

SYRMIA

Novi trendovi u razvoju hardvera i softvera za
modele dubokog učenja

Syrmia LLC

18-Maj-2023.

Matematički fakultet u Beogradu

SYRMIA

Sadržaj današnje prezentacije:

1. Par reči o kompaniji
2. CPU vs GPU
3. Nova rešenja i novi pristupi u razvoju AI HW
4. Tenstorrent
5. AMD
6. NeuralMagic



- SYRMIA je kompanija u oblasti IT specijalizovana za sistemski softver
 - Postojimo od 2018.
 - Danas imamo 200+ zaposlenih, od kojih je 93% inženjerski kadar.
 - Poslujemo u dva domena:
 - 70% - Potrošačka elektronika
 - 30% - Automobilaska industrija
- Naš fokus je sistemski softver u različitim poljima, prvenstveno:
 - softver za automobilsku industriju (audio sistemi, optimizacije)
 - softver za namenske čipove za mrežne uređaje
 - programski prevodioci i alati (LLVM, GCC, Valgrind, v8, ...)
 - platforme za mašinsko učenje (Tensorflow, PyTorch, ...)
 - softver za grafičke kartice (virtualizacija, prevodioci za GPU, TrustedOS)



- Naučno-tehnološki park u Nišu

- ML, Deep Learning, NLP, programiranje u PyTorch-u, Python-u, C++
- U Syrmiji od 2021. kao machine learning inženjeri
- Tenstorrent tima iz Niša
 - Četiri člana
 - TensTorrent projekat
 - Testiranje AI hardvera, softvera i kompajlera
 - Istraživanje novih alata i rešenja (MLops), automatizacija procesa...



SYRMIA

Sadržaj današnje prezentacije:

1. Par reči o kompaniji

2. CPU vs GPU

3. Nova rešenja i novi pristupi u razvoju AI HW

4. Tenstorrent

5. AMD

6. NeuralMagic



Zašto koristimo GPU u Neuronskim mrežama?

Pogled iz perspektive hardvera

SYRMIA

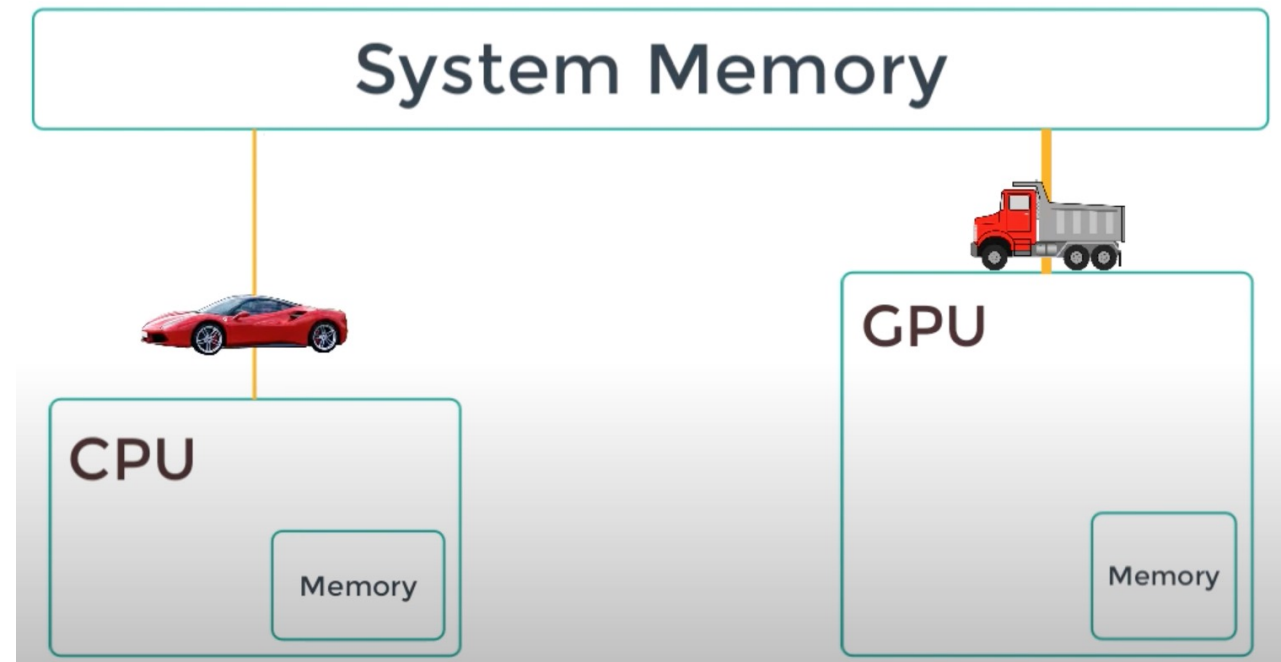
- Skalarna izračunavanja:

15 x 6

- Matrična izračunavanja:

