

Hardvér pre umelú inteligenciu

Lukáš Knížat

FPV UCM, APIN, 2. ročník, 2025/26

Obsah

- Úvod
- Tenzorová procesorová jednotka (TPU)
- Ako funguje TPU
- Prečo je TPU tak dôležité a príklady jeho využitia
- História TPU
- TPU7x (Ironwood)
- Neurónová procesorová jednotka (NPU)
- Kľúčové vlastnosti a výhody NPU
- Ako funguje NPU

TPU

- Tenzorová procesorová jednotka (TPU) je špecializovaný čip určený na urýchlenie úloh umelej inteligencie a strojového učenia (ML). Na rozdiel od tradičných počítačových procesorov (CPU) alebo grafických procesorov (GPU) sú TPU špeciálne navrhnuté na spracovanie zložitých výpočtov potrebných pre modely hlbokého učenia.
- TPU boli vyvinuté spoločnosťou Google v roku 2015 s cieľom zlepšiť výkon svojich aplikácií umelej inteligencie, ako sú Google vyhľadávanie, Google prekladač a Google fotografie. Odvtedy sa TPU stali kľúčovou súčasťou infraštruktúry umelej inteligencie a sú široko používané v dátových centrách a cloudových výpočtoch.

Ako funguje TPU

- Modely umelej inteligencie sa spoliehajú na typ matematickej operácie nazývanej tenzorové výpočty. Tensor je viacrozmerné pole čísel, podobné tabuľke údajov. Modely hlbokého učenia používajú tieto tenzory na spracovanie veľkého množstva informácií a vytváranie predpovedí.
- TPU sú optimalizované pre tenzorové výpočty, čo im umožňuje spracovávať veľké súbory údajov oveľa rýchlejšie ako CPU alebo GPU. Dosahujú to prostredníctvom:
- Masívneho paralelizmu – TPU dokážu vykonávať veľa výpočtov naraz, vďaka čomu sú vysoko efektívne.
- Nízkej spotreby energie – V porovnaní s GPU spotrebujú TPU menej energie a zároveň poskytujú vysoký výkon.
- Špecializované obvody – TPU majú obvody špeciálne navrhnuté pre pracovné zaťaženie umelej inteligencie, čím sa znižuje potreba zbytočných výpočtov.
- Zatiaľ čo CPU sú skvelé na všeobecné úlohy a GPU sú vynikajúcou voľbou pre hry a umelú inteligenciu, TPU sú špeciálne navrhnuté tak, aby modely umelej inteligencie pracovali rýchlejšie a efektívnejšie.

Prečo sú TPU tak dôležité

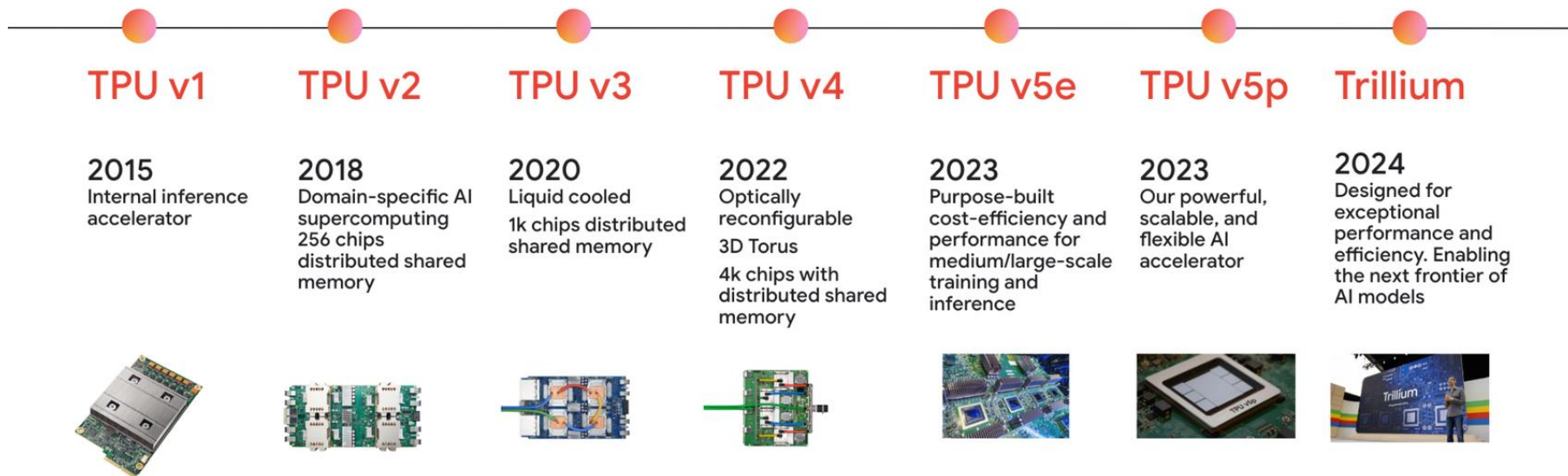
- TPU sú nevyhnutné pre rozvoj AI, pretože umožňujú vedcom a spoločnostiam rýchlejšie trénovať a nasadzovať modely hlbokého učenia. Dôvody prečo sú TPU dôležité:
- Rýchlejšie trénovanie AI – Trénovanie modelov AI môže trvať dni alebo dokonca týždne s CPU alebo GPU, ale TPU môžu tento čas výrazne skrátiť.
- Nižšie náklady – Hoci sú TPU drahé, znižujú celkové náklady na vývoj AI zrýchlením spracovania a znížením spotreby energie.
- Škálovateľnosť – TPU umožňujú spoločnostiam prevádzkovať aplikácie AI vo väčšom rozsahu, čím sa sprístupňujú služby ako rozpoznávanie hlasu, analýza obrázkov a preklad jazyka.
- Cloudové služby AI – Google ponúka TPU prostredníctvom svojej služby Google Cloud TPU, čo umožňuje firmám a vedcom používať vysokovýkonné výpočty AI bez nutnosti kupovať drahý hardvér.

