

HARDWARE AI



Igor Oborný, 2. ročník – Ext.Bc

OBSAH

- Úvod
 - Definícia AI hardvéru
 - História
- Architektonické limity klasických systémov
 - Von Neumannova architektúra
 - Von Neumannov Bottleneck
- Procesorové architektúry pre AI
 - Hardvérová akcelerácia výpočtov
 - Typy hardvérových akceleratorov
 - Maticové spracovanie
 - Spájanie operácií (FMA)
 - Systolické polia
- Pamäťové systémy a priepustnosť
 - HBM (High Bandwidth Memory)
 - In-memory Computing
 - on-chip SRAM
- Škálovanie a infraštruktúra
 - Konektivita čipov
- Prevádzkové výzvy a budúcnosť AI hardvéru
 - Energetická efektívnosť
 - Budúcnosť

DEFINÍCIA AI HARDVÉRU

- **Hardvér (HW)** sú fyzické a hmatateľné časti počítača
- umožňuje spracovanie dát, uchováva informácie a umožňuje interakciu s používateľom
- Hlavné časti:
 - CPU, základná doska, RAM pamäť, GPU, úložisko, zdroj

DEFINÍCIA AI HARDVÉRU

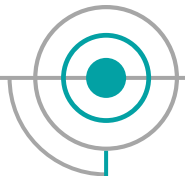
- špecializovaný hardvér pre výpočtové nároky AI
- prechod od sekvenčného spracovania (von Neumannova architektúra) k masívnemu paralelizmu
- AI akcelerátory (TPU, NPU, ASIC)
 - špecializované čipy pre rýchlejšie a úspornejšie AI operácie, ktorých hlavnou úlohou je akcelerácia maticových operácií a tenzorových výpočtov
- optimalizácia pamäťovej priepustnosti
- precíznosť výpočtov