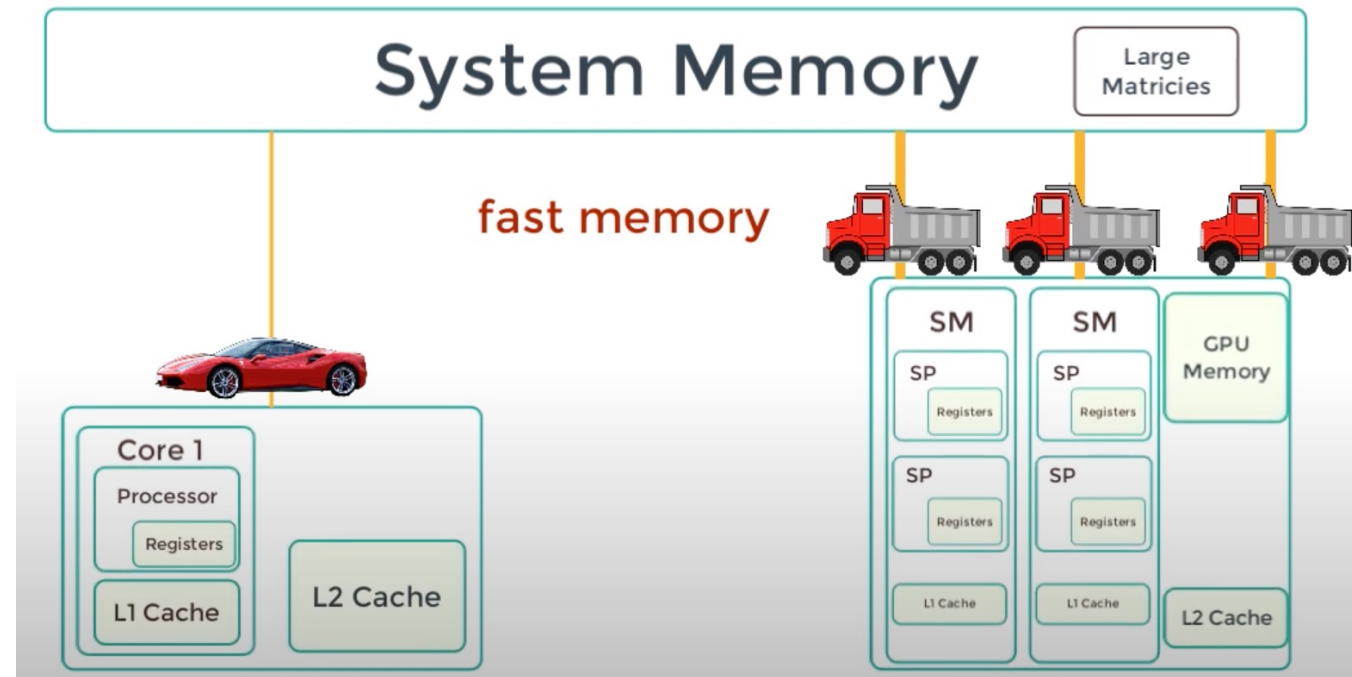
$$\begin{bmatrix} 15 & \dots & 6 \\ 21 & \dots & 8 \\ \vdots & \ddots & \vdots \\ 6 & \dots & 7 \end{bmatrix} \times \begin{bmatrix} 4 & \dots & 3 \\ -3 & \dots & 12 \\ \vdots & \ddots & \vdots \\ -6 & \dots & -9 \end{bmatrix}$$

- *Fetch operacija* - pribavljaju se podaci glavne iz memorije:
 - Brzi fetch – CPU
 - Sporiji fetch – GPU
- Zašto onda koristimo GPU?



Zašto koristimo GPU u Neuronskim mrežama?

- Slučaj matričnog množenja:
 - CPU mora da napravi hiljade pristupa memoriji kako bi sakupio sve potrebne podatke.
 - GPU može da pribavi više podataka odjednom, stoga nije neophodno da izvrši veliki broj fetch operacija.
- Šta se dešava dok GPU čeka na podatke?
 - Uvođenje paralelizacije!



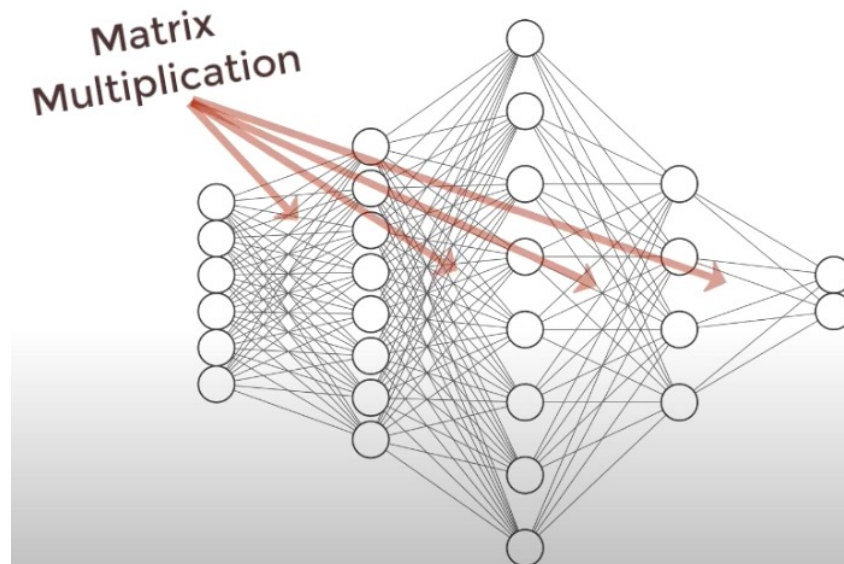
SP – Streamlined Procesor
SM – Streamlined Multiprocesori

Zašto koristimo GPU u Neuronskim mrežama?

Prednosti GPU kartica:

- **Veća memorijska propusnost** – može da pribavi veći broj podataka u jednom pristupu glavnoj memoriji.
- **Korišćenje paralelizacije** - veliki broj SM procesora koji poseduje veliki broj malih L1 keš memorija.
- **Veći broj pojedinačnih pristupa i brza memorija** – svaki SM procesor se sastoji iz nekoliko SP procesora pri čemu svaki SP dodatno poseduje registre.

Brža matrična množenja = brže neuronske mreže



Kako koristimo GPU?

- CUDA¹ – API koji omogućava da pristupimo komponentama GPU-a (memoriji, SP procesorima)
- Deep Learning radna okruženja - Pytorch, TF, ... (viši nivo apstrakcije, koristi se CUDA u pozadini)

¹ Compute Unified Device Architecture

- Da li neuronske mreže postaju prevelike?

- GPT3 (2020):

- 175 milijardi parametara koji se treniraju
- Trening koštao oko 10 miliona dolara