Prípady použitia TPU

- TPU sa používajú v širokej škále aplikácií umelej inteligencie vrátane:
- Spracovanie prirodzeného jazyka (NLP) – chatboty umelej inteligencie, preklad jazyka a rozpoznávanie reči.
- Počítačové videnie – Rozpoznávanie tváre, lekárske zobrazovanie a autonómne vozidlá.
- Systémy odporúčaní – Online nakupovanie, odporúčania hudby a videa.
- Vedecký výskum – Objavovanie liekov, modelovanie klímy a genetický výskum.

História TPU

TPU AI accelerators



TPU v1(2015)

- Google nasadil prvú Tenzorová procesorovú jednotku v roku 2015, aby vyriešil konkrétny problém: záťaž neurónových sietí pri inferencii hrozila zdvojnásobením priestoru potrebného v dátových centrách spoločnosti. Inžinieri navrhli TPU v1 výhradne pre inferenciu – trénovacie schopnosti úplne vynechali, aby maximalizovali výkon a energetickú efektivitu nasadených modelov. Čip obsahoval systolické pole o rozmeroch 256×256 s 8-bitovými jednotkami pre násobenie a akumuláciu, čo mu umožňovalo dosahovať 92 teraoperácií za sekundu pri tepelnom výkone len 28–40 wattov.
- Architektúra stavila na radikálnu jednoduchosť. Jediná jednotka pre násobenie matíc spracovávala operácie vo formáte INT8 pomocou tzv. weight-stationary dátového toku – váhy zostávali pevne uložené v systolickom poli, zatiaľ čo aktivácie ním prúdili horizontálne. Čiastkové súčty sa šírili vertikálne, čím sa pri celom násobení matíc úplne eliminovali medzizápisy do pamäte. Čip bol k hostiteľským systémom pripojený cez PCIe, ako externú pamäť využíval DDR3 DRAM a pracoval na frekvencii 700 MHz – zámerne konzervatívnej hodnote kvôli nízkej spotrebe energie.
- Výsledky prekonalí očakávania aj samotných inžinierov Googlu. TPU v1 dosiahol 30- až 80-násobné zlepšenie výkonu na watt v porovnaní s vtedajšími CPU a GPU pri produkčných inferenčných záťažach. Čip obsluhoval radenie výsledkov Google Vyhľadávania, prekladové služby spracúvajúce miliardu požiadaviek denne a odporúčania na YouTube pre dve miliardy používateľov. Tento úspech potvrdil kľúčový architektonický predpoklad: akcelerátory navrhnuté na mieru pre úzko vymedzené úlohy dokážu prekonať procesory všeobecného použitia.

TPU v2 (2017)

