Digital Design & Computer Arch.

Lecture 2a: Tradeoffs, Metrics, Mindset

Prof. Onur Mutlu

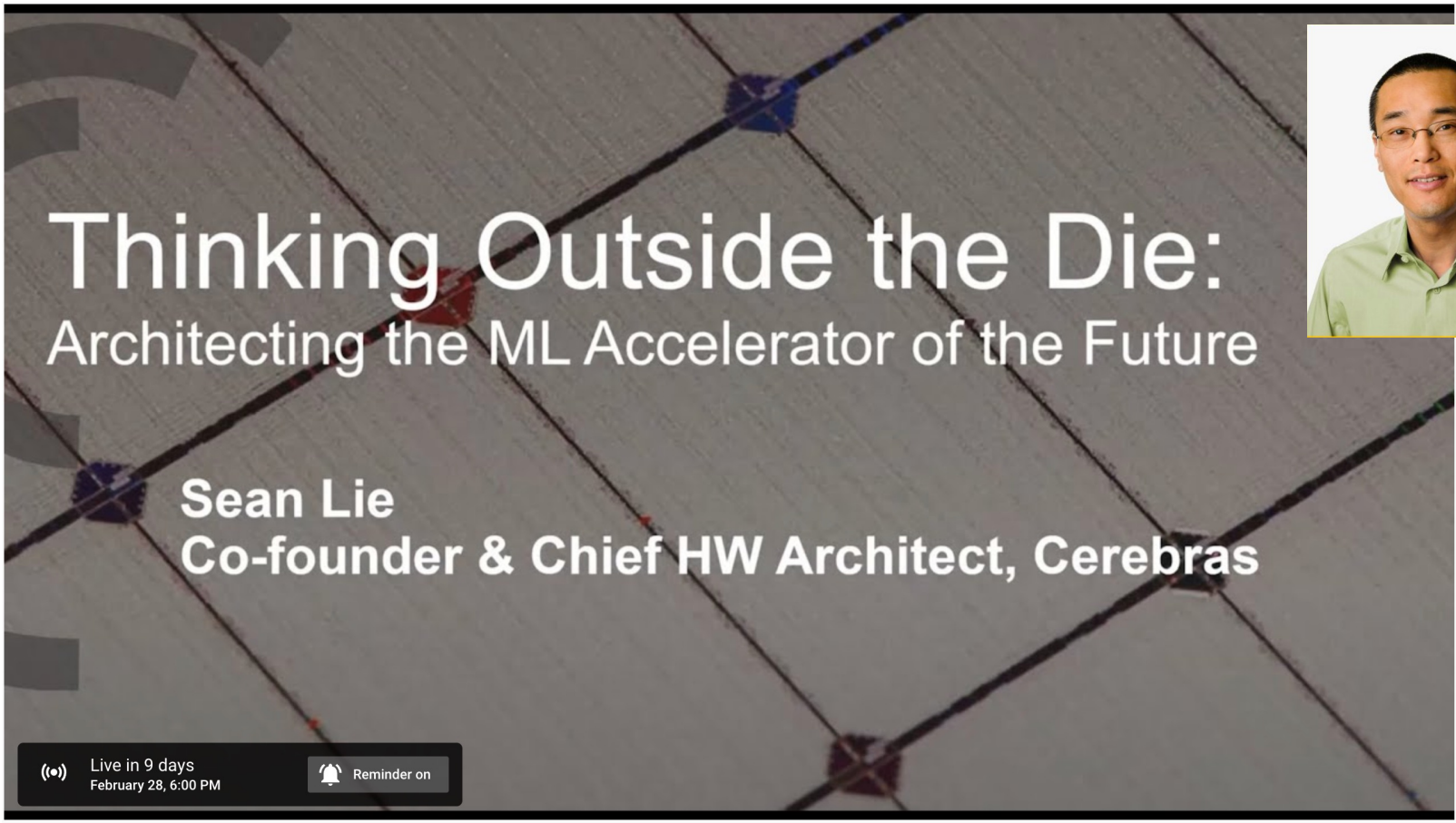
ETH Zürich

Spring 2022

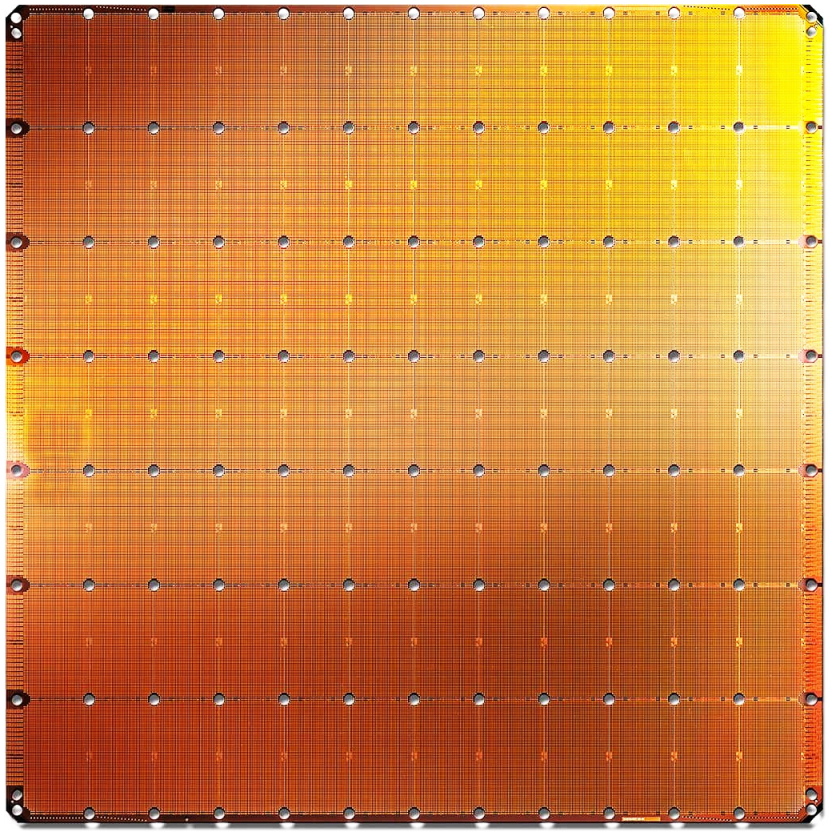
25 February 2022

Upcoming SAFARI Live Seminar (Feb 28)

https://www.youtube.com/watch?v=x2-qB0J7KHw



Cerebras's Wafer Scale ML Engine (2019)



- The largest ML accelerator chip
- 400,000 cores

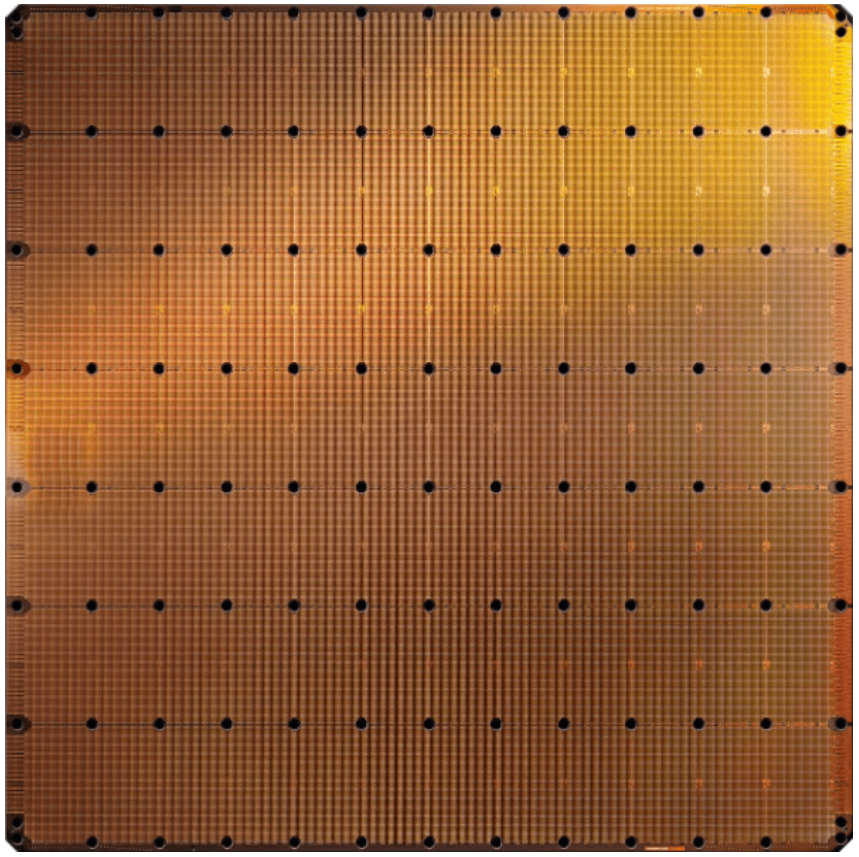
Cerebras WSE
1.2 Trillion transistors
46,225 mm^2

Largest GPU
21.1 Billion transistors
815 mm^2
NVIDIA TITAN V

https://www.anandtech.com/show/14758/hot-chips-31-live-blogs-cerebras-wafer-scale-deep-learning

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

Cerebras's Wafer Scale ML Engine-2 (2021)



- The largest ML accelerator chip (2021)
- 850,000 cores

Cerebras WSE-2
2.6 Trillion transistors
46,225 mm^2

Largest GPU
54.2 Billion transistors
826 mm^2
NVIDIA Ampere GA100

https://www.anandtech.com/show/14758/hot-chips-31-live-blogs-cerebras-wafer-scale-deep-learning

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

Extra Credit Assignment

- **Attend and watch Sean Lie's talk on Feb 28**
  - Either on Zoom or Youtube
  - https://safari.ethz.ch/safari-live-seminar-sean-lie-28-feb-2022/

- **Optional Assignment – for 1% extra credit**
  - **Write and submit a 1-page summary** of the talk
    - What are the key ideas used in the Cerebras system?
    - What are your key takeaways from the talk?
    - What did you learn?
    - What did you like or dislike?
    - Submit your summary to Moodle: https://moodle-app2.let.ethz.ch/mod/assign/view.php?id=722952

Many Interesting Things Are Happening Today in Computer Architecture

Performance and Energy Efficiency

Many Interesting Things Are Happening Today in Computer Architecture

Reliability
Safety
Security
Privacy

Many Interesting Things Are Happening Today in Computer Architecture

More Demanding Workloads

The Problem

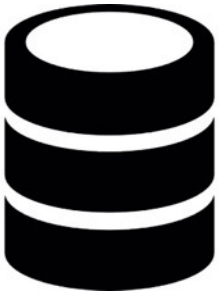
Computing

is Bottlenecked by Data

Data is Key for AI, ML, Genomics, …

- Important workloads are all data intensive
- They require rapid and efficient processing of large amounts of data
- Data is increasing
  - We can generate more than we can process

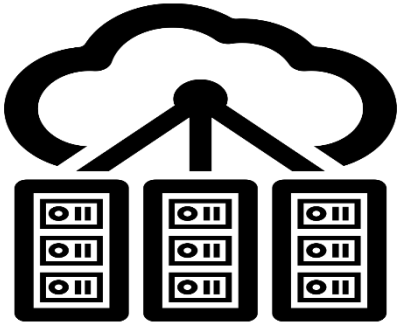
Data is Key for Future Workloads



In-memory Databases
[Mao+, EuroSys'12;
Clapp+ (**Intel**), IISWC'15]

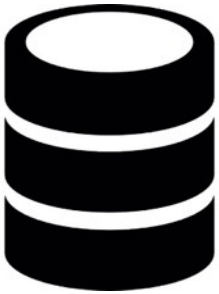
Graph/Tree Processing
[Xu+, IISWC'12; Umuroglu+, FPL'15]

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



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

Data Overwhelms Modern Machines



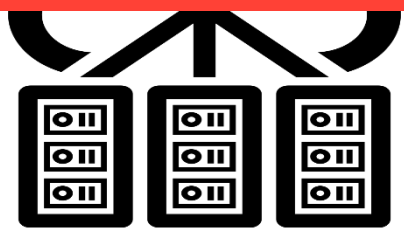
In-memory Databases

Graph/Tree Processing

Data → performance & energy bottleneck

In-Memory Data Analytics

[Clapp+ (**Intel**), IISWC'15; Awan+, BDCloud'15]



Datacenter Workloads

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

Data is Key for Future Workloads

Chrome

Google's web browser

TensorFlow Mobile

Google's machine learning framework

VP9 YouTube

Video Playback

Google's video codec

VP9 YouTube

Video Capture

Google's video codec

Data Overwhelms Modern Machines

Chrome

TensorFlow Mobile

Data → performance & energy bottleneck

VP9

Video Playback

Google's **video codec**

VP9

Video Capture

Google's **video codec**

Data Movement Overwhelms Modern Machines

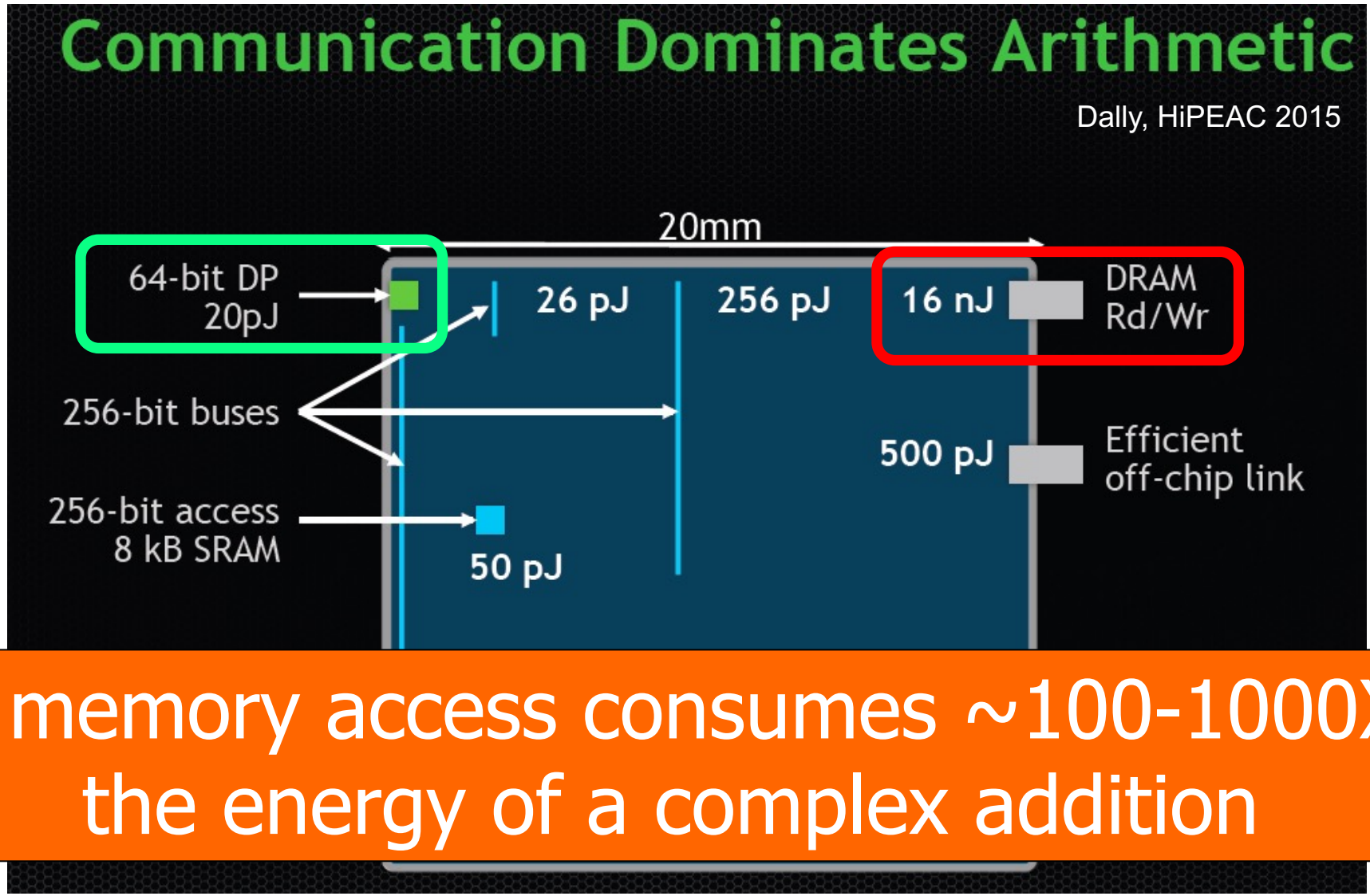
- Amirali Boroumand, Saugata Ghose, Youngsok Kim, Rachata Ausavarungnirun, Eric Shiu, Rahul Thakur, Daehyun Kim, Aki Kuusela, Allan Knies, Parthasarathy Ranganathan, and Onur Mutlu, **"Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks"** *Proceedings of the 23rd International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