Llama - 7B, 13B, 33B i 65B

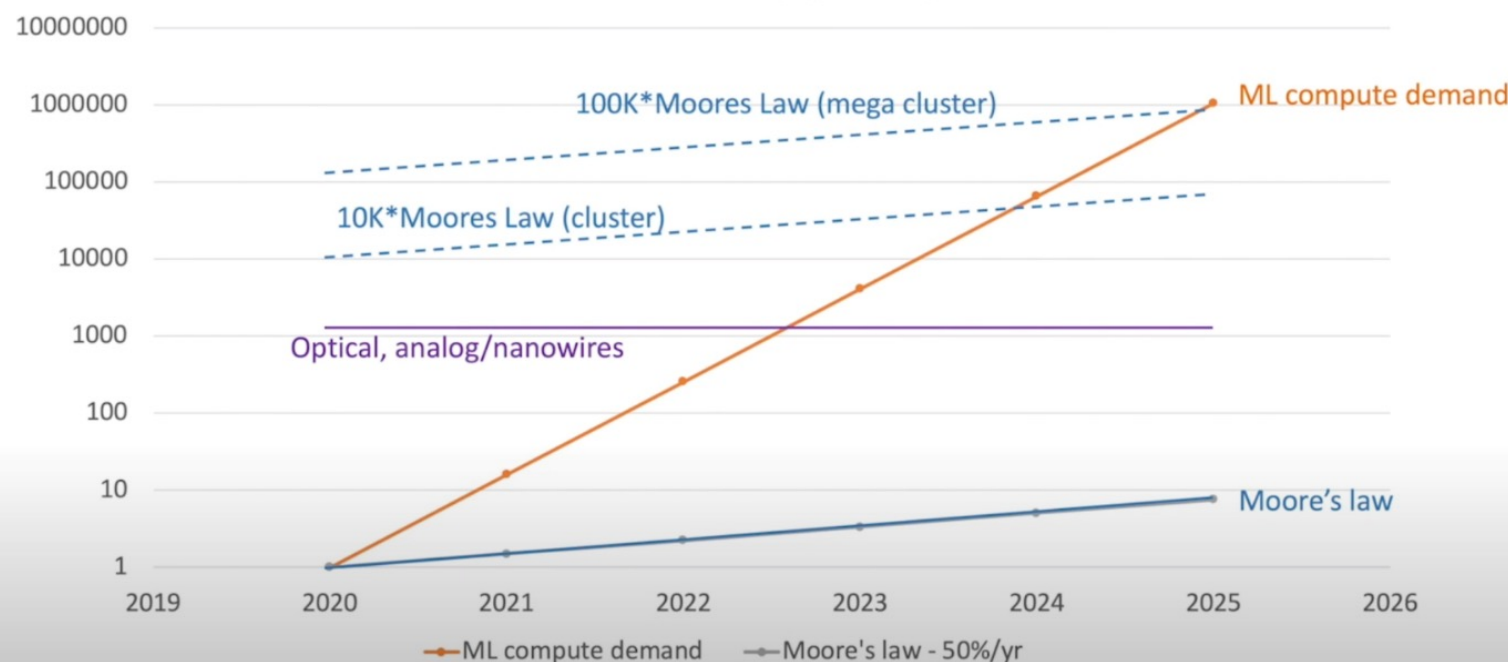
GPT-4 - 1000B

PanGu-Σ (Huawei) - 1085B

Stable Diffusion - trening koštao oko \$600 000

...

ML vs. Moore's Law (Optimistic)



- Potrebno je efikasno skaliranje za rešavanje većih AI problema
- Trka za pravljenje najmoćnijih AI mašina je već počela!

SYRMIA

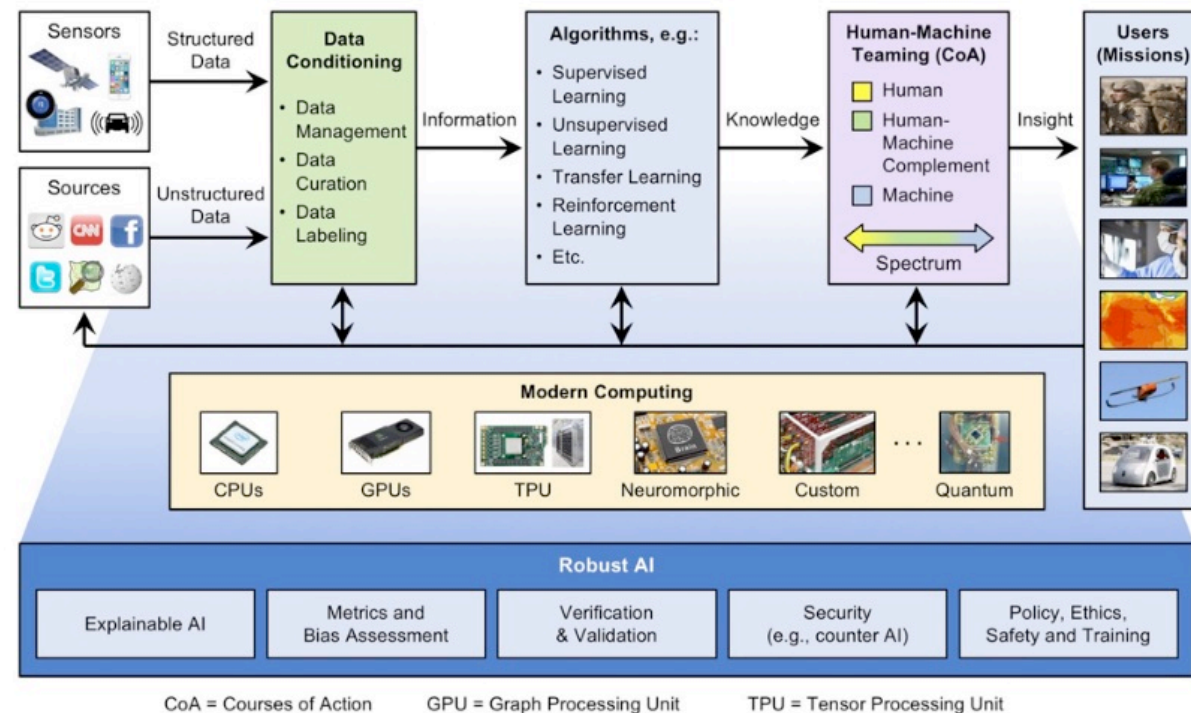
Sadržaj današnje prezentacije:

1. Par reči o kompaniji
2. CPU vs GPU
3. Nova rešenja i novi pristupi u razvoju AI HW
4. Tenstorrent
5. AMD
6. NeuralMagic



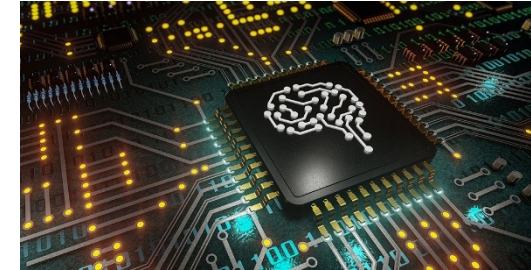
Kanonska AI arhitektura

- Pregled end-to-end AI rešenja i njihovih komponenti.
- Trendovi sistema na čipu (SoC) su prvi put uočeni u automobilskej industriji i u pametnim telefonima.
- Vojna industrija
- Marketing
- Medicina
- Robotika
- Finansije
- ...

