 GTX (2006)

Prvá architektúra s unified shader modelom

Vertex a pixel shadery boli zlúčené do univerzálnych jednotiek

Výrazný výkonový skok v hrách aj výpočtoch

Význam: základ pre flexibilnejšie GPU architektúry



GTX 1080 Ti – Pascal (2017)

Jedna z najúspešnejších herných GPU generácií

Vysoký výkon, energetická efektivita a VR pripravenosť

Silná pozícia pred nástupom RTX

Význam: vrchol klasického rasterizačného výkonu



RTX 4090 – Ada Lovelace (2022)

Extrémny výkon v hrách, AI aj tvorbe obsahu

DLSS 3 a Frame Generation

Masívny nárast výkonu oproti predchádzajúcim generáciám

Význam: moderné GPU ako hybrid grafiky a AI akcelerácie



**Ďakujem za
pozornosť**