62.7% of the total system energy is spent on data movement

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]

Data Movement vs. Computation Energy



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

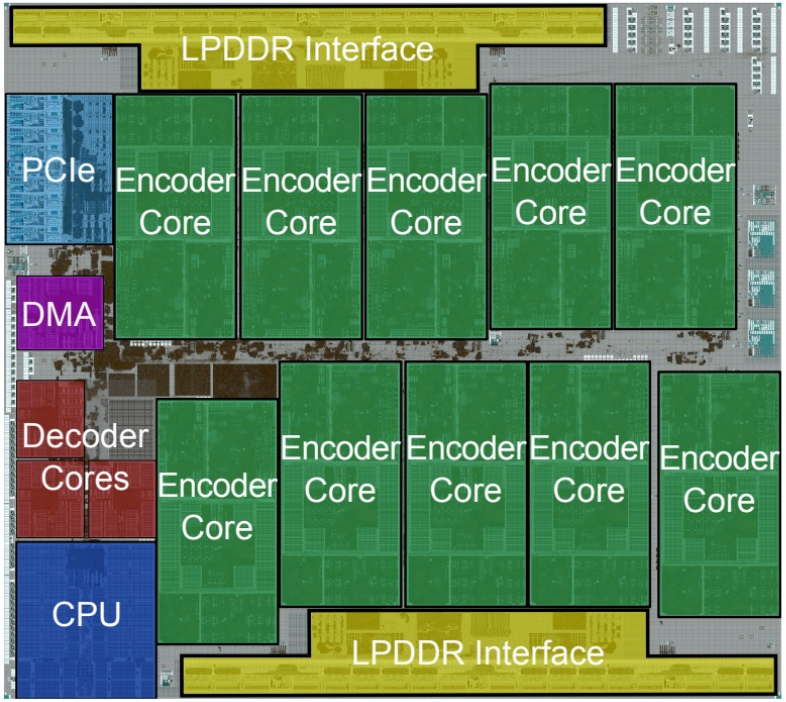
Many Interesting Things Are Happening Today in Computer Architecture

Many Novel Concepts Investigated Today

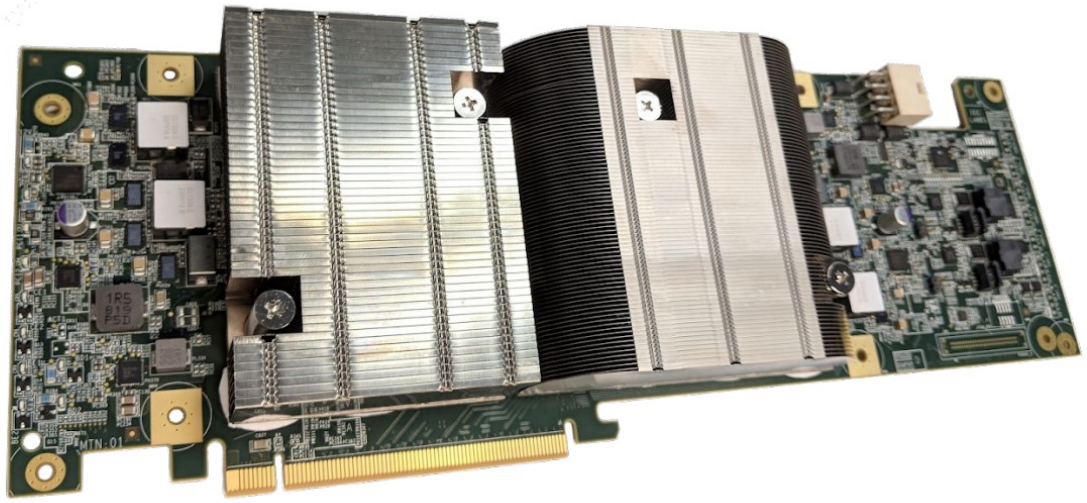
- New Computing Paradigms (Rethinking the Full Stack)
  - Processing in Memory, Processing Near Data
  - Neuromorphic Computing, Quantum Computing
  - Fundamentally Secure and Dependable Computers
- New Accelerators & Systems (Algorithm-Hardware Co-Designs)
  - Artificial Intelligence & Machine Learning
  - Graph & Data Analytics, Vision, Video
  - Genome Analysis
- New Memories, Storage Systems, Interconnects
  - Non-Volatile Main Memory, Intelligent Memory Systems
  - High-Speed Interconnects, Disaggregated Systems

Google's Video Coding Unit (2021)

Warehouse-Scale Video Acceleration: Co-design and Deployment in the Wild



(a) Chip floorplan



(b) Two chips on a PCBA

Figure 5: Pictures of the VCU

Source: https://dl.acm.org/doi/pdf/10.1145/3445814.3446723

Google's Video Coding Unit (2021)

ABSTRACT

Video sharing (e.g., YouTube, Vimeo, Facebook, TikTok) accounts for the majority of internet traffic, and video processing is also foundational to several other key workloads (video conferencing, virtual/augmented reality, cloud gaming, video in Internet-of-Things devices, etc.). The importance of these workloads motivates larger video processing infrastructures and – with the slowing of Moore's law – specialized hardware accelerators to deliver more computing at higher efficiencies. This paper describes the design and deployment, at scale, of a new accelerator targeted at warehouse-scale video transcoding. We present our hardware design including a new accelerator building block – the *video coding unit (VCU)* – and discuss key design trade-offs for balanced systems at data center scale and co-designing accelerators with large-scale distributed software systems. We evaluate these accelerators "in the wild" serving live data center jobs, demonstrating 20-33x improved efficiency over our prior well-tuned non-accelerated baseline. Our design also enables effective adaptation to changing bottlenecks and improved failure management, and new workload capabilities not otherwise possible with prior systems. To the best of our knowledge, this is the first work to discuss video acceleration at scale in large warehouse-scale environments.

Source: https://dl.acm.org/doi/pdf/10.1145/3445814.3446723

Google's Video Coding Unit (2021)

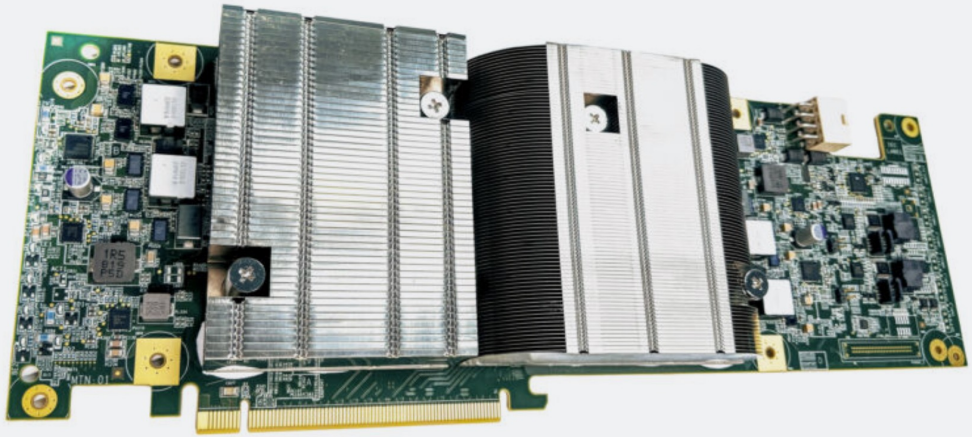
ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE STORE

I WONDER IF NETFLIX WANTS TO BUY SOME —

YouTube is now building its own video-transcoding chips

Google throws custom silicon at YouTube's massive video-transcoding workload.

RON AMADEO - 4/22/2021, 8:24 PM



Google

Enlarge / A Google Argos VCU. It transcodes video very quickly.

116

Google has decided that YouTube demands such a huge transcoding workload that it needs to build its own server chips. The company detailed its new "Argos" chips in a YouTube blog post, a CNET interview, and in a paper for ASPLOS, the Architectural Support for Programming Languages and Operating Systems Conference. Just as there are GPUs for graphics workloads and Google's TPU (tensor processing unit) for AI workloads, the YouTube infrastructure team says it has created the "VCU" or "Video (trans)Coding Unit," which helps YouTube transcode a single video into over a dozen versions that it needs to provide a smooth,

Table 1: Offline two-pass single output (SOT) throughput in VCU vs. CPU and GPU systems

System	Throughput [Mpix/s]		Perf/TCO[8]	
	H.264	VP9	H.264	VP9
Skylake	714	154	1.0x	1.0x
4xNvidia T4	2,484	—	1.5x	—
8xVCU	5,973	6,122	4.4x	20.8x
20xVCU	14,932	15,306	7.0x	33.3x

Encoding Throughput: Table 1 shows throughput and perf/TCO (performance per total cost of ownership) for the four systems and is normalized to the perf/TCO of the CPU system. The performance is shown for offline two-pass SOT encoding for H.264 and VP9. For H.264, the GPU has 3.5x higher throughput, and the 8xVCU and 20xVCU provide 8.4x and 20.9x more throughput, respectively. For VP9, the 20xVCU system has 99.4x the throughput of the CPU baseline. The two orders of magnitude increase in performance clearly demonstrates the benefits of our VCU system.

Source: https://dl.acm.org/doi/pdf/10.1145/3445814.3446723

Source: https://arstechnica.com/gadgets/2021/04/youtube-is-now-building-its-own-video-transcoding-chips/

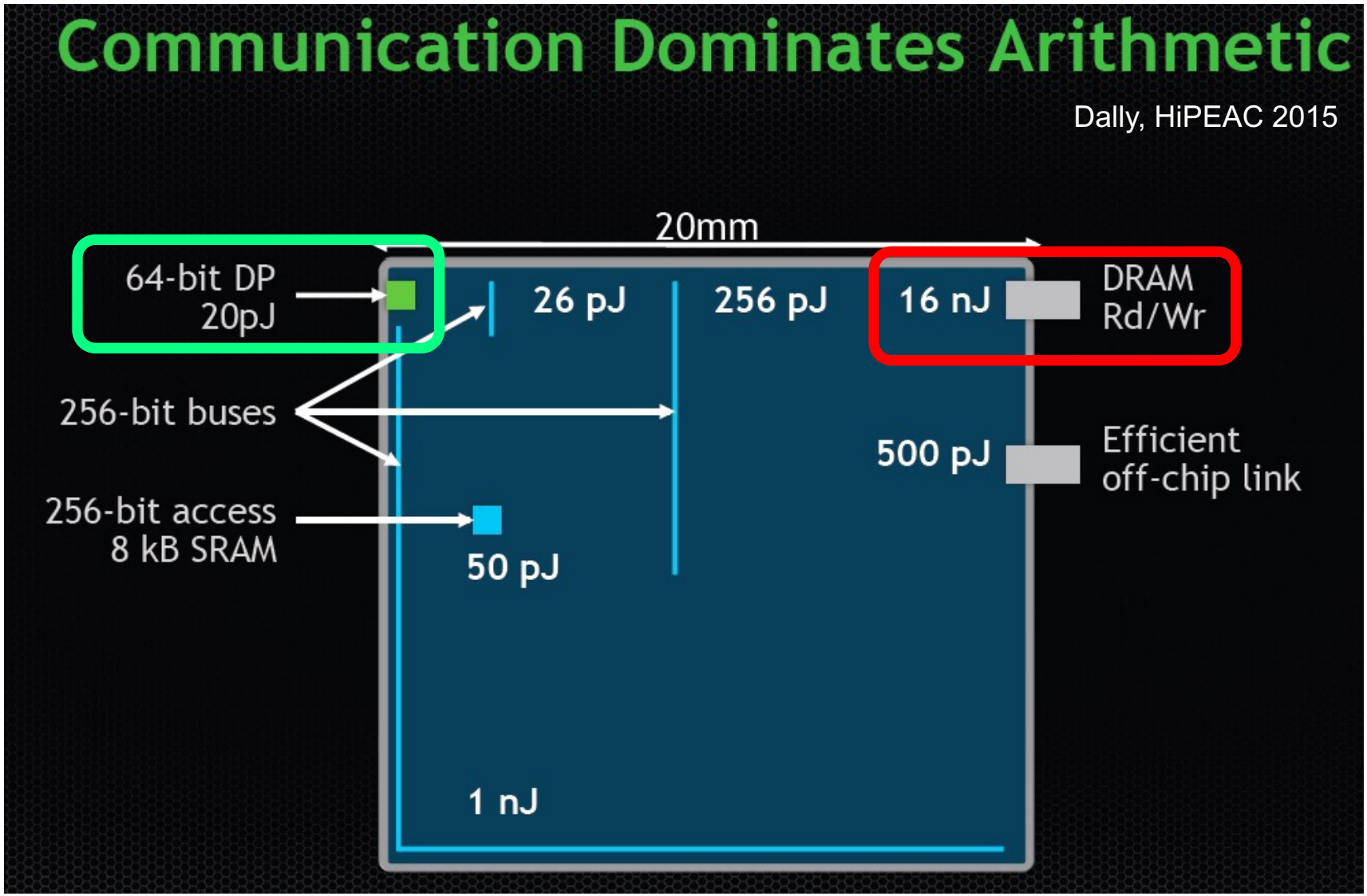
Increasingly Demanding Applications

Dream

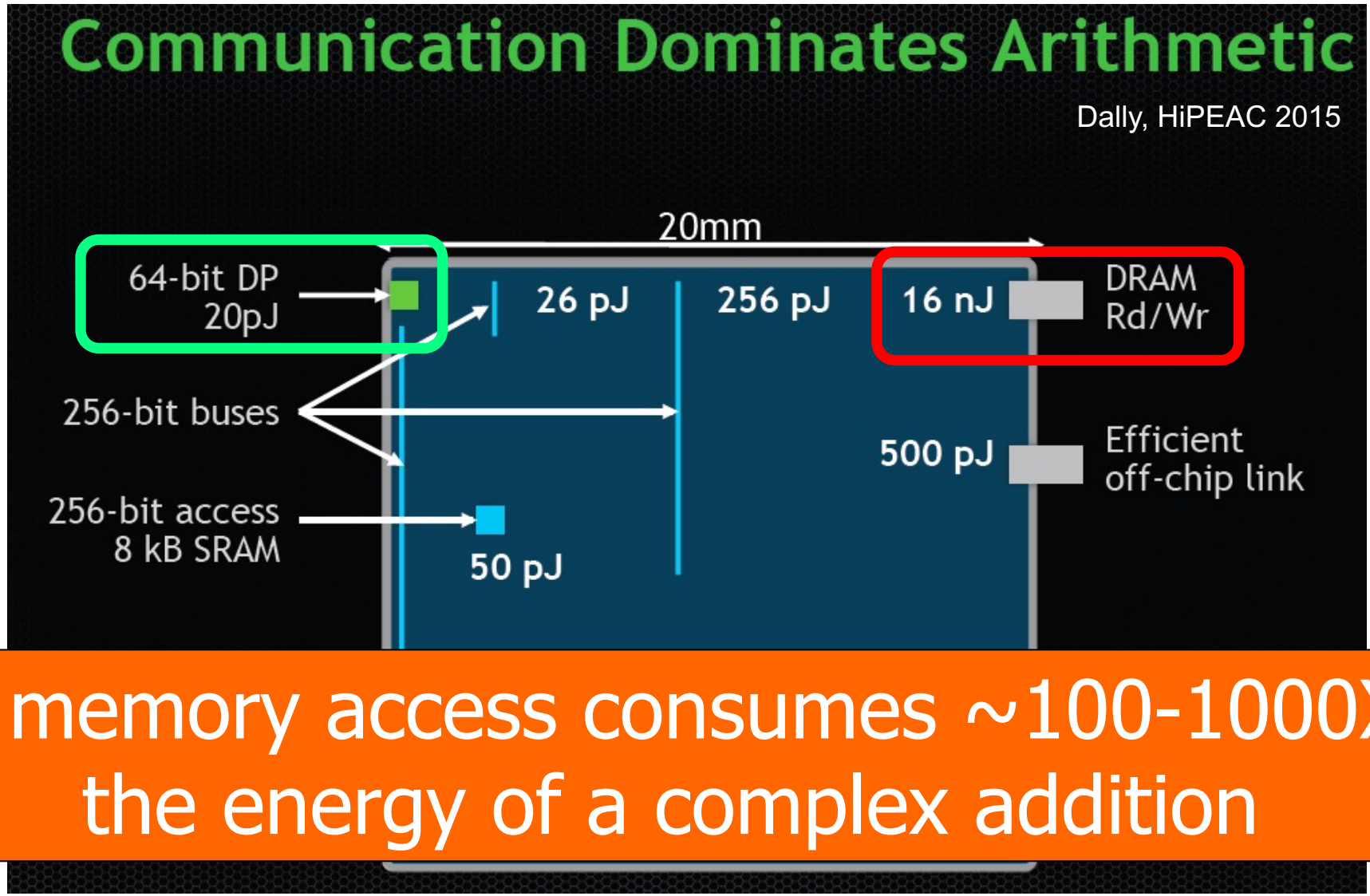
and, they will come

As applications push boundaries, computing platforms will become increasingly strained.

