EE/CS277: Data-Centric Computer Architecture

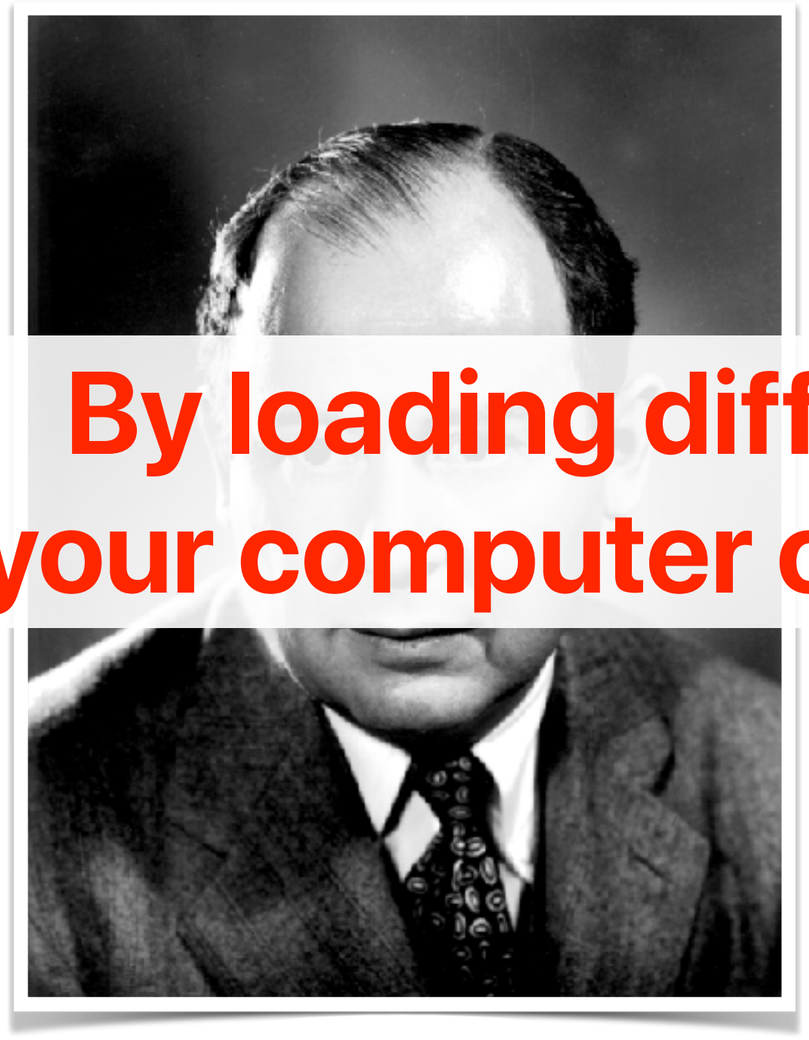
Hung-Wei Tseng

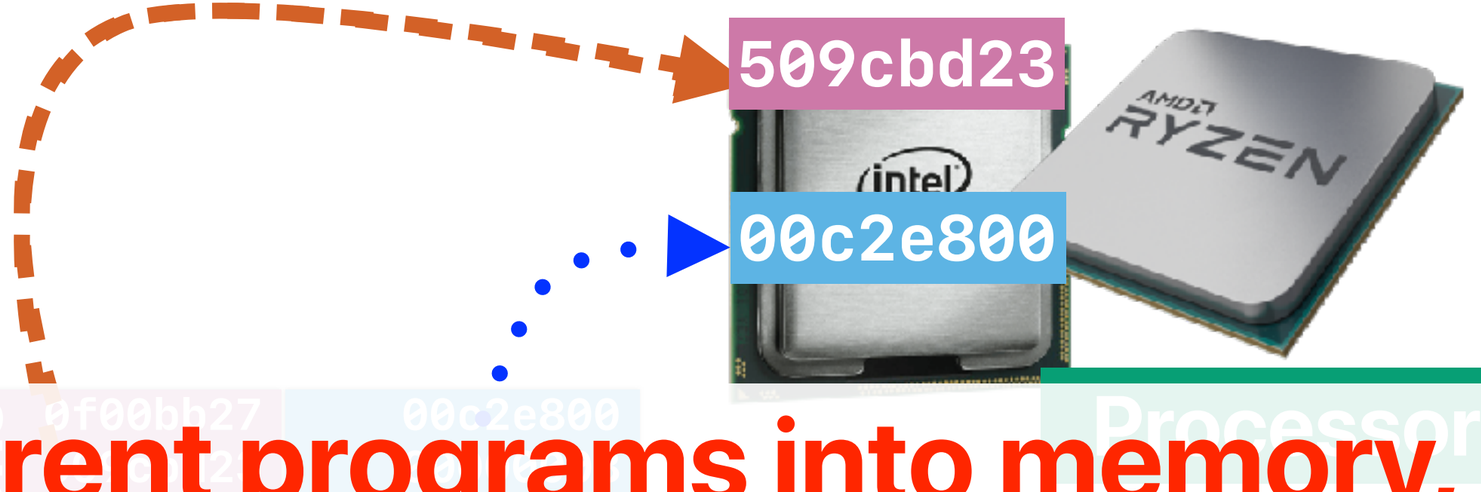
Brief introduce yourself & your new year resolution?