- Druhá generácia zmenila TPU z akcelérátora zameraného výlučne na inferenciu na plnohodnotnú trénovaciu platformu. Google prepracoval celú architektúru smerom k operáciám s pohyblivou rádovou čiarkou – pole INT8 s rozmermi 256×256 nahradil duálnymi multiplikátormi-akumulátormi bfloat16 s rozmermi 128×128 na jadro. Každý čip obsahoval dve jadrá TensorCore so zdieľanou pamäťou s vysokou šírkou pásma o kapacite 8 GB na jadro, čo predstavovalo zásadný posun oproti DDR3 a zabezpečovalo šírku pásma nevyhnutnú pre trénovanie neurónových sietí.
- Presnosť bfloat16 sa ukázala ako jeden z kľúčových faktorov úspechu TPU v2. Tento formát si zachováva rovnaký 8-bitový rozsah exponentov ako FP32, no mantisu skracuje na 7 bitov – vďaka tomu si uchováva dynamický rozsah dôležitý pri trénovaní a zároveň znižuje nároky na šírku pásma pamäte na polovicu. Inžinieri pritom zaznamenali zaujímavý vedľajší efekt: znížená presnosť mantisy v mnohých modeloch dokonca zlepšila schopnosť zovšeobecňovania, keďže fungovala ako istá forma regularizácie. Plný rozsah exponentov zhodný s FP32 zároveň eliminoval problémy s podtečením a pretečením, ktoré pri trénovaní komplikovali používanie formátu FP16.
- Skutočným architektonickým prelomom TPU v2 však bolo prepojenie medzi čipmi, známe ako Inter-Chip Interconnect (ICI). Staršie akcelérátory využívali pre komunikáciu medzi čipmi Ethernet alebo InfiniBand, čo prinášalo výrazné obmedzenia v oblasti latencie aj šírky pásma. Google navrhol vlastné vysokorýchlostné obojsmerné prepojenia, ktoré každý TPU spájali priamo so štyrmi susednými čipmi v topológii 2D torusu. Vďaka tomu mohli pody TPU v2 s až 256 čipmi fungovať ako jeden logický akcelérátor, pričom kolektívne operácie ako all-reduce prebehli výrazne rýchlejšie ako pri akýchkoľvek sieťových alternatívach.

TPU v3 (2018)

- Google v tretej generácii výrazne zvýšil frekvenciu hodín aj počet jadier, čím dosiahol výkon 420 teraflopov na čip – viac ako dvojnásobok oproti v2. Takáto hustota výkonu si však vynútila zásadnú architektonickú zmenu: prechod na kvapalinové chladenie. Každý pod TPU v3 vyžadoval infraštruktúru vodného chladenia, čo bol výrazný odklon od vzduchom chladených riešení predchádzajúcich generácií aj väčšiny ostatných akcelerátorov nasadzovaných v dátových centrách.
- Čip si zachoval duálnu architektúru MXU s rozmermi 128×128 , no zvýšil celkový počet jadier a zlepšil šírku pásma pamäte. Každý TPU v3 tvorili štyri čipy s dvoma jadrami, pričom celková kapacita zdieľanej pamäte HBM dosiahla 32 GB. Vektorové procesorové jednotky sa dočkali vylepšení v oblasti aktivačných funkcií, normalizačných operácií a výpočtov gradientov – práve tieto operácie totiž pri tréňovaní závislom výhradne na maticových jednotkách často tvorili úzke miesto.
- Nasadenia sa škálovali na pody s 2 048 čipmi, ktoré využívali rovnakú topológiu 2D torusu pre ICI ako v2, tentokrát však s vyššou šírkou pásma na linku. Google na podoch v3 trénoval čoraz rozsiahlejšie modely a zistil, že topológia torusu prirodzene znižuje priemerný priemer siete – maximálna vzdialenosť medzi ľubovoľnými dvoma čipmi rastie úmerne $N/2$ namiesto N , čo výrazne znižuje komunikačnú réžiu pri tréňovaní paralelnom po dátach aj paralelnom po modeli.

TPU v4 (2021)

- Štvrtá generácia predstavovala najvýznamnejší architektonický skok Googlu od vzniku pôvodného TPU. Inžinieri rozšírili pody na 4 096 čipov a zároveň zaviedli optické prepínanie obvodov (OCS) – technológiu prevzatú z telekomunikácií, ktorá zásadne zmenila spôsob budovania infraštruktúry strojového učenia v dátových centrách.
- Jadrová architektúra TPU v4 obsahovala štyri MXU s rozmermi 128×128 na TensorCore, doplnené o vylepšené vektorové a skalárne jednotky. Každý pár TensorCore zdieľal 128 MB spoločnej pamäte nad rámec vektorovej pamäte na jadro, čo umožnilo sofistikovanejšie postupy spracovania a opätovného využívania dát. Topológia prepojenia čipov sa posunula z 2D na 3D torus – každý TPU sa tak prepojil so šiestimi susedmi namiesto štyroch, čím sa ďalej znížil priemer siete a zlepšila šírka pásma bisekcie.
- Systém optického prepínania obvodov zmenil pravidlá hry pri rozsiahlych nasadeniach. Namiesto pevnej kabeláže medzi TPU Google nasadil programovateľné optické prepínače schopné dynamicky meniť, ktoré čipy sú navzájom prepojené. Zrkadlá MEMS fyzicky presmerúvajú svetelné lúče tak, aby spojili ľubovoľné dvojice TPU, pričom vnášajú prakticky nulovú latenciu nad rámec samotného prenosu optickým vláknom. Prepínače sa rekonfigurujú v časových oknách pod 10 nanosekúnd – rýchlejšie, ako väčšina sieťových protokolov stihne dokončiť handshake.
- Architektúra OCS otvorila možnosti, ktoré predtým neboli realizovateľné. Google mohol vhodným naprogramovaním optických prepínačov poskytnúť sekcie ľubovoľnej veľkosti – od štyroch čipov až po celý pod s 4 096 čipmi. Chybné čipy bolo možné elegantne obísť bez nutnosti rozoberať celé racky. Najpozoruhodnejšie bolo, že fyzicky vzdialené TPU v rôznych dátových centrách mohli byť v sieťovej topológii logicky susedné, čím sa fyzické a logické rozloženie infraštruktúry úplne oddelili.
- TPU v4 priniesol aj SparseCore – špecializovaný procesor navrhnutý pre operácie s vkladami (embeddings), ktoré denne využívajú odporúčacie systémy, rankingové modely a veľké jazykové modely s rozsiahlymi slovníkovými vkladami. SparseCore disponoval štyrmi dedikovanými procesormi na čip, každý s 2,5 MB scratchpad pamäťou a optimalizovaným dátovým tokom pre vzory riedkeho prístupu k pamäti. Modely s mimoriadne veľkými vkladami zaznamenali 5- až 7-násobné zrýchlenie, a to pri využití iba 5 % celkovej plochy čipu a energetického rozpočtu.