Increasingly Diverging/Complex Tradeoffs



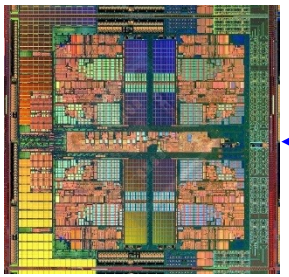
Data Movement vs. Computation Energy



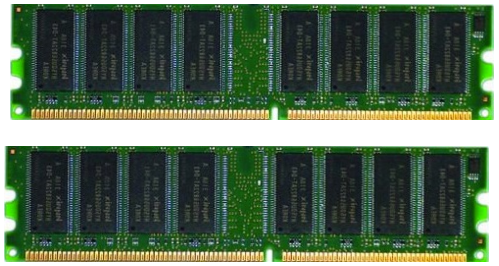
A memory access consumes ~100-1000X the energy of a complex addition

Increasingly Complex Systems

Past systems



Microprocessor

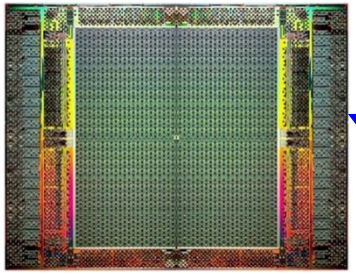


Main Memory

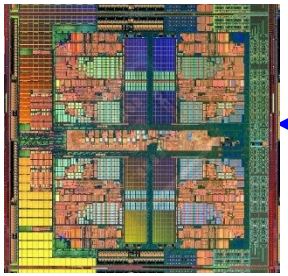
Storage (SSD/HDD)

Increasingly Complex Systems

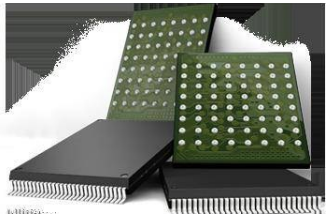
FPGAs



Modern systems

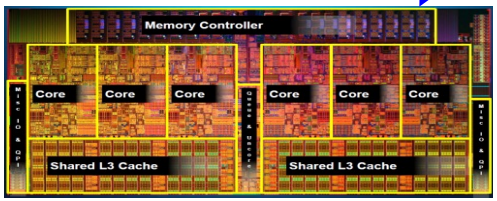


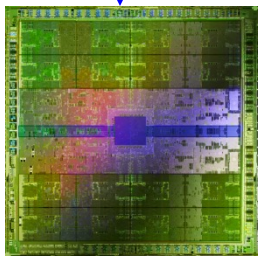
Hybrid Main Memory



Persistent Memory/Storage

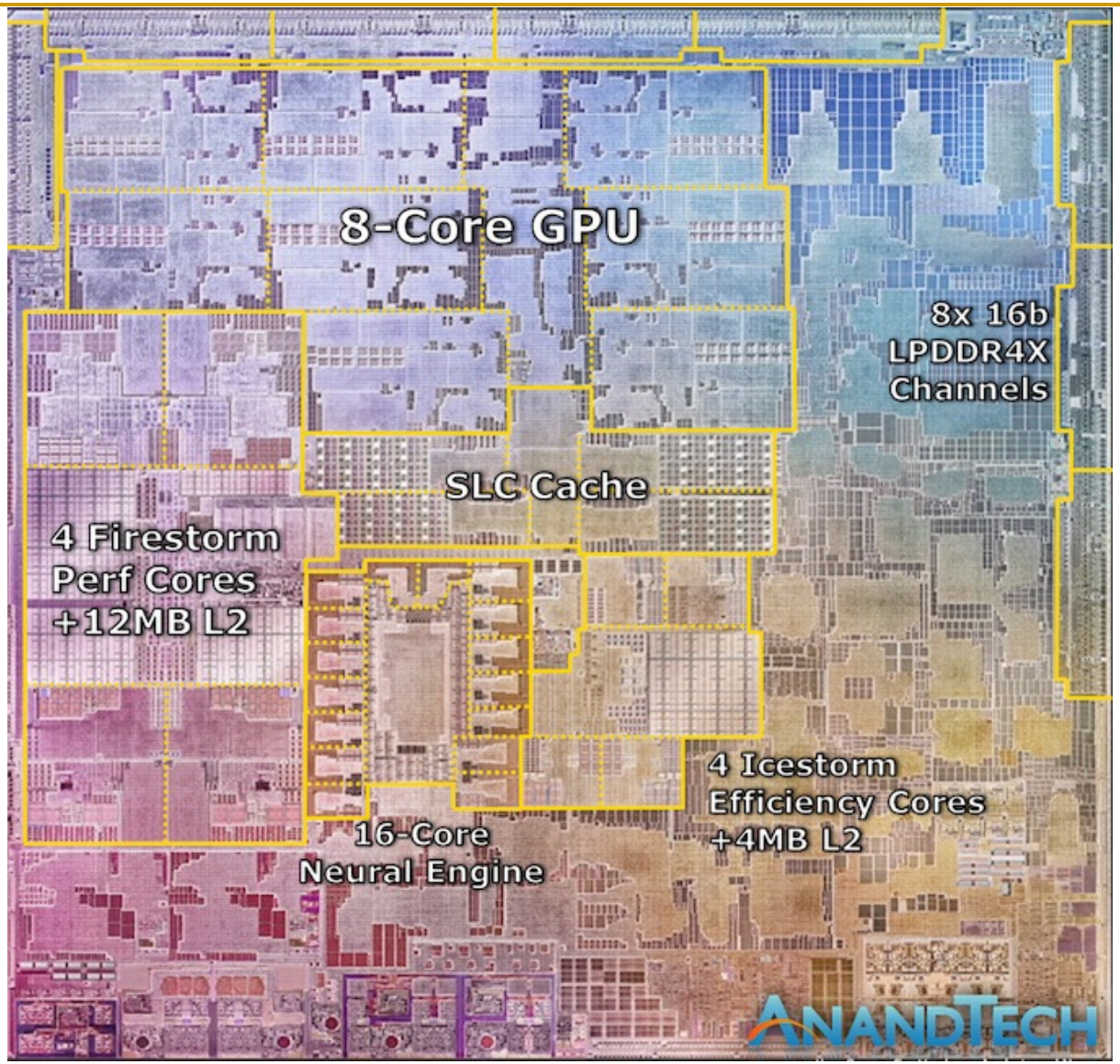
Heterogeneous Processors and Accelerators





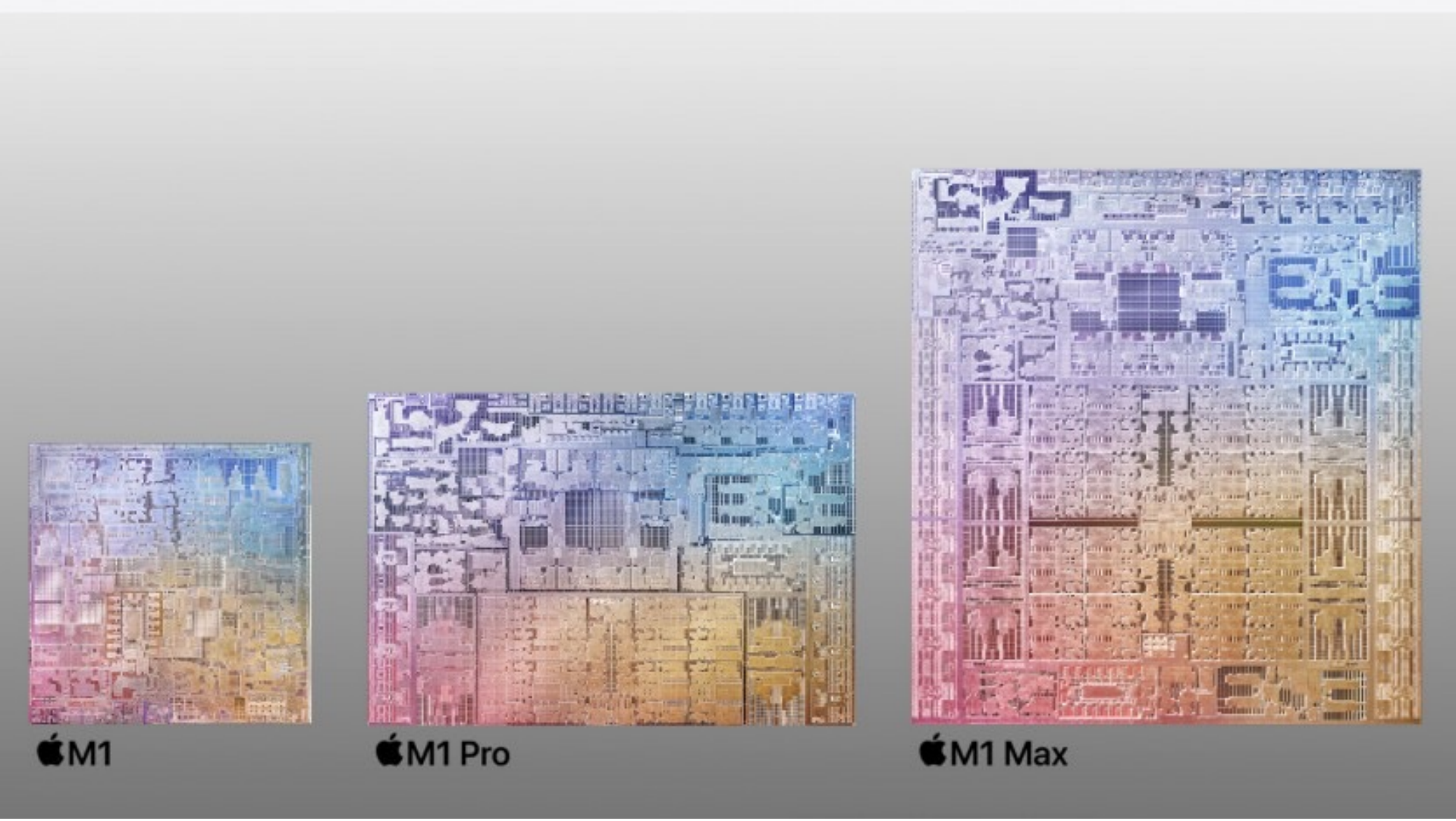
(General Purpose) GPUs

Increasingly Complex Systems on Chip



Source: https://www.anandtech.com/show/16252/mac-mini-apple-m1-tested

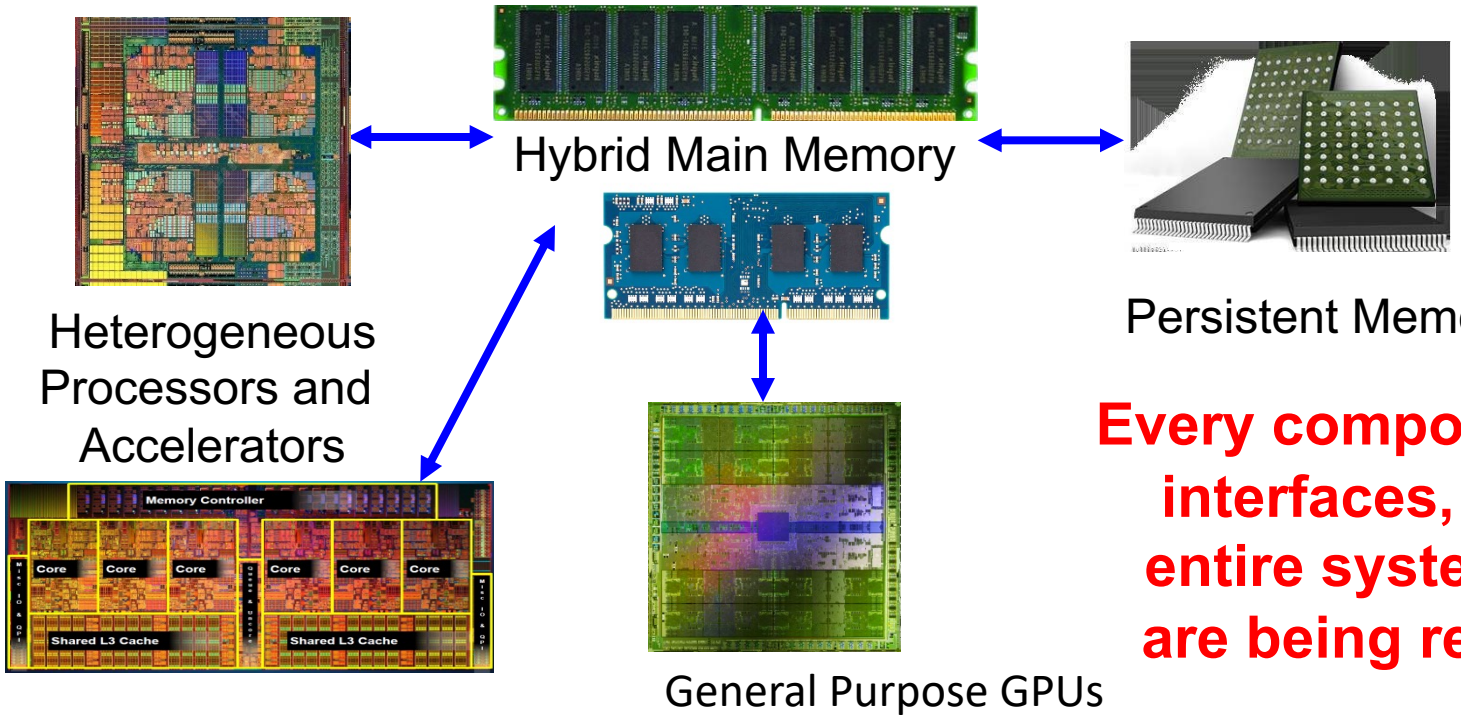
Bigger and More Powerful Systems (2021)



Source: https://www.golem.de/news/m1-pro-max-dieses-apple-silicon-ist-gigantisch-2110-160415.html

Computer Architecture Today

- Computing landscape is very different from 10-20 years ago
- Applications and technology both demand novel architectures



Every component and its interfaces, as well as entire system designs are being re-examined

Computer Architecture Today (II)

- You can revolutionize the way computers are built, if you understand both the hardware and the software (and change each accordingly)