HISTÓRIA

Sekvenčné spracovanie dát

NVIDIA CUDA

1945



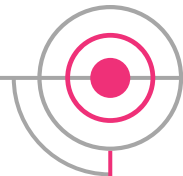
Von Neumannova
architektúra

1950s



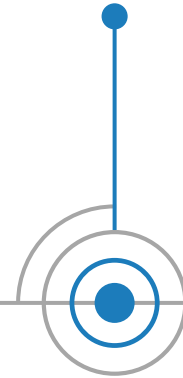
1980s

1990s

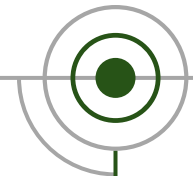


Nástup GPU

2006



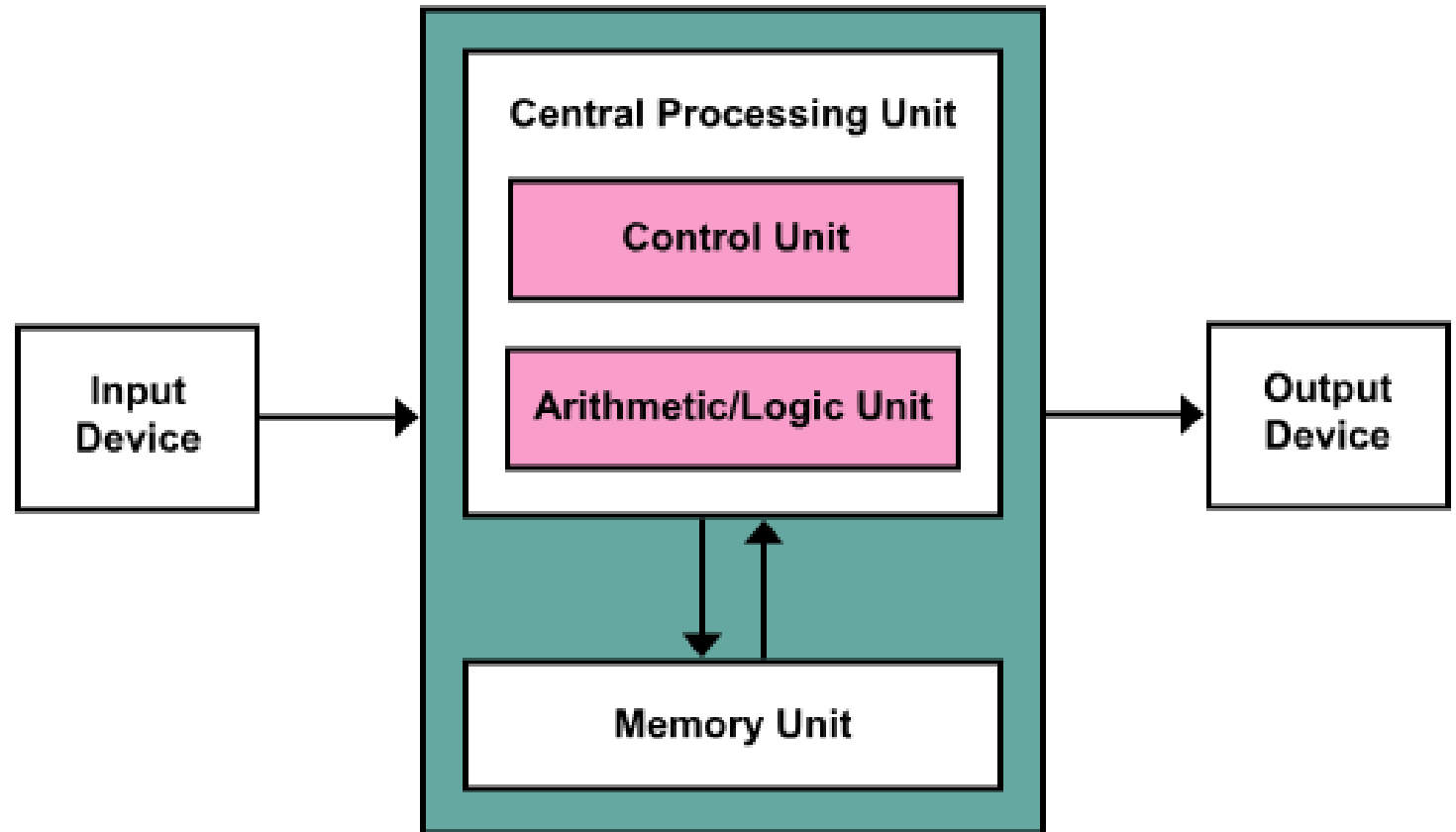
2016+



Príchod špecializovaného
kremíka

VON NEUMANN

- procesor (CPU) a pamäť sú samostatné jednotky prepojené spoločnou zbernicou
- dáta aj programy zdieľajú ten istý komunikačný kanál a pamäťový priestor
- **flexibilita** - oddelenie pamäte a výpočtových jednotiek znamená, že ich môžeme navrhovať samostatne a konfigurovať takmer ľubovoľným spôsobom



Obr.1 : Von Neumann architektura

VON NEUMANN BOTTLENECK

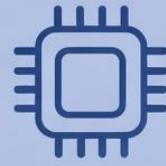
- najväčšia slabina tejto architektúry
- keďže procesor a pamäť sú oddelené a používajú spoločnú cestu (zbernicu), procesor je často oveľa rýchlejší než schopnosť pamäte dodávať mu dáta
- procesor tak musí „čakať“, kým mu dáta dorazia, čo obmedzuje celkový výkon
- AI hardvér sa to snaží obísť tak, že výpočty prebiehajú priamo v pamäti alebo veľmi blízko nej (tzv. *In-memory computing*), čím sa radikálne znižuje potreba neustáleho presúvania dát
- **limit zbernice – vznik Memory Wall**
 - výkon procesora rastie exponenciálne, ale priepustnosť a latencia pamätí (DRAM) stagnuje – procesor čaká na doručenie dát zo zbernice

HARDVÉROVÁ AKCELERÁCIA VÝPOČTOV

- presun špecifických výpočtových úloh (napr. násobenie matic) z procesora (CPU) na špecializované akcelerátory
- Rýchlosť – akcelerátory zvládnu konkrétne úlohy rýchlejšie ako CPU
- plynulosť – CPU ostáva voľné pre iné úlohy
- zníženie spotreby energie – špecializovaný čip spotrebuje menej energie na tú istú úlohu v porovnaní s univerzálnym procesorom

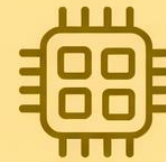
TYPY HARDVÉROVÝCH AKCELERÁTOROV

- Grafické procesory (GPUs)
- Programovateľné hradlové polia (FPGAs)
- Špecializované integrované obvody (ASICs)
- Tenzorové procesory (TPUs)



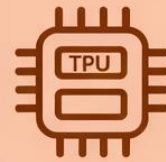
CPU

- Malé modely
- Malé dátové súbory
- Užitočný pre dizajnový priestor prieskumu



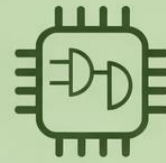
GPU

- Stredné až veľké modely, dátové súbory
- Spracovanie obrazu, videa
- Aplikácie cez CUDA alebo OpenCL



TPU

- Matricové výpočty
- Spracovanie hustých vektorov
- Bez vlastných operácií TensorFlow



FPGA

- Veľké dátové súbory, modely
- Výpočtovo náročné aplikácie
- Vysoký výkon, vysoký pomer výkon/cena

