

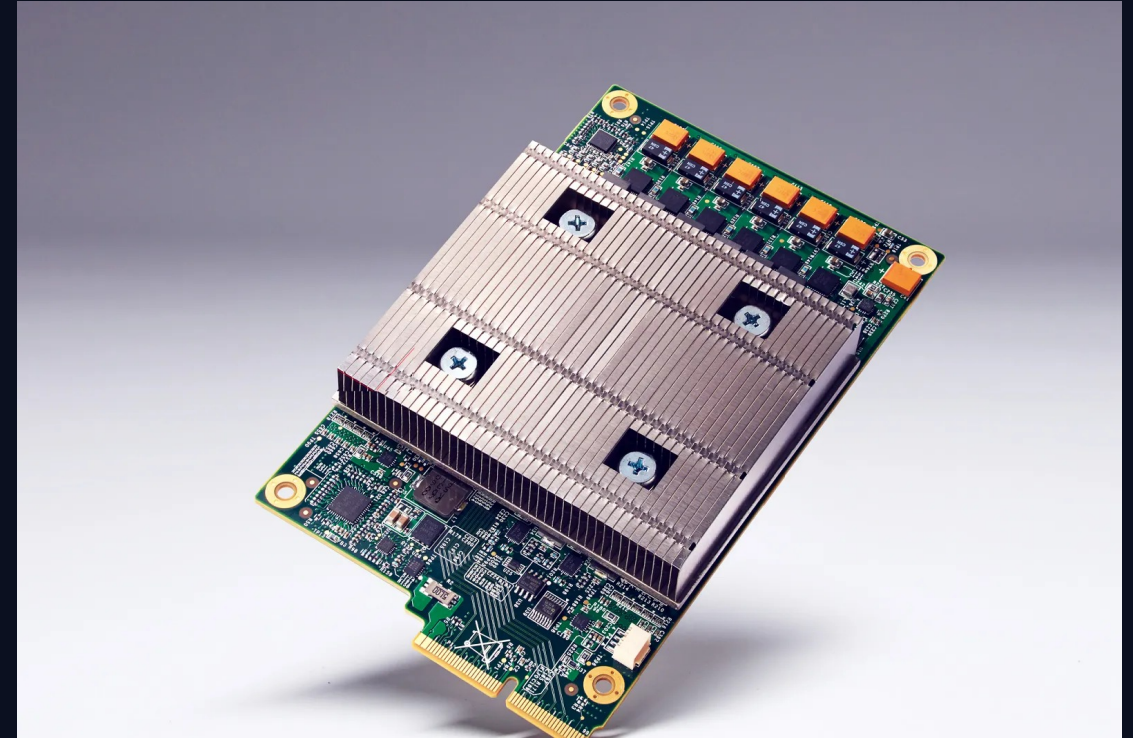
Hardvér pre umelú inteligenciu



Prečo je hardvér pre AI dôležitý?

AI modely rastú rýchlejšie ako tradičné procesory; rozhodujú výkon, pamäť a spotreba.

AI akcelerátor je čip alebo systém optimalizovaný na veľké množstvo paralelných násobení a sčítaní. Kľúčové obmedzenia: pamäťová priepustnosť, spotreba energie, chladenie a rýchle prepojenie medzi čipmi. Dobrý hardvér skracuje tréning, zrýchľuje inferenciu a znižuje cenu za výpočet.



Obr.: Google TPU modul. Zdroj: WIRED / Google, 2016.

Čo je vnútri moderného AI systému?

Výpočtový čip

GPU, TPU, NPU, FPGA alebo ASIC. Obsahuje tisíce jadier, vektorové jednotky a často jednotky pre nízku presnosť.

Pamäť

HBM alebo SRAM drží dáta čo najbližšie k výpočtu. Priepustnosť je často dôležitejšia ako samotný počet TOPS.

Prepojenie

PCIe, NVLink, Ethernet/InfiniBand spájajú akcelerátory do serverov a klastrov.

Napájanie a chladenie

AI servery majú vysoký príkon. Používa sa vzduchové, kvapalinové alebo hybridné chladenie.

Softvér a ovládače

Bez kompilátorov, runtime knižníc a plánovania úloh nie je hardvér efektívne využitý.

Hlavné medzníky v historickom vývoji

1950s-70s



CPU a prvé digitálne počítače

1999



GPU sa stáva programovateľným paralelným procesorom

2006-2012



CUDA a nástup tréningu neurónových sietí na GPU

2016



Google oznamuje TPU

2020s



HBM, čiplety, wafer-scale a edge NPU

Prvý výskyt špecializovaného komponentu pre AI v praxi: GPU sa masovo presadili najprv v grafike, potom v GPGPU výpočtoch; Google TPU bol verejne oznámený v roku 2016 ako ASIC pre TensorFlow a dátové centrá.