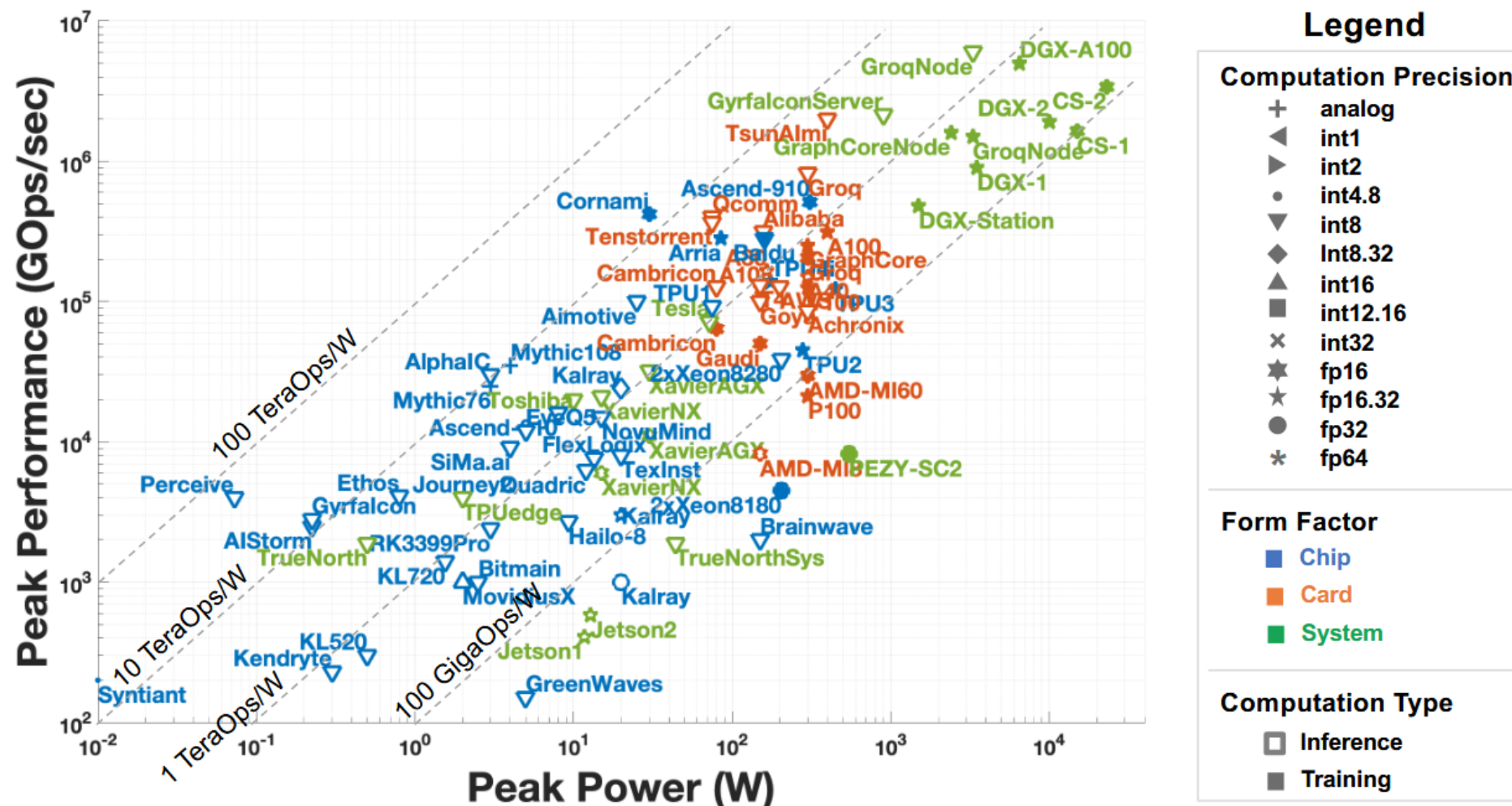


Kanonska AI arhitektura se sastoji od senzora, obrade podataka, algoritama, modernog računarstva, robusne veštačke inteligencije, interakcija između ljudi i mašina i korisnika (misije).

ML - Zašto nam trebaju grafički procesori i AI akceleratori?



- **Poboljšane performanse:** AI akceleratori su optimizovani za matrične operacije i paralelnu obradu, omogućavajući im da izvršavaju proračune mnogo brže od CPU-a ili GPU-a opšte namene.
- **Energetska efikasnost:** AI akceleratori su dizajnirani sa fokusom na energetska efikasnost, omogućavajući više proračuna po vatu u poređenju sa tradicionalnim procesorima.
- **Specijalizovana arhitektura:** AI akceleratori imaju arhitekture posebno prilagođene za AI zadatke. Oni koriste tehnike kao što su aritmetika smanjene preciznosti, specijalizovane memorijske hijerarhije...
- **Skalabilnost:** Oni nude opcije skalabilnosti koje omogućavaju organizacijama da ispune računarske zahteve svojih AI radnih opterećenja. Pogodni za edge uređaje, kao i AI u oblaku.
- **Zaključivanje u realnom vremenu:** AI akceleratori omogućavaju brže i efikasnije zaključivanje u realnom vremenu, što je ključno za aplikacije kao što su autonomna vozila, obrada prirodnog jezika, kompjuterski vid i robotika.



Dijagram maksimalnih performansi u odnosu na potrošnju energije javno objavljenih AI akceleratora i procesora [1].

- Potencijalno rešenje: Inteligentnija hardverska rešenja za obuku i zaključivanje dubokih mreža kao i mogućnost linearnog skaliranja takvih čipova unutar klastera
- Više od 50 kompanija proizvodi čipove posebno za AI sa procenjenih 9,9 milijardi dolara kapitala samo u 2021.
 - SambaNova Systems - 1,1 milijarda dolara
 - Cerebras Systems - 750 miliona dolara
 - Tenstorrent - 234 miliona dolara
 - SiMa - 200+ miliona dolara + X (2023. godine)
 - ...