MATICOVÉ SPRACOVANIE

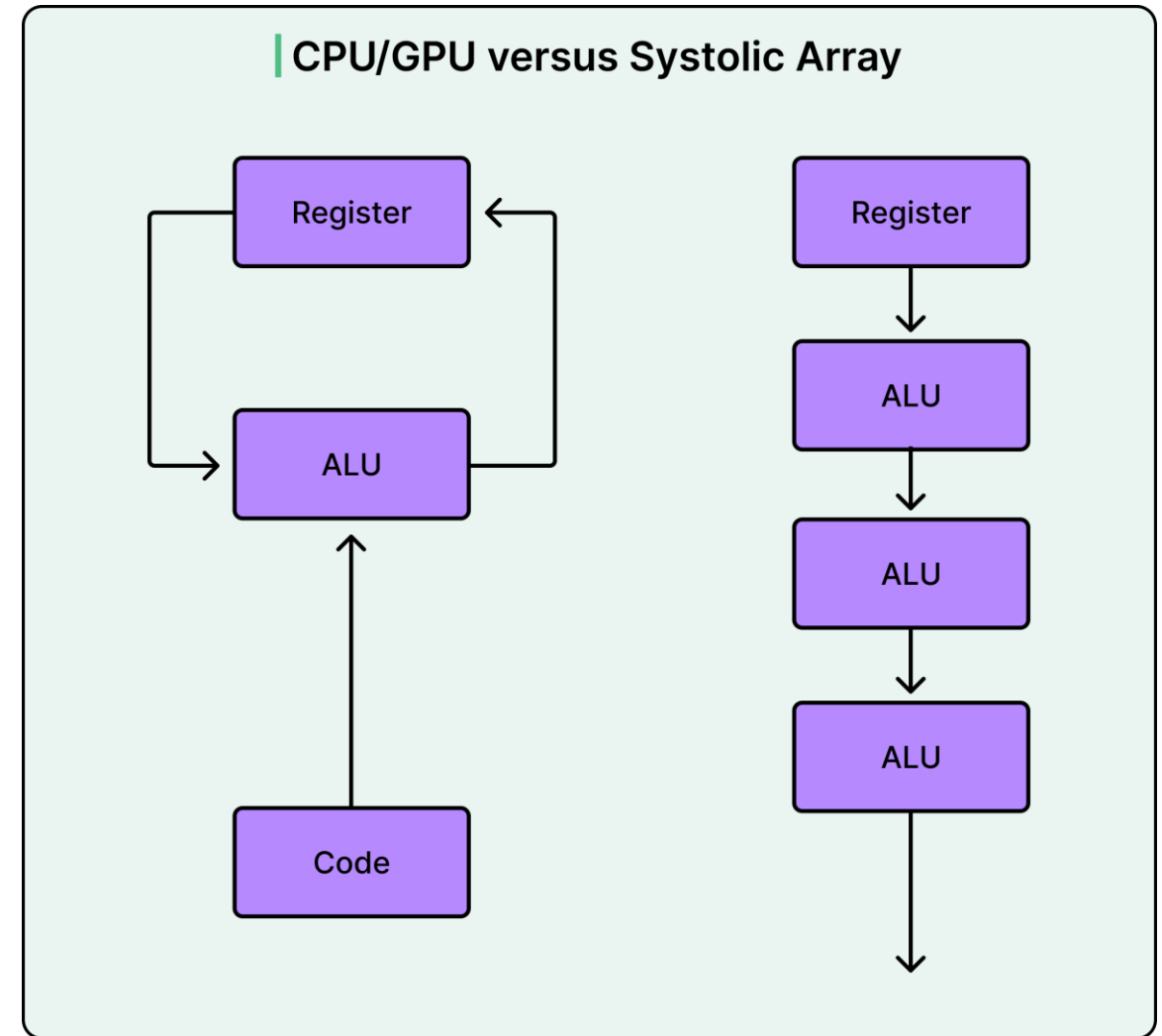
- Základ AI a grafiky
- vykonávanie viacerých matematických operácií naraz
- AI je postavená na násobení matic
- výpočty matic realizujú akcelerátory
- základným prvkom násobenia matic je tzv. **skalárny súčin**
$$(\mathbf{x}_1 \cdot \mathbf{y}_1) + (\mathbf{x}_2 \cdot \mathbf{y}_2) + \dots$$
- Bez FMA
 - Procesor potrebuje 2 inštrukcie na každý pár čísel
- S FMA
 - Procesor potrebuje len 1 inštrukciu

SPÁJANIE OPERÁCIÍ (FMA)

- namiesto dvoch samostatných krokov (vynásob a potom pripočítaj) vykoná operáciu $D = A \times B + C$ ako jeden nedeliteľný úkon
- **Výhody**
 - **Vyššia presnosť**
 - výpočet prebieha s extrémnou vnútornou presnosťou a zaokrúhľuje sa až úplne finálny výsledok, to je kritické pre stabilitu hlbokých neurónových sietí
 - **Rýchlosť a latencia**
 - odstránenie čakania medzi násobením a sčítaním zrýchľuje celý výpočtový reťazec
 - **Energetická efektivita**
 - menej tranzistorov a menej presunov dát znamená, že čip spotrebuje menej elektriny a menej sa zahrieva