TPU v5p a v5e (2022-2023)

- Google rozdelil piatu generáciu na dva odlišné produkty zacielené na rozdielne prípady použitia. TPU v5p staval na maximálnom výkone pre rozsiahle tréningové úlohy, zatiaľ čo v5e bol optimalizovaný pre nákladovo efektívnu inferenciu a menšie tréningové záťaže.
- TPU v5p dosiahol približne 4,45 exaflopov za sekundu naprieč podmi s 8 960 čipmi, čím viac ako zdvojnásobil maximálnu veľkosť podu štvrtej generácie. Šírka pásma prepojenia dosiahla 4 800 Gb/s na čip a 3D torusová topológia spájala čipy v masívnych superpodoch s rozmermi 16×20×28. Optická prepínacia štruktúra spravovala 13 824 optických portov naprieč 48 jednotkami OCS, aby vytvorila kompletný superpod v5p – jedno z najväčších nasadení optického prepínania v celej histórii výpočtovej techniky.
- TPU v5e zvolil opačný prístup. Znížením počtu jadier a frekvencie hodín sa Google zameral na agresívne ciele v oblasti výkonu na jednotku nákladov. Čipy optimalizované pre inferenciu obsahovali jediné TPU jadro namiesto dvoch a vrátili sa k topológii 2D torusu, ktorá pre menšie veľkosti podov plne postačovala. Toto architektonické zjednodušenie umožnilo Googlu ponúknuť v5e za konkurencieschopnú cenu pre záťaže, kde absolútny výkon ustupoval do pozadia a rozhodujúci bol výkon na dolár.

TPU v6e Trillium (2024)

- Trillium predstavoval ďalší architektonický inflexný bod – základná jednotka pre násobenie matíc sa rozšírila z rozmerov 128×128 na 256×256 multiplikačných akumulátorov. Väčšie pole štvornásobne zvýšilo počet operácií na cyklus pri rovnakej taktovacej frekvencii, pričom v kombinácii so zvýšenými taktovacími frekvenciami dosiahol Trillium 4,7-násobok špičkového výpočtového výkonu TPU v5e.
- Rovnako výrazné zmeny sa dotkli pamäťového subsystému. Kapacita HBM sa zdvojnásobila na 32 GB na čip a šírka pásma vzrástla vďaka kanálom HBM novej generácie na dvojnásobok pôvodnej hodnoty. Šírka pásma Inter-Chip Interconnect sa takisto zdvojnásobila, čo umožnilo podom s 256 čipmi udržať vysokú priepustnosť aj pri modeloch zaťažujúcich súčasne výpočtový výkon aj medzičipovú komunikáciu.
- Trillium priniesol aj SparseCore tretej generácie s rozšírenými schopnosťami pre mimoriadne veľké vkladania využívané pri rankingových a odporúčacích úlohách. Prepracovaný dizajn zlepšil vzory prístupu k pamäti a zvýšil efektívnu šírku pásma medzi SparseCores a HBM, čo ocenili najmä modely, v ktorých dominovali vyhľadávané vkladania namiesto klasického násobenia matíc.
- Energetická účinnosť vzrástla o 67 % oproti v5e, a to napriek výraznému nárastu výkonu. Google toto zlepšenie dosiahol kombináciou pokročilejšieho výrobného procesu, architektonických optimalizácií znižujúcich zbytočne vykonávanú prácu a starostlivého riadenia napájania neaktívnych jednotiek počas operácií, ktoré nezaťažovali všetky časti čipu súčasne.

