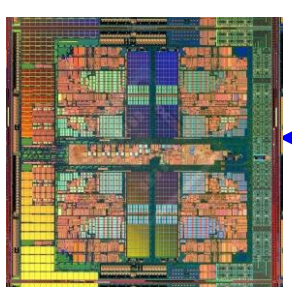
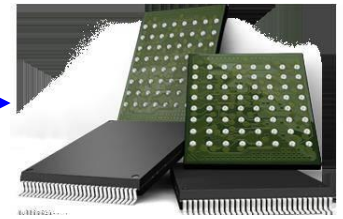
Računalna arhitektura

Osnovni pojmovi i ciljevi

Trenutno najaktivnija područja istraživanja u računalnim arhitekturama

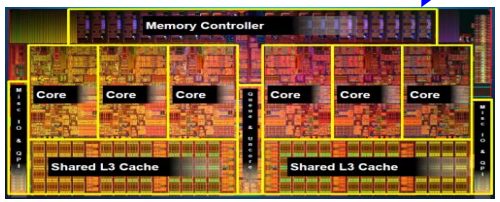


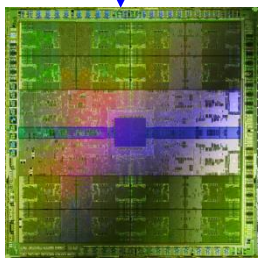
Hibridna memorija



Perzistentna memorija I pohrana

Heterogeni procesori I akceleratori





Grafičko procesiranje

Želimo izgraditi bolje arhitekture

Četiri glavna smjera razvoja

- Sigurne/pouzdane arhitekture
- Energetski efikasne arhitekture
- Arhitekture sa niskom latencijom, predvidljivih performansi
- Arhitekture za AI/ML, analizu genoma, medicinu

Što ćemo naučiti na ovom kolegiju?

Kako računala rade

(od “dolje” prema “gore”)

… I zašto nam je to bitno

Zašto imamo računala

Zašto koristimo računalnu znanost

Da rješimo probleme

Da dobijemo uvid u način rada računala

Hamming, "Numerical Methods for Scientists and Engineers," 1962.

Da omogućimo bolji život I budućnost

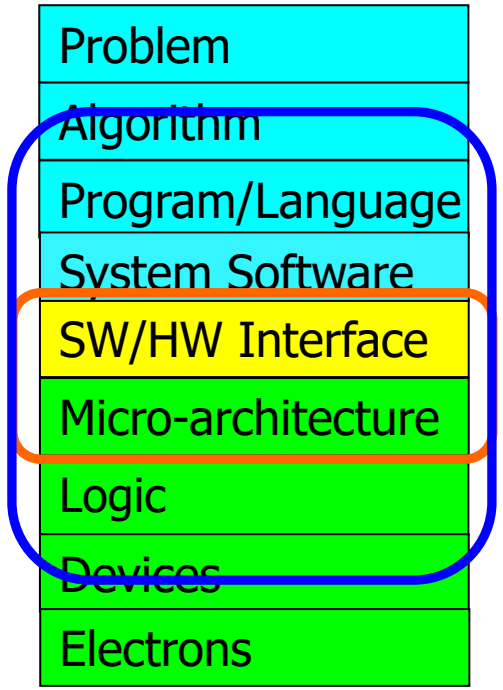
Kako računalo rješava probleme?

Kroz “orkestraciju” elektrona

Kako elektroni rješavaju probleme?

Transformacijska hijerarhija

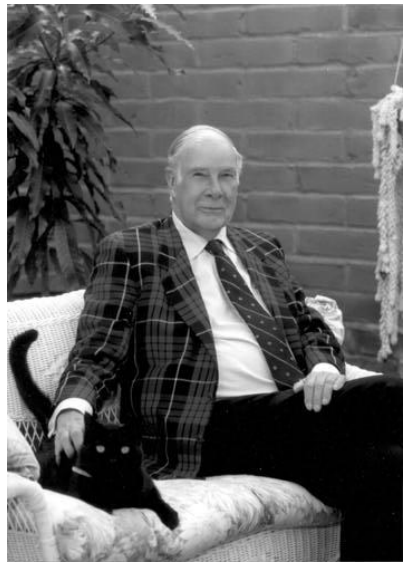
Arhitektura (prošireni pogled)



Arhitektura (“uski” pogled)

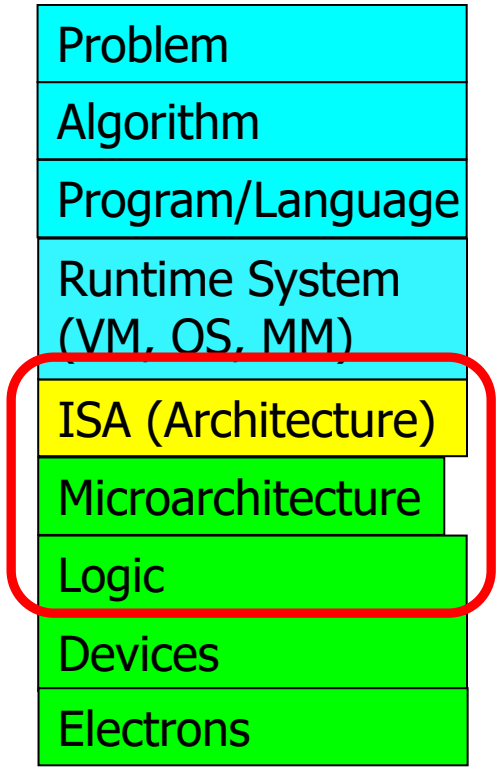
Nivoi transformacije

"The purpose of computing is [to gain] insight" (*Richard Hamming*)
We gain and generate insight by solving problems
How do we ensure problems are solved by electrons?



Algoritam

Korak-po-korak procedura koja sigurno ima svoj kraj, u kojoj je svaki korak precizno određen i računalo ga može odraditi.
Many algorithms for the same problem



ISA
(Instruction Set Architecture)

Sučelje/ugovor između softvera I hardvera. Ono za što programer pretpostavlja da hardver može odraditi.

Mikroarhitektura
Implementacija ISA

Digitalni logički krugovi
Gradivni blokovi arhitekture

Građa računala

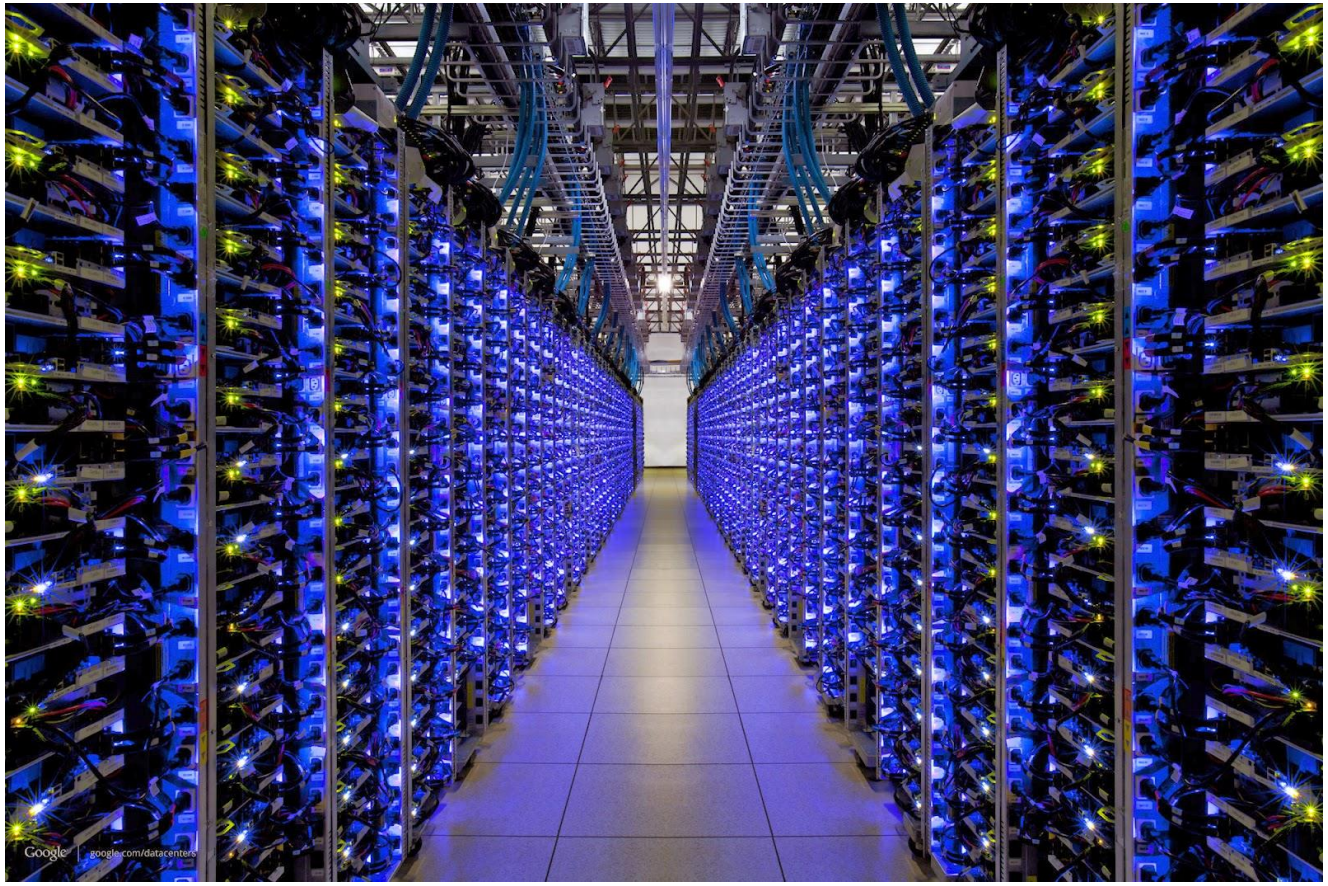
- Znanost i umjetnost dizajniranja računalnih platformi (hardvera, sučelja, sistemskog softvera, programerskog modela)
- Da bi se dostigao skup ciljeva dizajna
  - Npr. najviše performance za specifične zadatke
  - Npr. najduži vijek rada baterije
  - Npr. najbolje prosječne perfromanse u poznatim zadacima uz najbolji odnos cijene I performansi
  - …