SYSTOLICKÉ POLIA

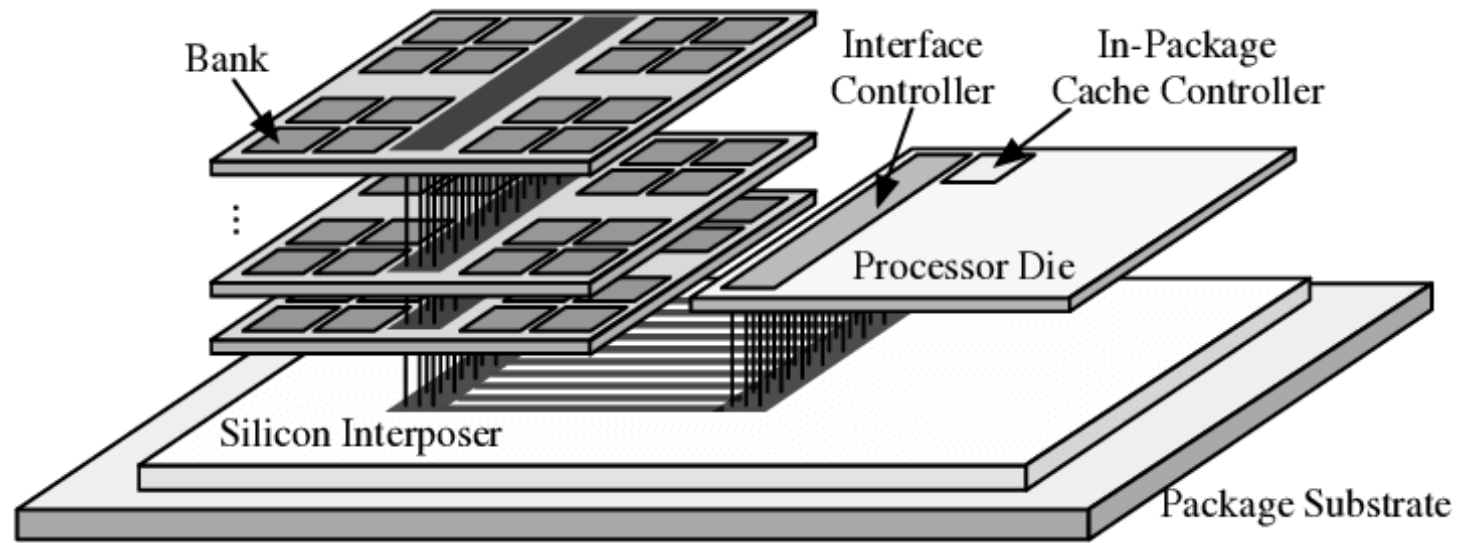
- riešia problém maticových výpočtov tak, že namiesto presúvania dát do procesora sa dáta pohybujú cez mriežku výpočtových jednotiek
- tým sa výrazne zvyšuje paralelizmus (stovky výpočtov naraz), zlepšuje sa prístup do pamäte a vykonáva sa viac operácií
- využitie pri maticovom násobení (GEMM), neurónových sieťach



Obr.3 : systolické polia

HBM (HIGH BANDWIDTH MEMORY)

- **HBM (High Bandwidth Memory)**
 - pamäťové čipy sú vrstvené vertikálne na seba, čo šetrí miesto a zvyšuje kapacitu
 - pamäť sa nachádza v rovnakom čipe ako procesor (GPU/TPU)
 - bežná DDR5 má 64-bitovú zbernicu, HBM využíva tisíce spojov, čím dosahuje priepustnosť v (TB/s), DDR5 (GB/s)