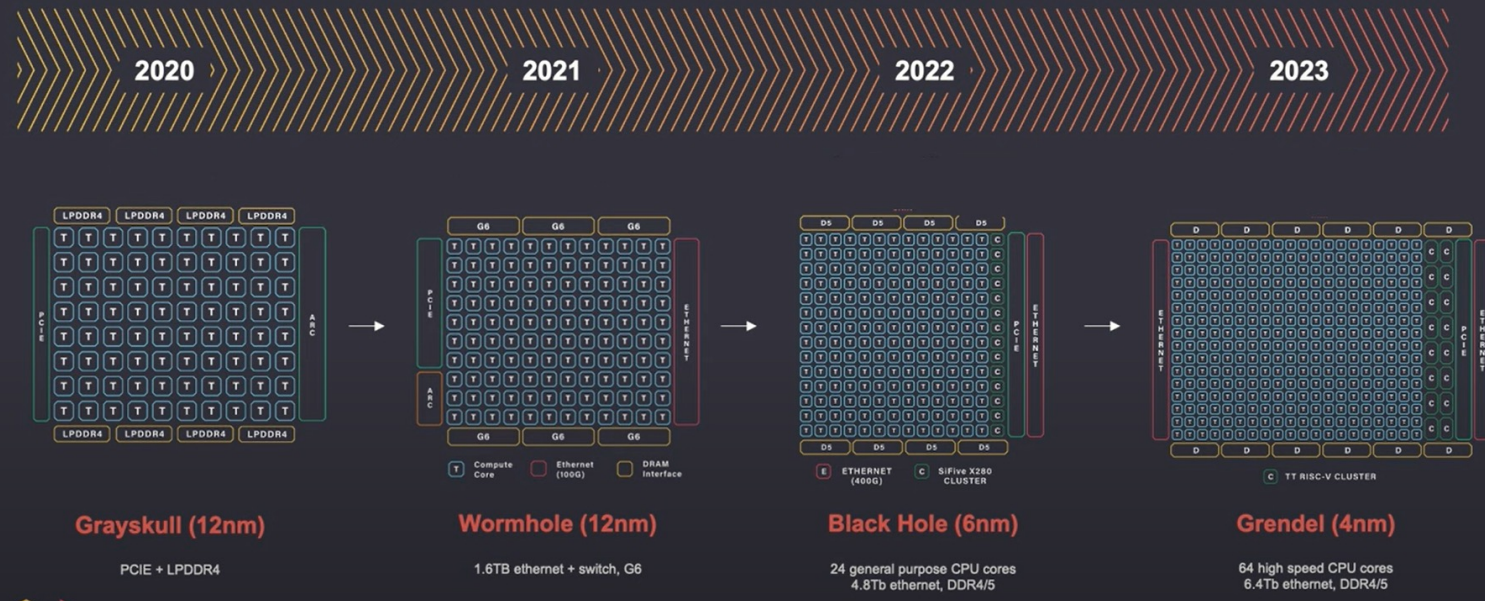


SYRMIA

Sadržaj današnje prezentacije:

1. Par reči o kompaniji
2. CPU vs GPU
3. Nova rešenja i novi pristupi u razvoju AI HW
4. Tenstorrent
5. AMD
6. NeuralMagic

Chip Roadmap



ML - Grafički procesori i heterogeni sistemi

Novi pristup

SYRMIA



- Pravi skalabilan i efikasan hardver za duboko učenje. Ciljevi:
 - Brži čipovi
 - Naprednije skaliranje paralelnih algoritama
 - Dinamičko izvršenje - smanjenje opterećenja izračunavanja
- Prva generacija AI procesora - napravljen je za pokretanje AI inference-a:
 - Pametni mehanizam za množenje matrica
 - Tenzorske manipulacije
 - Može da pokrene inference za bilo koju neuronsku mrežu - graf



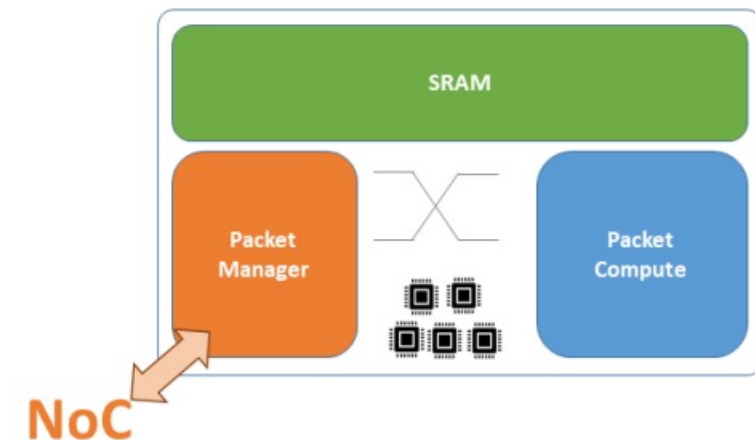
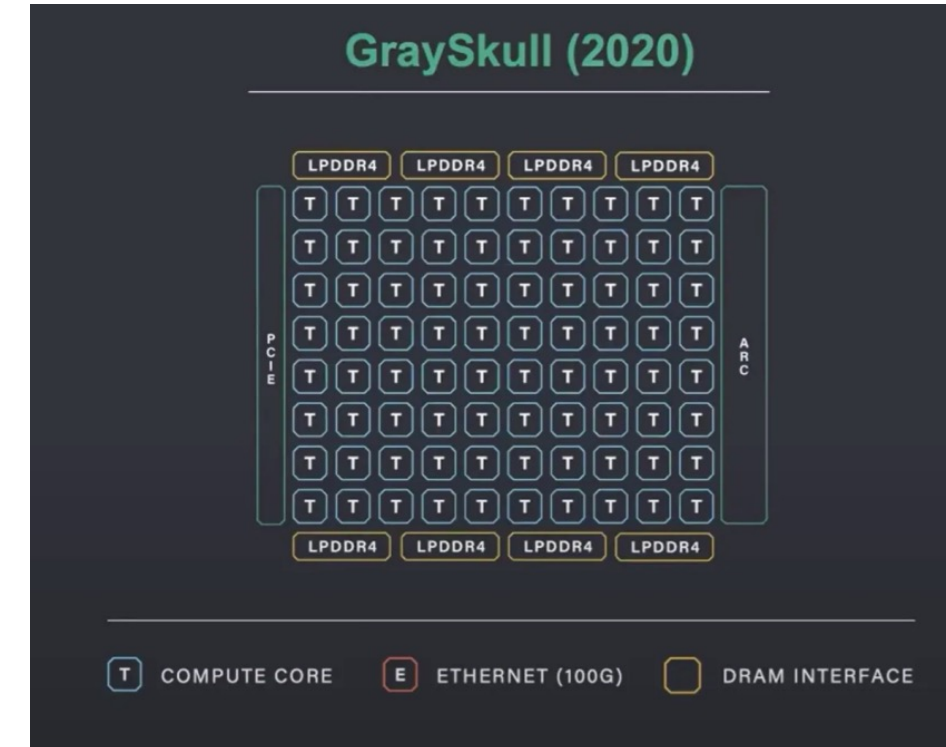
Grayskull™ AI processor