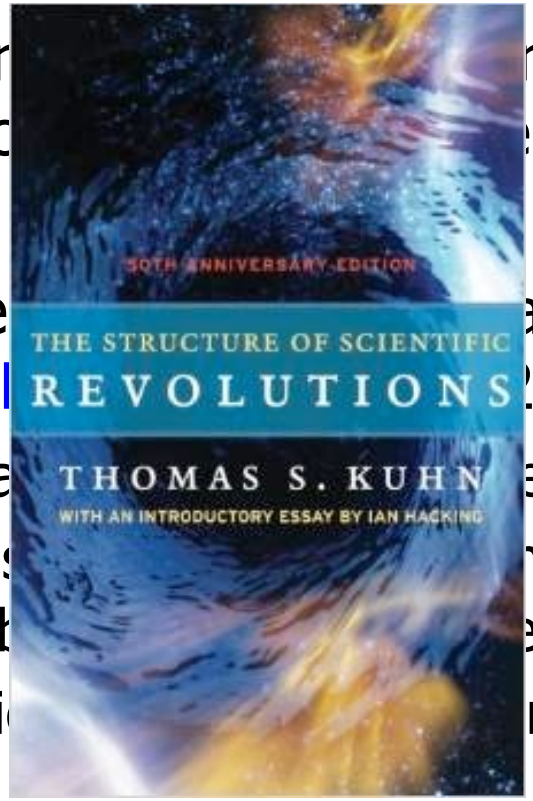
- You can invent new paradigms for computation, communication, and storage

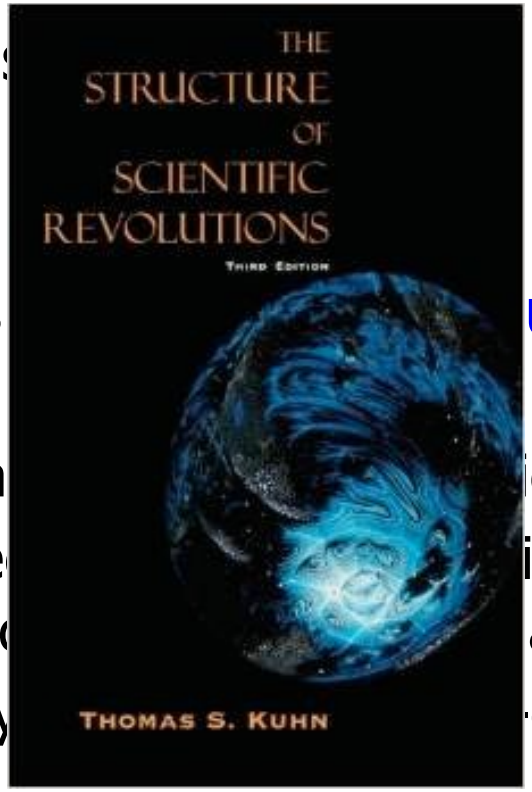
- Recommended book: Thomas Kuhn, "The Structure of Scientific Revolutions" (1962)
  - Pre-paradigm science: no clear consensus in the field
  - Normal science: dominant theory used to explain/improve things (business as usual); exceptions considered anomalies
  - Revolutionary science: underlying assumptions re-examined

Computer Architecture Today (II)

- You can revolutionize the way computers are built, if you understand both the hardware and the software (and change each accordingly)

- You can ir … m… communic … e
- Recomme … as … ure of Scientific … ?)
  - Pre-para … ea … ield
  - Normal … ne … improve things (b … ex … anomalies
  - Revoluti … rly … examined





Takeaways

- It is an exciting time to be understanding and designing computing architectures
- Many challenging and exciting problems in platform design
  - That no one has tackled (or thought about) before
  - That can have huge impact on the world's future
- Driven by huge hunger for data (Big Data), new applications (ML/AI, graph analytics, genomics), ever-greater realism, …
  - We can easily collect more data than we can analyze/understand
- Driven by significant difficulties in keeping up with that hunger at the technology layer
  - Five walls: Energy, reliability, complexity, security, scalability

Let's Start with Some Puzzles

a.k.a. Computer Architecture resembles Building Architecture

What Is This?

Source: https://www.flickr.com/photos/tambako/2286064777/in/photostream/

What About This?

Source: By Toni_V, CC BY-SA 2.0, https://commons.wikimedia.org/w/index.php?curid=4087256

What Do the Following Have in Common?

Gare do Oriente, Lisbon

Source: By Martín Gómez Tagle - Lisbon, Portugal, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=13764903

Milwaukee Art Museum

Source: By Andrew C. from Flagstaff, USA - Flickr, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=379223

Athens Olympic Stadium

Source: By Spyrosdrakopoulos - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=16172519

City of Arts and Sciences, Valencia

Source: CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=172107

Florida Polytechnic University (I)

Source: http://www.architectmagazine.com/design/buildings/florida-polytechnic-university-designed-by-santiago-calatrava_o

Oculus, New York City

Source: https://www.dezeen.com/2016/08/29/santiago-calatrava-oculus-world-trade-center-transportation-hub-new-york-photographs-hufton-crow/

What do All Those Have in Common with Bahnhof Stadelhofen?

Answer: All Designed by a Famous Architect

- ETH Alumnus, PhD Civil Engineering
- "The train station has several of the features that became signatures of his work; straight lines and right angles are rare."

Santiago Calatrava Valls (born 28 July 1951) is a Spanish architect, structural engineer, sculptor and painter, particularly known for his bridges supported by single leaning pylons, and his railway stations, stadiums, and museums, whose sculptural forms often resemble living organisms.[1] His best-known works include the Milwaukee Art Museum, the Turning Torso tower in Malmo, Sweden, the Margaret Hunt Hill Bridge in Dallas, Texas, and the Museum of Tomorrow in Rio de Janeiro,

Source: By 準建築人手札網站 Forgemind ArchiMedia - Flickr: IMG_2489.JPG, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=31493356, https://en.wikipedia.org/wiki/Santiago_Calatrava

Your First Comp. Architecture Assignment

- Go and find the closest Calatrava building to this classroom
  - For those who like a challenge, find the furthest building that was designed by Calatrava to his classroom ☺
- Appreciate the beauty & out-of-the-box and creative thinking
- Think about tradeoffs in the design
  - Strengths, weaknesses, goals of design
- Derive principles on your own for good design and innovation
- Due date: **Any time during or after this course**
  - Later during the course is better
  - Apply what you have learned in this course
  - Think out-of-the-box

But First, Today's First Assignment

Find The Differences of This and That

This

Source: By Toni_V from Zurich, Switzerland - Stadelhofen2, CC BY-SA 2.0, https://commons.wikimedia.org/w/index.php?curid=4087256

That

Source: http://cookiemagik.deviantart.com/art/Train-station-207266944 - Göttingen, DE

Many Tradeoffs Between Two Designs

- You can list them after you complete the first assignment…

Aside: Evaluation Criteria for the Designs

- Functionality (Does it meet the specification?)
- Reliability
- Space requirement
- Cost
- Expandability
- Comfort level of users
- Happiness level of users
- Aesthetics
- Security
- …

- How to evaluate goodness of design is always a critical question → "Performance" evaluation and metrics

A Key Question

- How was Calavatra able to design especially his key buildings?
- Can have many guesses
  - (Very) hard work, perseverance, dedication (over decades)
  - Experience
  - Creativity, Out-of-the-box thinking
  - A good understanding of past designs
  - Good judgment and intuition
  - Strong skill combination (math, architecture, art, engineering, …)
  - Funding ($$$$), luck, initiative, entrepreneurialism
  - Strong understanding of and commitment to fundamentals
  - Principled design
  - …
- You will be exposed to and hopefully develop/enhance many of these skills in this course

Principled Design

- "To me, there are two overriding principles to be found in nature which are most appropriate for building:
  - one is the optimal use of material,
  - the other the capacity of organisms to change shape, to grow, and to move."
  - *Santiago Calatrava*

- "Calatrava's constructions are inspired by natural forms like plants, bird wings, and the human body."

Source: http://www.arcspace.com/exhibitions/unsorted/santiago-calatrava/

Gare do Oriente, Lisbon, Revisited

Source: By Martín Gómez Tagle - Lisbon, Portugal, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=13764903

Source: http://www.arcspace.com/exhibitions/unsorted/santiago-calatrava/

A Principled Design

Zoomorphic architecture

From Wikipedia, the free encyclopedia

Zoomorphic architecture is the practice of using animal forms as the inspirational basis and blueprint for architectural design. "While animal forms have always played a role adding some of the deepest layers of meaning in architecture, it is now becoming evident that a new strand of biomorphism is emerging where the meaning derives not from any specific representation but from a more general allusion to biological processes."[1]

Some well-known examples of Zoomorphic architecture can be found in the TWA Flight Center building in New York City, by Eero Saarinen, or the Milwaukee Art Museum by Santiago Calatrava, both inspired by the form of a bird's wings.[3]

What Does This Remind You Of?

Source: https://www.dezeen.com/2016/08/29/santiago-calatrava-oculus-world-trade-center-transportation-hub-new-york-photographs-hufton-crow/

The Architect's Answer

Design [edit]

Calatrava said that the Oculus resembles a bird being released from a child's hand. The roof was originally designed to mechanically open to increase light and ventilation to the enclosed space. Herbert Muschamp, architecture critic of *The New York Times*, compared the design to the Bethesda Terrace and Fountain in Central Park, and wrote in 2004:

Source: https://en.wikipedia.org/wiki/World_Trade_Center_station_(PATH)

Strengths and Praise

> Santiago Calatrava's design for the World Trade Center PATH station should satisfy those who believe that buildings planned for ground zero must aspire to a spiritual dimension. Over the years, many people have discerned a metaphysical element in Mr. Calatrava's work. I hope New Yorkers will detect its presence, too. With deep appreciation, I congratulate the Port Authority for commissioning Mr. Calatrava, the great Spanish architect and engineer, to design a building with the power to shape the future of New York. It is a pleasure to report, for once, that public officials are not overstating the case when they describe a design as breathtaking.[43]

Source: https://en.wikipedia.org/wiki/World_Trade_Center_station_(PATH)

Design Constraints and Criticism

However, Calatrava's original soaring spike design was scaled back because of security issues. The *New York Times* observed in 2005:

> In the name of security, Santiago Calatrava's bird has grown a beak. Its ribs have doubled in number and its wings have lost their interstices of glass.... [T]he main transit hall, between Church and Greenwich Streets, will almost certainly lose some of its delicate quality, while gaining structural expressiveness. It may now evoke a slender stegosaurus more than it does a bird.[45]

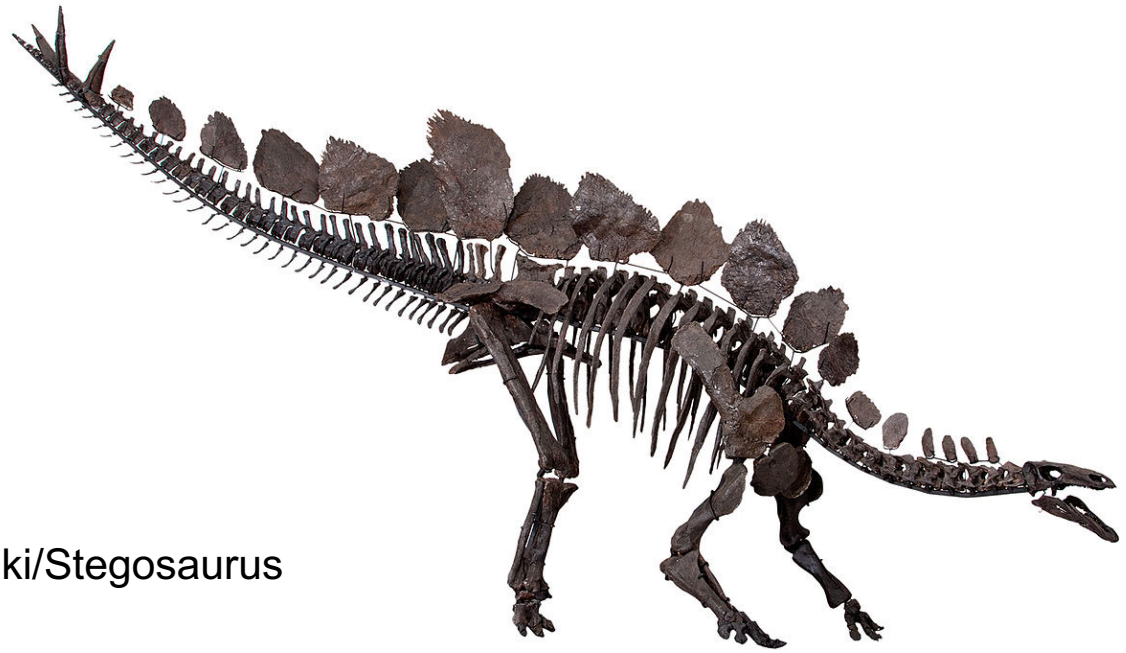
Source: https://en.wikipedia.org/wiki/World_Trade_Center_station_(PATH)

Stegosaurus

From Wikipedia, the free encyclopedia

For the pachycephalosaurid of a similar name, see Stegoceras.

Stegosaurus (/ˌstɛgəˈsɔːrəs/[1]) is a genus of armored dinosaur. Fossils of this genus date to the Late Jurassic period, where they are found in Kimmeridgian to early Tithonian aged strata, between 155 and 150 million years ago, in the western United States and Portugal. Several



Source: https://en.wikipedia.org/wiki/Stegosaurus

Susannah Maidment et al. & Natural History Museum, London - Maidment SCR, Brassey C, Barrett PM (2015) The Postcranial Skeleton of an Exceptionally Complete Individual of the Plated Dinosaur Stegosaurus stenops (Dinosauria: Thyreophora) from the Upper Jurassic Morrison Formation of Wyoming, U.S.A. PLoS ONE 10(10): e0138352. doi:10.1371/journal.pone.0138352

Design Constraints: Noone is Immune

However, Calatrava's original soaring spike design was scaled back because of security issues. The *New York Times* observed in 2005:

> In the name of security, Santiago Calatrava's bird has grown a beak. Its ribs have doubled in number and its wings have lost their interstices of glass.... [T]he main transit hall, between Church and Greenwich Streets, will almost certainly lose some of its delicate quality, while gaining structural expressiveness. It may now evoke a slender stegosaurus more than it does a bird.[45]

The design was further modified in 2008 to eliminate the opening and closing roof mechanism because of budget and space constraints.[46]

The Transportation Hub has been dubbed "the world's most expensive transportation hub" for its massive cost for reconstruction—$3.74 billion dollars.[48][58] By contrast, the proposed two-mile PATH extension

Source: https://en.wikipedia.org/wiki/World_Trade_Center_station_(PATH)

The Lecture Was Slightly Different When I Was at CMU

What Is This?

Source: https://roadtrippers.com/stories/falling-water

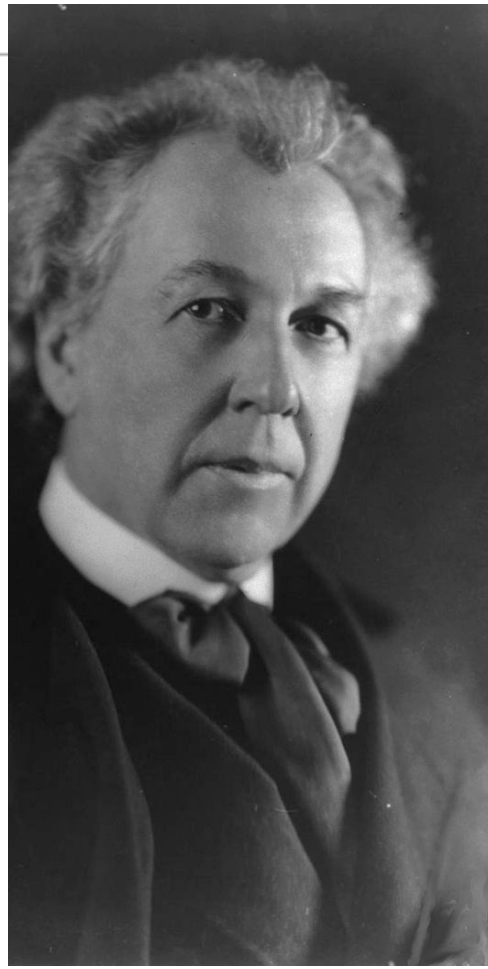
Answer: Masterpiece of A Famous Architect

Fallingwater

From Wikipedia, the free encyclopedia

Fallingwater or **Kaufmann Residence** is a house designed by architect Frank Lloyd Wright in 1935 in rural southwestern Pennsylvania, 43 miles (69 km) southeast of Pittsburgh.[4] The home was built partly over a waterfall on Bear Run in the Mill Run section of Stewart Township, Fayette County, Pennsylvania, in the Laurel Highlands of the Allegheny Mountains.

