CPU a GPU

Osnova

- úvod
- CPU
- GPU
- hlavní rozdíly
- paralelní vs sekvenční zpracování
- CPU a GPU v AI
- vývoj CPU a GPU
- výhody a nevýhody

- dvě klíčové výpočetní jednotky moderního počítače
- Obě ke zpracování dat - navrženy pro rozdílné typy úloh
- **CPU (Central Processing Unit)** – centrální procesor, „mozek počítače“
- **GPU (Graphics Processing Unit)** – grafický procesor, specializovaný na paralelní výpočty
- **CPU – Centrální procesor**
    - hlavní **řídicí a výpočetní** jednotka počítače
    - **vykonávání instrukcí** operačního systému i aplikací
    - **Jak CPU funguje?**
        - **CPU pracuje v cyklu**:

            1. Fetch – **načtení** instrukce z paměti
            2. Decode – **dekódování** instrukce (instrukční sada)
            3. Execute – **vykonání** instrukce
            4. Store – **uložení** výsledku

        - Tento cyklus se opakuje miliardkrát za sekundu
    - **Vnitřní struktura CPU**
        - **Řadič (Control Unit)**
            - řídí tok instrukcí
            - koordinuje ostatní části procesoru
        - **ALU (Aritmeticko-logická jednotka)**
            - provádí matematické operace
            - logické operace (AND, OR, NOT)
        - **registry** -paměti uvnitř CPU, velmi malé (jen to s čím aktuálně pracuje), extrémně rychlé
        - **Cache paměť** - uchovává často používaná data (L1, L2, L3 - pro všechna jádra)
    - **Pipelining** - jedna instrukce se načítá, další se dekóduje a jiná se už provádí (probíhají najednou)
        - Uvnitř každého jádra
    - **Parametry CPU**
        - frekvence (GHz) - jak rychle procesor provádí operace (kolik miliard cyklů za sekundu)
        - počet jader (cores) - každé jádro zpracovává úlohy nezávisle
        - počet vláken (threads) - kolik menších úkolů může jádro najednou řešit (nejčastěji 2)
        - velikost cache
        - architektura (např. x86, ARM) - určuje jaký set instrukcí počítač používá (programy musí odpovídat instrukční sadě dané architektury)
    - Moderní CPU mají obvykle **4–16 jader**
    - vysoká **univerzálnost** (komunikuje s OS, programy, provádí výpočty)
    - **rychlá reakce na změny** instrukcí
    - nízký počet jader, ale vysoký výkon na jádro
- **GPU – Grafický procesor**
    - specializovaný procesor - určený **primárně pro zpracování grafiky**
    - Dnes se používá nejen pro grafiku, ale i pro:
        - umělou inteligenci

        - vědecké výpočty
        - kryptografii
        - rendering videa
    - **Jak GPU funguje?**
        - GPU je navrženo pro **paralelní zpracování**
        - To znamená, že dokáže provádět **tisíce stejných operací současně**
        - Například:
            - výpočet barvy milionů pixelů
            - trénování neuronových sítí
    - **Struktura GPU**
        - tisíce menších výpočetních jednotek (jader)
        - vlastní grafická paměť (VRAM)
        - optimalizováno pro paralelismus
        - Tensorová jádra - operace s maticemi (AI)
    - **Typy GPU**
        - **Integrované GPU**
            - součást procesoru
            - nižší výkon
            - energeticky úsporné
        - **Dedikované GPU**
            - samostatná grafická karta
            - vysoký výkon
            - vlastní paměť
- **HLAVNÍ ROZDÍLY MEZI CPU A GPU**

CPU	GPU
má méně výkonných jader	má tisíce jednoduchých jader
univerzální výpočty	paralelní výpočty
vyšší výkon na jedno jádro	nižší výkon na jádro
řídí celý systém	specializované výpočty
pracuje s RAM	používá VRAM

- **PARALELNÍ VS SEKVENČNÍ ZPRACOVÁNÍ**
    - **Sekvenční zpracování (CPU)**
        - Instrukce jsou vykonávány postupně
        - Vhodné pro: operační systém, běžné aplikace, složité logické operace
    - **Paralelní zpracování (GPU)**
        - Mnoho instrukcí je vykonáváno současně
        - Vhodné pro: grafiku, výpočty matic, neuronové sítě
- **CPU A GPU V UMĚLÉ INTELIGENCI**
    - Trénování neuronových sítí vyžaduje:
        - velké množství paralelních výpočtů
        - operace s maticemi
    - **GPU** je proto **mnohem efektivnější** než CPU
    - Například: trénování modelu na CPU může trvat dny, na GPU jen hodiny
    - Moderní GPU obsahují specializované jednotky:

        - Tensor Cores
        - AI akcelerátory
- **VÝVOJ CPU A GPU**
    - **CPU** - důraz na vyšší frekvenci
        - přechod na vícejádrové architektury
        - energetická efektivita
    - **GPU** - masivní nárůst počtu jader
        - rozvoj paralelního programování (CUDA)
        - využití mimo grafiku (GPU)
- **VÝHODY A NEVÝHODY**
    - **CPU: Výhody:** univerzálnost, vysoká flexibilita, dobré řízení systému

Nevýhody: omezený paralelní výkon

    - **GPU: Výhody:** extrémní paralelní výkon, vhodné pro grafiku a AI

Nevýhody: méně flexibilní, vyšší spotřeba energie