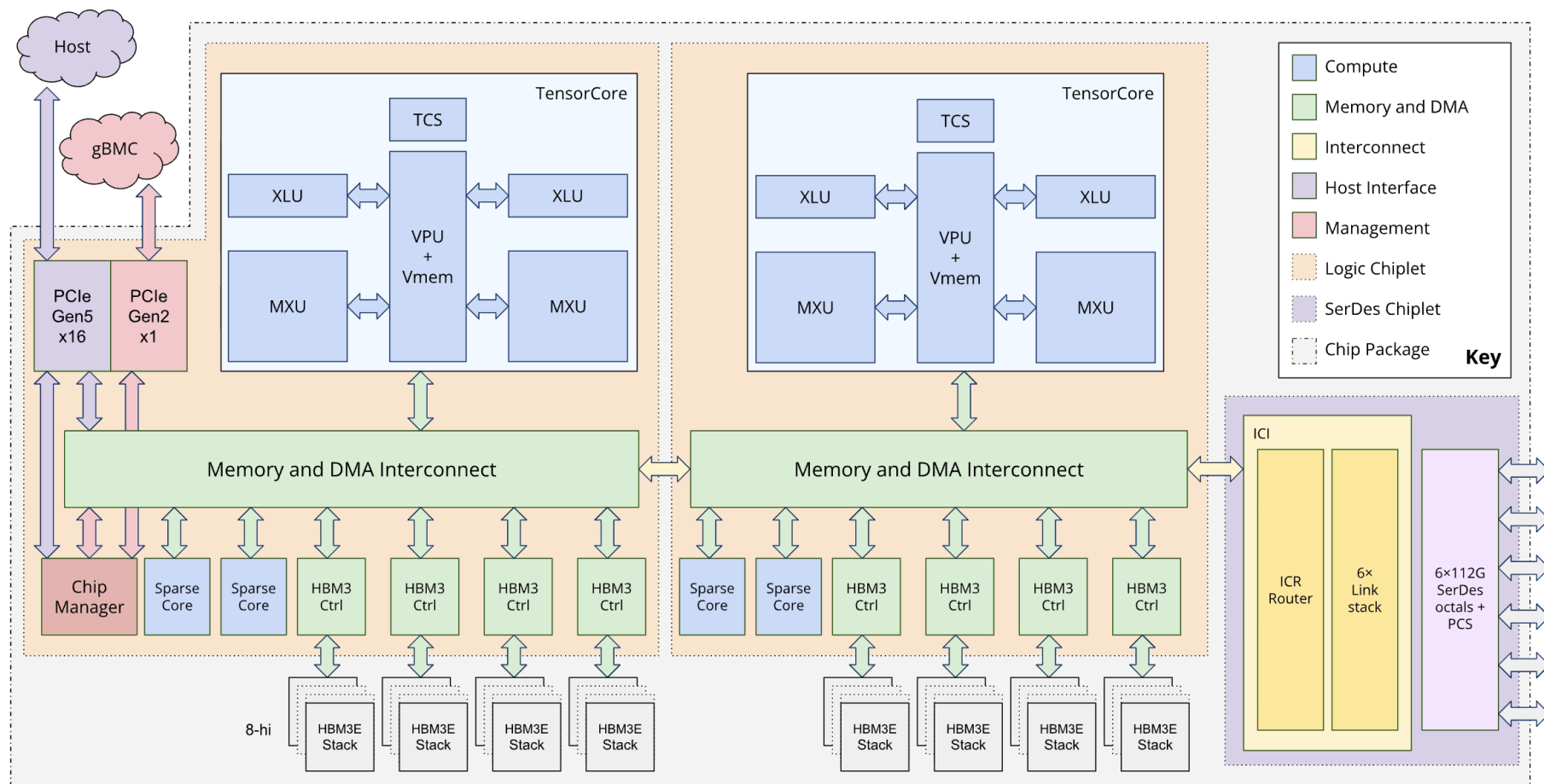
TPU v7 Ironwood (2025)

