EE260: Trends in Computer Engineering

Hung-Wei Tseng

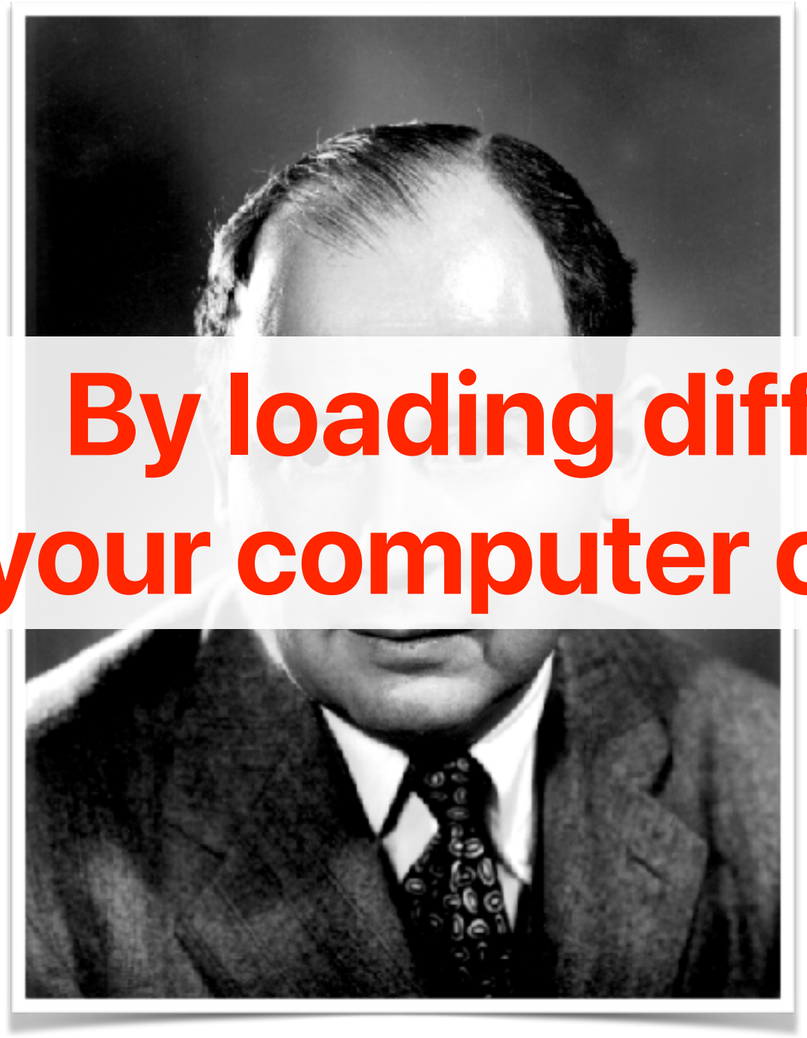
EE260: Let’s say something!

What’s your name?

What you interested in research

Why’re you interested in EE260?