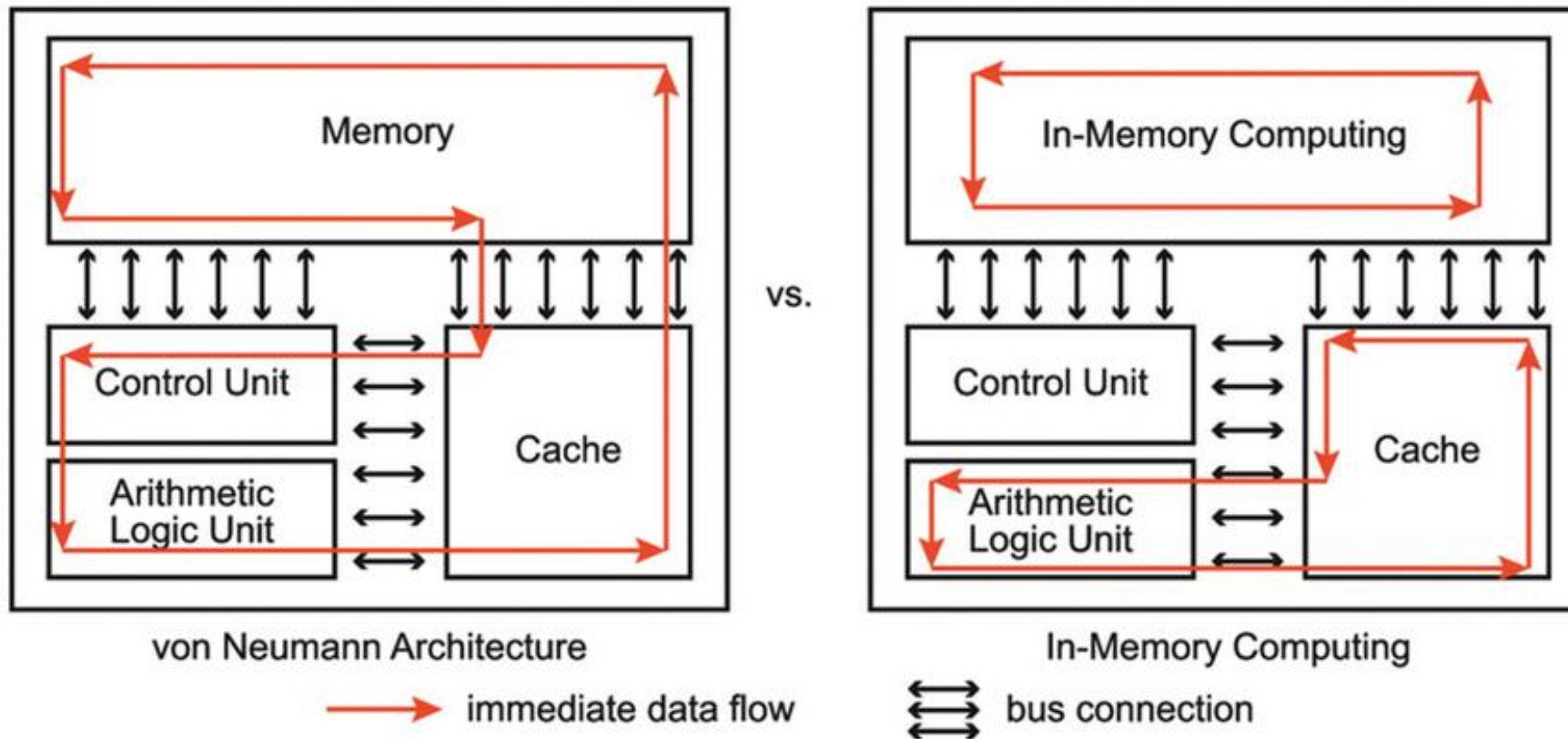


Obr.4 : HBM

IN-MEMORY COMPUTING

- Von Neumannova architektúra presúva dáta k procesoru, zatiaľ čo in-memory computing sa snaží integrovať výpočtové jednotky priamo tam, kde sú dáta uložené
- základom je mriežka vodičov, kde riadky slúžia ako vstupy a stĺpce ako výstupy, pričom v každom priesečníku sa nachádza pamäťový prvok, ktorý má určitú elektrickú vodivosť
- vykonáva logické a aritmetické operácie (napr. násobenie matíc pre AI) priamo v poli pamäťových buniek, čím umožňuje masívny paralelizmus