ML - Grafički procesori i heterogeni sistemi

Arhitektura čipa

SYRMIA

- Procesori se sastoje od grida *Tensix* core-ova.
- Mrežni komunikacioni hardver je prisutan u svakom procesoru i oni razgovaraju jedni sa drugima direktno preko mreže, umesto preko DRAM-a.
- Tenstorrentova procesorska jezgra su poznata kao Tensix jezgra.
 - Svako Tensix jezgro uključuje 5 RISC procesora
 - Array-Math jedinicu za tenzorske operacije
 - SIMD jedinicu za vektorske operacije
 - Hardver za ubrzavanje operacija mrežnih paketa i kompresiju/dekompresiju i manipulaciju tenzorima
 - Paketi sadrže instrukcije i delove tenzora modela
 - SRAM
- 120 Tensix core-a po čipu



Uvodjenje novih, inteligentnijih metoda za izvršavanje DL aplikacija:

- Dinamičko izvršenje - smanjuje izračunavanje potrebno za obuku i izvršenje NN
 - Uslovno izvršenje - rano okončanje inference-a
 - Dinamička retka izvršavanja (eng. sparsity)
 - Runtime kompresija - smanjenje veličine skupa podataka, transfera, potrošnje energije
 - Dinamička preciznost - procesiranje najmanjom preciznošću, fino podešavanje između više int i float formata tako da se dobije zadovoljavajuća tačnost.
- Google-ov BERT nakon primene uslovnog izvršenja i dinamičke preciznosti

Workload	Score
BERT BASE, SQuAD 1.1, fp16 – no conditionals	2,830
BERT BASE, SQuAD 1.1, fp16 + light conditional execution	10,150
BERT BASE, SQuAD 1.1, mixed precision, moderate conditional execution	23,345 *

* Work in progress, BERT model modified with conditional execution

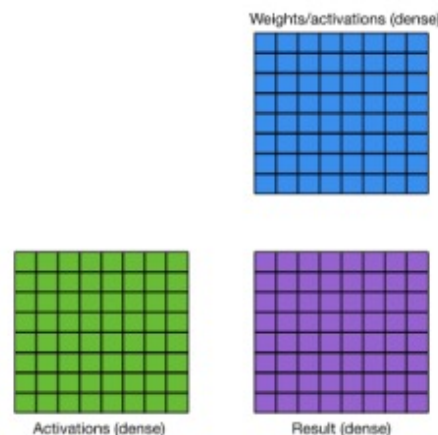
Source: Tenstorrent

ML - Grafički procesori i heterogeni sistemi

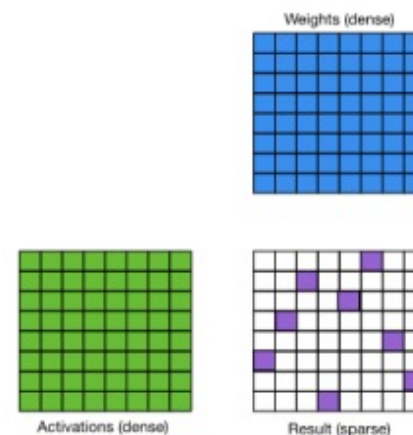
Dinamička retkost - Dynamic sparsity

SYRMIA

- Linearna poboljšanja performansi nisu dovoljna
- AI modeli će se povećati za šest redova veličine u narednih pet godina!
- Ideja: Ne vršiti izračunavanje onih stvari koje nisu potrebne - uštedeti na broju neophodnih izračunavanja
- Primer: dynamic sparsity prilikom množenja aktivacija i težina modela

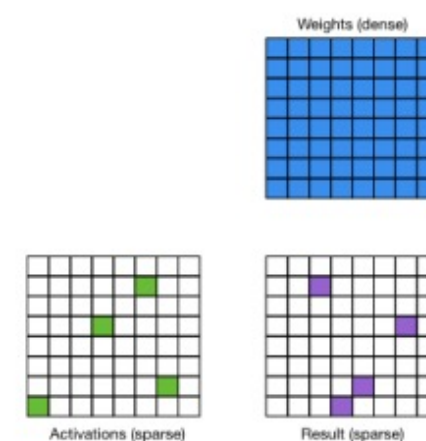


Dense: $O(n^3)$



Sparse: linear speedup

Sparsity	Max boost
50%	2X
90%	10X



Chained sparse MM: quadratic speedup

Sparsity	Max boost
50%	4X
90%	100X

Source: Tenstorrent

- Glavna ideja - Pametna obrada podataka, skaliranje i dinamičko izvršenje
- Kompajliranje ulaza svake neuronske mreže do nivoa optimizovanog grafa
- Implementacija dinamičke retkosti:
 - Podaci se prenose do sledećeg elementa obrade (**push**) umesto da budu izvučeni od strane procesnog elementa iz RAM-a (**pull** pristup).
 - Ako se podaci ne push-uju, onda se ne dešava obrada, što štedi računanje.
- Jezgra komuniciraju koristeći mrežni interfejs - ista tehnika će se koristiti za niz jezgara, niz čipova i niz rekova.
- Data format: bfloat16

SYRMIA

Sadržaj današnje prezentacije:

1. Par reči o kompaniji
2. CPU vs GPU
3. Nova rešenja i novi pristupi u razvoju AI HW
4. Tenstorrent
5. **AMD**
6. NeuralMagic



AMD ROCm

AMD-ov opensource ekosistem

SYRMIA

- ROCm - ekosistem za kreiranje inovativnih soft. aplikacija koje koriste prednost ubrzanog hardvera
- Nastao zvanično 2016 godine.
- Kreiran za širok spektar programera
- Ubrzanje za ML i High Performance Computing aplikacije
- ROCm sadrži program za otklanjanje grešaka, alate za analizu performansi, validaciju sistema i upravljanje sistemom...
- Cilj izvući maksimum iz hardvera koji imamo!