Time cited it after its completion as Wright's "most beautiful job";[5] it is listed among *Smithsonian*'s Life List of 28 places "to visit before you die."[6] It was designated a National Historic Landmark in 1966.[3] In 1991, members of the American Institute of Architects named the house the "best all-time work of American architecture" and in 2007, it was ranked twenty-ninth on the list of America's Favorite Architecture according to the AIA.



Source: https://en.wikipedia.org/wiki/Fallingwater

Find The Differences of This and That

This

Source: http://www.fallingwater.org/

This

That

A Key Question

- How was Wright able to design his masterpiece?
- Can have many guesses
  - (Very) hard work, perseverance, dedication (over decades)
  - Experience
  - Creativity, Out-of-the-box thinking
  - A good understanding of past designs
  - Good judgment and intuition
  - Strong skill combination (math, architecture, art, engineering, …)
  - Funding ($$$$), luck, initiative, entrepreneurialism
  - Strong understanding of and commitment to fundamentals
  - Principled design
  - …
- You will be exposed to and hopefully develop/enhance many of these skills in this course

A Quote from The Architect Himself

- "architecture [...] based upon principle, and not upon precedent"

Source: http://www.fallingwater.org/

A Principled Design

Organic architecture

From Wikipedia, the free encyclopedia

Organic architecture is a philosophy of architecture which promotes harmony between human habitation and the natural world through design approaches so sympathetic and well integrated with its site, that buildings, furnishings, and surroundings become part of a unified, interrelated composition.

A well-known example of organic architecture is Fallingwater, the residence Frank Lloyd Wright designed for the Kaufmann family in rural Pennsylvania. Wright had many choices to locate a home on this large site, but chose to place the home directly over the waterfall and creek creating a close, yet noisy dialog with the rushing water and the steep site. The horizontal striations of stone masonry with daring cantilevers of colored beige concrete blend with native rock outcroppings and the wooded environment.

A Key Question

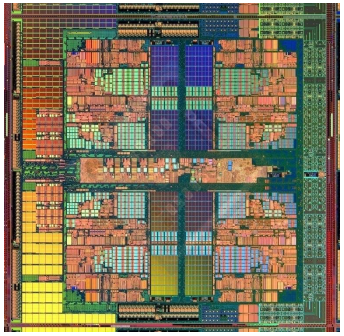
- How was Wright able to design his masterpiece?
- Can have many guesses
  - (Very) hard work, perseverance, dedication (over decades)
  - Experience
  - Creativity, Out-of-the-box thinking
  - A good understanding of past designs
  - Good judgment and intuition
  - Strong skill combination (math, architecture, art, engineering, …)
  - Funding ($$$$), luck, initiative, entrepreneurialism
  - Strong understanding of and commitment to fundamentals
  - Principled design
  - …
- You will be exposed to and hopefully develop/enhance many of these skills in this course

Takeaways

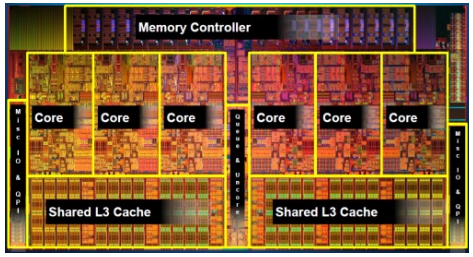
- It all starts from the basic building blocks and design principles
- And, knowledge of how to use, apply, enhance them
- Underlying technology might change (e.g., steel vs. wood)
  - but methods of taking advantage of technology bear resemblance
  - methods used for design depend on the principles employed

The Same Applies to Processor Chips

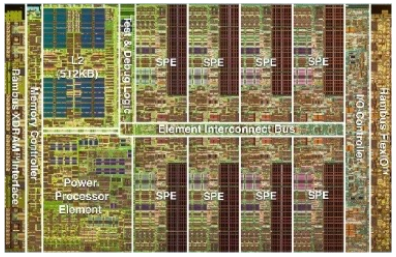
- There are basic building blocks and design principles



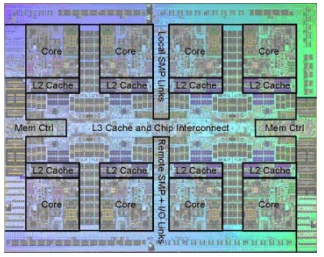
AMD Barcelona
4 cores



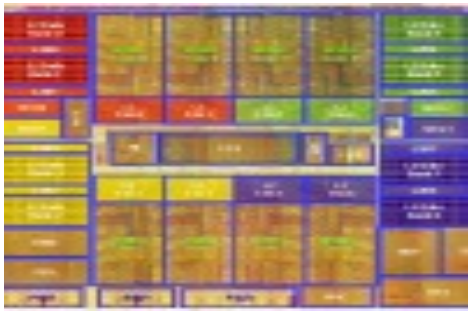
Intel Core i7
8 cores



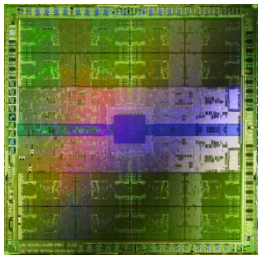
IBM Cell BE
8+1 cores



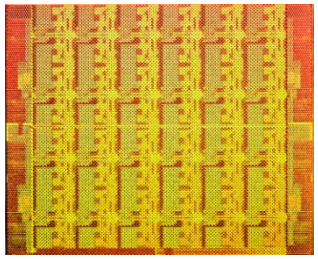
IBM POWER7
8 cores



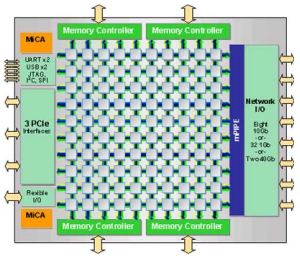
Sun Niagara II
8 cores



Nvidia Fermi
448 “cores”



Intel SCC
48 cores, networked



Tilera TILE Gx
100 cores, networked

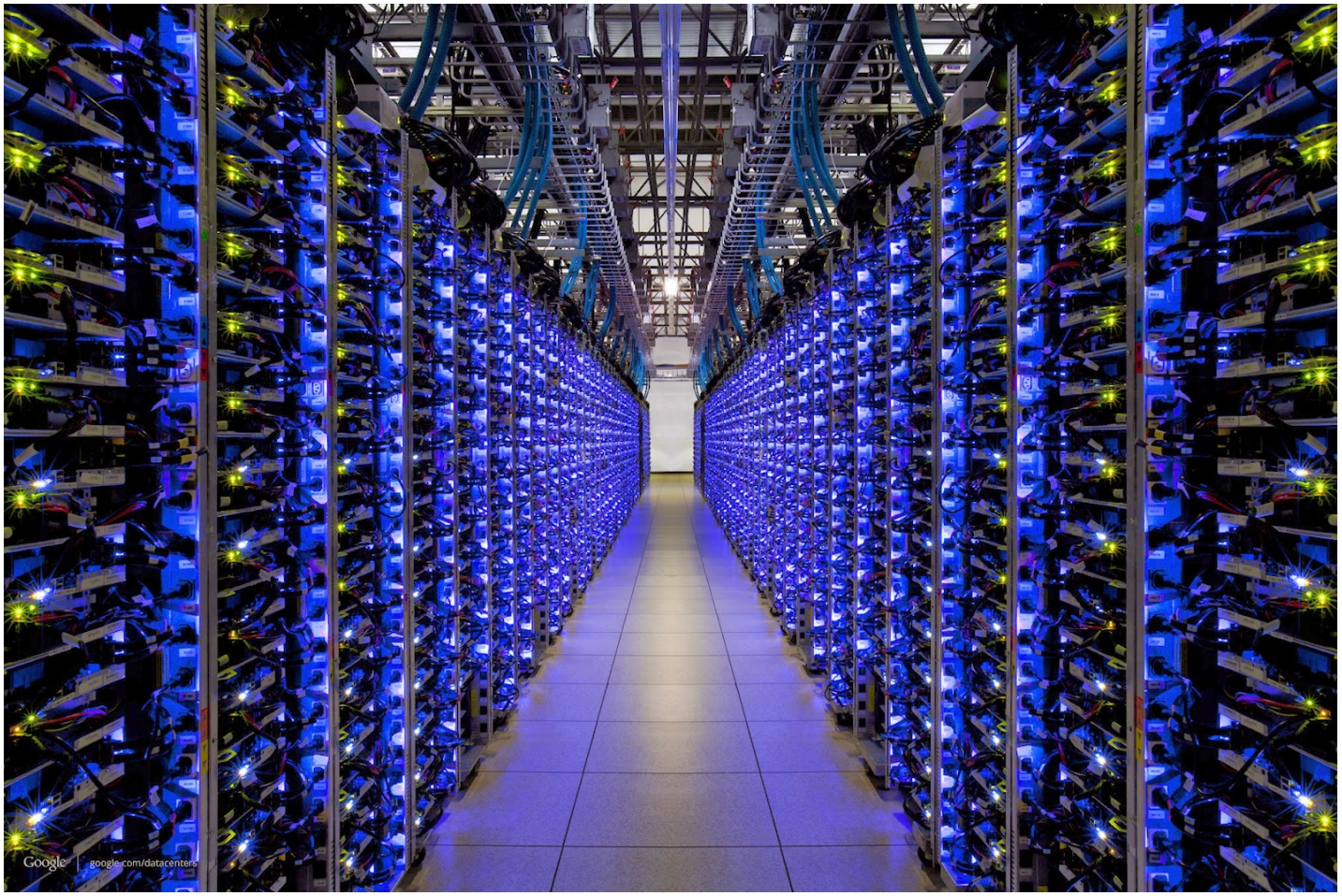
The Same Applies to Computing Systems

- There are basic building blocks and design principles

source: http://www.sia-online.org (semiconductor industry association)

The Same Applies to Computing Systems

- There are basic building blocks and design principles



Source: http://datacentervoice.com/wp-content/uploads/2015/10/data-center.jpg

Different Platforms, Different Goals

Source: https://iq.intel.com/5-awesome-uses-for-drone-technology/

Different Platforms, Different Goals

Source: https://taxistartup.com/wp-content/uploads/2015/03/UK-Self-Driving-Cars.jpg

Different Platforms, Different Goals

Source: http://sm.pcmag.com/pcmag_uk/photo/g/google-self-driving-car-the-guts/google-self-driving-car-the-guts_dwx8.jpg

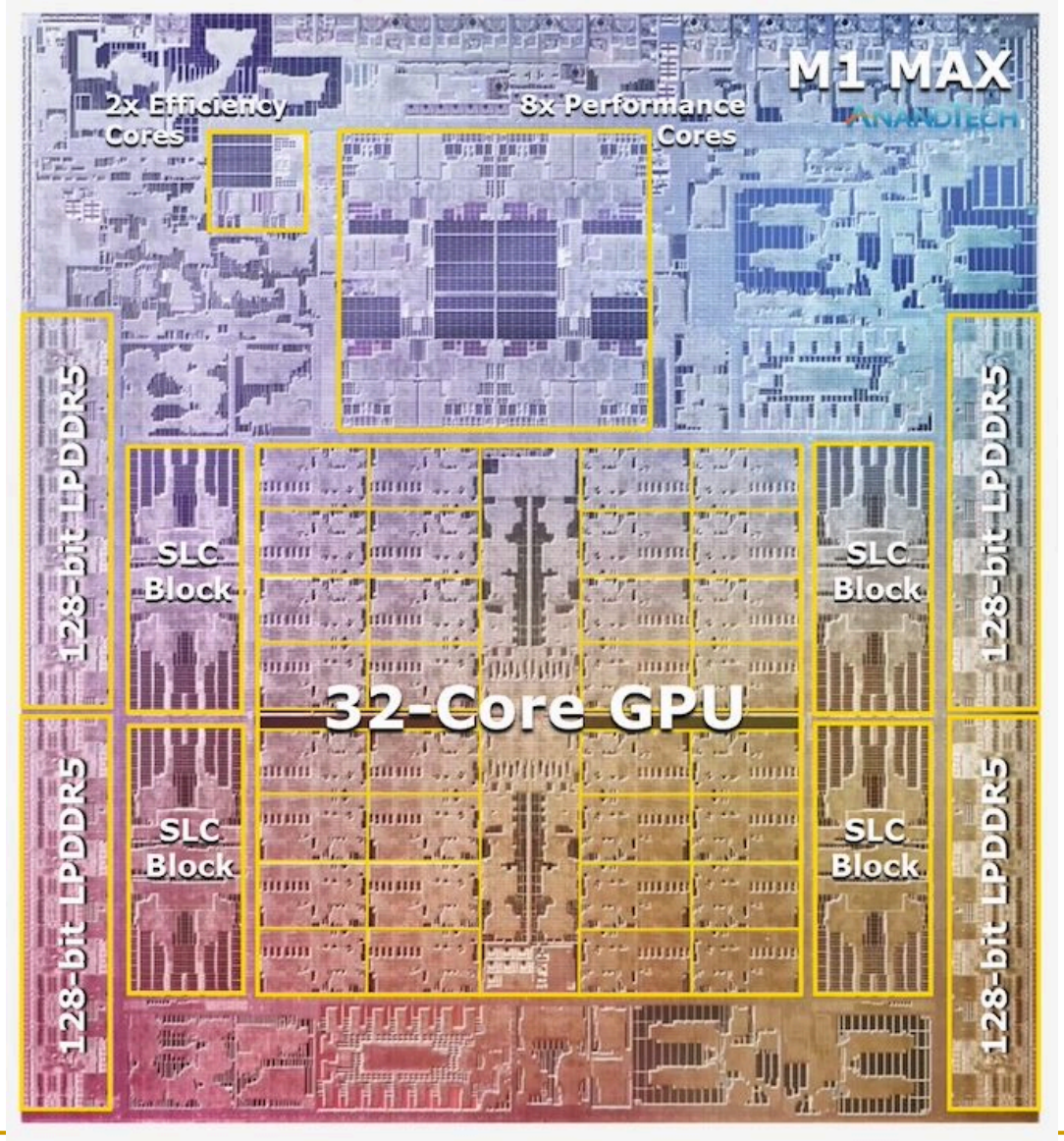
Different Platforms, Different Goals

Source: https://fossbytes.com/wp-content/uploads/2015/06/Supercomputer-TIANHE2-china.jpg

Different Platforms, Different Goals

Source: https://www.itmagazine.ch/artikel/72401/Fugaku_Der_schnellste_Supercomputer_der_Welt.html

Apple M1 Max System on Chip (2021)



Source: https://www.anandtech.com/show/17024/apple-m1-max-performance-review

Google Tensor Processing Unit (~2016)

