TPU (Tensor Processing Unit)

TPU je integrovaný obvod specifický pro danou aplikaci (ASIC), který Google vyvinul od základu pro svůj framework TensorFlow. Jeho hlavním cílem je poskytnout řádově vyšší výkon při nižší spotřebě energie pro trénování a inferenci (provoz) velkých modelů AI, jako jsou LLM nebo difuzní modely.

1. Architektura: Proč je TPU jiný?

Klíčem k výkonu TPU je architektura **Systolic Array** (systolické pole).

- **CPU/GPU:** Při každém výpočtu musí procesor přistupovat k paměti (registru), provést operaci a výsledek uložit. To vytváří úzké hrdlo (tzv. von Neumannovo hrdlo).
- **TPU:** Data protékají polem aritmetických jednotek jako vlna (podobně jako krev v srdci – odtud „systolické“). Tisíce násobení a sčítání proběhnou v jednom taktu bez neustálého zápisu do paměti. To extrémně zrychluje **násobení matic**, což je 90 % práce při trénování AI.

2. Srovnání: CPU vs. GPU vs. TPU

Procesor	Charakteristika	Přirovnání
CPU	Flexibilní, zvládne jakýkoli kód, ale pomalu.	Švýcarský nůž.
GPU	Tisíce jader pro paralelní výpočty (grafika, AI).	Rychlá dodávka (uveze hodně balíků najednou).
TPU	Extrémně rychlý, ale pouze pro specifické AI operace.	Nákladní vlak na vyhrazené trati.

3. Generace TPU

Google neustále vyvíjí nové verze, které jsou dostupné skrze **Google Cloud Platform (GCP)**:

- **v1:** Pouze pro inferenci (používání již hotových modelů).
- **v2 a v3:** Přidána podpora pro trénování modelů a možnost zapojení do tzv. **TPU Podů** (superpočítačů).
- **v4 a v5p:** Současná špička, optimalizovaná pro trénování obřích modelů jako Gemini nebo PaLM.

4. Cloud TPU a TPU Pods

TPU se běžně neprodávají jako samostatné karty do PC (jako NVIDIA GPU). Jsou dostupné jako cloudová služba.

- **TPU Pod:** Seskupení stovek až tisíců TPU čipů propojených ultra-rychlou sítí. To umožňuje trénovat modely, které by na běžných počítačích trvaly roky, během několika dnů.

5. Výhody a omezení

Výhody:

- **Výkon na watt:** Mnohem úspornější než GPU při stejném výkonu.
- **Rychlost:** Ideální pro obří maticové operace v hlubokém učení.
- **Ekosystém:** Perfektní integrace s Google Cloud a frameworky TensorFlow/JAX.

Omezení:

- **Specializace:** Špatně si poradí s kódem, který není založen na tenzorech (např. klasické větvení programu).
- **Uzavřenost:** Hardware vlastní a provozuje výhradně Google.
- **Cena:** Pronájem výkonných TPU verzí může být pro menší projekty velmi nákladný.

> **Zajímavost:** TPU čipy byly použity k poražení světového šampiona ve hře Go systémem **AlphaGo**. Právě díky TPU mohl systém propočítat miliony tahů v reálném čase během zápasu.

Zpět na Hardware