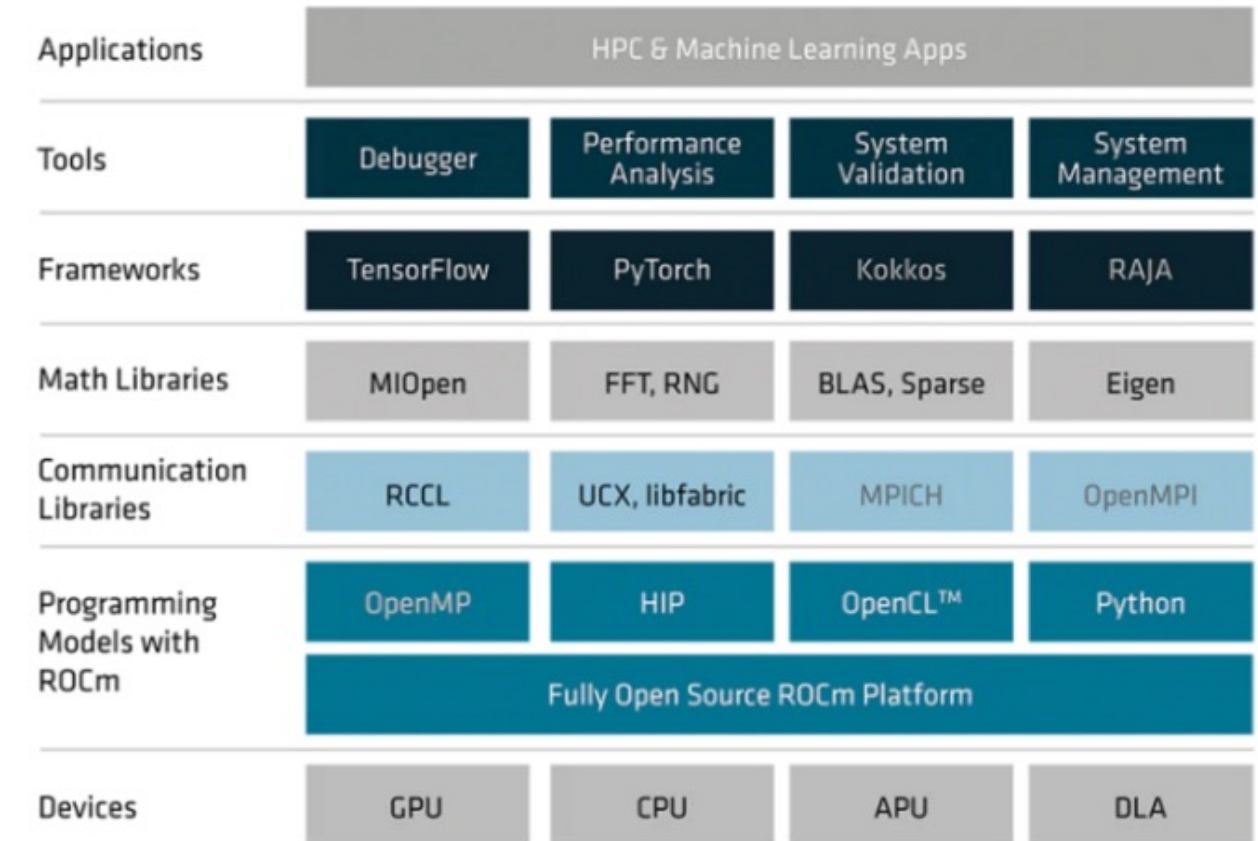
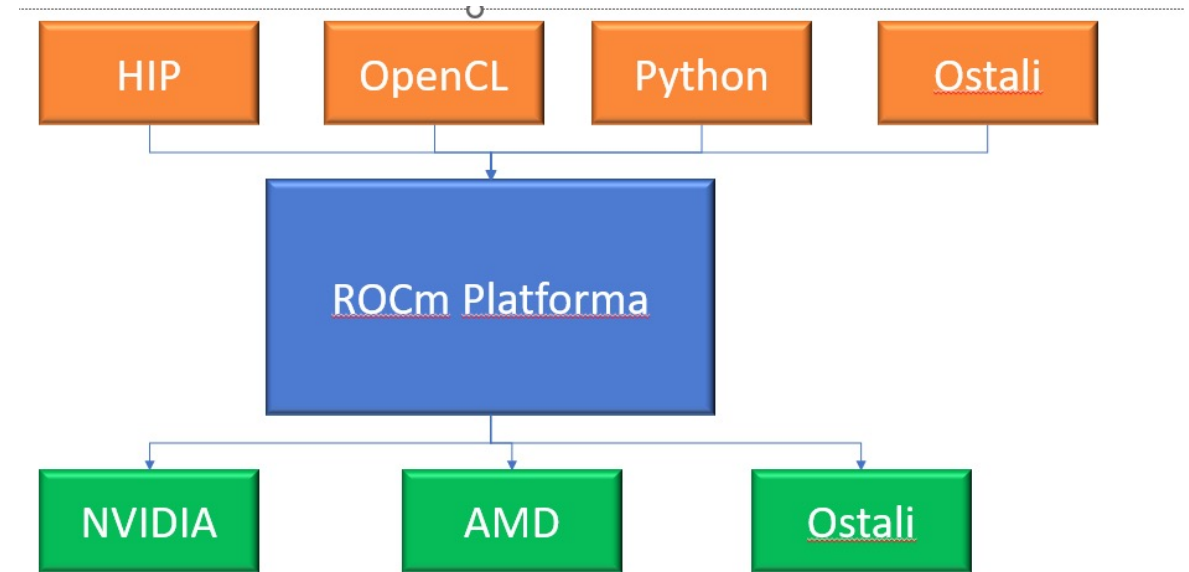