IN-MEMORY COMPUTING

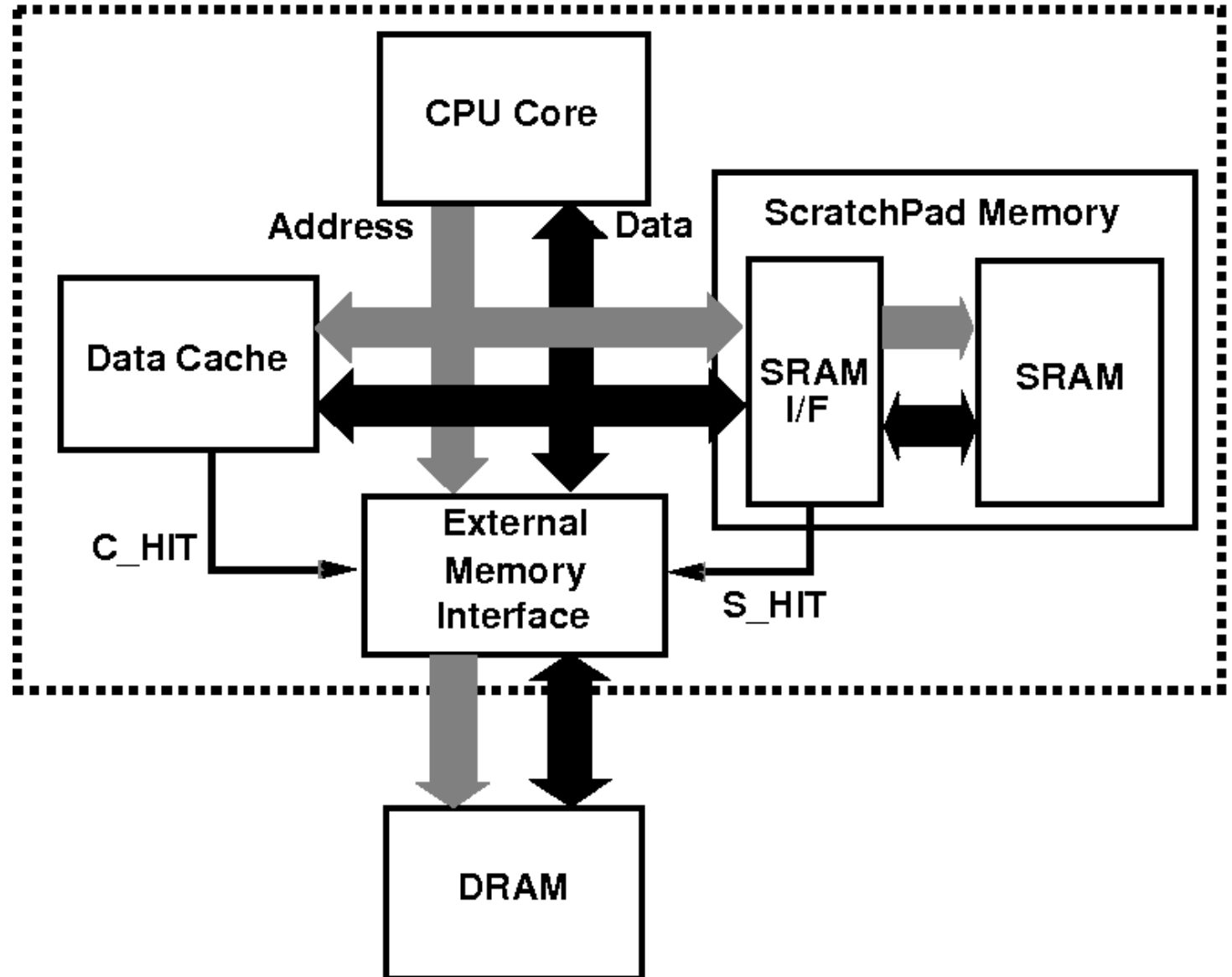


ON-CHIP SRAM

on-chip SRAM

- integrácia veľkých blokov pamäte typu SRAM (Static RAM) priamo na čip k výpočtovým jadrám
- pamäť nie je v samostatnom slote na základnej doske, ale je priamo na tom istom kúsku kremíka ako výpočtové jadrá
- SRAM je typ pamäte, ktorá je fyzicky postavená z tranzistorov (6 tranzistorov na 1 bit)
- na rozdiel od klasickej DRAM (v RAM), ktorá používa tranzistor a kondenzátor, SRAM nepotrebuje neustály refresh a je extrémne rýchla (desiatky nanosekúnd vs stovky nanosekúnd)
- nachádza sa v blízkosti výpočtových jednotiek (ALU/Tensor Cores)
- v klasických procesoroch ju poznáme ako L1, L2 a L3 Cache
- Eliminuje čakanie na pomalú externú RAM (prístup 100-200ns) a znižuje energetickú náročnosť presunu dát

ON-CHIP vs. OFF-CHIP



Obr.6 : on-chip vs. off-chip