3:00

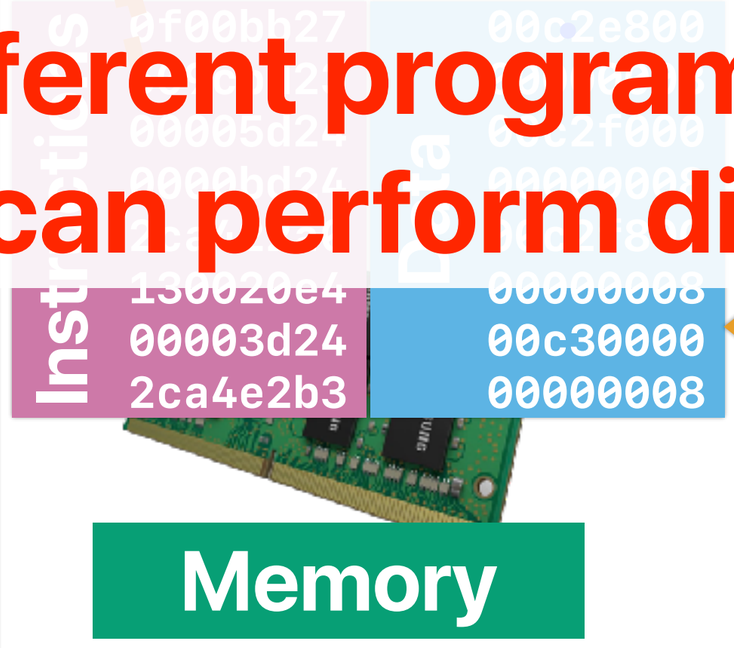
How does a computer process & store data?

von Neumman Architecture



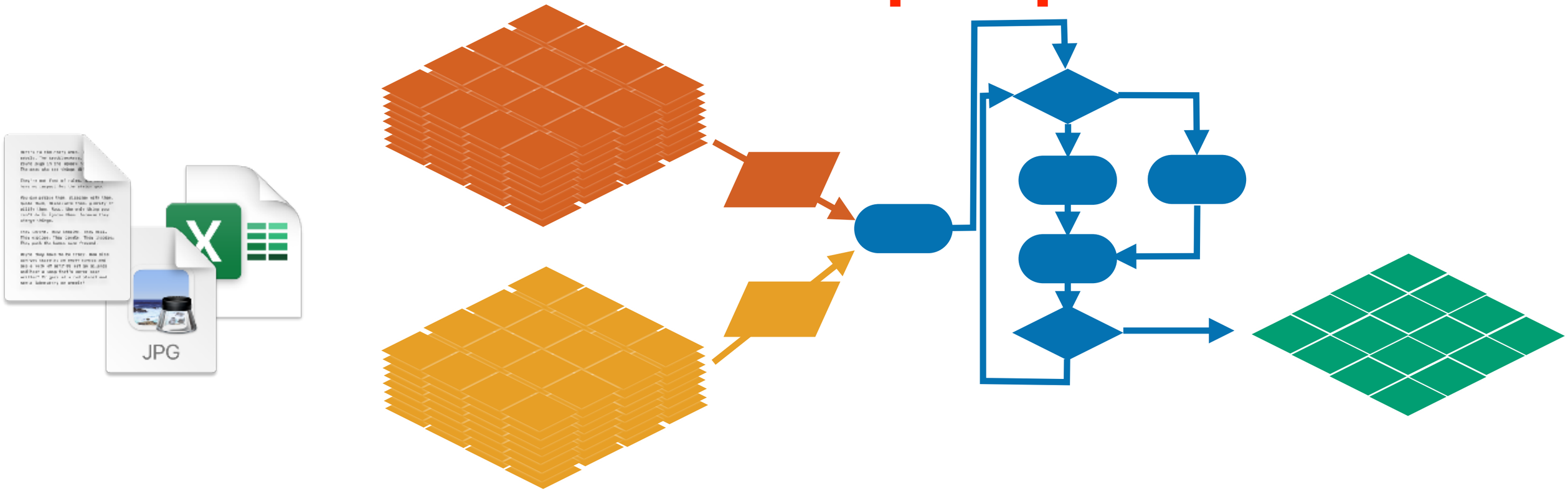


By loading different programs into memory, your computer can perform different functions



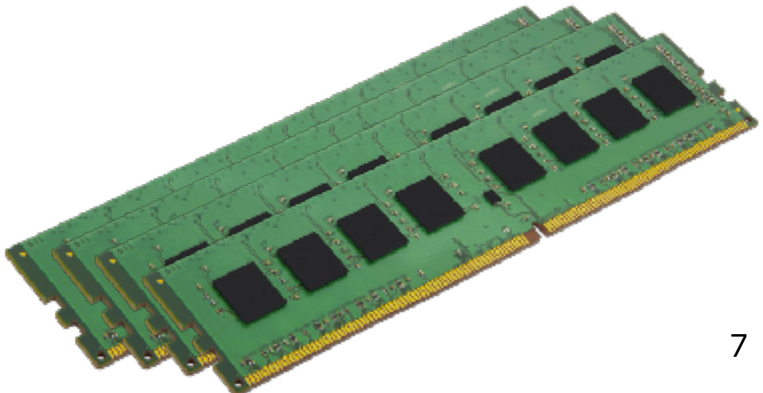
Let’s take a look from a program

From the software perspective

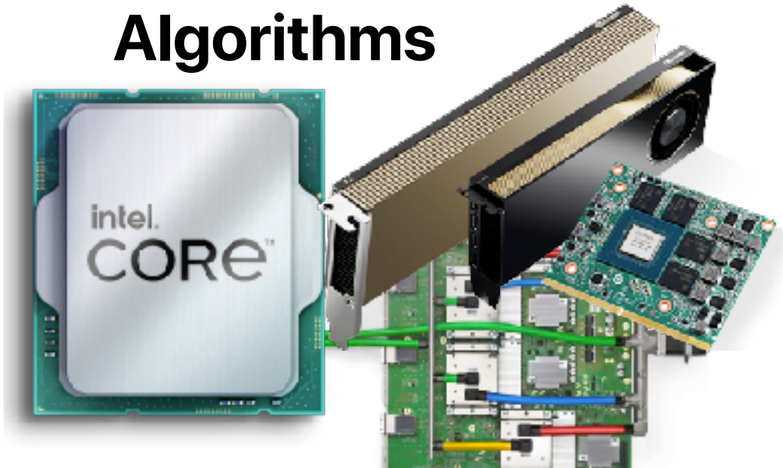


Source “data"