  - Dizajn superračunala uključuje potpuno drugačije ciljeve dizajna u usporedbi sa pametnim telefonom (iako je dosta osnovnih pretpostavki slično)

Različite platforme, različiti ciljevi

Različite platforme, različiti ciljevi

Različite platforme, različiti ciljevi



Različite platforme, različiti ciljevi

Različite platforme, različiti ciljevi

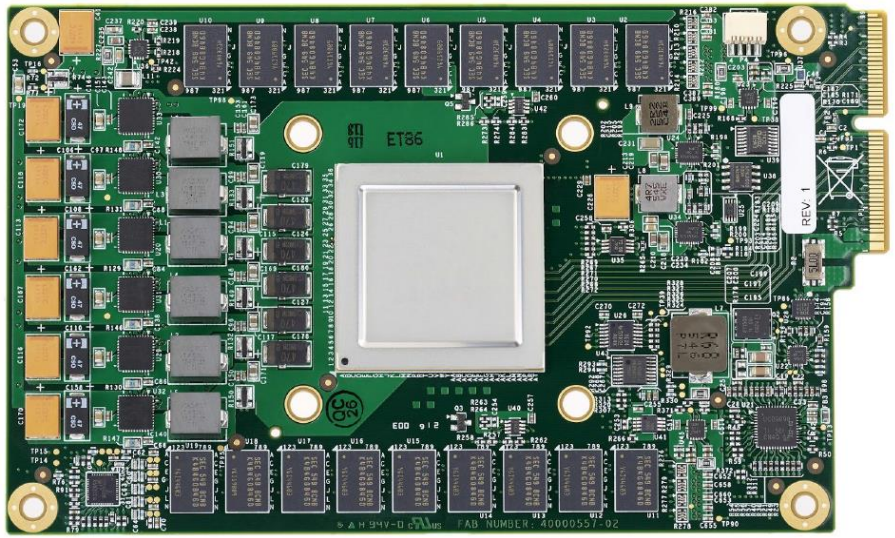


Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

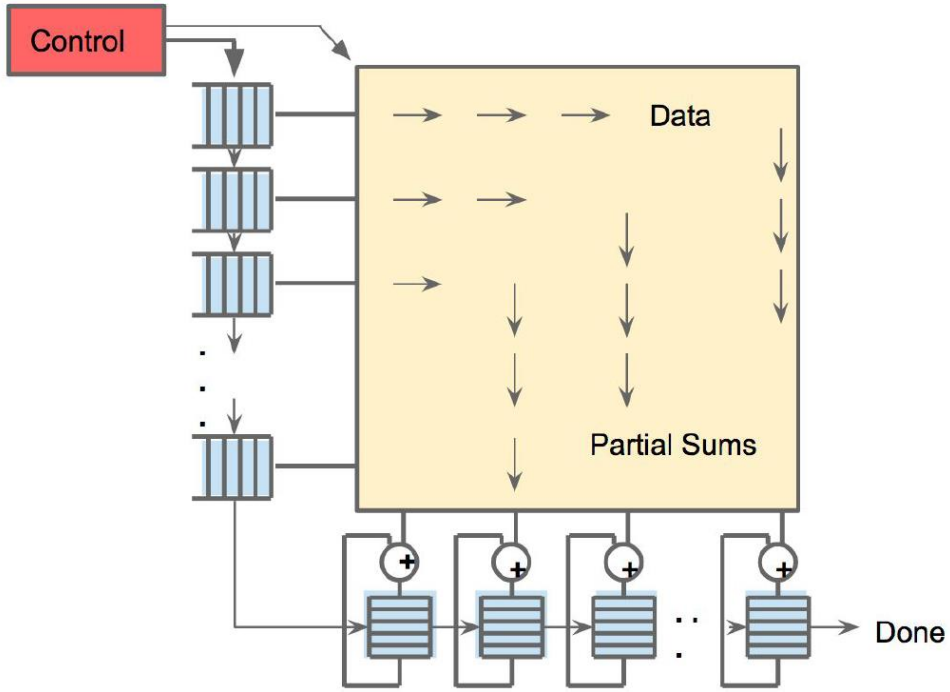
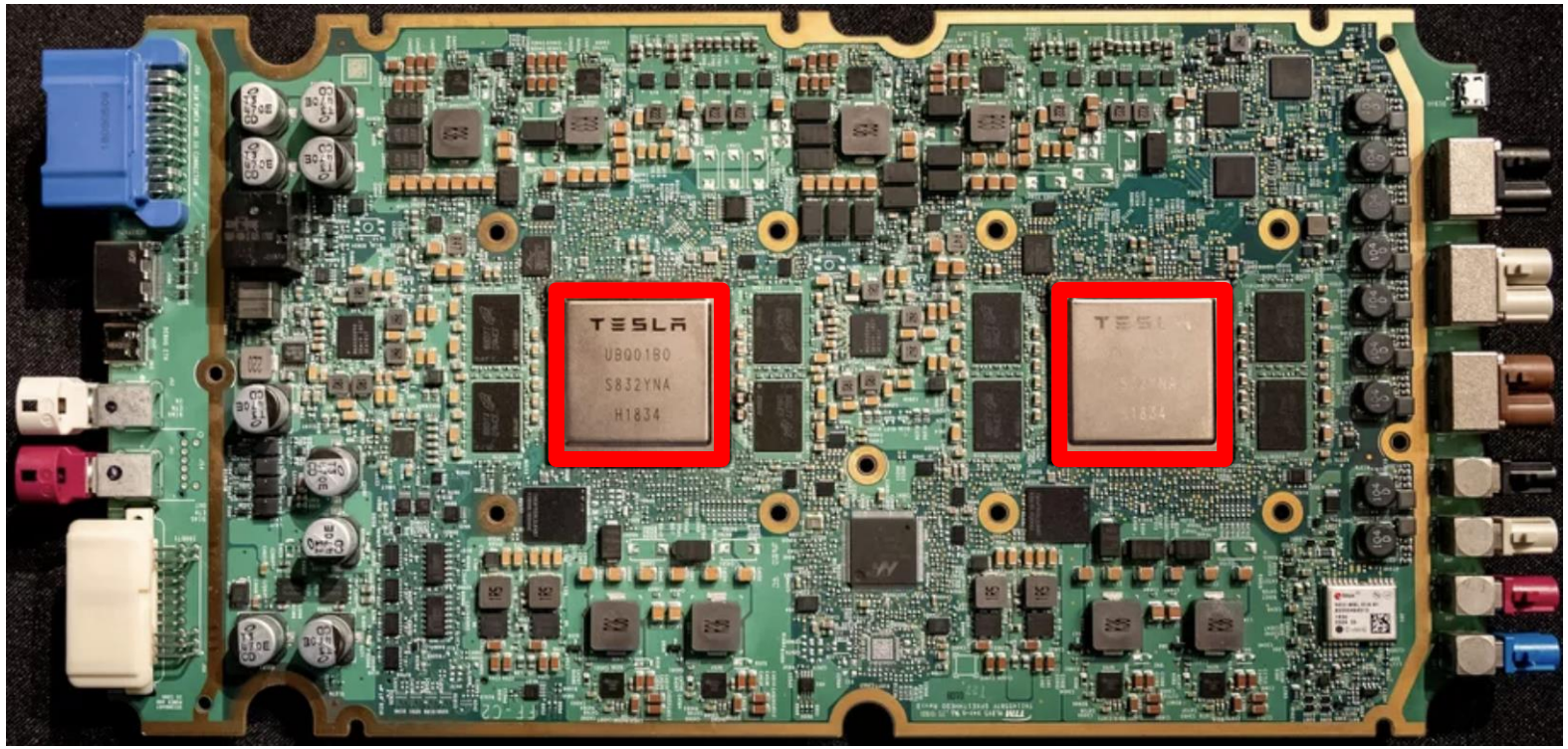


Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit", ISCA 2017.