KONEKTIVITA ČIPOV

- **limitácia PCIe zbernice**

- pôvodne bola navrhnutá pre komunikáciu medzi CPU a perifériami (disk, sieťovka, grafická karta)
- PCIe je univerzálna zbernica navrhnutá pre širokú škálu zariadení, čo spôsobuje značné softvérové a hardvérové zdržanie (latenciu) pri každej komunikácii
- nízka priepustnosť
 - tréning veľkých jazykových modelov (LLM) vyžaduje neustálu synchronizáciu miliárd parametrov medzi desiatkami GPU

KONEKTIVITA ČIPOV

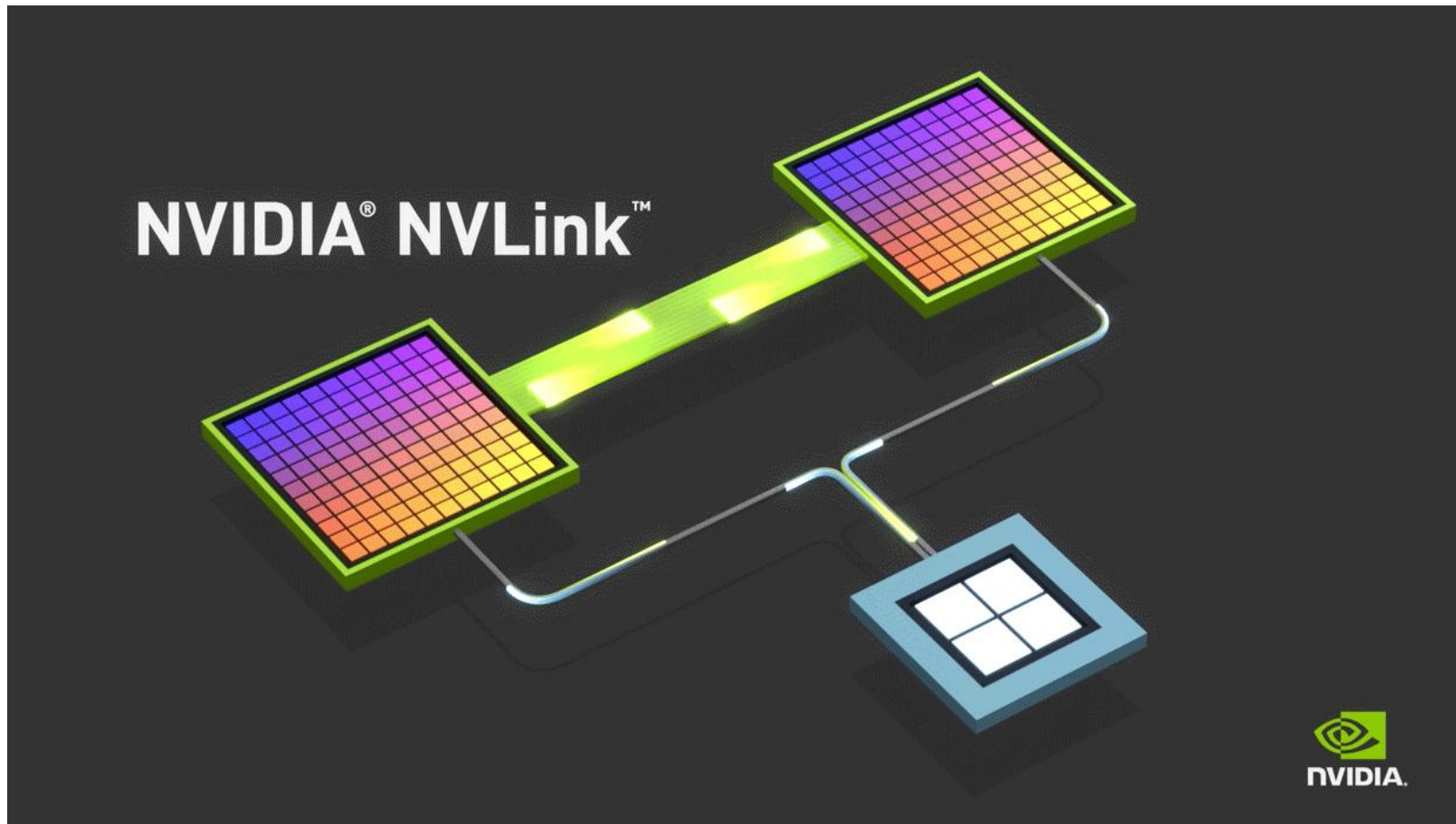
SPEC	X1	X2	X4	X8	X16
PCIe 1.x (2.5 GT/s)	500 MB/s	1 GB/s	2GB/s	4GB/s	8GB/s
PCIe 2.x (5.0 GT/s)	1GB/s	2GB/s	4GB/s	8GB/s	16GB/s
PCIe 3.x (8.0 GT/s)	2GB/s	4GB/s	8GB/s	16GB/s	32GB/s
PCIe 4.x (16 GT/s)	4GB/s	8GB/s	16GB/s	32GB/s	64GB/s
PCIe 5.x (32 GT/s)	8GB/s	16GB/s	32GB/s	64GB/s	128GB/s
PCIe 6.x (64 GT/s)	16GB/s	32GB/s	64GB/s	128GB/s	256GB/s