“Data” structures



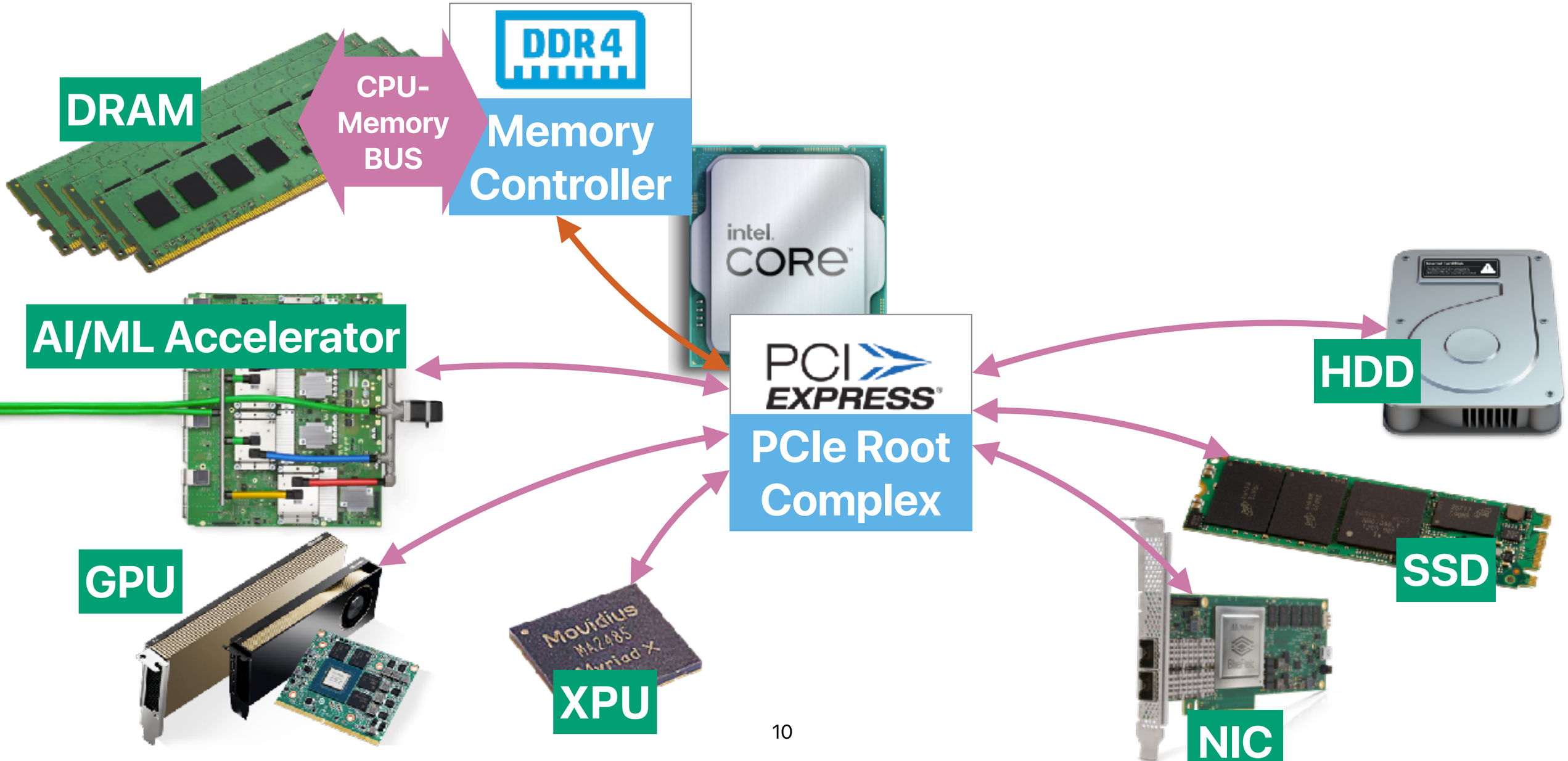
Algorithms



Result

What's the "center" of current computer architectures?