Različite platforme, različiti ciljevi

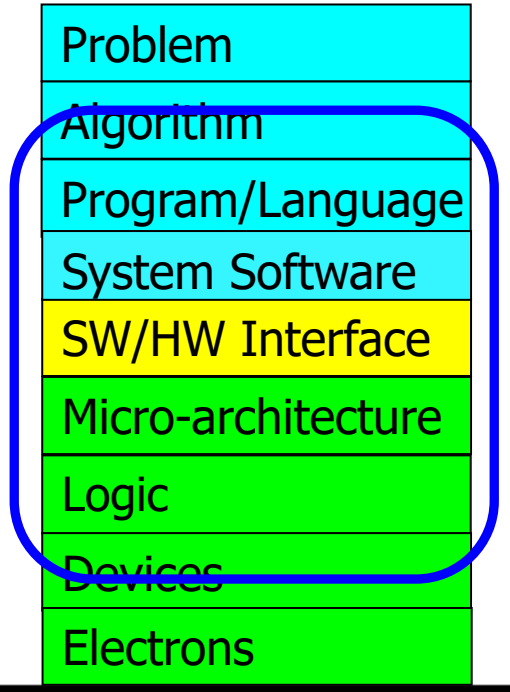
- ML accelerator: 260 mm^2, 6 billion transistors, 600 GFLOPS GPU, 12 ARM 2.2 GHz CPUs.
- Dva redundantna čipa radi sigurnosti.



Aksiom

Da bismo postigli najbolju energetsku efikasnost I performanse,

Moramo koristiti prošireni pogled na računalo.



Ko-dizajnirati kroz hijerarhiju algoritme prema uređajima

Specijalizirati koliko je moguće unutar ciljeva dizajna

Što je građa računala/računalna arhitektura?

- Znanost I umjetnost dizajniranja, odabiranja I povezivanja hardverskih komponenti I dizajniranje hardver/softver sučelja radi kreiranja računalnog sustava koji je u skladu sa zahtjevima – funkcionalnim, zahtjevima za performanse, potrošnje, cijene I sličnih ciljeva

Zašto treba upoznati građu računala?

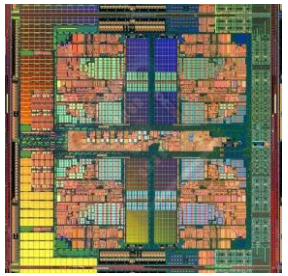
- Razvoj boljih sustava: bržih, jeftinijih, manjih, pouzdanijih
  - Korištenjem napretka i promjena u proizvodnim tehnologijama
- Razvoj boljih aplikacija
  - Gdje je 3D bio prije 20 godina? VR? Samovozeći automobile?
  - Analiza genoma? Personalizirana medicina?
- Razvoj boljih rješenja za probleme
  - Softverske inovacije su izgrađene na trendovima i promjenama u računalnoj arhitekturi
    - Moore-ov zakon o napretku performansi omogućava ovakav napredak
- Da razumijemo zašto računala rade kako rade

Računalna arhitektura danas

- Vrlo dobro vrijeme za učenje o računalnoj arhitekturi I građi računala
- Industrija je u fazi velikih promjena (nove arhitekture) – velika količina različitih potencijalnih dizajna
- Velika količina kompleksnih problema motiviraju, ali su I uzrokovane promjenama
  - Glad za podacima I aplikacijama koje intenzivno rade s podacima
  - Ograničenja u dizajnu po pitanju potrošnje, energije, zahtjeva za hlađenjem
  - Kompleksnost dizajna
  - Problemi u skaliranju tehnologije
  - Usko grlo memorije
  - Problemi sa pouzdanošću
  - Problemi sa programibilnošću
  - Problemi sa sigurnosti
- Na ova pitanja nemamo nužno jednoznačne, definitivne odgovore

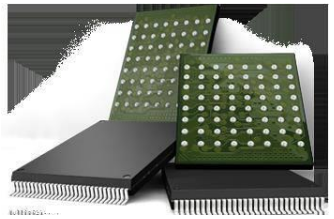
Računalna arhitektura danas

- Velika razlika u usporedbi sa 90-im ili 2000-im
- Aplikacije I tehnologije traže nove arhitekture

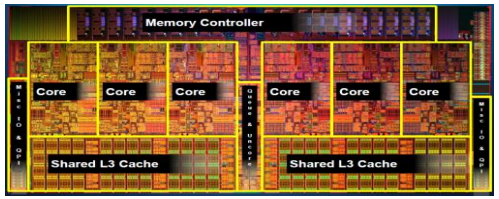


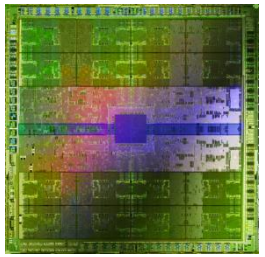
Heterogeni procesori I akceleratori

Hibridna memorija



Perzistentna memorija I pohrana





Grafičko procesiranje

Sve komponente I sučelja prolaze kroz fazu ponovnog pregleda

Iz povijesti: “Prilike na dnu”

There's Plenty of Room at the Bottom

From Wikipedia, the free encyclopedia

"**There's Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics**" was a lecture given by physicist Richard Feynman at the annual American Physical Society meeting at Caltech on December 29, 1959.[1] Feynman considered the possibility of direct manipulation of individual atoms as a more powerful form of synthetic chemistry than those used at the time. Although versions of the talk were reprinted in a few popular magazines, it went largely unnoticed and did not inspire the conceptual beginnings of the field. Beginning in the 1980s, nanotechnology advocates cited it to establish the scientific credibility of their work.

https://en.wikipedia.org/wiki/There%27s_Plenty_of_Room_at_the_Bottom

Važne postavke “prilika na dnu”

- Moore-ov zakon
- Kako “produljiti” vijek trajanja?
  - Proizvodnja manjih tranzistora – već sada su na razini “par atoma”
  - Pronalaženje materijala sa boljim svojstvima – npr. bakar umjesto aluminija, hafnium oksid I zrak za izolaciju, kompatibilnost materijala (veliki izazov)
  - Preciznijom proizvodnjom (ExtremeUV za <10nm nodove)
  - Kreiranjem novih tehnologija – FinFET, Gate All Around tranzistori, Single Electron transistor, …
- Zašto? Zato što smo često previše lijeni da mijenjamo “na vrhu”

Iz povijesti: “Prilike na vrhu”

REVIEW

There’s plenty of room at the Top: What will drive computer performance after Moore’s law?

Charles E. Leiserson[1], Neil C. Thompson[1,2,*], Joel S. Emer[1,3], Bradley C. Kuszmaul[1,†], Butler W. Lampson[1,4], ...

+ See all authors and affiliations

Science 05 Jun 2020:
Vol. 368, Issue 6495, eaam9744
DOI: 10.1126/science.aam9744

Much of the improvement in computer performance comes from decades of miniaturization of computer components, a trend that was foreseen by the Nobel Prize–winning physicist Richard Feynman in his 1959 address, “There’s Plenty of Room at the Bottom,” to the American Physical Society. In 1975, Intel founder Gordon Moore predicted the regularity of this miniaturization trend, now called Moore’s law, which, until recently, doubled the number of transistors on computer chips every 2 years.

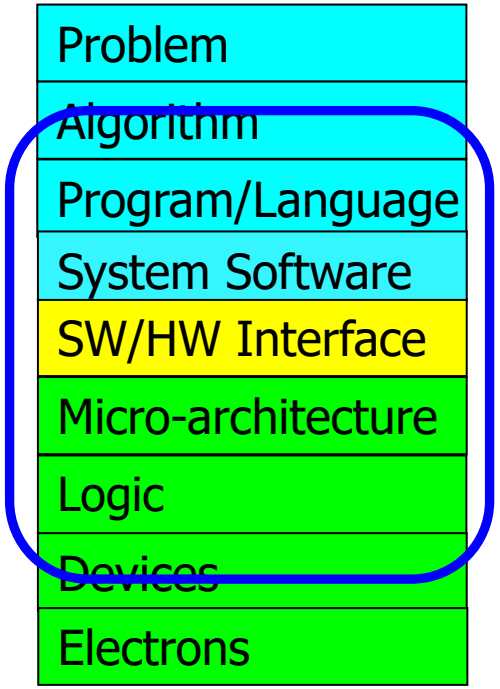
Unfortunately, semiconductor miniaturization is running out of steam as a viable way to grow computer performance—there isn’t much more room at the “Bottom.” If growth in computing power stalls, practically all industries will face challenges to their productivity. Nevertheless, opportunities for growth in computing performance will still be available, especially at the “Top” of the computing-technology stack: software, algorithms, and hardware architecture.

https://science.sciencemag.org/content/368/6495/eaam9744

Axiom, još jednom

Ima puno prostora I “na vrhu” I na dnu”, ali još više prostora I prilika imamo ako kvalitetno komuniciramo između “vrha” I dna” I ako optimiziramo kroz cijeli stack