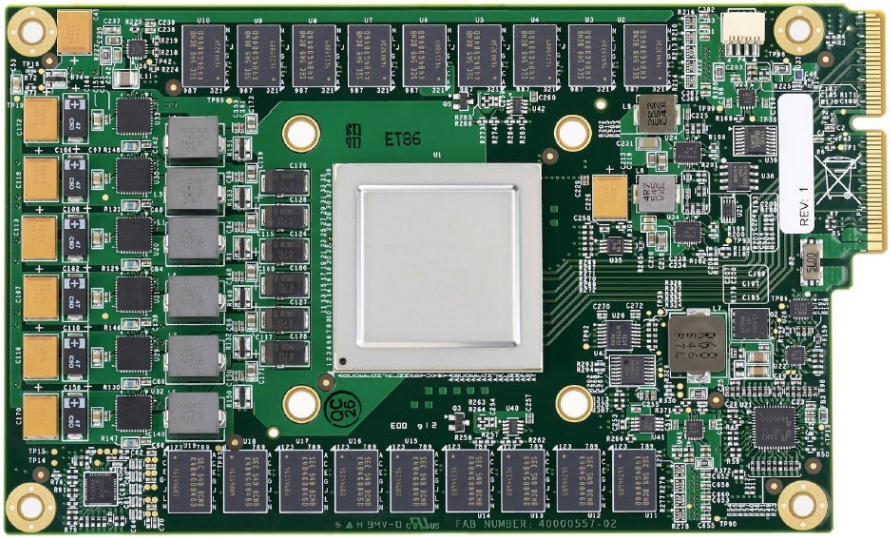


Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

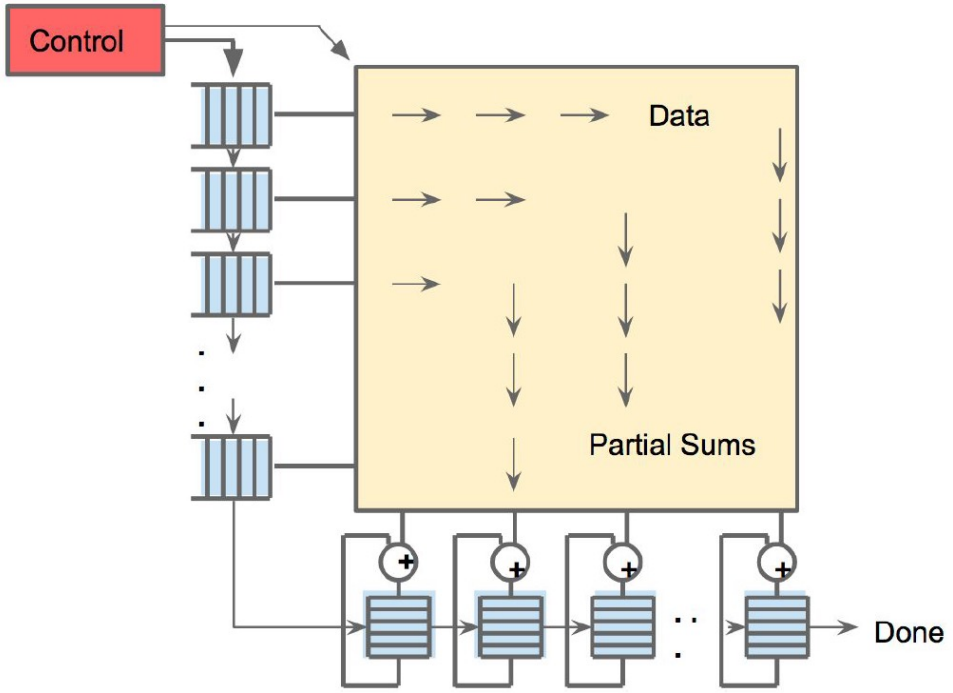
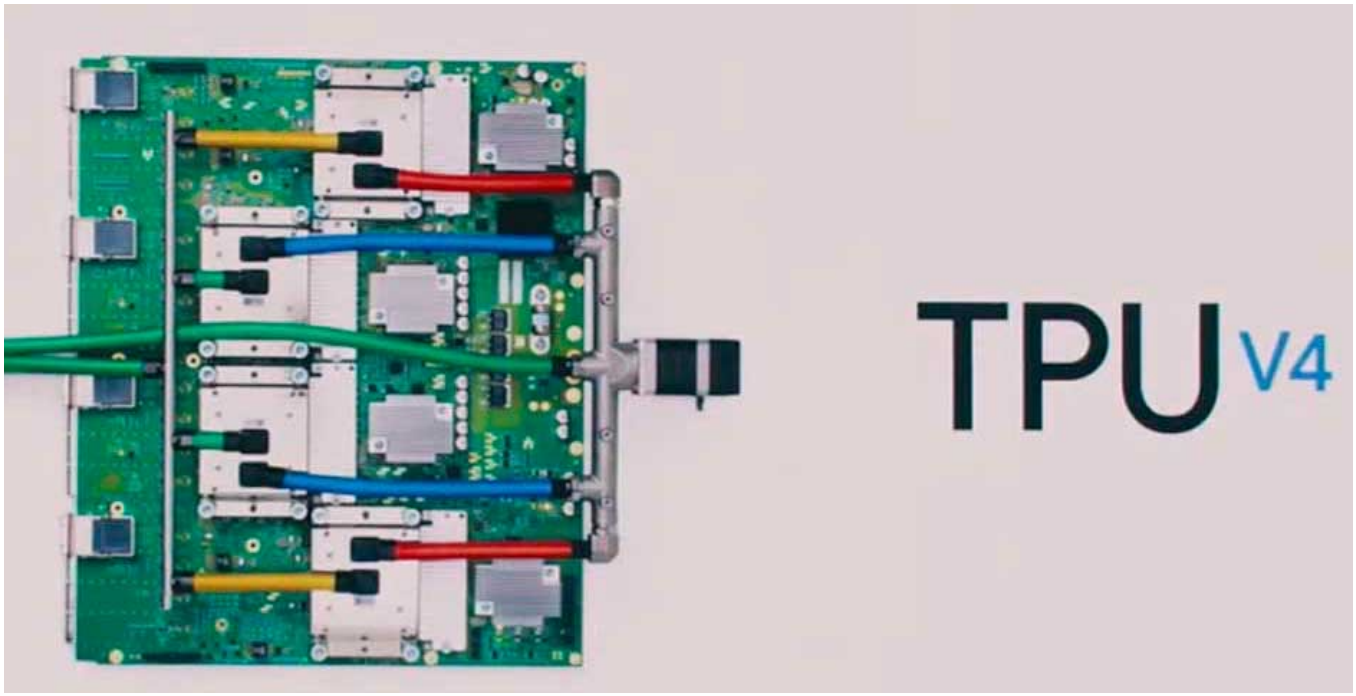


Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

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

Google TPU Generation IV (2021)



New ML applications (vs. TPU3):

- Computer vision
- Natural Language Processing (NLP)
- Recommender system
- Reinforcement learning that plays Go

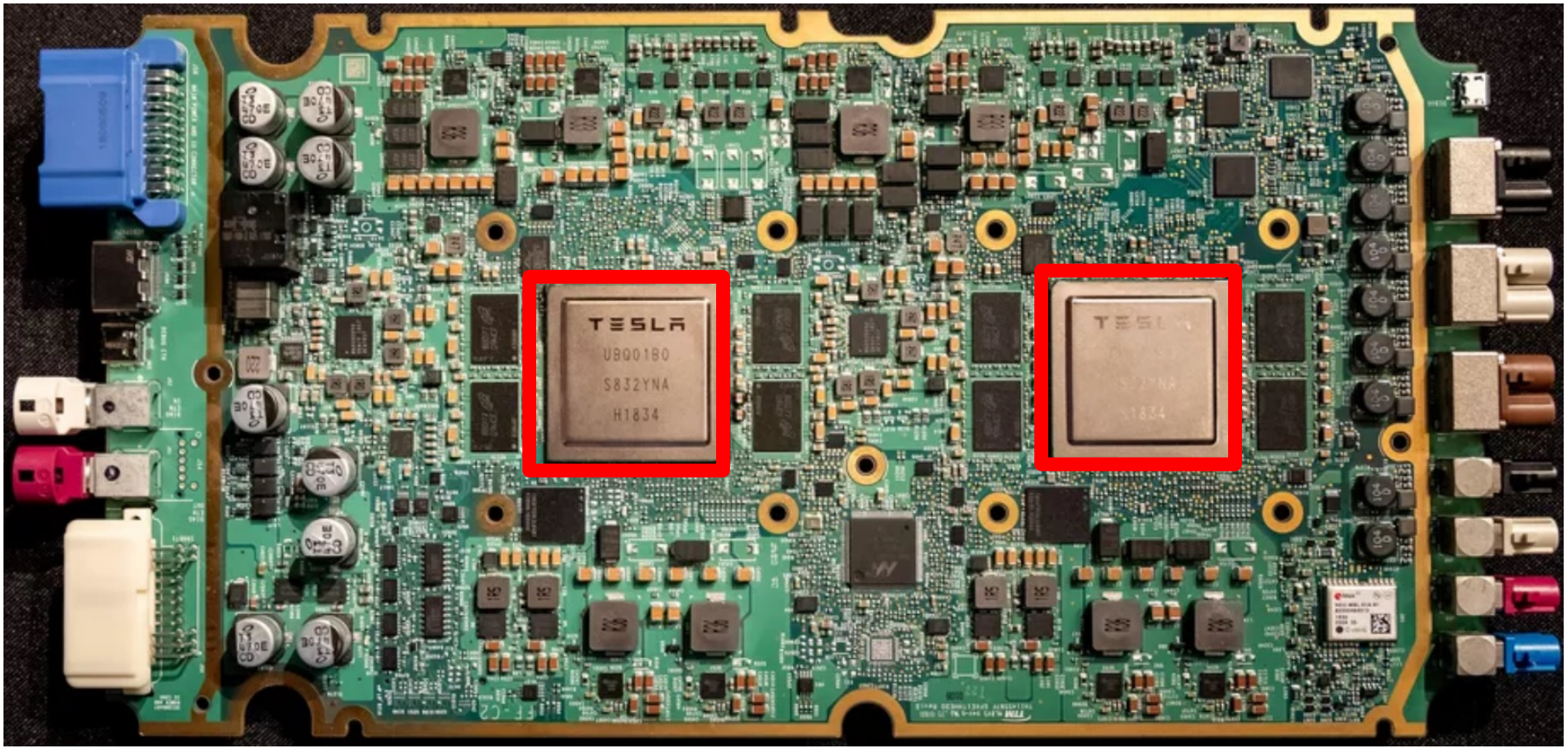
250 TFLOPS per chip in 2021
vs 90 TFLOPS in TPU3

1 ExaFLOPS per board

https://spectrum.ieee.org/tech-talk/computing/hardware/heres-how-googles-tpu-v4-ai-chip-stacked-up-in-training-tests

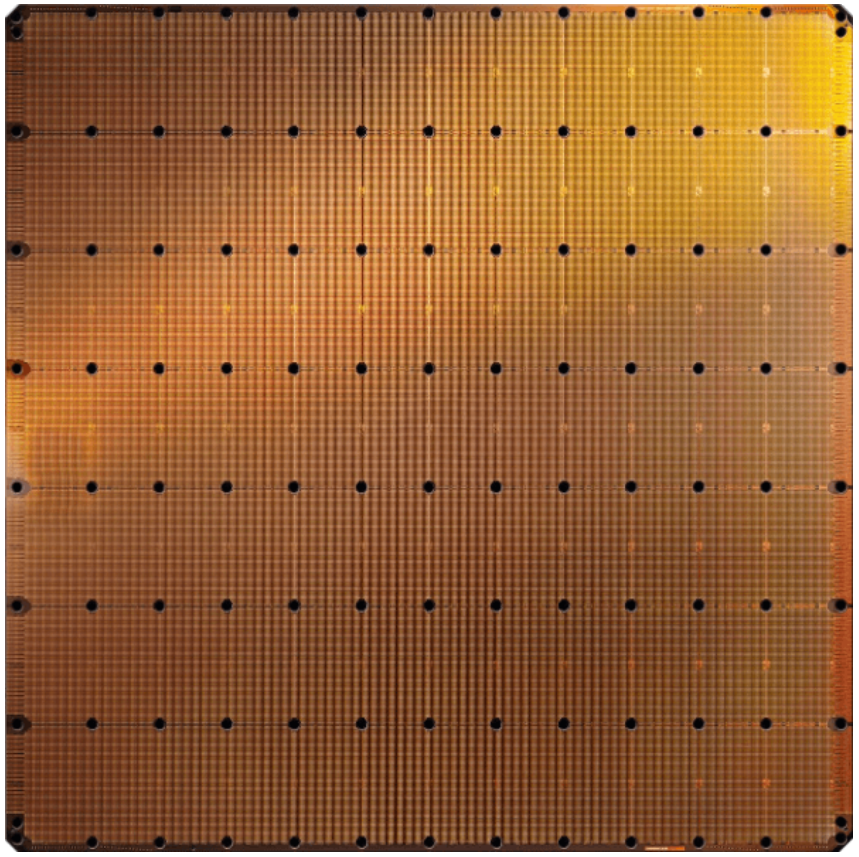
TESLA Full Self-Driving Computer (2019)

- ML accelerator: 260 mm^2, 6 billion transistors, 600 GFLOPS GPU, 12 ARM 2.2 GHz CPUs.
- Two redundant chips for better safety.



https://youtu.be/Ucp0TTmvqOE?t=4236

Cerebras's Wafer Scale ML Engine-2 (2021)



- The largest ML accelerator chip (2021)
- 850,000 cores

Cerebras WSE-2
2.6 Trillion transistors
46,225 mm^2

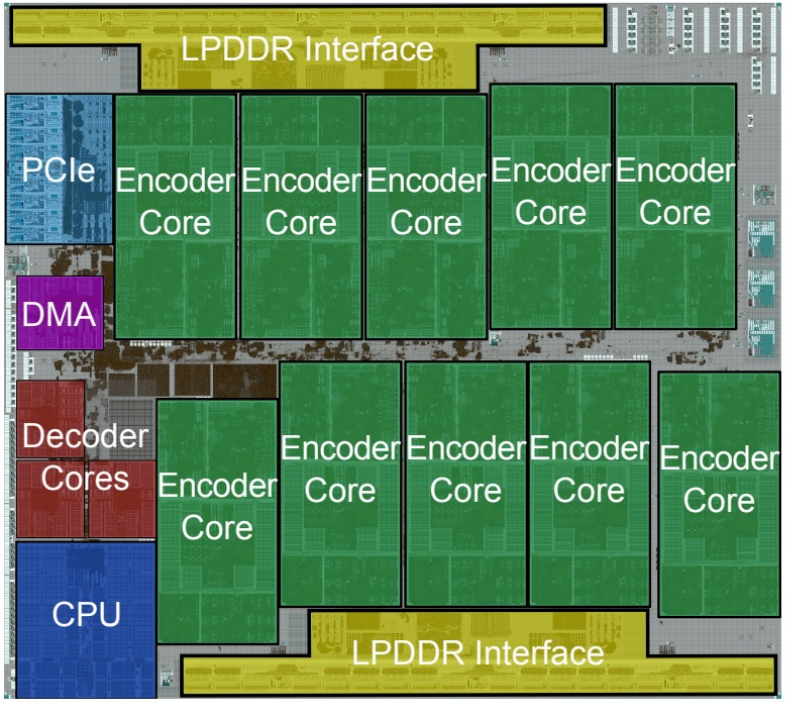
Largest GPU
54.2 Billion transistors
826 mm^2
NVIDIA Ampere GA100

https://www.anandtech.com/show/14758/hot-chips-31-live-blogs-cerebras-wafer-scale-deep-learning