The “real” datapath

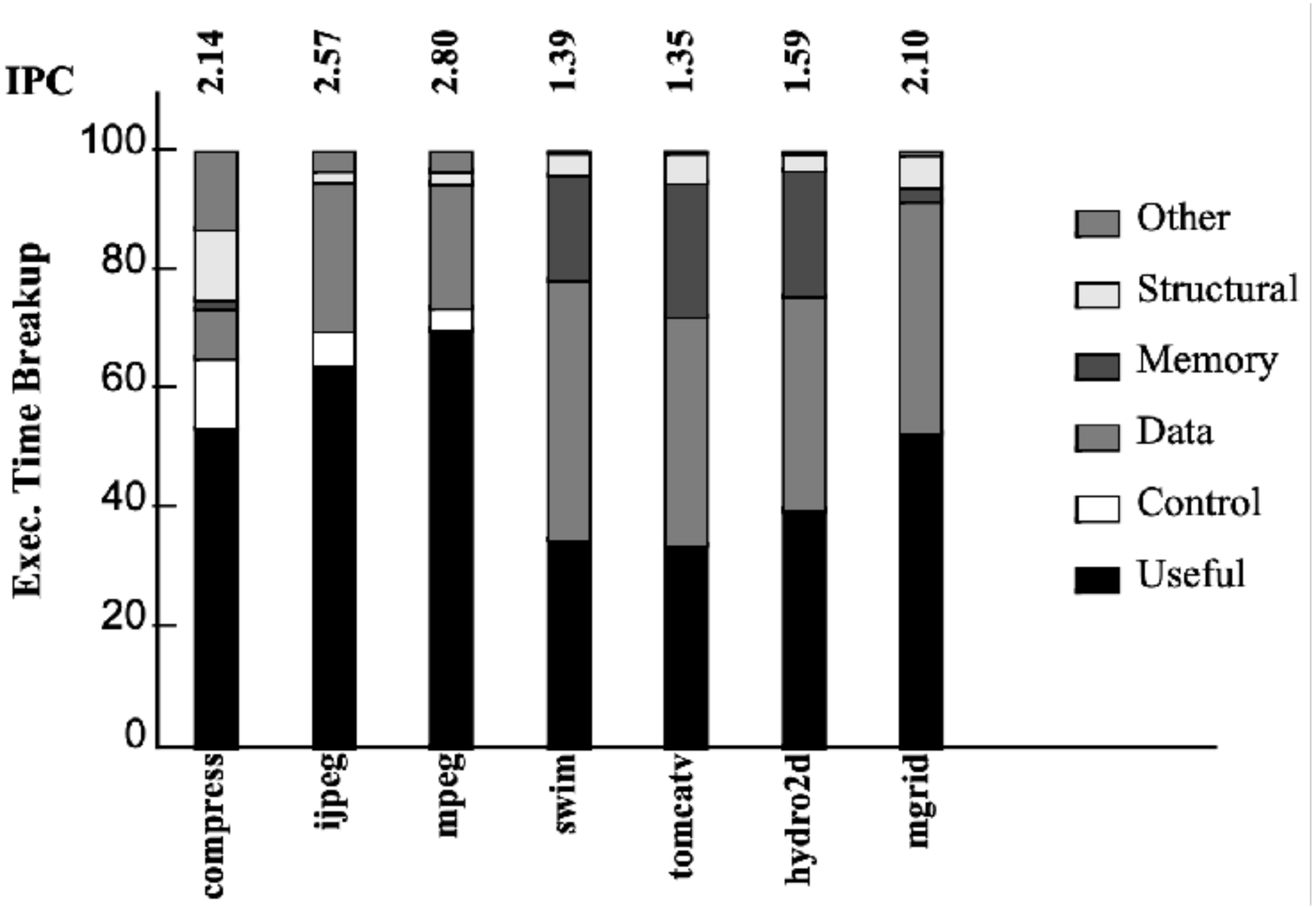


The current architecture is “CPU-centric”

3:00

Is it an “efficient” computing model?

Why is traditional architecture “CPU-centric”



V. Krishnan and J. Torrellas, "A direct-execution framework for fast and accurate simulation of superscalar processors," Proceedings. 1998 International Conference on Parallel Architectures and Compilation Techniques

The capability of CPU's PCIe root complex

Essentials

Export sp

Product Collection	5th Generation Intel® Xeon® Scalable Processors
Code Name	Products formerly SAPPHIRE RAPIDS
Vertical Segment	Server
Processor Number	4510
Lithography	Intel 7
Use Conditions	Server/Enterprise

Expansion Options

Scalability	2S
PCI Express Revision	5
Max # of PCI Express Lanes	80

Essentials

Download Specific

Product Collection	13th Generation Intel® Core™ i9 Processors
Code Name	Products formerly Raptor Lake
Vertical Segment	Desktop
Processor Number	i9-13900K
Lithography	Intel 7
Use Conditions	PC/Client/Tablet, Workstation