Zato nam je bitan prošireni pogled



Jako puno je noviteta u razvoju računala

Jako puno je noviteta u razvoju računala

Performanse

I

Energetska efikasnost

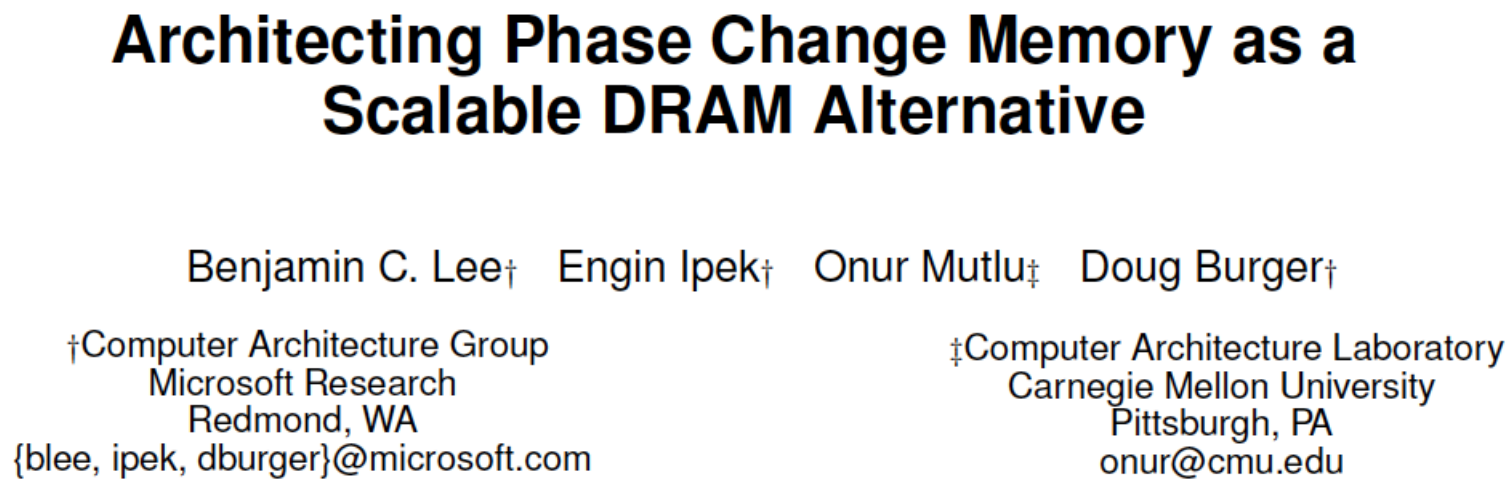
Intel Optane Persistent Memory (2019)

- Postojana podatkovna memorija
- Bazirana na 3D-XPoint tehnologiji

https://www.storagereview.com/intel_optane_dc_persistent_memory_module_pmm

PCM kao glavna memorija: Ideja iz 2009

- Benjamin C. Lee, Engin Ipek, Onur Mutlu, and Doug Burger, **"Architecting Phase Change Memory as a Scalable DRAM Alternative"** *Proceedings of the 36th International Symposium on Computer Architecture (**ISCA**)*, pages 2-13, Austin, TX, June 2009. Slides (pdf)



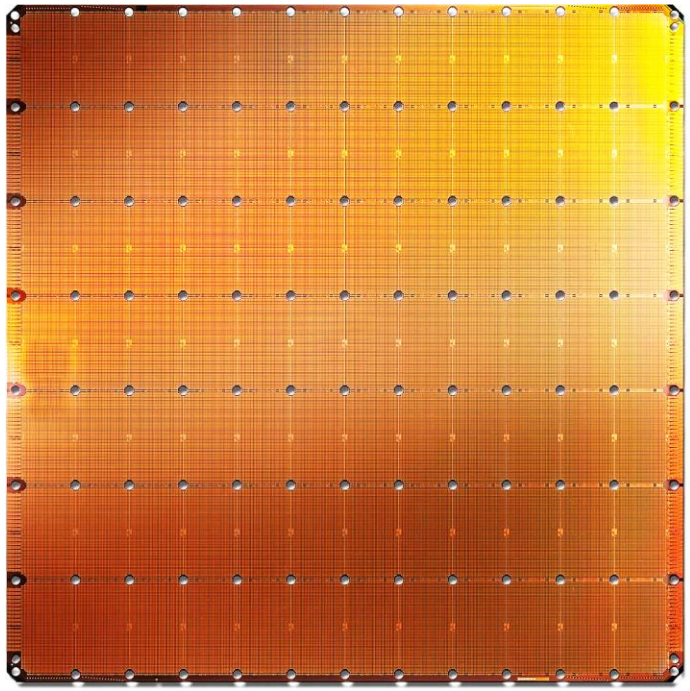
Architecting Phase Change Memory as a Scalable DRAM Alternative

Benjamin C. Lee† Engin Ipek† Onur Mutlu‡ Doug Burger†

†Computer Architecture Group
Microsoft Research
Redmond, WA
{blee, ipek, dburger}@microsoft.com

‡Computer Architecture Laboratory
Carnegie Mellon University
Pittsburgh, PA
onur@cmu.edu

Cerebras's Wafer Scale Engine (2019)



Cerebras WSE
1.2 Trillion transistors
46,225 mm^2

- Najveći ML akcelerator na čipu
- 400,000 jezgri

Largest GPU
21.1 Billion transistors
815 mm^2

NVIDIA TITAN V

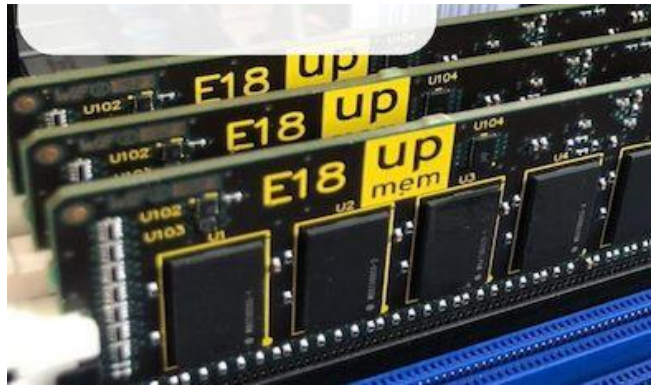
https://www.cerebras.net/cerebras-wafer-scale-engine-why-we-need-big-chips-for-deep-learning/

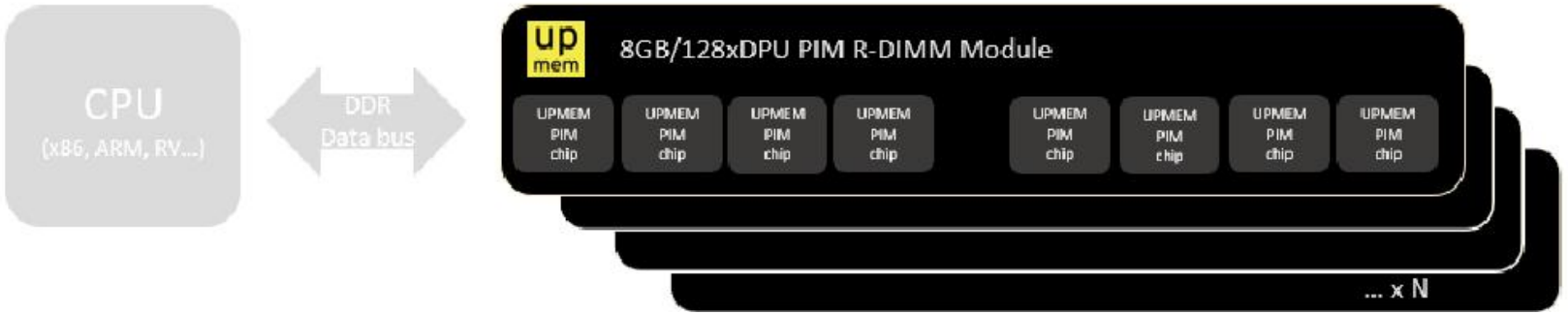
UPMEM Processing-in-DRAM Engine (2019)

- Procesiranje u DRAM-u
- Uključuje standardne DIMM module, sa velikim brojem

DPU procesora koji su kombinirani sa DRAM čipovima

- Zamjena za standardne DIMM-ove
  - DDR4 R-DIMM modules
    - 8GB+128 DPUs (16 PIM chips)
    - Standard 2x-nm DRAM process
  - **Ogromne rezerve** compute & memory bandwidtha