Obr.7 : PCIe zbernice

KONEKTIVITA ČIPOV

- **NVLink (Direct GPU-to-GPU)**
 - špecializovaný komunikačný protokol a fyzické rozhranie od spoločnosti NVIDIA
 - umožňuje spojenie napriamo (PTP)
 - umožňuje grafikám (GPU) rozprávať sa priamo medzi sebou, úplne vynechávajúc CPU a pomalú PCIe zbernicu
 - najnovšie verzie (NVLink 4. generácie) dosahujú priepustnosť až 900 GB/s, čo je takmer 10x viac než PCIe
 - vďaka tejto rýchlosti sa prepojené grafiky správajú tak, ako keby mali jednu zdieľanú pamäť

KONEKTIVITA ČIPOV



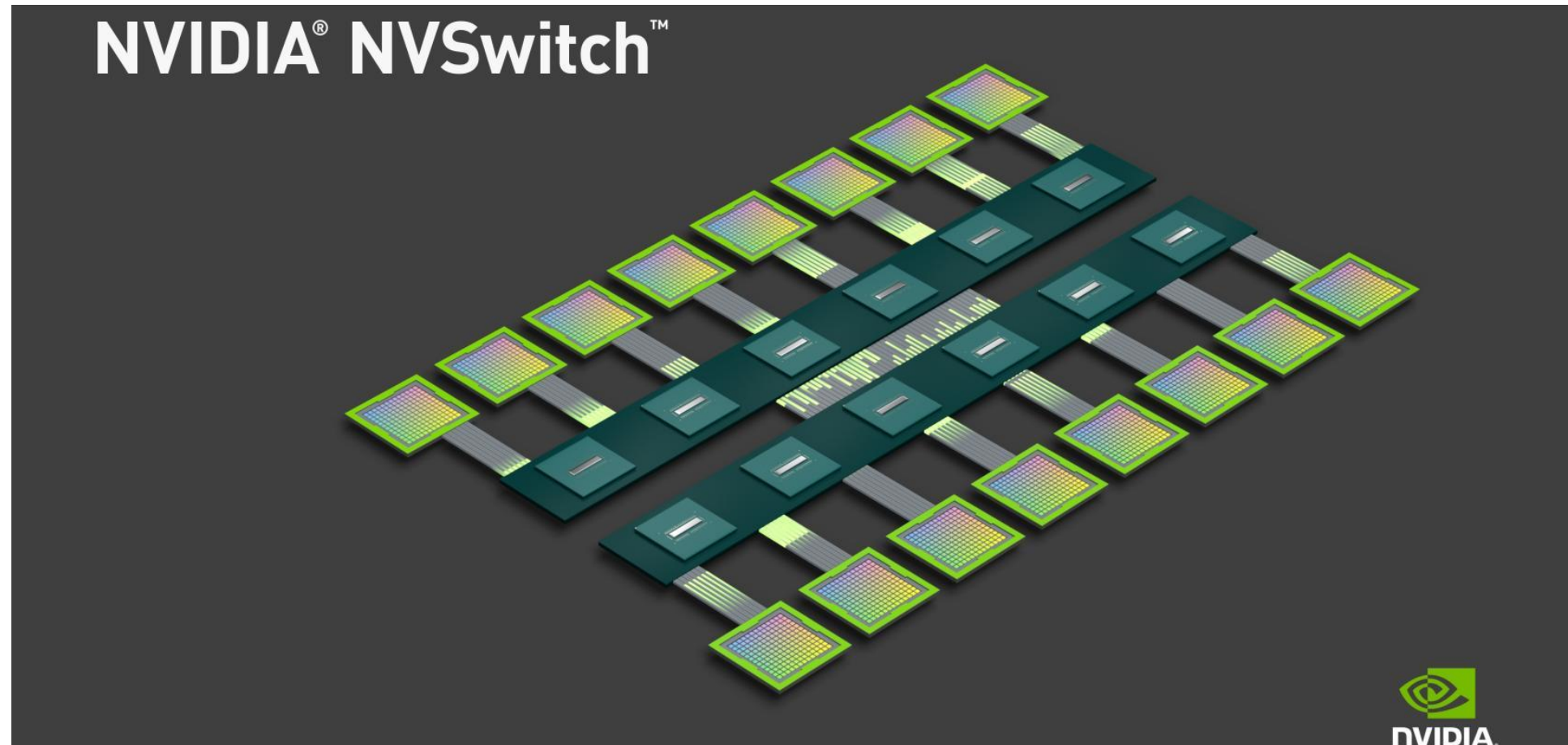
Obr.8 : NVLink

KONEKTIVITA ČIPOV

- **NVSwitch**

- rozšírenie možností NVLinku
- samostatný kremíkový čip
- každá grafická karta (GPU) má jeden alebo viac NVLink výstupov, ktoré vedú priamo do NVSwitchu
- dáta sa vo forme paketov prenášajú do NVSwitchu, ktorý na základe smerovacej logiky identifikuje cieľovú adresu a prepojí ich na zodpovedajúci port, ku ktorému je pripojený cieľový akcelerátor (GPU)

KONEKTIVITA ČIPOV



Obr.9 : NVSwitch

ENERGETICKÁ EFEKTIVITA

- **optimalizácia dizajnu a architektúry**
 - namiesto celého modelu sa aktivujú len tie časti, ktoré sú potrebné pre danú úlohu, čím sa šetrí výpočtový výkon
- **kompresia a optimalizácia prevádzky**
 - trend malých modelov
 - Prechod od gigantických modelov k menším, špecializovaným verziám, ktoré spotrebujú zlomok energie
 - **kvantizácia**
 - zníženie presnosti parametrov, čo znižuje nároky na pamäť a procesor
- **zníženie nárokov na chladenie**
 - efektívnejšie modely generujú menej tepla, čím sa nepriamo šetrí aj voda a energia potrebná na chladenie dátových centier

BUDÚCNOSŤ

- **neuromorfné inžinierstvo**

- zameriava sa na neuromorfné čipy, ktoré sa snažia napodobniť fungovanie ľudského mozgu

- **optické čipy**

- využívajú svetlo namiesto elektriny, čo môže dramaticky zrýchliť výpočty

- **edge AI**

- miniaturizácia hardvéru - presun z cloudových serverov priamo do zariadení, ako sú smartfóny, autá alebo rôzne inteligentné zariadenia čo umožňuje rýchlejšie reakcie, nižšiu spotrebu dát a väčšie súkromie

- **kvantové procesory**

- kvantové procesory využívajú princípy kvantovej fyziky a umožňujú riešiť extrémne zložité problémy, ktoré by klasickým počítačom trvali nepredstaviteľne dlho alebo by ich nedokázali vyriešiť vôbec

ZDROJE

- Wikipedia. Hardware for artificial intelligence [online]. Dostupné z: https://en.wikipedia.org/wiki/Hardware_for_artificial_intelligence
- Emergent Mind. Matrix Processing Units (MPUs) [online]. Dostupné z: <https://www.emergentmind.com/topics/matrix-processing-units-mpus>
- IBM. What is AI hardware? [online]. Dostupné z: <https://www.ibm.com/think/topics/ai-hardware>
- Allied. CPU vs GPU vs TPU: The ultimate guide to choosing the right accelerator for AI and ML [online]. Dostupné z: <https://www.allied.vc/guides/cpu-vs-gpu-vs-tpu-the-ultimate-guide-to-choosing-the-right-accelerator-for-ai-and-ml>
- Skymod Tech. Energy efficiency in AI models: Strategies for a sustainable future [online]. Dostupné z: <https://skymod.tech/energy-efficiency-in-ai-models-strategies-for-a-sustainable-future/>
- Inaccel. CPU, GPU, FPGA or TPU: Which one to choose for my machine learning training? [online]. Dostupné z: <https://inaccel.medium.com/cpu-gpu-fpga-or-tpu-which-one-to-choose-for-my-machine-learning-training-948902f058e0>
- ScienceDirect. Hardware Acceleration [online]. Dostupné z: <https://www.sciencedirect.com/topics/computer-science/hardware-acceleration>

ĎAKUJEM ZA POZORNOST