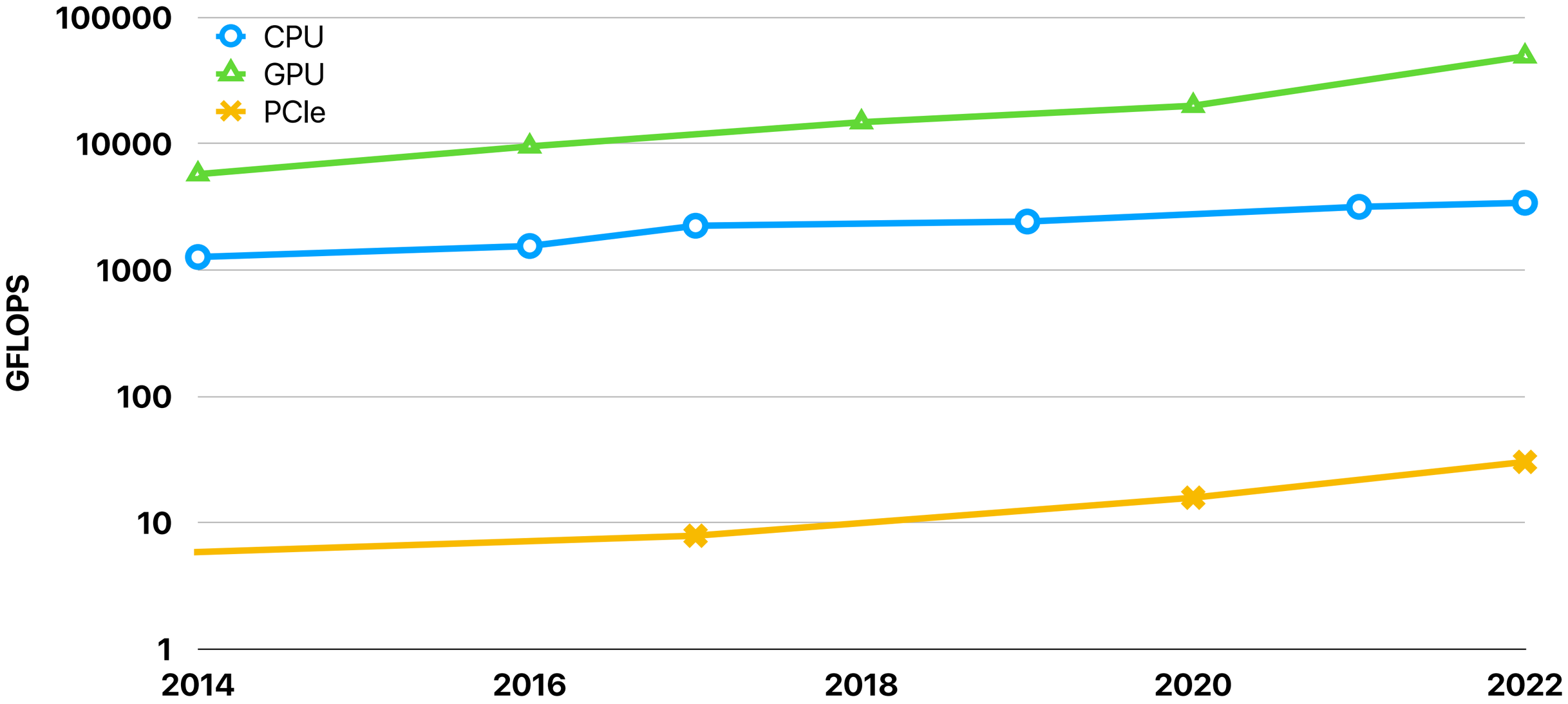
Expansion Options

Direct Media Interface (DMI) Revision	4.0
Max # of DMI Lanes	8
Scalability	1S Only
PCI Express Revision	5.0 and 4.0
PCI Express Configurations	Up to 1x16+4, 2x8+4
Max # of PCI Express Lanes	20

AMD Ryzen™ 7 7700X

Native SATA Ports:	0
PCI Express® Version:	PCIe® 5.0
Native PCIe® Lanes (Total/Usable):	28 / 24

The “speed” of PCIe compared to computing



大改造!! 劇的ビフォーアフター

SEASON II

That'S a DRAMATIC CHANGE!

毎週日曜ごご4:00～5:00放送


Living room centric

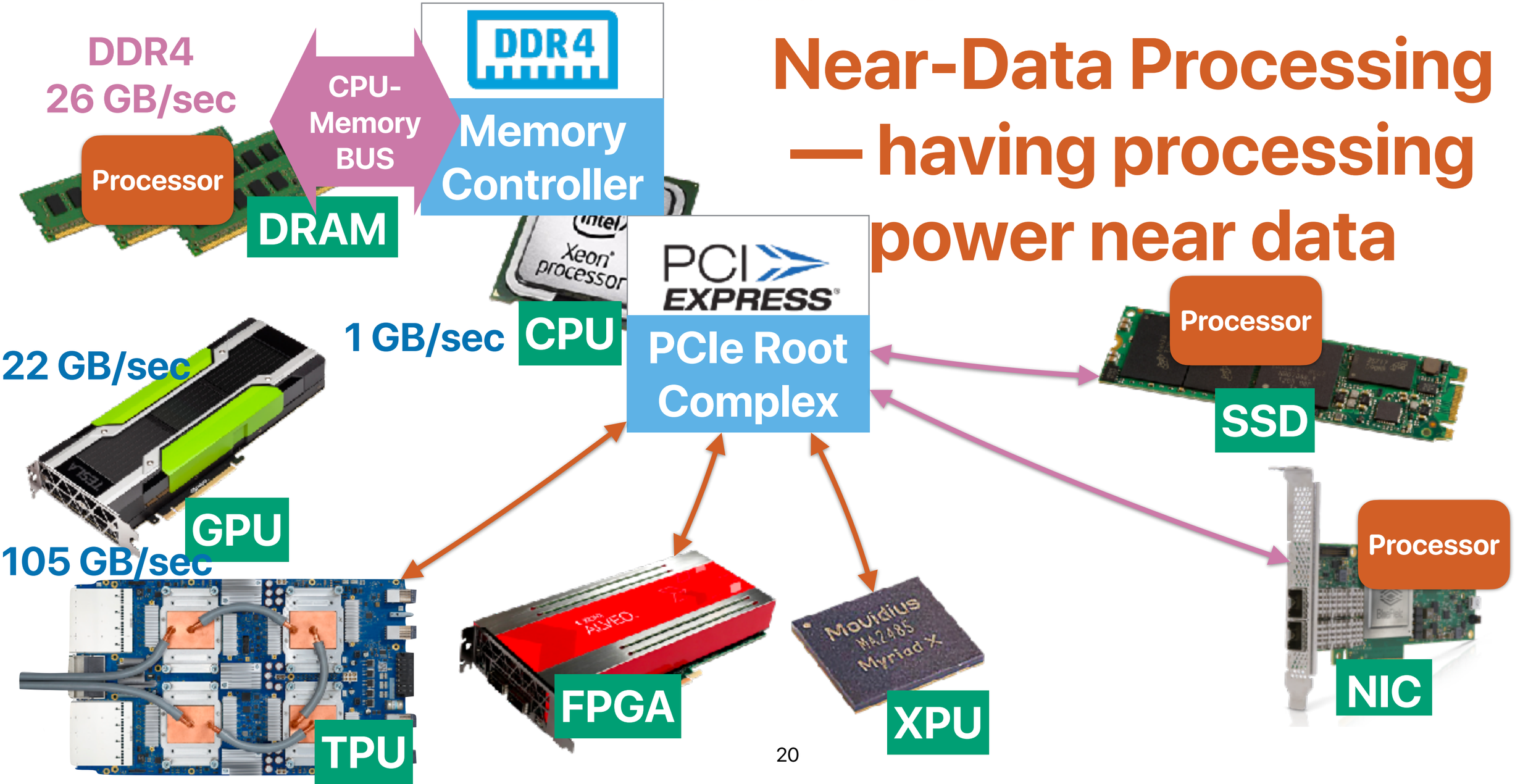

Kitchen-centric

3:00

What will you do differently if the application spends more time interacting with data/storage/memory?