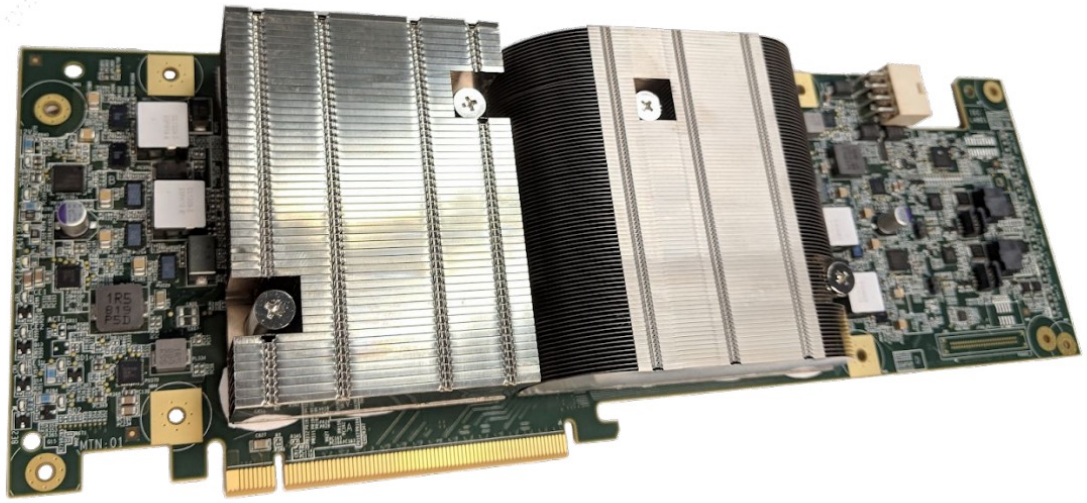
https://www.cerebras.net/cerebras-wafer-scale-engine-why-we-need-big-chips-for-deep-learning/

Google's Video Coding Unit (2021)

Warehouse-Scale Video Acceleration: Co-design and Deployment in the Wild



(a) Chip floorplan



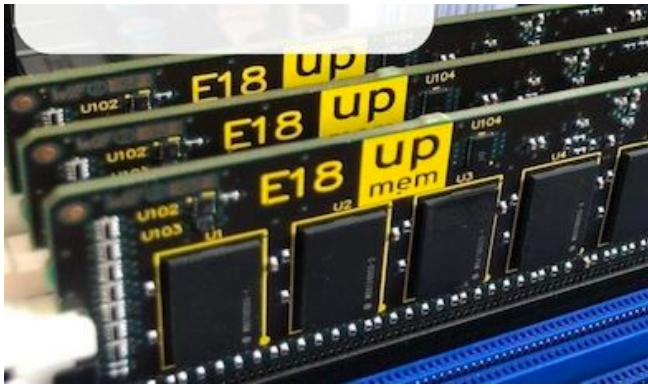
(b) Two chips on a PCBA

Figure 5: Pictures of the VCU

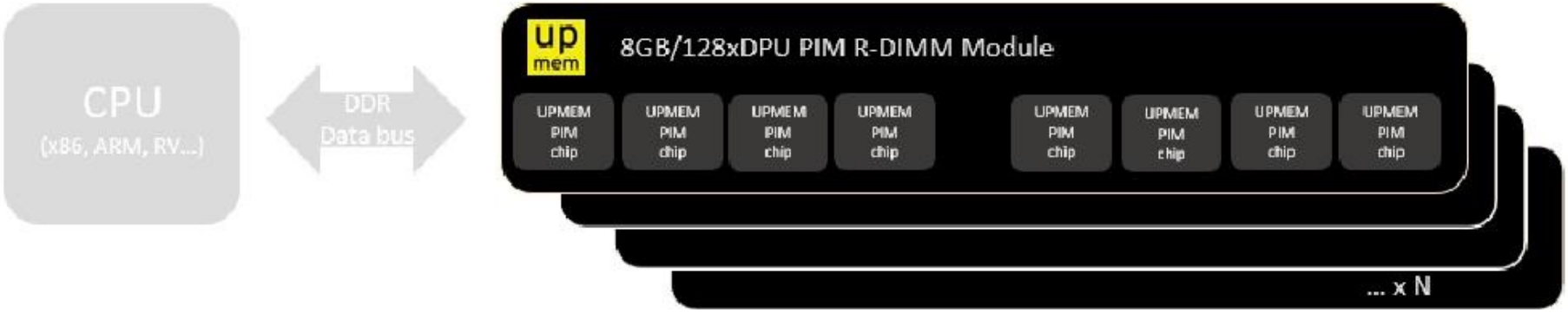
Source: https://dl.acm.org/doi/pdf/10.1145/3445814.3446723

UPMEM Processing-in-DRAM Engine (2019)

- Processing in DRAM Engine
- Includes **standard DIMM modules**, with a **large number of DPU processors** combined with DRAM chips.

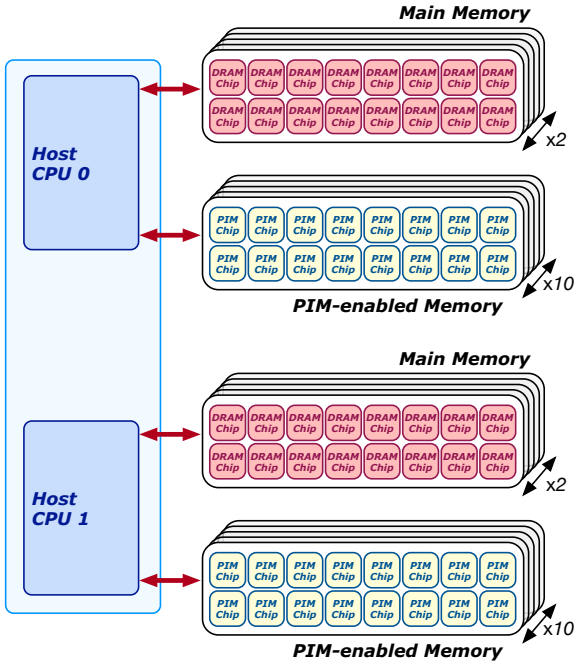


- Replaces **standard** DIMMs
  - DDR4 R-DIMM modules
    - 8GB+128 DPUs (16 PIM chips)
    - Standard 2x-nm DRAM process
  - **Large amounts of** compute & memory bandwidth



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/

Different Platforms, Different Goals



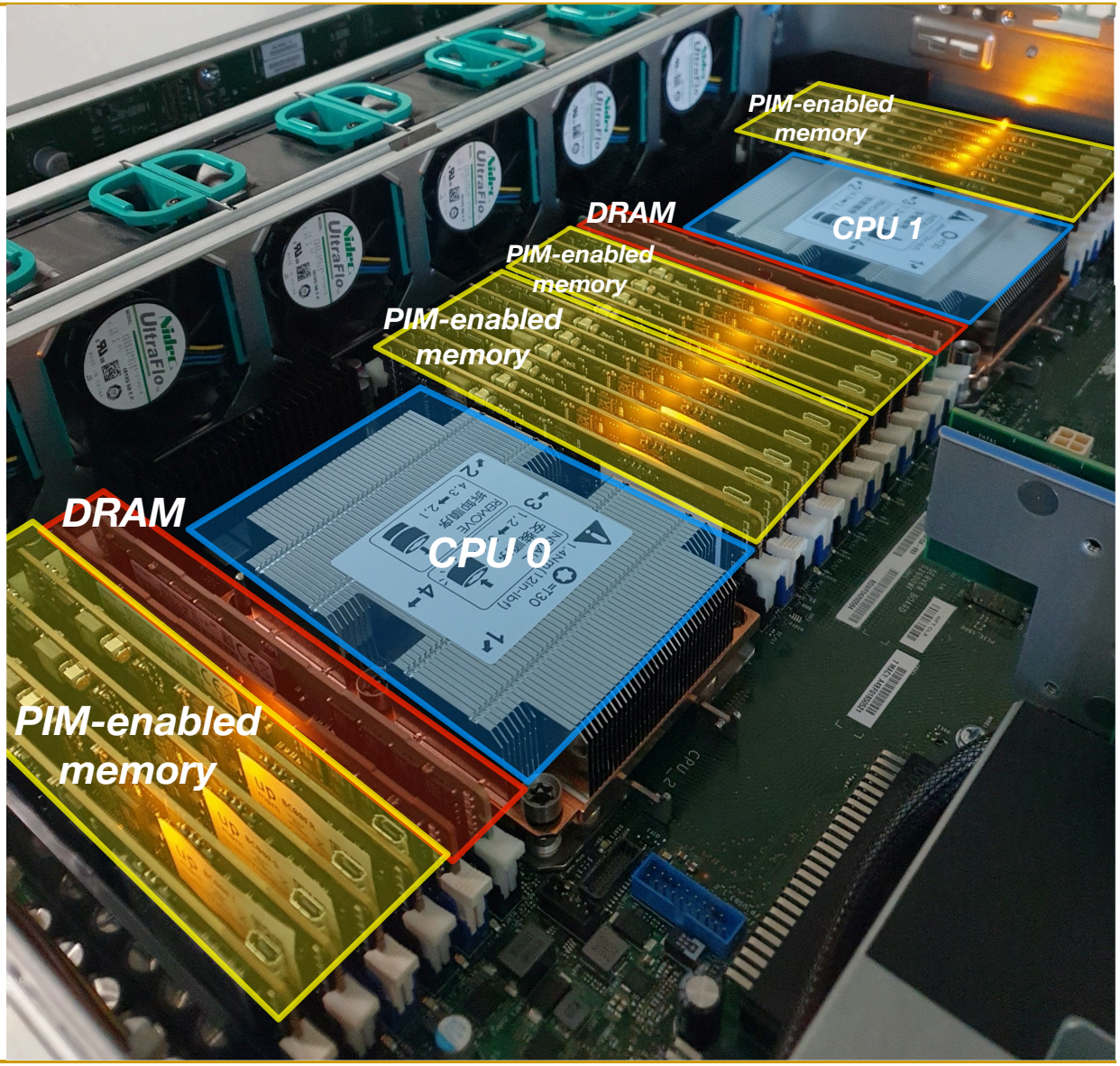
Benchmarking a New Paradigm: An Experimental Analysis of a Real Processing-in-Memory Architecture

JUAN GÓMEZ-LUNA, ETH Zürich, Switzerland
IZZAT EL HAJJ, American University of Beirut, Lebanon
IVAN FERNANDEZ, ETH Zürich, Switzerland and University of Malaga, Spain
CHRISTINA GIANNOULA, ETH Zürich, Switzerland and NTUA, Greece
GERALDO F. OLIVEIRA, ETH Zürich, Switzerland
ONUR MUTLU, ETH Zürich, Switzerland

Many modern workloads, such as neural networks, databases, and graph processing, are fundamentally memory-bound. For such workloads, the data movement between main memory and CPU cores imposes a significant overhead in terms of both latency and energy. A major reason is that this communication happens through a narrow bus with high latency and limited bandwidth, and the low data reuse in memory-bound workloads is insufficient to amortize the cost of main memory access. Fundamentally addressing this *data movement bottleneck* requires a paradigm where the memory system assumes an active role in computing by integrating processing capabilities. This paradigm is known as *processing-in-memory* (*PIM*).

Recent research explores different forms of PIM architectures, motivated by the emergence of new 3D-stacked memory technologies that integrate memory with a logic layer where processing elements can be easily placed. Past works evaluate these architectures in simulation or, at best, with simplified hardware prototypes. In contrast, the UPMEM company has designed and manufactured the first publicly-available real-world PIM architecture. The UPMEM PIM architecture combines traditional DRAM memory arrays with general-purpose in-order cores, called *DRAM Processing Units* (*DPUs*), integrated in the same chip.

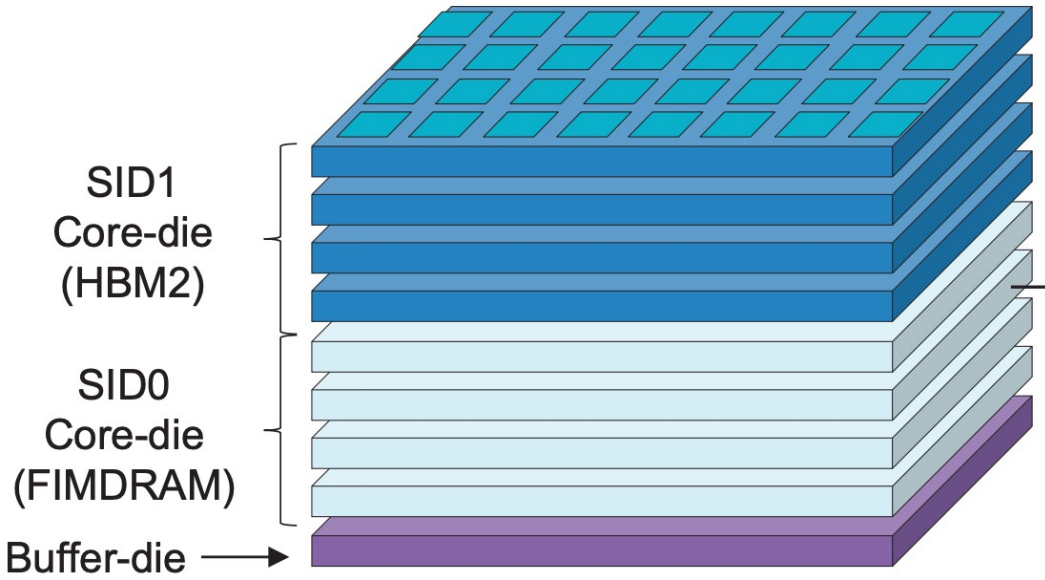
This paper provides the first comprehensive analysis of the first publicly-available real-world PIM architecture. We make two key contributions. First, we conduct an experimental characterization of the UPMEM-based PIM system using microbenchmarks to assess various architecture limits such as compute throughput and memory bandwidth, yielding new insights. Second, we present *PrIM* (*Processing-In-Memory benchmarks*), a benchmark suite of 16 workloads from different application domains (e.g., dense/sparse linear algebra, databases, data analytics, graph processing, neural networks, bioinformatics, image processing), which we identify as memory-bound. We evaluate the performance and scaling characteristics of PrIM benchmarks on the UPMEM PIM architecture, and compare their performance and energy consumption to their state-of-the-art CPU and GPU counterparts. Our extensive evaluation conducted on two real UPMEM-based PIM systems with 640 and 2,556 DPUs provides new insights about suitability of different workloads to the PIM system, programming recommendations for software designers, and suggestions and hints for hardware and architecture designers of future PIM systems.



https://arxiv.org/pdf/2105.03814.pdf

Samsung Function-in-Memory DRAM (2021)

- FIMDRAM based on HBM2



[3D Chip Structure of HBM with FIMDRAM]

Chip Specification
128DQ / 8CH / 16 banks / BL4
32 PCU blocks (1 FIM block/2 banks)
1.2 TFLOPS (4H)
FP16 ADD / Multiply (MUL) / Multiply-Accumulate (MAC) / Multiply-and- Add (MAD)

ISSCC 2021 / SESSION 25 / DRAM / 25.4

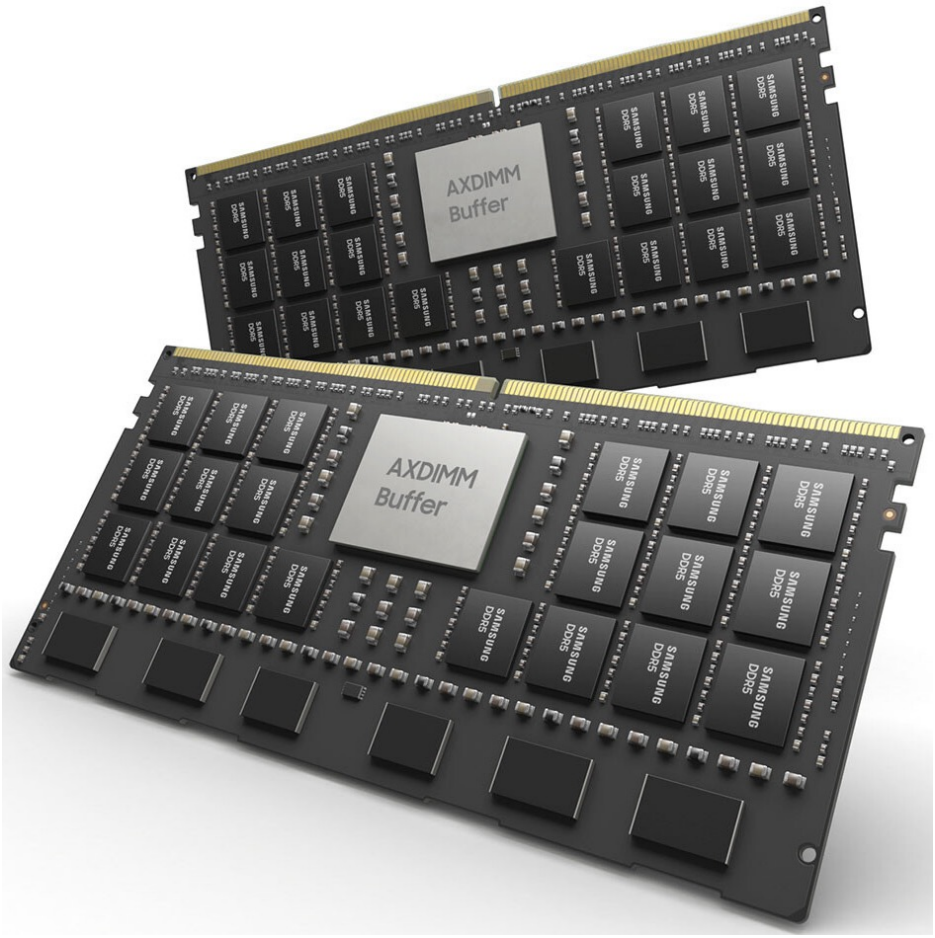
25.4 A 20nm 6GB Function-In-Memory DRAM, Based on HBM2 with a 1.2TFLOPS Programmable Computing Unit Using Bank-Level Parallelism, for Machine Learning Applications