- Siedma generácia TPU s kódovým označením Ironwood je prvou TPU navrhnutou s natívnou podporou FP8 a optimalizovanou špeciálne pre „age of inference“, pričom si zachováva špičkový tréningový výkon. Každý čip Ironwood dosahuje 4,6 petaFLOPS hustého výpočtového výkonu FP8 – čo mierne prevyšuje konkurenčný NVIDIA B200 s výkonom 4,5 petaFLOPS – a zároveň pracuje pri tepelnom výkone 600 W.
- Pamäťový systém sa rozšíril na 192 GB HBM3e na čip, čo je šesťnásobok kapacity Trillium, so šírkou pásma dosahujúcou 7,4 TB/s. Takýto nárast pamäte umožňuje obsluhovať mimoriadne veľké modely s vyrovnanými pamäťami kľúč-hodnota, ktoré predtým vyžadovali komplexný tenzorový paralelizmus naprieč viacerými akcelerátormi. Google pamäťovú kapacitu zámerné navrhol s ohľadom na nastupujúce multimodálne modely a aplikácie s dlhým kontextom blížiacim sa k oknám s miliónmi tokenov.
- Prepojenie Ironwood poskytuje 9,6 Tbps celkovej obojsmernej šírky pásma cez štyri ICI linky, čo zodpovedá maximálnej šírke pásma 1,2 TB/s na čip. Architektúra sa škáluje od 256-čipových podov pre menšie nasadenia až po masívne superpody s 9 216 čipmi dosahujúce 42,5 exaflopov výpočtového výkonu FP8. Sieťová technológia dátového centra Jupiter od Googlu by teoreticky mohla prepojiť až 43 takýchto superpodov v jedinom klastri – čo predstavuje zhruba 400 000 akcelerátorov a takmer nepredstaviteľnú škálu výpočtového výkonu.
- Natívna podpora FP8 znamená zásadný posun v stratégii presnosti. Predchádzajúce generácie TPU emulovali 8-bitové operácie softvérovými technikami, čo prinášalo nezanedbateľnú réžiu. Ironwood implementuje natívne multiplikačné a akumulčné jednotky FP8 s podporou dvoch formátov – E4M3 so 4-bitovým exponentom a 3-bitovou mantisou, ako aj E5M2 s 5-bitovým exponentom a 2-bitovou mantisou. Podpora oboch formátov umožňuje ich kombinovanie: E4M3 pre priechody vpred, kde nároky na presnosť nie sú kritické, a E5M2 pre priechody vzad, kde udržiavanie veľkosti gradientov zabraňuje nestabilite trénovania.
- Závazok Anthropicu nasadiť od roku 2026 viac ako milión čipov Ironwood svedčí o produkčnej zrelosti tejto architektúry. Spoločnosť plánuje využiť viac ako gigawatt kapacity TPU – dostatok na napájanie malého mesta – výhradne na trénovanie a obsluhu modelov Claude. Tento rozsah prevyšuje aj najvýznamnejšie známe nasadenia GPU a predstavuje odvážnu stávkou na architektúru TPU pri vývoji modelov na hraniciach možností súčasnej výpočtovej techniky.

Špecifikácie

Specification	v5p	v6e (Trillium)	TPU7x (Ironwood)
Number of chips per pod	8960	256	9216
Peak compute per chip (BF16) (TFLOPs)	459	918	2307
Peak compute per chip (FP8) (TFLOPs)	459	918	4614
HBM capacity per chip (GiB)	95	32	192
HBM bandwidth per chip (GiBps)	2575	1638	7380
Number of vCPUs (4-chip VM)	208	180	224
RAM (GB) (4-chip VM)	448	720	960
Number of TensorCores per chip	2	1	2
Number of SparseCores per chip	4	2	4
Bidirectional inter-chip interconnect (ICI) bandwidth per chip (GBps)	1200	800	1200
Data center network (DCN) bandwidth per chip (Gbps)	50	100	100

TPU7x architektúra



NPU

- Neurónová procesorová jednotka (NPU) je špecializovaný počítačový mikroprocesor navrhnutý tak, aby napodobňoval funkciu ľudského mozgu. Je optimalizovaný pre úlohy a aplikácie umelej inteligencie (AI), neurónové siete, hlboké učenie a strojové učenie.
- Na rozdiel od univerzálnych centrálnych procesorových jednotiek (CPU) alebo grafických procesorových jednotiek (GPU) sú NPU prispôbované na urýchlenie úloh a pracovných záťaží AI, ako je výpočet vrstiev neurónovej siete zložených zo skalárnej, vektorovej a tenzorovej matematiky.
- NPU, známe aj ako AI čip alebo AI akcelerátor, sa zvyčajne používajú v heterogénnych výpočtových architektúrach, ktoré kombinujú viacero procesorov (napríklad CPU a GPU). Veľké dátové centrá môžu používať samostatné NPU pripojené priamo k základnej doske systému; väčšina spotrebiteľských aplikácií, ako sú smartfóny, mobilné zariadenia a notebooky, však kombinuje NPU s inými koprocormi na jednom polovodičovom mikročipe známom ako systém na čipe (SoC).
- Integráciou špecializovanej NPU sú výrobcovia schopní ponúkať generatívne aplikácie umelej inteligencie priamo na zariadení, ktoré dokážu spracovať aplikácie umelej inteligencie, pracovné zaťaženie umelej inteligencie a algoritmy strojového učenia v reálnom čase s relatívne nízkou spotrebou energie a vysokou priepustnosťou.