Komponenty používané v súčasnosti

GPU

univerzálny akcelérátor pre tréning aj inferenciu

TPU/ASIC

vysoká efektívnosť pre konkrétne operácie

NPU

nízka spotreba v mobiloch, PC a edge zariadeniach

FPGA

preprogramovateľný čip pre špecifické nasadenie

HBM/SRAM

kritická pamäťová vrstva blízko výpočtu

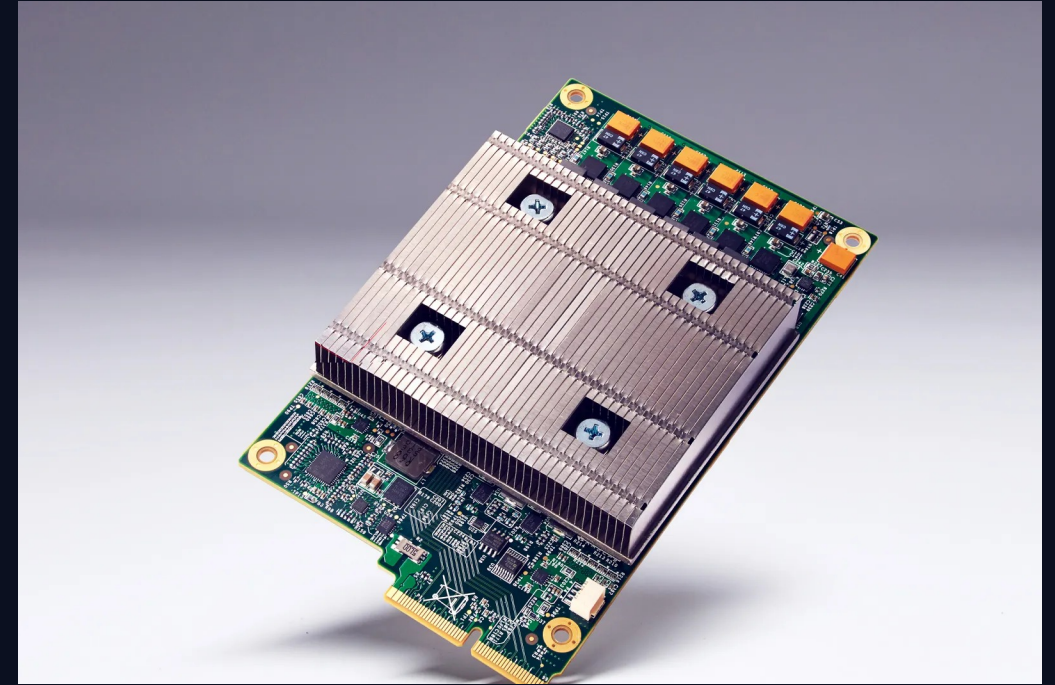
Sieťové čipy

NVLink, InfiniBand, Ethernet pre klastre

GPU a TPU: najznámejšie AI akcelerátory



Obr.: NVIDIA H100/DGX zostava. Zdroj: NVIDIA Blog, 2022.

