CUDA (Compute Unified Device Architecture)

CUDA je paralelní výpočetní platforma a programovací model vytvořený společností **NVIDIA**. Umožňuje softwarovým vývojářům využívat výkon grafických procesorů (GPU) pro výpočty pro obecné účely (GPGPU), které byly dříve vyhrazeny pouze pro CPU.

Zatímco CPU je skvělé v sekvenční logice, CUDA využívá tisíce jader GPU k paralelnímu řešení složitých matematických problémů.

1. Jak CUDA funguje?

CUDA poskytuje vývojářům přímý přístup k instrukční sadě a paměti paralelních výpočetních prvků v GPU. Programovací model je založen na jazyce **C/C++**, ale existují obaly (wrappers) pro Python (PyCUDA), Fortran a další.

- **Kernely:** Funkce, které běží na GPU. Jeden kernel je spuštěn mnohokrát paralelně v různých vláknech.
- **Hierarchie vláken:** Vlákna jsou organizována do **bloků** a bloky do **mřížek** (grids). To umožňuje efektivní škálování na různé modely karet.
- **Paměťová hierarchie:** CUDA rozlišuje mezi globální pamětí (VRAM), sdílenou pamětí (velmi rychlá v rámci bloku) a registry.

2. Proč je CUDA standardem v AI?

Dominance CUDA v oblasti umělé inteligence a hlubokého učení (Deep Learning) není náhodná:

- **Knihovny:** NVIDIA vytvořila extrémně optimalizované knihovny jako **cuDNN** (pro neuronové sítě) a **cuBLAS** (pro lineární algebru).
- **Ekosystém:** Populární frameworky jako **PyTorch** a **TensorFlow** jsou primárně vyvíjeny a optimalizovány pro CUDA.
- **Hardware-Software integrace:** NVIDIA navrhuje specializovaná jádra (Tensor Cores) přímo pro operace, které CUDA knihovny využívají.

3. Oblasti využití

Obor	Konkrétní aplikace
Umělá inteligence	Trénování LLM (např. GPT-4), rozpoznávání obrazu, generování videa.
Věda a výzkum	Simulace dynamiky kapalin, modelování proteinů, astrofyzika.
Finance	Algoritmické obchodování, analýza rizik (Monte Carlo simulace).
Kryptografie	Těžba kryptoměn (zejména algoritmy náročné na paměť a výpočty).
Zpracování obrazu	Rendering (V-Ray, Octane), střih videa, lékařské zobrazování (MRI).

4. CUDA vs. konkurence

Přestože je CUDA nejrozšířenější, je to uzavřený (proprietary) systém, který funguje **pouze na kartách NVIDIA**. To vedlo ke vzniku alternativ:

- **OpenCL:** Otevřený standard podporovaný různými výrobci (AMD, Intel, Apple). Je však náročnější na programování a často méně optimalizovaný pro AI.
- **ROCm (AMD):** Snaha společnosti AMD vytvořit přímou konkurenci pro CUDA, která umožňuje snadný převod CUDA kódu pro karty Radeon.
- **Apple Metal:** Nízkoúrovňové API pro hardware Apple (čipy řady M).

5. Budoucnost: Tensor Cores

V novějších generacích architektury (Ampere, Hopper, Blackwell) CUDA úzce spolupracuje s **Tensor Cores**. Tato jádra jsou navržena speciálně pro násobení matic v nízké přesnosti (FP16/INT8), což je klíčová operace pro inferenci neuronových sítí.

Související články:

- Grafické procesory (GPU)
- Deep Learning a neuronové sítě
- Ovladače zařízení (VOD)

Tagy: hw gpu cuda nvidia ai computing programming