Ako funguje NPU

- Neurónové procesorové jednotky (NPU), založené na neurónových sieťach mozgu, fungujú simuláciou správania ľudských neurónov a synapsí na úrovni obvodov. To umožňuje spracovanie inštrukčných sád hlbokého učenia, v ktorých jedna inštrukcia dokončí spracovanie sady virtuálnych neurónov.
- Na rozdiel od tradičných procesorov nie sú NPU určené na presné výpočty. Namiesto toho sú NPU určené na riešenie problémov a časom sa môžu zlepšovať, učiť sa z rôznych typov údajov a vstupov. Vďaka strojovému učeniu môžu systémy umelej inteligencie obsahujúce NPU poskytovať prispôbené riešenia rýchlejšie a bez potreby manuálneho programovania.
- Výnimočnou vlastnosťou je, že NPU ponúkajú vynikajúce paralelné spracovanie a sú schopné urýchliť operácie umelej inteligencie prostredníctvom zjednodušených vysokokapacitných jadier, ktoré sú oslobodené od vykonávania viacerých typov úloh. NPU obsahuje špecifické moduly pre násobenie a sčítanie, aktivačné funkcie, operácie s 2D údajmi a dekompresiu. Špecializovaný modul pre násobenie a sčítanie sa používa na vykonávanie operácií relevantných pre spracovanie aplikácií neurónových sietí, ako je výpočet násobenia a sčítania matíc, konvolúcia, skalárny súčin a ďalšie funkcie.
- Zatiaľ čo tradičné procesory vyžadujú na dokončenie tohto typu spracovania neurónov tisíce inštrukcií, NPU by mohla byť schopná dokončiť podobnú operáciu iba s jednou. NPU bude tiež integrovať úložisko a výpočet prostredníctvom synaptických váh, čo je fluidná výpočtová premenná priradená sieťovým uzlom, ktorá indikuje pravdepodobnosť „správneho“ alebo „požadovaného“ výsledku, ktorá sa môže časom upravovať alebo „učiť“, čo vedie k zlepšeniu prevádzkovej efektívnosti.
- Hoci sa vývoj NPU neustále vyvíja, testovanie ukázalo, že výkon niektorých NPU je viac ako 100-krát lepší ako u porovnateľnej GPU pri rovnakej spotrebe energie.