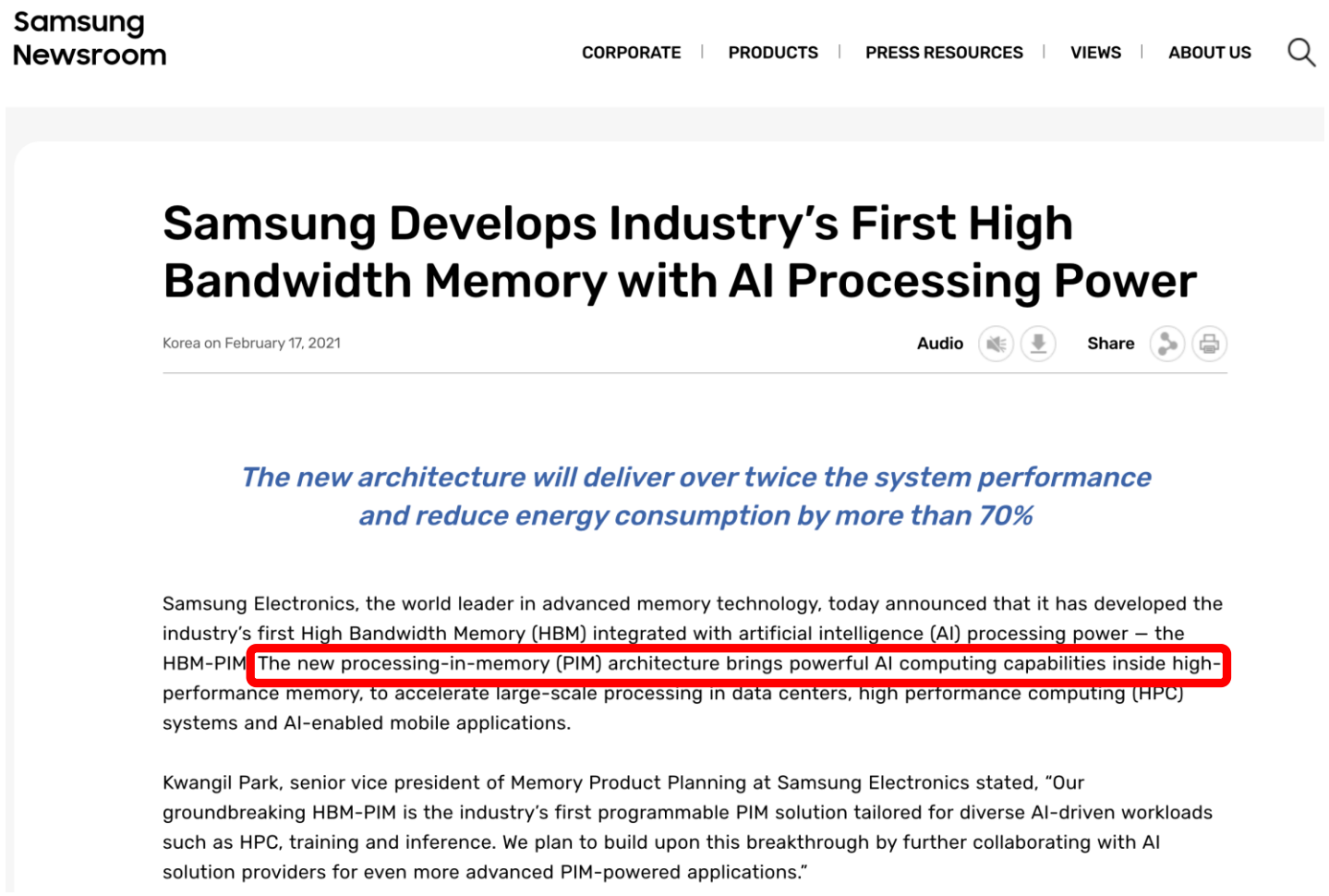




https://www.anandtech.com/show/14750/hot-chips-31-analysis-inmemory-processing-by-upmem

https://www.upmem.com/video-upmem-presenting-its-true-processing-in-memory-solution-hot-chips-2019/

Samsung Function-in-Memory DRAM (2021)



Samsung Newsroom

CORPORATE | PRODUCTS | PRESS RESOURCES | VIEWS | ABOUT US

Samsung Develops Industry's First High Bandwidth Memory with AI Processing Power

Korea on February 17, 2021 Audio Share

The new architecture will deliver over twice the system performance and reduce energy consumption by more than 70%

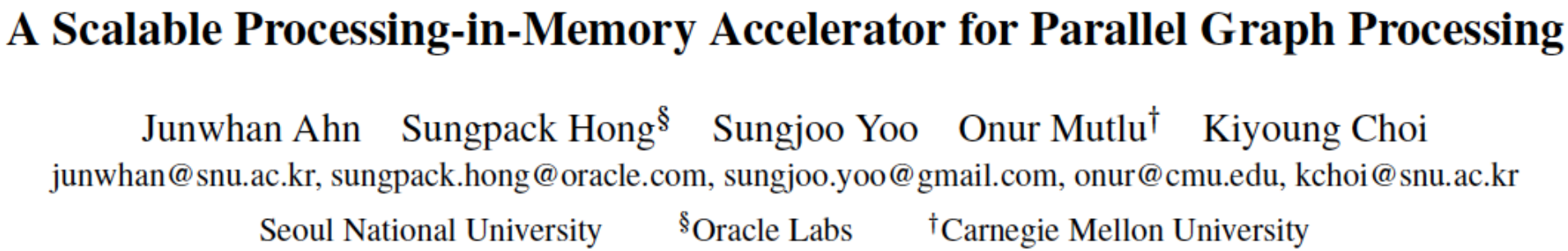
Samsung Electronics, the world leader in advanced memory technology, today announced that it has developed the industry's first High Bandwidth Memory (HBM) integrated with artificial intelligence (AI) processing power – the HBM-PIM. The new processing-in-memory (PIM) architecture brings powerful AI computing capabilities inside high-performance memory, to accelerate large-scale processing in data centers, high performance computing (HPC) systems and AI-enabled mobile applications.

Kwangil Park, senior vice president of Memory Product Planning at Samsung Electronics stated, "Our groundbreaking HBM-PIM is the industry's first programmable PIM solution tailored for diverse AI-driven workloads such as HPC, training and inference. We plan to build upon this breakthrough by further collaborating with AI solution providers for even more advanced PIM-powered applications."

https://news.samsung.com/global/samsung-develops-industrys-first-high-bandwidth-memory-with-ai-processing-power

Specialized Processing in Memory (2015)

- Junwhan Ahn, Sungpack Hong, Sungjoo Yoo, Onur Mutlu, and Kiyoung Choi,
  "A Scalable Processing-in-Memory Accelerator for Parallel Graph Processing"
  Proceedings of the 42nd International Symposium on Computer Architecture (***ISCA***), Portland, OR, June 2015.
  [Slides (pdf)] [Lightning Session Slides (pdf)]



A Scalable Processing-in-Memory Accelerator for Parallel Graph Processing

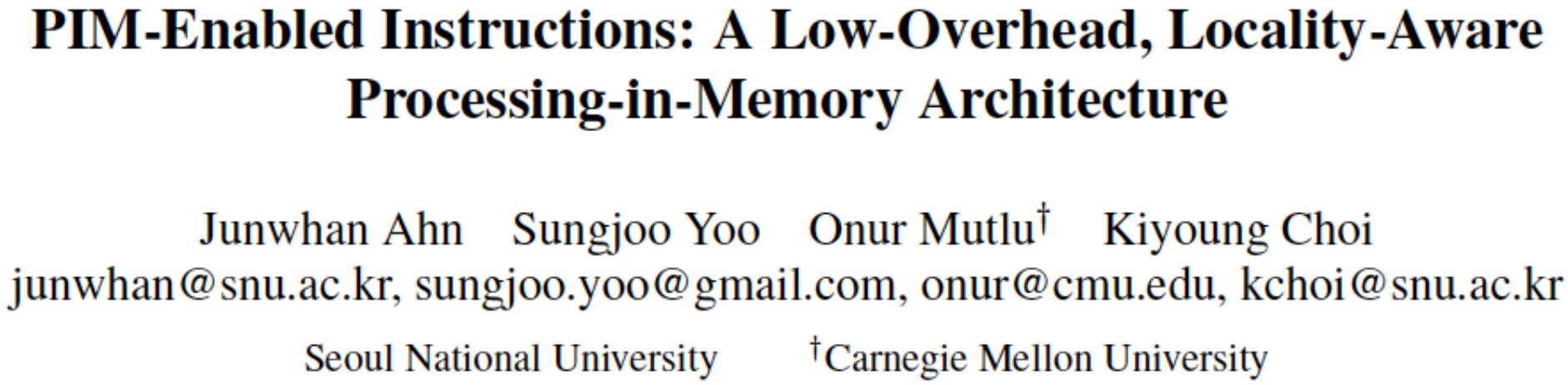
Junwhan Ahn Sungpack Hong[§] Sungjoo Yoo Onur Mutlu[†] Kiyoung Choi

junwhan@snu.ac.kr, sungpack.hong@oracle.com, sungjoo.yoo@gmail.com, onur@cmu.edu, kchoi@snu.ac.kr

Seoul National University [§]Oracle Labs [†]Carnegie Mellon University

Simple Processing in Memory (2015)

- Junwhan Ahn, Sungjoo Yoo, Onur Mutlu, and Kiyoung Choi, **"PIM-Enabled Instructions: A Low-Overhead, Locality-Aware Processing-in-Memory Architecture"**
  Proceedings of the 42nd International Symposium on Computer Architecture ***(ISCA)***, Portland, OR, June 2015.
  [Slides (pdf)] [Lightning Session Slides (pdf)]



PIM-Enabled Instructions: A Low-Overhead, Locality-Aware Processing-in-Memory Architecture

Junwhan Ahn Sungjoo Yoo Onur Mutlu† Kiyoung Choi
junwhan@snu.ac.kr, sungjoo.yoo@gmail.com, onur@cmu.edu, kchoi@snu.ac.kr
Seoul National University †Carnegie Mellon University

Processing in Memory on Mobile Devices

- Amirali Boroumand, Saugata Ghose, Youngsok Kim, Rachata Ausavarungnirun, Eric Shiu, Rahul Thakur, Daehyun Kim, Aki Kuusela, Allan Knies, Parthasarathy Ranganathan, and Onur Mutlu,
  "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks"
  *Proceedings of the 23rd International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS**)*, Williamsburg, VA, USA, March 2018


Google Workloads for Consumer Devices:
Mitigating Data Movement Bottlenecks

Amirali Boroumand[1] Saugata Ghose[1] Youngsok Kim[2]
Rachata Ausavarungnirun[1] Eric Shiu[3] Rahul Thakur[3] Daehyun Kim[4,3]
Aki Kuusela[3] Allan Knies[3] Parthasarathy Ranganathan[3] Onur Mutlu[5,1]

In-DRAM Processing (2013)

- Vivek Seshadri et al., "**Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology**," MICRO 2017.

Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology

Vivek Seshadri[1,5] Donghyuk Lee[2,5] Thomas Mullins[3,5] Hasan Hassan[4] Amirali Boroumand[5]
Jeremie Kim[4,5] Michael A. Kozuch[3] Onur Mutlu[4,5] Phillip B. Gibbons[5] Todd C. Mowry[5]

[1]Microsoft Research India [2]NVIDIA Research [3]Intel [4]ETH Zürich [5]Carnegie Mellon University

Google TPU Generation I (~2016)

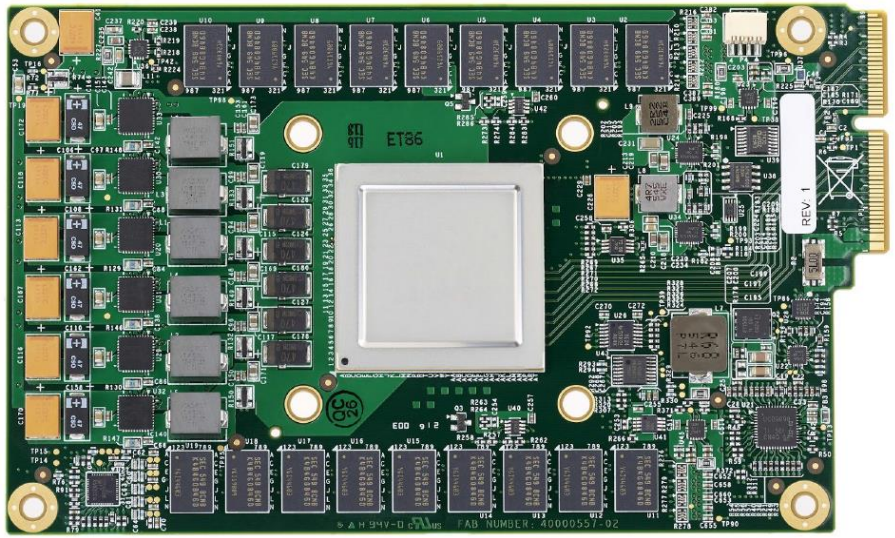


Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

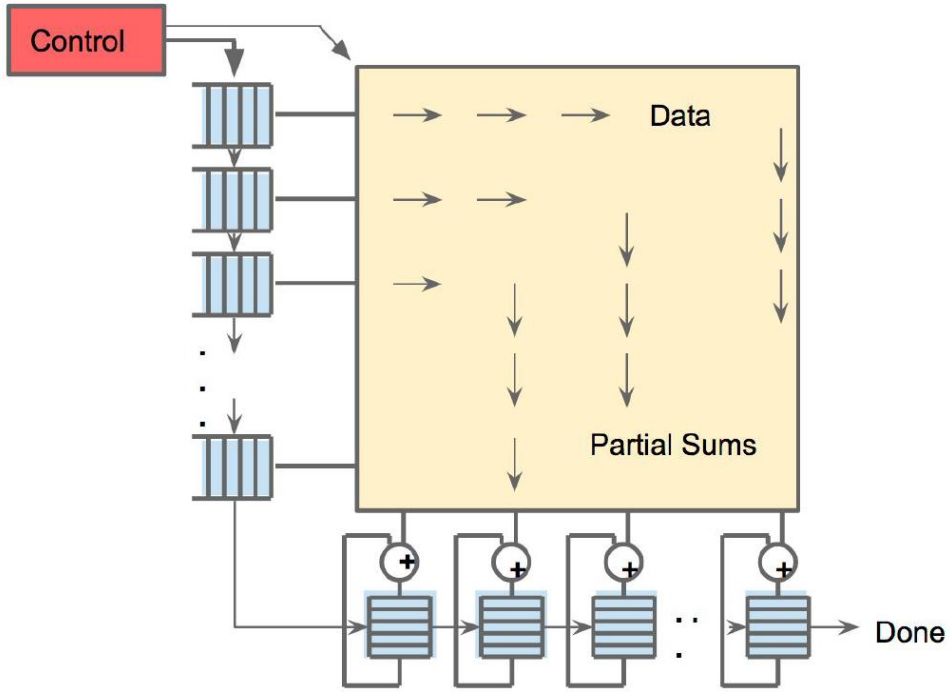
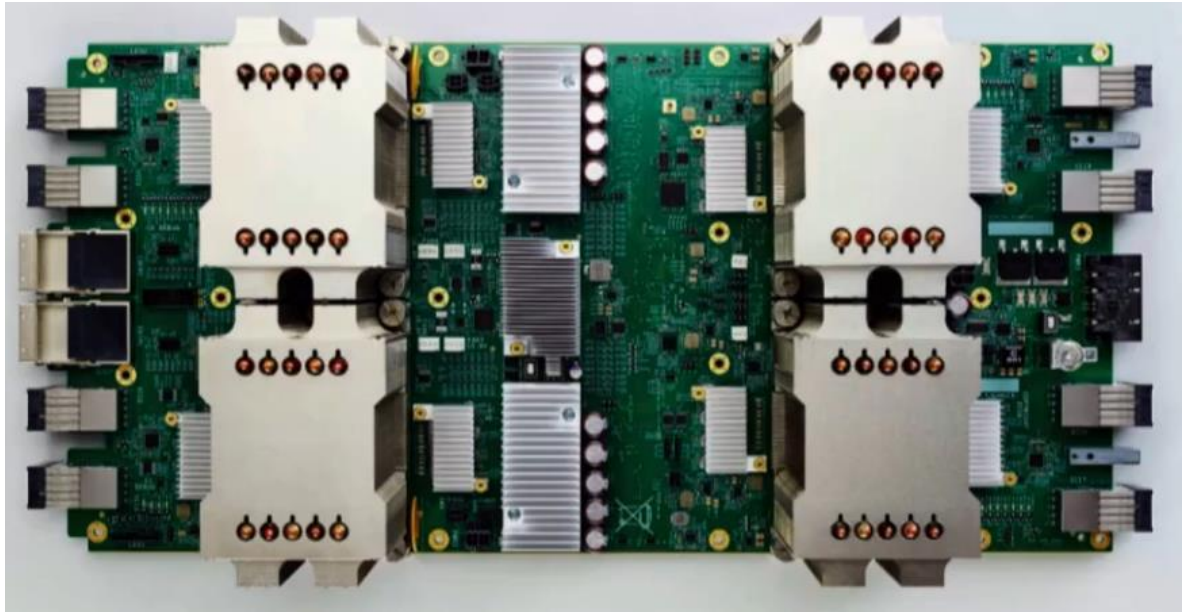


Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit", ISCA 2017.

Google TPU Generation II (2017)



https://www.nextplatform.com/2017/05/17/first-depth-look-googles-new-second-generation-tpu/

4 TPU chips
vs 1 chip in TPU1

High Bandwidth Memory
vs DDR3

Floating point operations
vs FP16

45 TFLOPS per chip
vs 23 TOPS

Designed for training
and inference
vs only inference

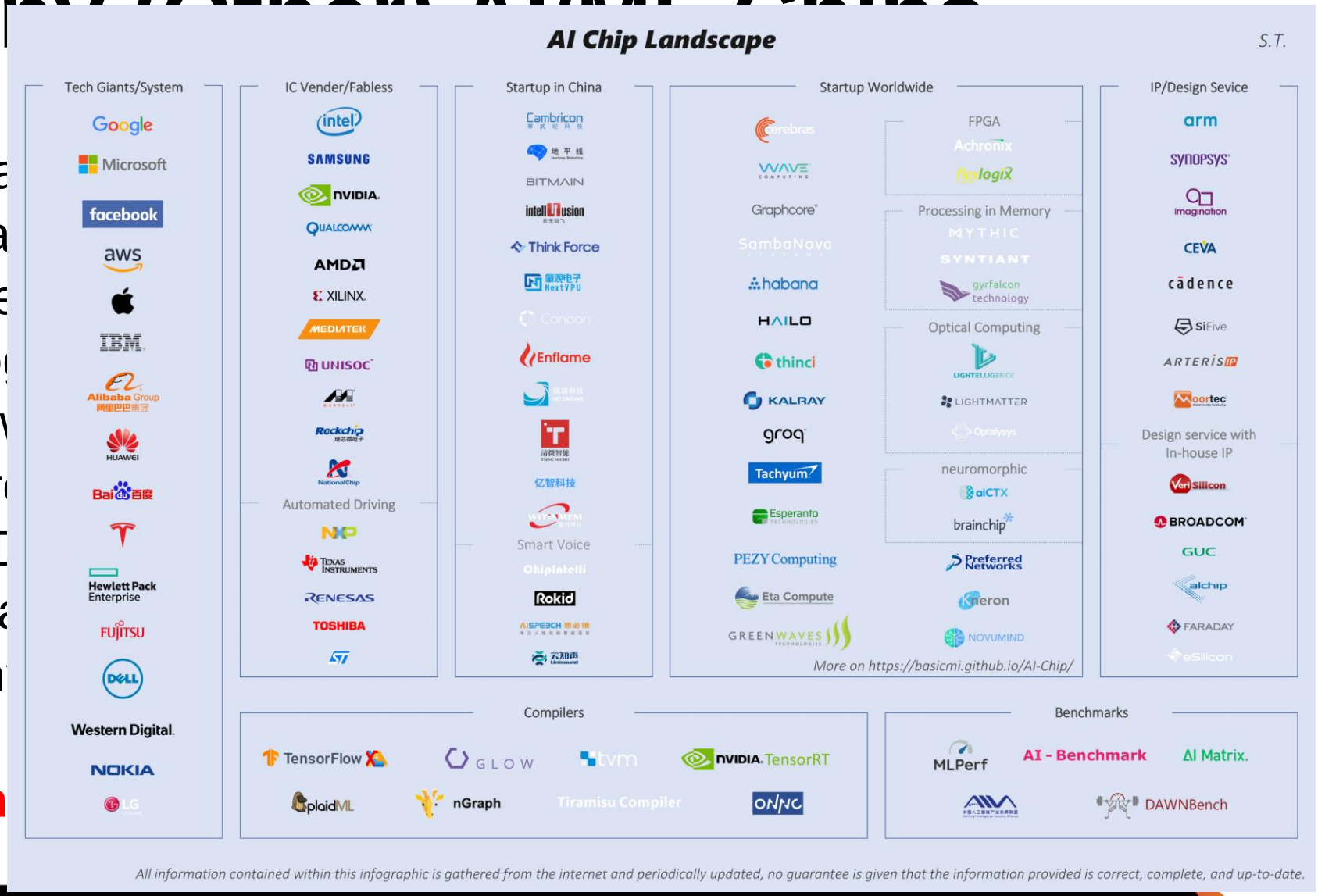
Ogromna količina AI/ML čipova drugih tvrtki