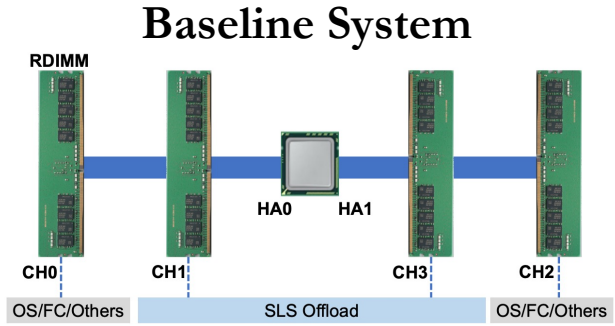
Young-Cheon Kwon[1], Suk Han Lee[1], Jaehoon Lee[1], Sang-Hyuk Kwon[1], Je Min Ryu[1], Jong-Pil Son[1], Seongil O[1], Hak-Soo Yu[1], Haesuk Lee[1], Soo Young Kim[1], Youngmin Cho[1], Jin Guk Kim[1], Jongyoon Choi[1], Hyun-Sung Shin[1], Jin Kim[1], BengSeng Phuah[1], HyoungMin Kim[1], Myeong Jun Song[1], Ahn Choi[1], Daeho Kim[1], SooYoung Kim[1], Eun-Bong Kim[1], David Wang[2], Shinhaeng Kang[1], Yuhwan Ro[3], Seungwoo Seo[3], JoonHo Song[3], Jaeyoun Youn[1], Kyomin Sohn[1], Nam Sung Kim[1]

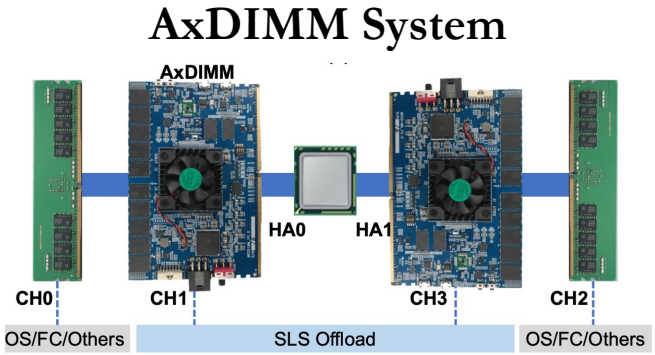
[1]Samsung Electronics, Hwaseong, Korea
[2]Samsung Electronics, San Jose, CA
[3]Samsung Electronics, Suwon, Korea

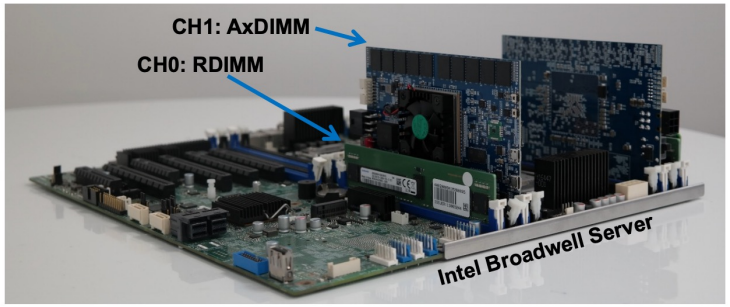
Samsung AxDIMM (2021)

- DDRx-PIM
  - Deep learning recommendation system









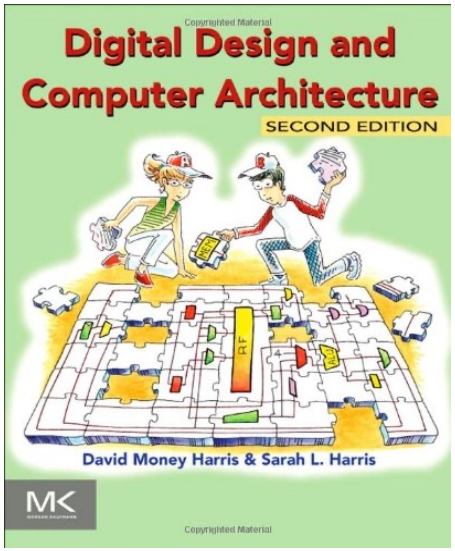
Ke et al. "Near-Memory Processing in Action: Accelerating Personalized Recommendation with AxDIMM", IEEE Micro (2021)

Basic Building Blocks

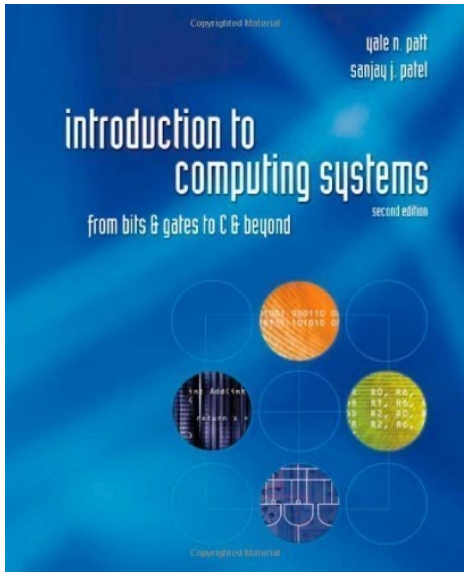
- Electrons
- Transistors
- Logic Gates
- Combinational Logic Circuits
- Sequential Logic Circuits
  - Storage Elements and Memory
- …
- Cores
- Caches
- Interconnect
- Memories
- …

Reading Assignments for This Week

- Chapter 1 in Harris & Harris



- Chapters 1-2 in Patt and Patel



- Supplementary Lecture Slides on Binary Numbers

Major High-Level Goals of This Course

- In Digital Circuits & Computer Architecture

- Understand the basics
- Understand the principles (of design)
- Understand the precedents

- Based on such understanding:
  - learn how a modern computer works underneath
  - evaluate tradeoffs of different designs and ideas
  - implement a principled design (a simple microprocessor)
  - learn to systematically debug increasingly complex systems
  - Hopefully enable you to develop novel, out-of-the-box designs

- The focus is on basics, principles, precedents, and how to use them to create/implement good designs

Why These Goals?

- Because you are here for a Computer Science degree

- Regardless of your future direction, learning the principles of digital design & computer architecture will be useful to
  - design better hardware
  - design better software
  - design better systems
  - make better tradeoffs in design
  - understand why computers behave the way they do
  - solve problems better
  - think "in parallel"
  - think critically
  - …

Course Info and Logistics

Brief Self Introduction

- Onur Mutlu
  - Full Professor @ ETH Zurich ITET (INFK), since Sept 2015
  - Strecker Professor @ Carnegie Mellon University ECE (CS), 2009-2016, 2016-…
  - Started the Comp Arch Research Group @ Microsoft Research, 2006-2009
  - Worked @ Google, VMware, Microsoft Research, Intel, AMD
  - PhD in Computer Engineering from University of Texas at Austin in 2006
  - BS in Computer Engineering & Psychology from University of Michigan in 2000
  - https://people.inf.ethz.ch/omutlu/ omutlu@gmail.com

- Research and Teaching in:
  - **Computer architecture, systems, hardware security, bioinformatics**
  - Memory and storage systems
  - Robust & dependable hardware systems: security, safety, predictability, reliability
  - Hardware/software cooperation
  - New computing paradigms; architectures with emerging technologies/devices
  - Architectures for bioinformatics, genomics, health, medicine, AI/ML
  - …

Course Info: Lecturers & PhD Assistants

- Head Assistant
  - Dr. Juan Gómez Luna
- Vice-Head Assistants
  - Dr. Mohammad Sadrosadati
  - Hasan Hassan
  - Ataberk Olgun
- Lecturer
  - Dr. Frank Gurkaynak
- (Other) Key Assistants and Guest Lecturers
  - Dr. Jisung Park
  - Dr. Mohammed Alser
  - Dr. Gagandeep Singh

Course Info: PhD Assistants

- (Other) Key Assistants and Guest Lecturers (cont.)
  - Dr. Behzad Salami
  - Giray Yaglikci
  - Can Firtina
  - Geraldo De Oliveira Junior
  - Rahul Bera
  - Konstantinos Kanellopoulos
  - Nika Mansouri Ghiasi
  - Rakesh Nadig
  - Joel Lindegger

Course Info: Student Assistants

- Joao Dinis Ferreira
- Roberto Starc
- Aditya Manglik
- Banu Cavlak
- Haocong Luo
- Lukas Gygi
- Marc Rettenbacher
- Amos Herz
- Yumi Kim
- Marie-Louise Dugua
- Leander Diaz-Bone
- Cashen Adkins
- Alexander Schlieper

Course Info: Lab Assistants (I)

- Tuesday 16-18
  - TBD
- Wednesday 16-18
  - TBD

Course Info: Lab Assistants (II)

- Friday 8-10
  - TBD
- Friday 10-12
  - TBD

If You Need Help

- Post your question on Moodle Q&A Forum
  - https://moodle-app2.let.ethz.ch/course/view.php?id=16852
  - We will create a forum on Moodle for each activity
  - **Preferred** for **technical** questions

- Write an e-mail to:
  - digitaltechnik@lists.inf.ethz.ch
  - The instructor and all assistants will receive this e-mail

- Come to office hours
  - We will provide office locations & Zoom links
  - TBD

Where to Get Up-to-date Course Info?

- Website:
  - **https://safari.ethz.ch/digitaltechnik/spring2022/**
  - Lecture slides and videos
  - Readings
  - Lab information
  - Course schedule, handouts, FAQs
  - Software
  - Plus other useful information for the course
  - Check frequently for announcements and due dates
  - This is your single point of access to all resources
- Your ETH Email
- Lecturers and Teaching Assistants

Lecture and Lab Times and Policies

- Lectures:
  - Thursday and Fridays, 14:00-16:00
  - YouTube livestream playlist: https://youtube.com/playlist?list=PL5Q2soXY2Zi97Ya5DEUpMpO2bbAoaG7c6
  - Zoom link provided via Moodle
  - Attendance is for your benefit and is therefore important
  - Some days, we may have guest lectures and exercise sessions
- Lab sessions:
  - See online
  - You should definitely attend the lab sessions
    - In-class evaluation (70%) and mandatory lab reports (30%)
  - Labs will start on March 8th
  - Lab information and handouts are here:
    - https://safari.ethz.ch/digitaltechnik/spring2022/doku.php?id=labs

Lab Organization (I)

- Groups
  - Choose your preferred group in Moodle
    - https://moodle-app2.let.ethz.ch/mod/choicegroup/view.php?id=716892
    - Due 02.03.2022 at 11:59pm
  - Choose your partner
    - https://moodle-app2.let.ethz.ch/mod/feedback/view.php?id=716899
    - Due 03.03.2022 at 11:59pm
  - Choose onsite or online
    - https://moodle-app2.let.ethz.ch/mod/choice/view.php?id=716917
    - Due 03.03.2022 at 11:59pm

Lab Organization (II)

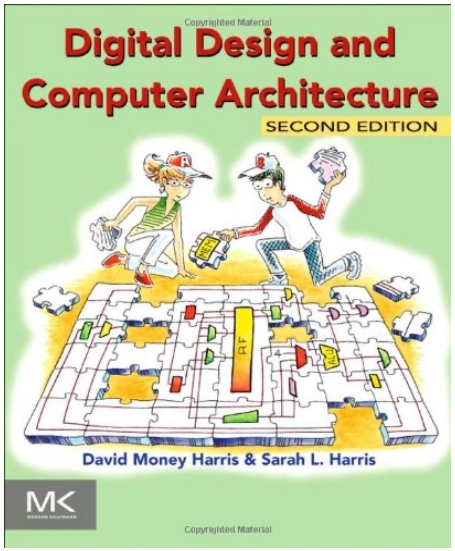
- Lab grades from previous years
  - https://moodle-app2.let.ethz.ch/mod/choice/view.php?id=716908
  - Choose one of the options (due 28.02.2022 at 11:59pm):
    - 1) I will use my lab grades from previous years, and I won't do the labs this year
    - 2) I will use my lab grades from previous years, but I will do the labs this year
    - 3) I won't use my lab grades from previous years. I will do the labs this year

Final Exam

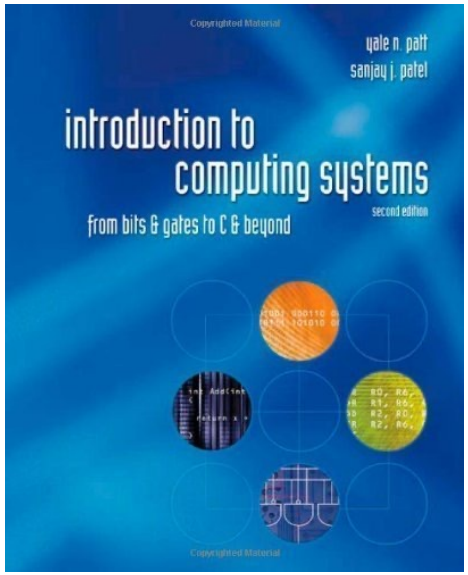
- 180-minute written exam
  - Find examination rules in Course Catalogue
    - http://www.vvz.ethz.ch/Vorlesungsverzeichnis/lerneinheit.view?semkez=2022S&ansicht=LEISTUNGSKONTROLLE&lerneinheitId=159117&lang=en
  - Also: study the first pages of previous exams
    - https://safari.ethz.ch/digitaltechnik/spring2022/doku.php?id=exams
  - Some exam questions are similar to questions in Optional HWs and Past Exams
    - Optional HWs are not graded, but highly recommended to solve
    - Solving past exams could also be useful

Reading Assignments for This Week

- Chapter 1 in Harris & Harris



- Chapters 1-2 in Patt and Patel



- Supplementary Lecture Slides on Binary Numbers

Reading Assignments for Next Week

- Combinational Logic chapters from both books
  - Patt and Patel, Chapter 3
  - Harris and Harris, Chapter 2

- Check course website for all future readings
  - Required
  - Recommended
  - Mentioned

Future Lectures and Assignments

- You can also anticipate (and study) future lectures and assignments based on Spring 2021 schedule:
  - https://safari.ethz.ch/digitaltechnik/spring2021/doku.php?id=schedule
  - https://www.youtube.com/playlist?list=PL5Q2soXY2Zi_uej3aY39YB5pfW4SJ7LlN

- An example of "Last Time Prediction"
  - Speculative Execution
    - The concept of doing something before knowing it is needed
  - A key concept we will cover in the design of microprocessors

Digital Design & Computer Arch.

Lecture 2a: Tradeoffs, Metrics, Mindset

Prof. Onur Mutlu

ETH Zürich

Spring 2022

25 February 2022