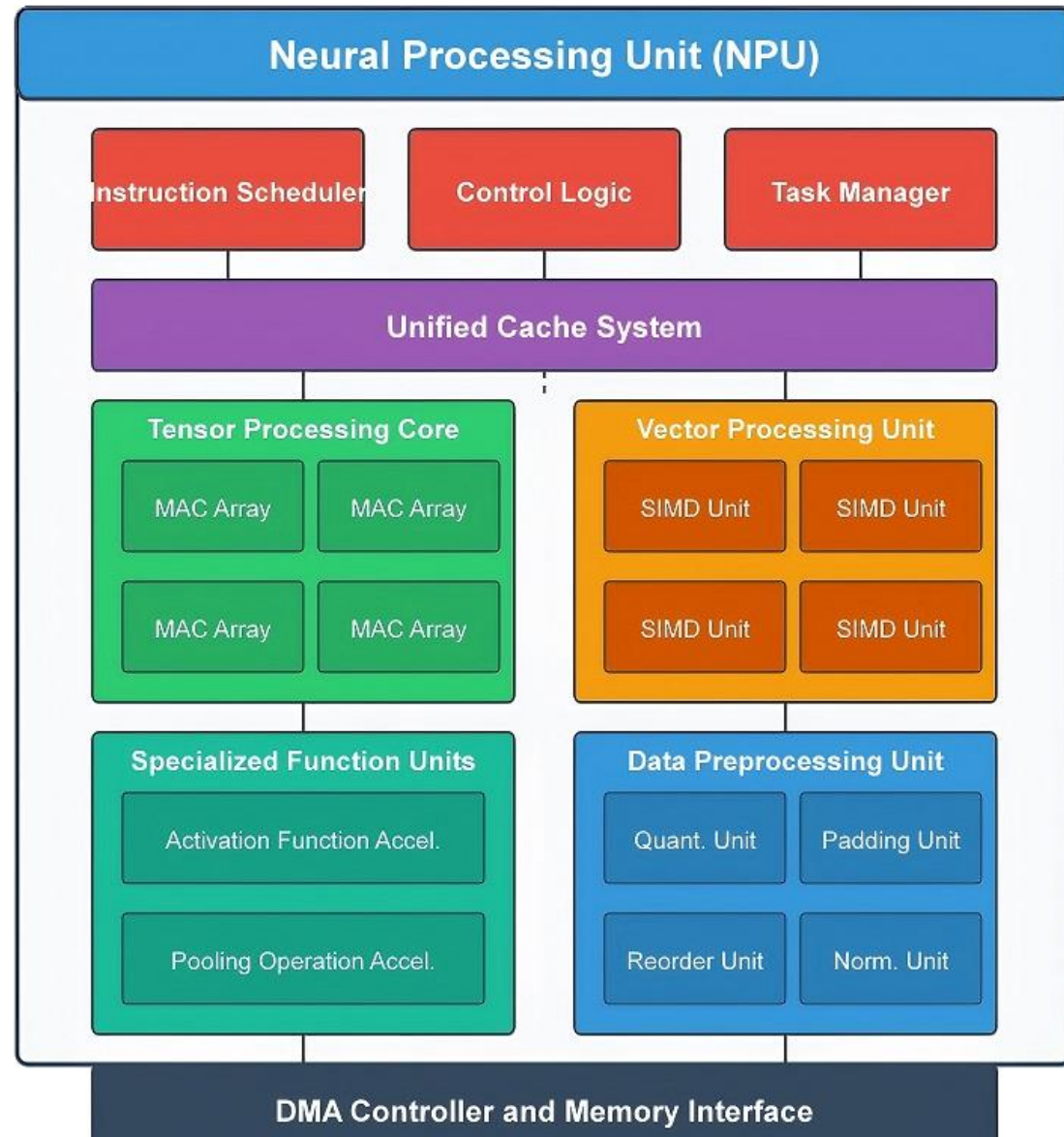
Klíčové vlastnosti NPU

- Neurónové procesorové jednotky (NPU) sú vhodné na úlohy, ktoré vyžadujú paralelné výpočty s nízkou latenciou, ako je spracovanie algoritmov hlbokého učenia, rozpoznávanie reči, spracovanie prirodzeného jazyka, spracovanie fotografií a videa a detekcia objektov.
- Klíčové vlastnosti NPU zahŕňajú:
- Paralelné spracovanie: NPU dokážu rozdeliť väčšie problémy na komponenty pre riešenie problémov s multitaskingom. To umožňuje procesoru spúšťať viacero operácií neurónovej siete súčasne.
- Aritmetika s nízkou presnosťou: NPU často podporujú 8-bitové (alebo nižšie) operácie, aby sa znížila výpočtová zložitosť a zvýšila energetická účinnosť.
- Pamäť s vysokou šírkou pásma: Mnohé NPU majú pamäť s vysokou šírkou pásma na čipe, aby efektívne vykonávali úlohy spracovania umelej inteligencie vyžadujúce veľké súbory údajov.
- Hardvérová akcelerácia: Pokroky v dizajne NPU viedli k začleneniu techník hardvérovej akcelerácie, ako sú architektúry systolických polí alebo vylepšené spracovanie tenzorov.

Kľúčové výhody NPU

- Paralelné spracovanie: NPU dokážu rozdeliť väčšie problémy na komponenty pre riešenie problémov s multitaskingom. Kľúčom je, že hoci GPU vynikajú aj v paralelnom spracovaní, jedinečná štruktúra NPU môže prekonať ekvivalentnú GPU so zníženou spotrebou energie a menšou fyzickou zastavanosťou.
- Zvýšená účinnosť: Zatiaľ čo GPU sa často používajú na vysokovýkonné výpočty a úlohy umelej inteligencie, NPU dokážu vykonávať podobné paralelné spracovanie s oveľa lepšou energetickou účinnosťou. Keďže umelá inteligencia a iné vysokovýkonné výpočty sa stávajú čoraz bežnejšími a vyžadujú viac energie, NPU ponúkajú cenné riešenie na zníženie kritickej spotreby energie.
- Spracovanie multimediálnych údajov v reálnom čase: NPU sú navrhnuté tak, aby lepšie spracovávali a reagovali na širšiu škálu dátových vstupov vrátane obrázkov, videa a reči. Rozšírenie aplikácií, ako je robotika, zariadenia internetu vecí (IoT) a nositeľná elektronika, o NPU môže poskytovať spätnú väzbu v reálnom čase, čím sa znižuje prevádzkové trenie a poskytuje kritickú spätnú väzbu a riešenia, keď je najdôležitejší čas odozvy.

Architektúra NPU



Prípady použitia NPU

Umelá
inteligencia a
LLM

Internet vecí
(IoT)

Data centrá

Autonómne
vozidlá

Edge
computing a
edge AI

Robotika

Zdroje

- <https://www.rcrwireless.com/20250404/fundamentals/tensor-processing-unit>
- <https://introl.com/blog/google-tpu-architecture-complete-guide-7-generations>
- <https://docs.cloud.google.com/tpu/docs/system-architecture-tpu-vm>
- <https://docs.cloud.google.com/tpu/docs/tpu7x>
- <https://www.ibm.com/think/topics/neural-processing-unit>
- <https://www.utmel.com/blog/categories/integrated%20circuit/neural-processing-unit-npu-explained>

Ďakujem za pozornosť