- Alibaba
- Amazon
- Facebook
- Google
- Huawei
- Intel
- Microsoft
- NVIDIA
- Tesla

- **I još puno drugih koji tek dolaze…**

Many (Other) AI/ML Chips

- Aliba
- Ama
- Face
- Goo
- Huav
- Micr
- NVID
- Tesl
- Man
- **Man**



https://basicmi.github.io/AI-Chip/

Puno zanimljivosti u građi računala danas

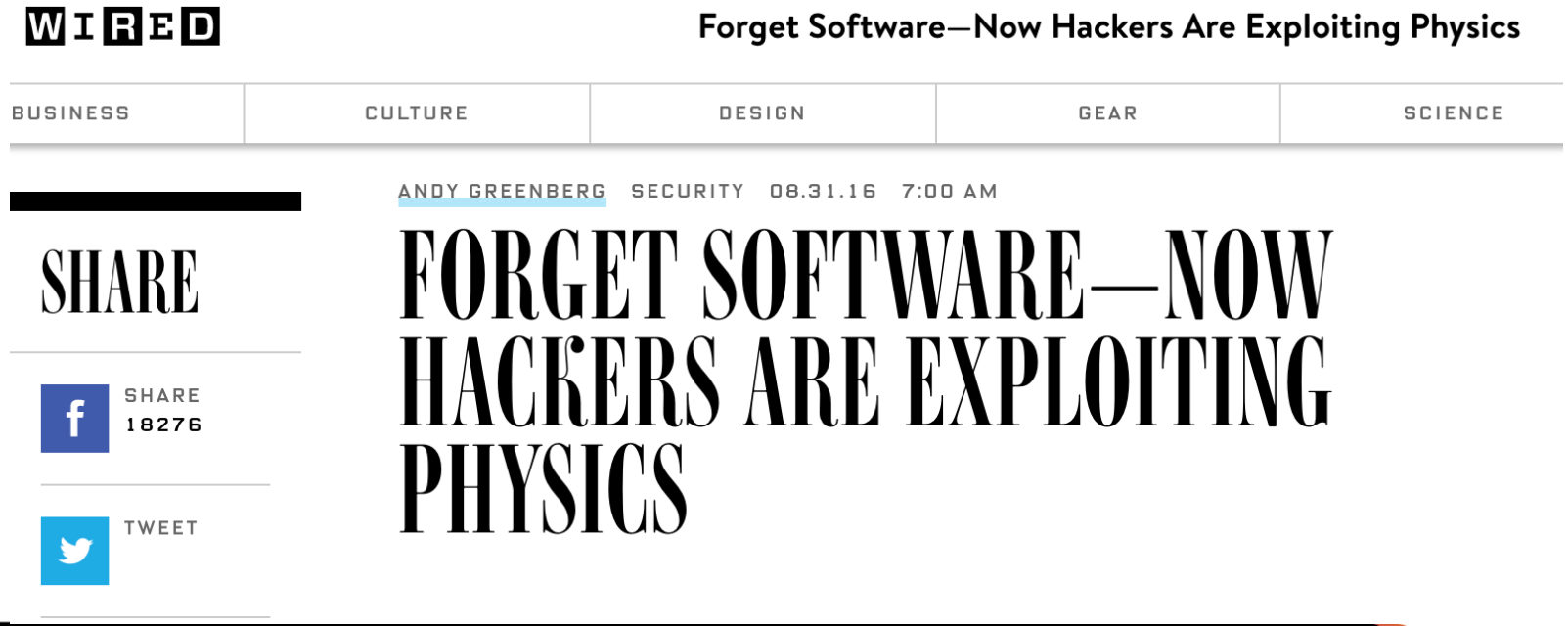
Puno zanimljivosti u građi računala danas

Pouzdanost I sigurnost

Sigurnost: RowHammer (2014)

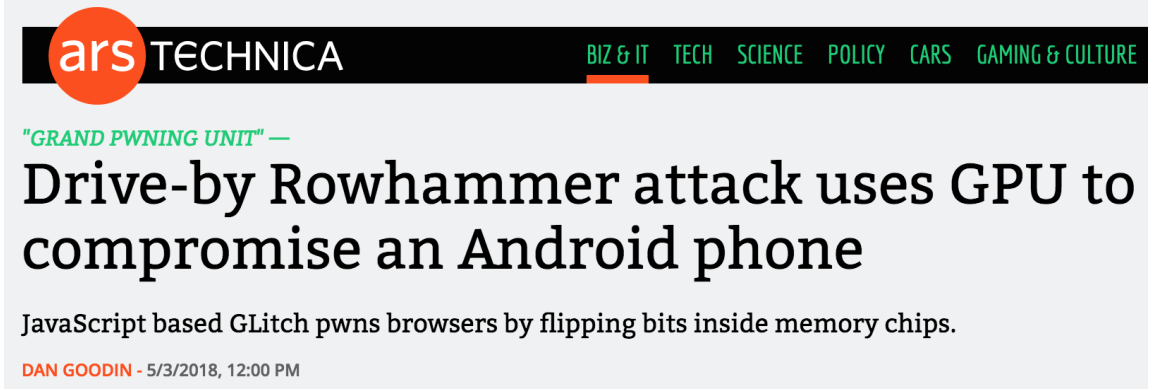
Priča o RowHammer napadu

- Predvidiva promjena bitova u uobičajenim DRAM čipovima
  - > 80% testiranih DRAM čipova je “ranjiva”
- Jedan u nizu primjera kako jednostavan hardverski problem može značiti veliku ranjivost na razini cijelog sustava



Još sigurnosnih problema

- Korištenje integriranog GPU-a u mobilnom sustavu za eskalaciju privilegija kroz WebGL sučelje



ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE

"GRAND PWNING UNIT" —

Drive-by Rowhammer attack uses GPU to compromise an Android phone

JavaScript based GLitch pwns browsers by flipping bits inside memory chips.

DAN GOODIN - 5/3/2018, 12:00 PM

Grand Pwning Unit: Accelerating Microarchitectural Attacks with the GPU

Pietro Frigo
Vrije Universiteit
Amsterdam
p.frigo@vu.nl

Cristiano Giuffrida
Vrije Universiteit
Amsterdam
giuffrida@cs.vu.nl

Herbert Bos
Vrije Universiteit
Amsterdam
herbertb@cs.vu.nl

Kaveh Razavi
Vrije Universiteit
Amsterdam
kaveh@cs.vu.nl

Sigurnost: Meltdown i Spectre (2018)

Meltdown i Spectre

- Netko nam može ukrasti tajne podatke sa sustava, iako:
  - Aplikacije I podaci rade korektno
  - Hardver radi prema specifikaciji
  - Nemamo nikakvih problema sa softverskim ranjivostima ili bugovima
- Zašto?
  - Spekulativno izvršavanje ostavlja tragove tajnih podataka u cache memoriji procesora
    - Podaci koji ne bi trebali biti tu da nema spekulativnog izvršavanja
  - Maliciozni program može pregledati sadržaj cache memorije da pokupi tajne podatke do kojih ne bi trebao imati pristup
  - Maliciozni program može prisiliti drugi program da spekulativno izvrši kod koji ostavlja tragove tajnih podataka

Više o Meltdown/Spectre ranjivostima

Project Zero

News and updates from the Project Zero team at Google

Wednesday, January 3, 2018

Reading privileged memory with a side-channel

Posted by Jann Horn, Project Zero

We have discovered that CPU data cache timing can be abused to efficiently leak information out of mis-speculated execution, leading to (at worst) arbitrary virtual memory read vulnerabilities across local security boundaries in various contexts.

https://googleprojectzero.blogspot.ch/2018/01/reading-privileged-memory-with-side.html