What if …

Near-Data Processing — having processing power near data



Should every computer be data-centric?

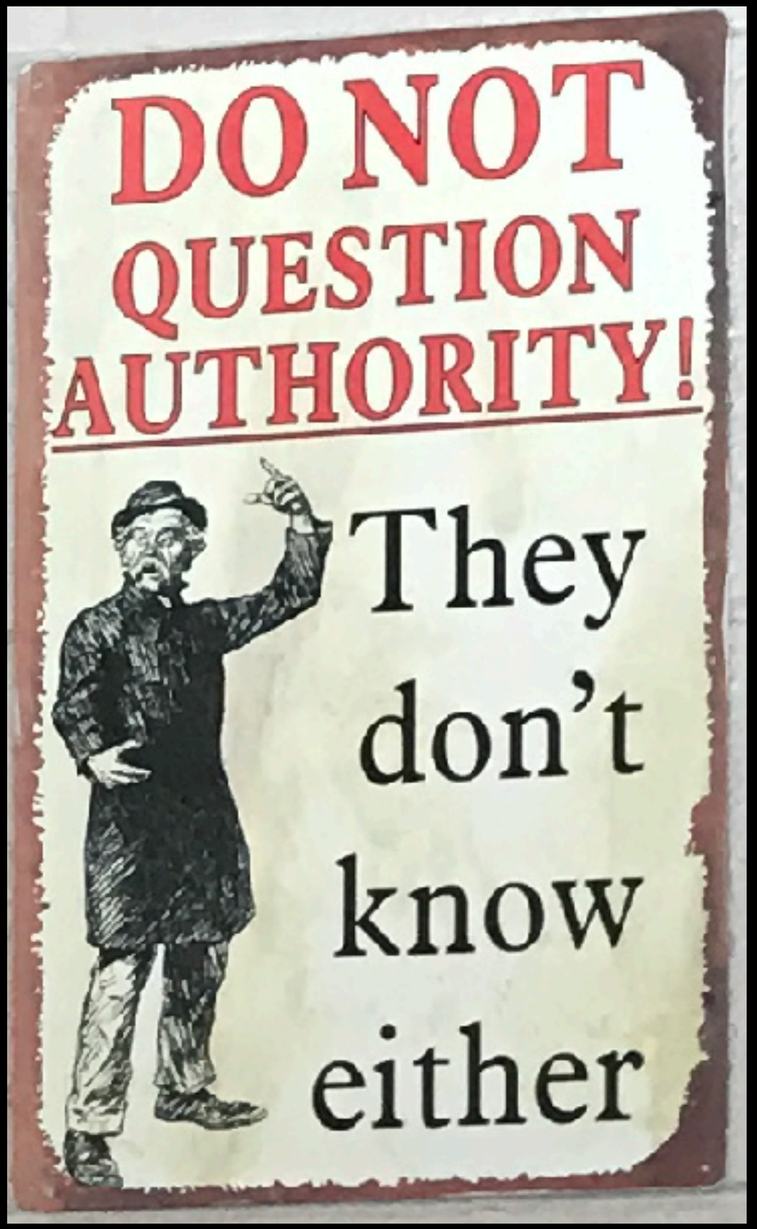
It depends

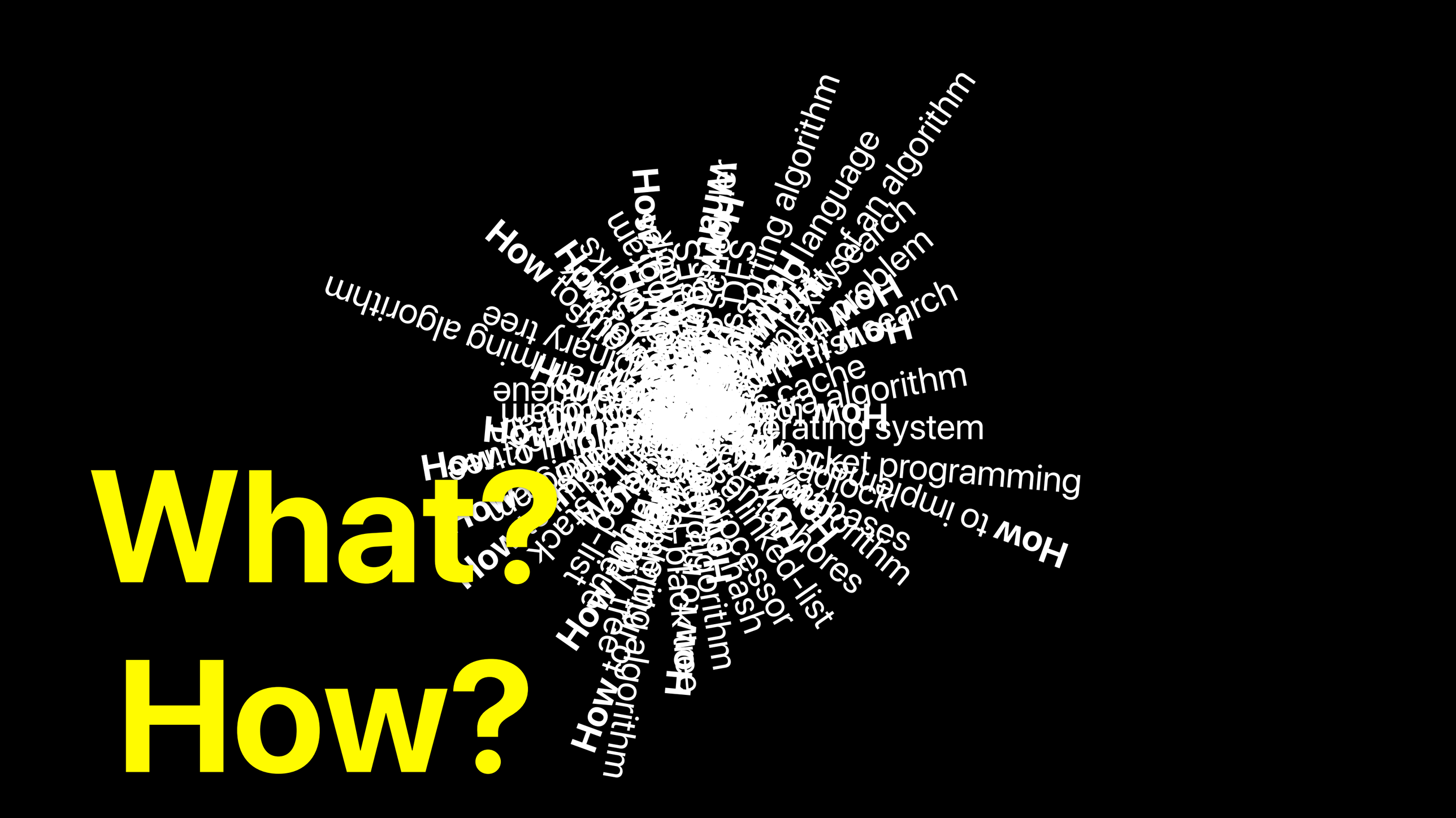
But how do I know what should do when it's "it depends"?