Figure 1 – ROCm Eco-system

- GPU kernel programiranje - HIP
 - Otvorena i prenosiva platforma za heterogeno računarstvo
 - moćan i fleksibilan API koji omogućava programeru da kreira aplikacije koje će raditi na **AMD-u i konkurentskim akceleratorima**
- Programiranje uz pomoć direktiva - podrška za OpenMP i Message Passing Interface
- OpenCL
- Python - razvoj ML

Ideja - pisanje različitih programa koji se mogu kompajlirati i izvršiti na različitim platformama



Pogled na ROCm platformu

- Podrška za različite programske jezike
- Portovanje na različite hardverske platforme

- AMD ROCm System Runtime je **nezavisan od jezika**, u velikoj meri koristi Heterogene Sistemske Arhitekture (HSA).
- ROCm napravljen za **skaliranje**.
- Podržava **multi GPU izračunavanja** (koristi Remote Direct Memory Access, RDMA, za komunikaciju unutar i van serverskih čvorova)
- AMD ROCm pruža developerima **fleksibilnost** prilikom izbora hardvera i pomaže prilikom razvoja računarski zahtevnih aplikacija
- ROCm **izbegava vendor lock-in**:
 - ML okviri (PyTorch, Tensorflow) se oslanjaju na različite CUDA implementacije
 - AMD HIP kao odgovor na problem

Najinovativnija komponenta ROCm ekosistema

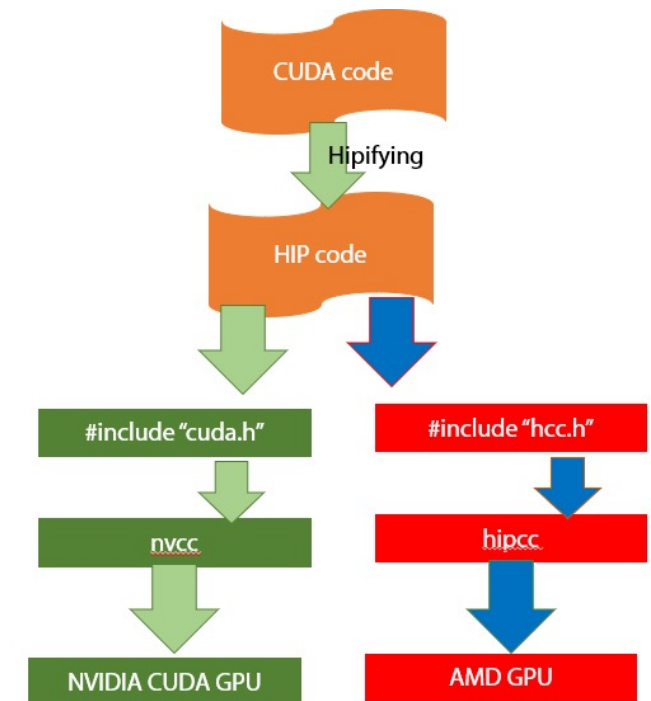
- GPU programsko okruženje za kreiranje kernela visokih performansi na GPU hardveru.
- HIP poseduje C++ runtime API
- Omogućava programerima da kreiraju **prenosive aplikacije** na različitim platformama.

Izbegnut *vendor lock-in*

Glavne karakteristike

- Veoma tanak sloj HIP-a, ne postoje penali po performanse sistema
- Pretvaranje Nvidia CUDA poziva u prenosivi C++ kod
- Izvorna podrška za AMD GPU-ove sa hipcc-om
- Omogućava kodiranje u single-source C++ programskom jeziku uključujući templejte iz C++11, lambda i drugo.

- HIP nije direktna zamena za CUDA, već teži da omogući konverziju CUDA koda u portabilni C++ kod
- Zamišljen je tako da programeri portuju iz CUDA u HIP, a zatim održavaju HIP verziju.
- Slična sintaksa CUDA i HIP programa.
- HIP pruža slične performanse kao *native* CUDA
- HIP API se ne mapira direktno na neku specifičnu verziju CUDA, već pruža podskup funkcionalnosti različitih verzija.
- Posедуje bogat skup alata i biblioteka
- Glavna prednost
 - Obezbeđuje da kod radi i na CUDA i AMD platformama
 - Redukuje troškove



SYRMIA

Sadržaj današnje prezentacije:

1. Par reči o kompaniji
2. CPU vs GPU
3. Nova rešenja i novi pristupi u razvoju AI HW
4. Tenstorrent
5. AMD
6. **NeuralMagic**



**NEURAL
MAGIC**

NeuralMagic koristi CPUs

- **Grafičke kartice nisu optimalne**

Zaključivanje mašinskog učenja evoluiralo je tokom godina vođeno napretkom GPU-a. GPU-ovi su brzi i moćni, ali mogu biti skupi, imaju kratak životni vek i zahtevaju mnogo električne energije.

- **CPU sami ne ispunjavaju standard**

CPU-i su fleksibilni u primeni i češće su dostupni. Ali oni su generalno zapostavljeni u svetu ML-a, zbog sporih performansi kako modeli rastu.

- **Kvantizacija i pruning**

- **DeepSparse biblioteka:**

- Koristi retkost da smanji broj operacija sa pokretnim zarezom.
- Koristi velike brze keš memorije CPU-a da pristupa referencama, izvršavajući mrežu po dubini i asinhrono.

SYRMIA

Reference

[1] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi and J. Kepner, "AI Accelerator Survey and Trends," *2021 IEEE High Performance Extreme Computing Conference (HPEC)*, Waltham, MA, USA, 2021, pp. 1-9, doi: 10.1109/HPEC49654.2021.9622867.

[2] Introduction to Quantization on PyTorch
<https://pytorch.org/blog/introduction-to-quantization-on-pytorch/>

[3] Pruning tutorial
https://pytorch.org/tutorials/intermediate/pruning_tutorial.html

[4] Model optimization
https://www.tensorflow.org/lite/performance/model_optimization

[5] Hsieh, Cheng-Yu, et al. "Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes." arXiv preprint arXiv:2305.02301 (2023).

SYRMIA

Kako se testiraju novi AI
akceleratori?

Pitanja?

Pridružite nam se!

Mail: branko.arsic@syrmia.com

Prijave za posao i program stručne prakse: jobs@syrmia.com

Tel. +381 11 4501 200

Web adresa: www.syrmia.com