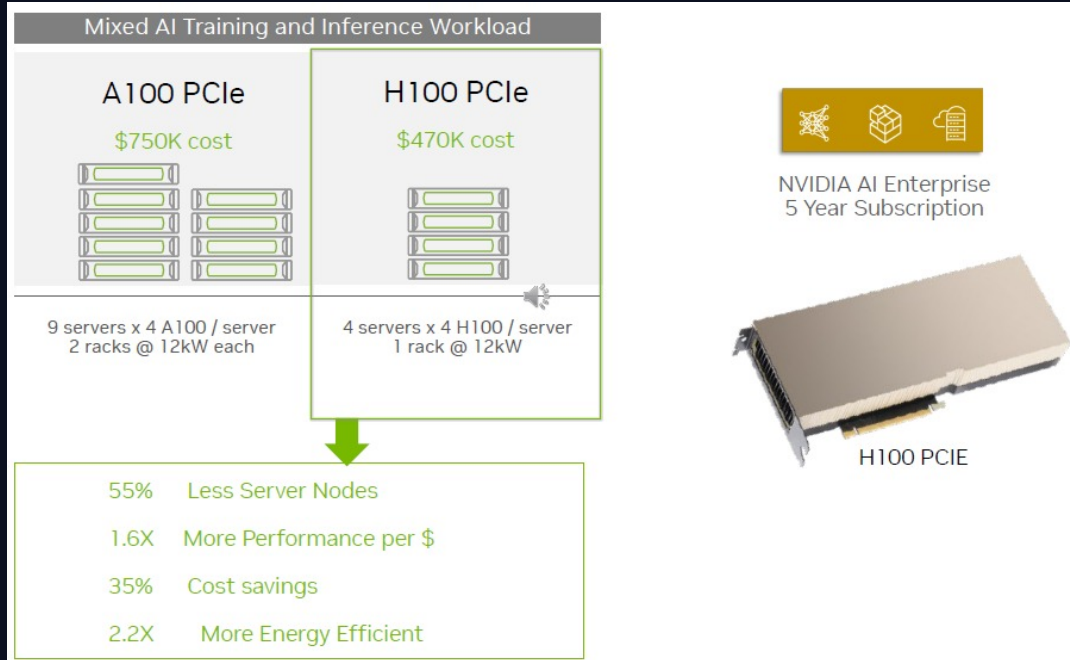


Obr.: Google TPU modul. Zdroj: WIRED / Google, 2016.

GPU: flexibilné knižnice, silný ekosystém, rýchle prepojenie medzi kartami.
TPU: ASIC navrhnutý pre tensorové operácie a vysoký výkon na watt v dátovom centre.

Pamäť a prepojenie

Výkon AI servera často limituje „presun dát“, nie samotná aritmetika.



HBM je pamäť s veľmi vysokou priepustnosťou uložená blízko výpočtového čipu.

Chladienie sa posúva ku kvapalinovým riešeniam, pretože hustota výkonu rastie.

Obr.: H100 PCIe a porovnanie TCO. Zdroj: PNY Blog, 2022.

AI hardvér na okraji siete (edge computing)

Smartfóny a notebooky

NPU zrýchľuje lokálne rozpoznávanie obrazu, hlasu a generatívne funkcie pri nízkej spotrebe.

Automobily a robotika

Dôležitá je latencia, spoľahlivosť a spracovanie senzorov v reálnom čase.

IoT a priemysel

Malé akcelerátory šetria prenos dát a môžu fungovať aj bez stáleho pripojenia.

Edge hardvér nerieši najväčší výkon, ale najlepší pomer výkon / watt / cena / latencia.

Vývoj do budúcnosti - vízie

Čiplety: viac výpočtu a pamäte v jednom module.
Nižšia presnosť výpočtov: FP8, FP4 a špeciálne formáty pre lepší výkon na watt.

Pamäť bližšie k výpočtu: HBM, on-chip SRAM a výpočty v/pri pamäti.
Kvapalinové chladenie a efektívnejšie dátové centrá.
Rozšírenie lokálnych NPU v PC, mobiloch, autách a priemyselných zariadeniach.

Hlavná vízia

AI nebude bežať len v cloude.
Tréning zostane prevažne v dátových centrách, ale inferencia sa bude presúvať čo najbližšie k používateľovi a senzorom.

Záver

Hardvér rozhoduje, aké AI aplikácie sú prakticky a ekonomicky možné.

AI akcelerátory sú špecializované na paralelné maticové operácie a rýchly presun dát.

Súčasnosť ovládajú GPU a ASIC/TPU, ale rastie význam NPU, pamätí HBM a rýchlych prepojení.

Budúcnosť smeruje k čipletom, wafer-scale riešeniam, neuromorfným smerom a vyššej energetickej efektívnosti.



Obr.: NVIDIA H100/DGX zostava. Zdroj: NVIDIA Blog, 2022.

Zoznam použitej literatúry

NVIDIA. NVIDIA Blackwell Architecture [online]. NVIDIA, 2024 [cit. 2026-04-27]. Dostupné z: nvidia.com/en-us/data-center/technologies/blackwell-architecture/

GOOGLE CLOUD. An in-depth look at Google's first Tensor Processing Unit [online]. Google, 2017 [cit. 2026-04-27]. Dostupné z: cloud.google.com/blog/products/ai-machine-learning/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

CEREBRAS SYSTEMS. Cerebras announces third-generation Wafer-Scale Engine [online]. Cerebras, 2024 [cit. 2026-04-27]. Dostupné z: cerebras.ai/press-release/cerebras-announces-third-generation-wafer-scale-engine

INTEL. Taking Neuromorphic Computing to the Next Level with Loihi 2 [online]. Intel, 2021 [cit. 2026-04-27]. Dostupné z: intel.com/content/www/us/en/research/neuromorphic-computing-loihi-2-technology-brief.html

AKOPYAN, F. et al. TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip [online]. IBM Research, 2015 [cit. 2026-04-27]. Dostupné z: research.ibm.com/publications/truenorth-design-and-tool-flow-of-a-65-mw-1-million-neuron-programmable-neurosynaptic-chip

WIRED. Google Built Its Very Own Chips to Power Its AI Bots [online]. WIRED, 2016 [cit. 2026-04-27]. Dostupné z: wired.com/2016/05/google-tpu-custom-chips/