Do I really fully utilize the hardware? How to make more efficient use of them?

Learning eXperience

Why?
What?
How?




What?
How?

Lecture

Why?

What?

How?

Project

Papers give you insights!

- Papers contain **design principles** that are missing in your textbook or online documents
  - You can apply these design principles and the skills of analyzing these principles to anywhere
  - They provide case-studies to those “it depends” answers
- You can learn those **whys** for those proposed work
- https://forms.gle/cZyvnctiaFB9Rsmv8

Industry cares

[redacted]@intel.com> 2011/2/15

寄給 h1tseng

Hi Hung-Wei,

I am very interested in your topic you presented yesterday. If possible, may I get a copy of

Best Regards,

[redacted]

[redacted]freescale.com 透過 cs.ucsd.edu 2012/1/10

寄給 h1tseng

Hung-Wei

I just finished reading your paper "Understanding the Impact of Power Loss on Flash Memory", very interesting information, do you have a PowerPoint presentation that goes along this paper?

[redacted]fb.com> 2019年7月9日 週二 下午2:12

寄給 Hung-Wei [redacted]

Hung-wei

[redacted]

Given we are also working on in-memory and near-memory computing at my Boston team, I would like to see how do we work more closely to churn out even more useful results and applications for Facebook's ML models/workloads in both datacenters and edge devices and instigate new research directions.

[redacted]

[redacted]sap.com 透過 cs.ucsd.edu 2012/11/12

寄給 h1tseng

Hi Tseng,

I have read your paper titled "Understanding the Impact of Power Loss on Flash Memory". It work. I would like to understand what specific tools did you use to observe the page-read and the FTL level. Did you use some sort of Flash simulator to get all the statistics about the num and the energy consumption? My second question would be regarding FTL algorithms. Did y real SSD or you used some kind of simulator and simulated the FTL algorithm?

Thanks.

[redacted]

SAP Research

[redacted]@huawei.com> 2016/6/24

寄給 h1tseng

Hi, Hung-Wei,

[redacted] from Huawei, and I am impressed by your ISCA 2016 presentation in Seoul. Near-data processing in ssds may be a promising solution for future data centers. Would you mind sending me your slides presented in the conference? I really appreciate your kindness. Thank you!

Best regards,

[redacted]

Technical perspectives

- Performance analysis
  - Evaluate the maximum potential of applications/algorithms using the roofline model
  - Design the “right”/most efficient architecture for your applications
  - Maximize the utilization of your system resources
- Emerging computer architectures
  - Heterogeneous computer architectures
  - Hardware accelerators
  - Near-data processing
- Performance programming
  - New programming models (e.g., TF Lite!)
  - New programming platforms (e.g., smartSSDs)
- Efficient data storage
  - Data representations
  - Storage system architectures

Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

Soft skills

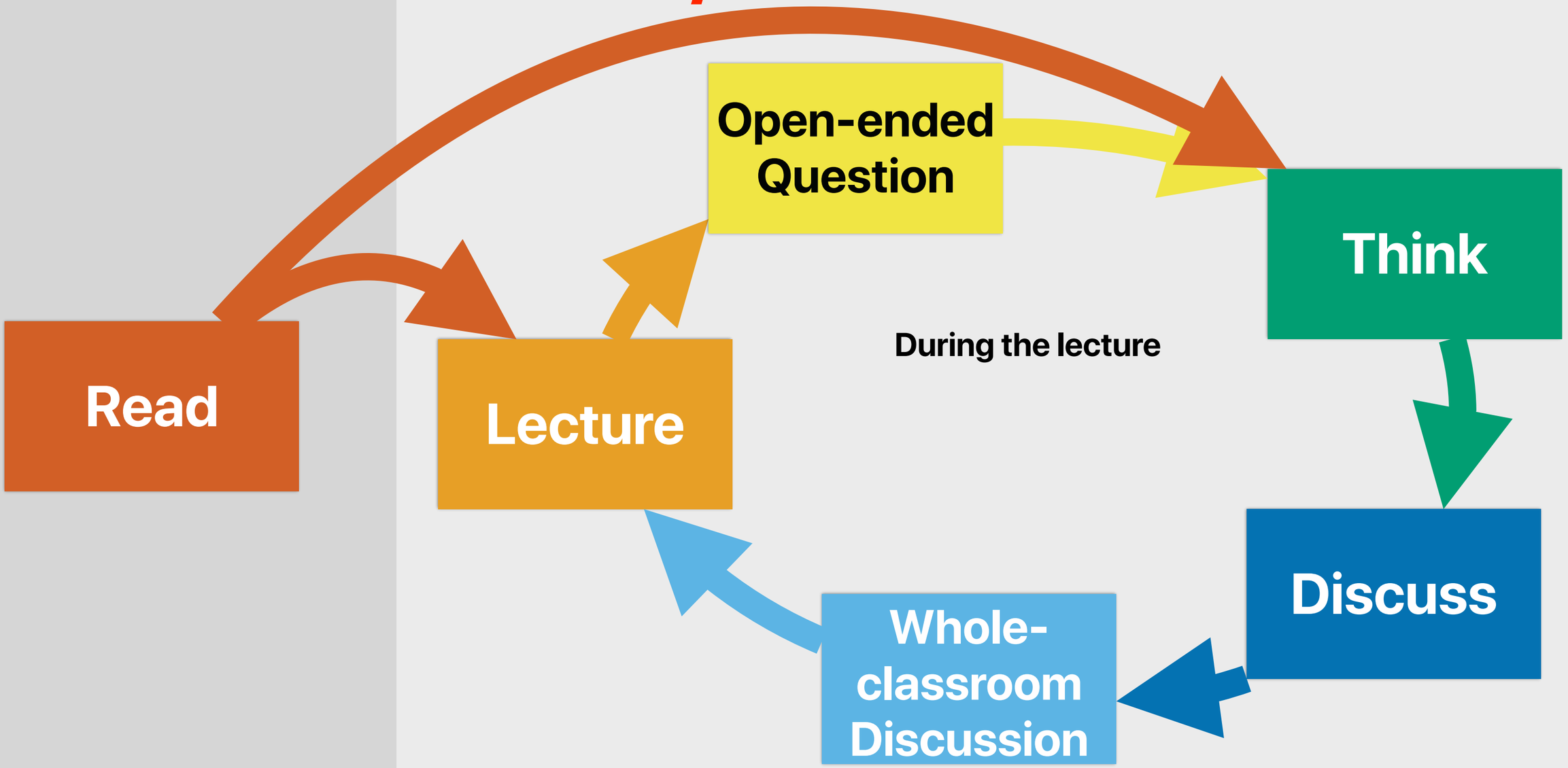
- Identifying “big” problems in our community
- Reading papers and pick up new ideas efficiently
- Coming up with new research ideas
- Learning the way of drafting a solution to an idea
- Working in a group

Traditional lectures


Me
Peer Instructions
You


What kind of show is ours?

EE/CS277


Read
Lecture
Open-ended Question
Think
During the lecture
Discuss
Whole-classroom Discussion

Your tasks — 50%

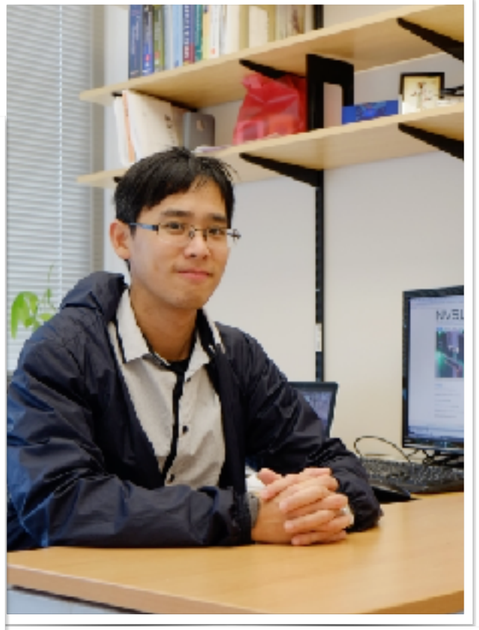
- Read papers!
  - We will read papers reflecting the most trending data-centric computing
  - I will try my best to teach you how to read papers quickly
  - I will try my best to lead you come up with research ideas naturally
  - Fill out this form with the “why”s and what”s behind papers https://forms.gle/cZyvnctiaFB9Rsmv8
- Discuss
  - Participate in the discussion questions during the lectures

Your tasks (cont.) — 50%

- Project ideas
  - Accelerating applications through AI/ML accelerators
  - Accelerating applications through intelligent storage devices
  - Accelerating applications through innovative parallel programming models that hardware accelerators enable
  - Anything related to what we discussed in this class!
- Working on escalab.org/datahub
  - Containerized environment with TF, TF Lite, CUDA
  - GPUs, Edge TPUs, smartSSD available
- Presentations
  - One on the 4th week to present your project ideas
  - The other on the 10th week to present your current progress/outcome

Instructor — Hung-Wei Tseng

- Associate Professor @ UC Riverside, 05/2019—
- Website: https://intra.engr.ucr.edu/~htseng/
- E-mail: cs203ucr@googlegroups.com
- Visiting Researcher @ Google, 01/2023—03/2023
  - Working for TensorFlow Lite
- PhD in **Computer Science**, University of California, San Diego, 2014
- Research Interests
  - General-purpose computing on AI/ML/NN accelerators
  - Intelligent storage devices & near-data processing
  - Or anything else fun — we have an OpenUVR project recently
- Fun fact: Hung-Wei was once considering a career path as a singer but went back to academia due to the unsuccessful trial



Course materials

- Course webpage
  - https://www.escalab.org/classes/eecs277-2024wi/
- Discussion
  - Google Space — you should be invited
- Recording
  - Prof Usagi's Youtube Channel — https://www.youtube.com/profusagi

Let’s act today

- Form your group!
  - At most 4
  - Discuss your research interests
  - Having our first milestone meeting between you and I next week
- Check out our Google Spaces
- Read two papers listed on the website and submit your paper summary

Electrical
Computer Science
Engineering

277

つづく