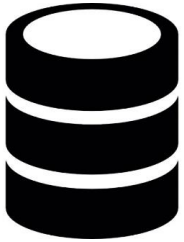
Sve zahtjevnije aplikacije

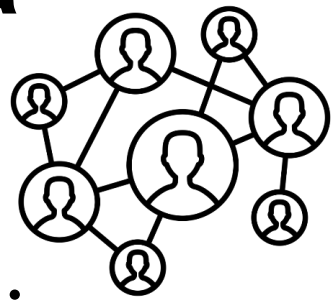
Zamisli,

I ostvariti će se

Kako aplikacije pomiču granice, računala sve teže odrađuju potrebne poslove

Podaci zatrpavaju moderna računala





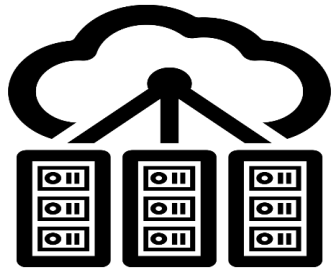
Graph/Tree Processing
[Xu+, IISWC'12; Umuroglu+, FPL'15]

In-memory Databases

Data → performance & energy bottleneck

In-Memory Data Analytics
[Clapp+ (**Intel**), IISWC'15; Awan+, BDCloud'15]

Datacenter Workloads
[Kanev+ (**Google**), ISCA'15]



Podaci zatrpavaju moderna računala

Chrome

TensorFlow Mobile

Data → performance & energy bottleneck

VP9

Video Playback

Google's video codec

VP9

Video Capture

Google's video codec

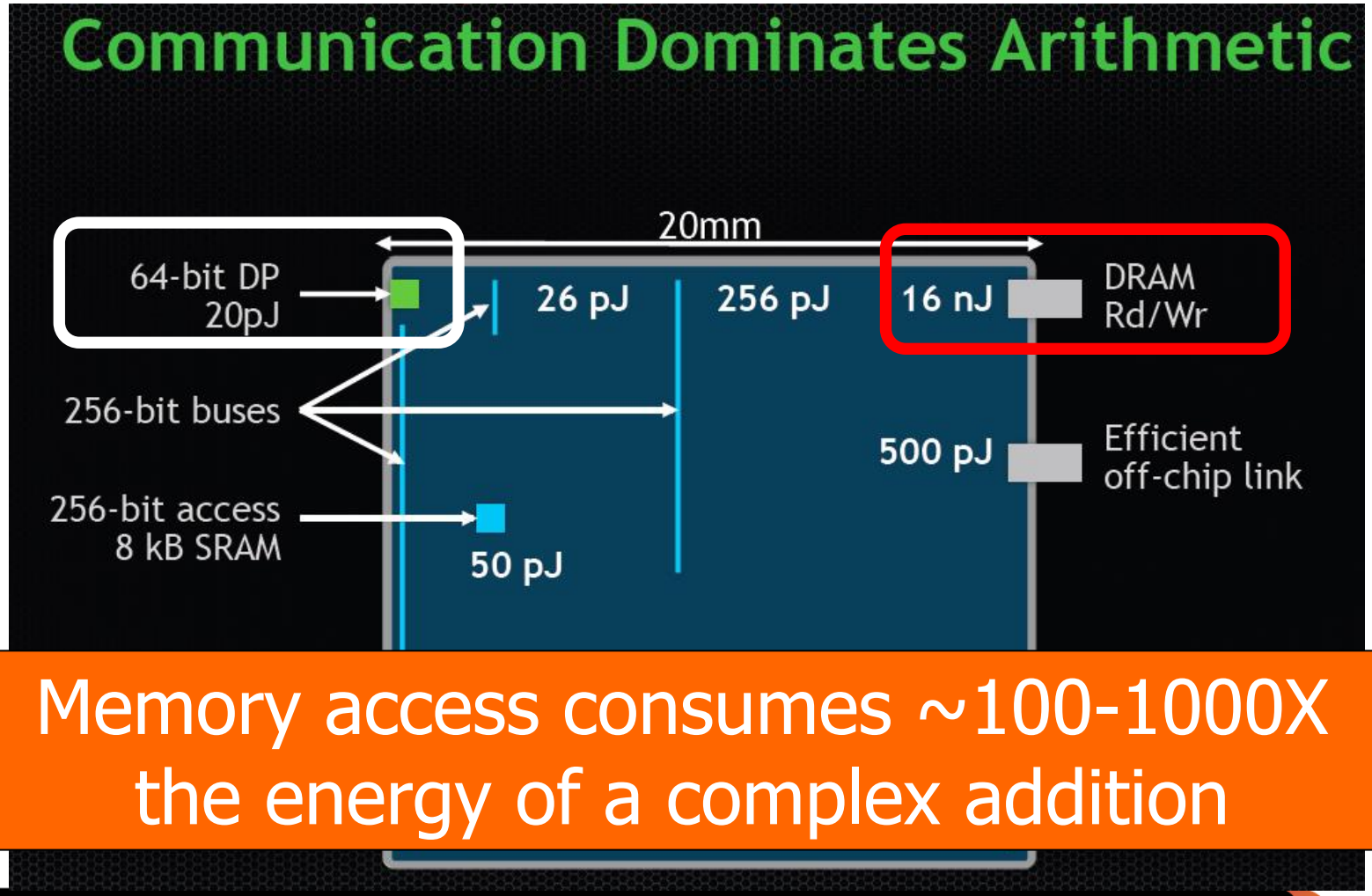
Premještanje podataka “ubija” rad računala

62.7% ukupne energije koju sustav troši se troši na premještanje podataka

Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks

Amirali Boroumand[1] Saugata Ghose[1] Youngsok Kim[2]
Rachata Ausavarungnirun[1] Eric Shiu[3] Rahul Thakur[3] Daehyun Kim[4,3]
Aki Kuusela[3] Allan Knies[3] Parthasarathy Ranganathan[3] Onur Mutlu[5,1]

Premještanje podataka vs. Potrošnja za računanje



Dally, HiPEAC 2015

Memory access consumes ~100-1000X the energy of a complex addition

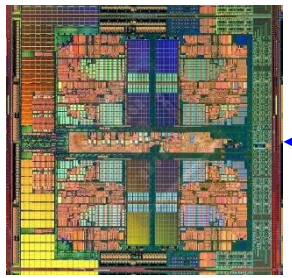
Puno zanimljivosti u građi računala danas

Istražuje se velika količina novih ideja

- Nove paradigme computinga (Full Stack)
  - Procesiranje u memoriji, procesiranje “blizu” podataka
  - Neuromorfno racunarstvo
  - Sigurno I pouzdano računarstvo
- Novi akceleratori (Algoritam-Hardver ko-dizajn)
  - Artificial Intelligence & Machine Learning
  - Graph Analytics
  - Genome Analysis
- Nove vrste memorija I sustava za pohranu
  - Postojana podatkovna memorija
  - Procesiranje u memoriji, inteligentna memorija

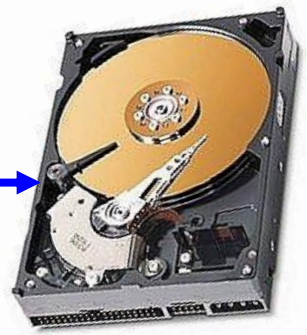
Sve kompleksniji sustavi

Past systems



Microprocessor

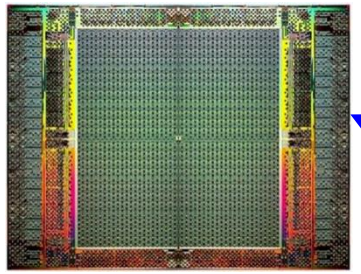
Main Memory



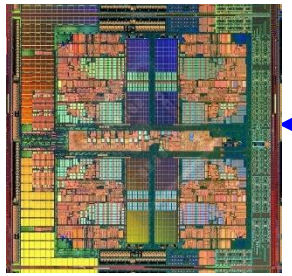
Storage (SSD/HDD)

Sve kompleksniji sustavi

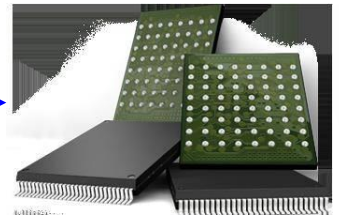
FPGA

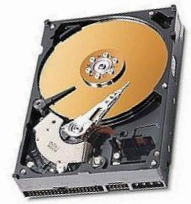


Modern systems



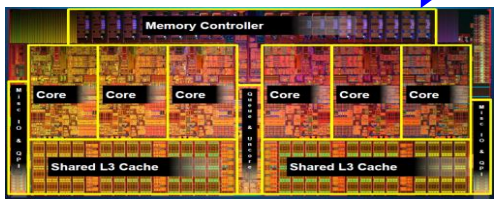
Hibridna glavna memorija

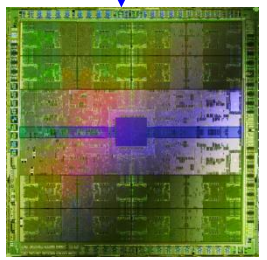




Perzistentna memorija/pohrana

Heterogeni procesori
I akceleratori





(General Purpose) GPUs

Građa računala danas

- Revolucija dolazi iz razumijevanja kako hardver I softver rade (I kako promijeniti obje tehnologije)
- Postoji puno prostora za izum novih paradigm za računanje, komunikaciju I pohranu
- Preporučeno štivo za pročitati: Thomas Kuhn, “The Structure of Scientific Revolutions” (1962)
  - Znanost “prije paradigme” – nema knosenzusa oko nove thnologije
  - “Normalna” znanost – nešto postaje dominantna teorija koju koristimo za objašnjavanje I unaprjeđivanje (postaju “business as usual), iznimke se smatraju anomalijama
  - Revolucionarna znanost – prethodne pretpostavke se ponovno pregledavaju

Hvala na pažnji!