The basics: von Neumman Architecture



509cbd23

00c2e800

Processor

Program

Instruction

Data

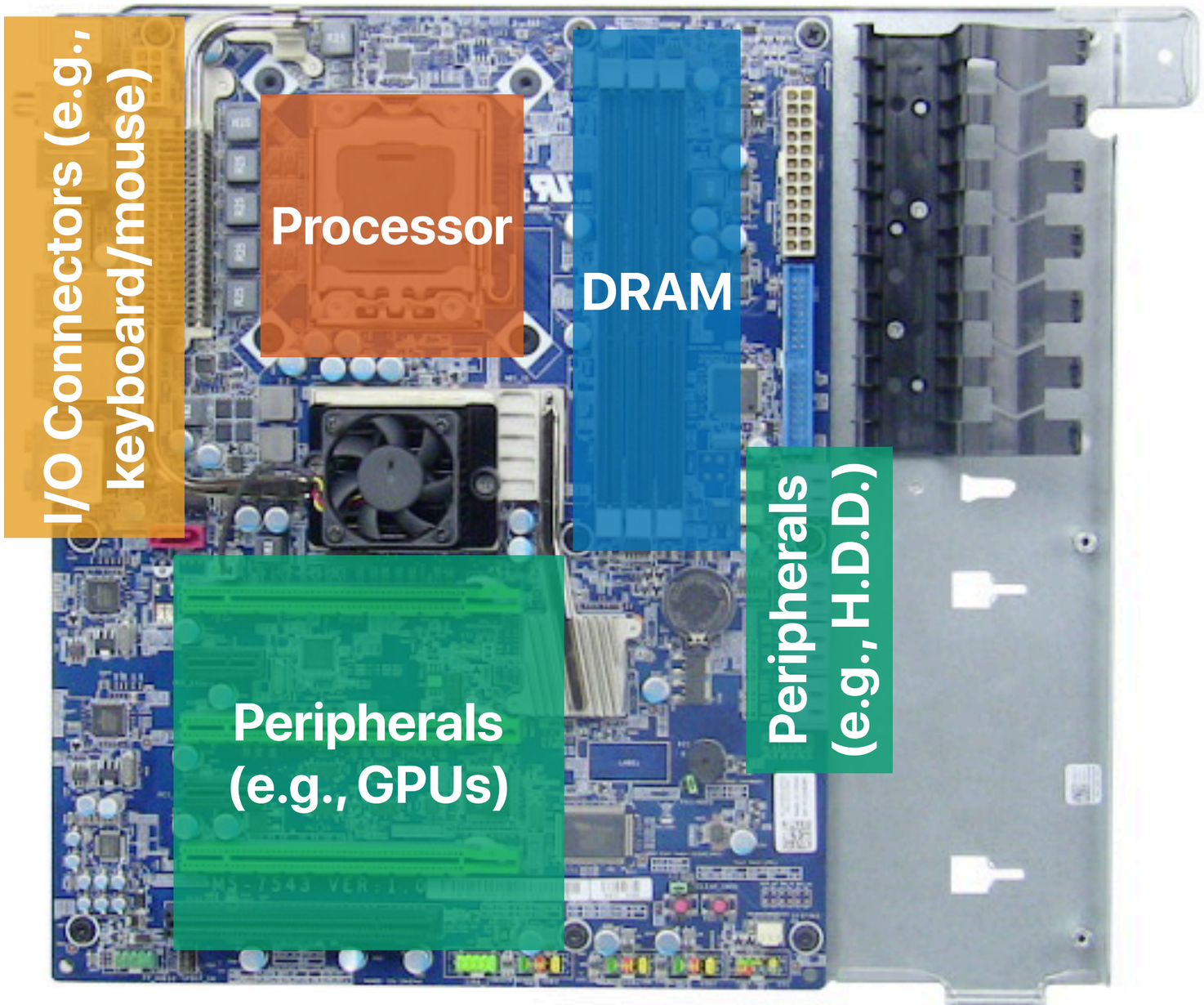
Instruction	Data
509cbd23	00000008
00005d24	00c2f000
0000bd24	00000008
2ca422a0	00c2f800
130020e4	00000008
00003d24	00c30000
2ca4e2b3	00000008

Memory

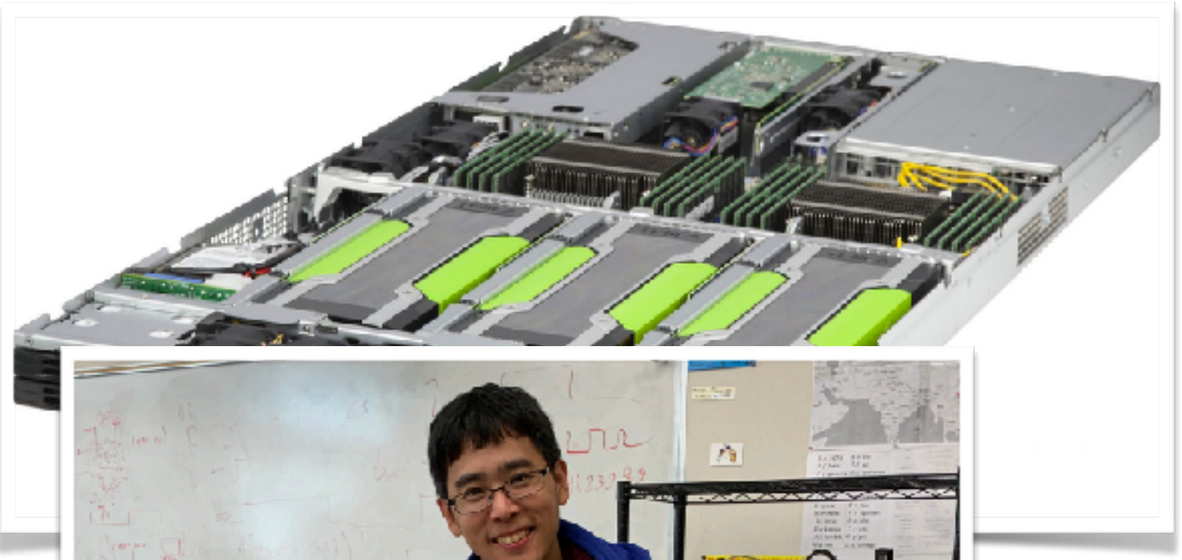
Storage

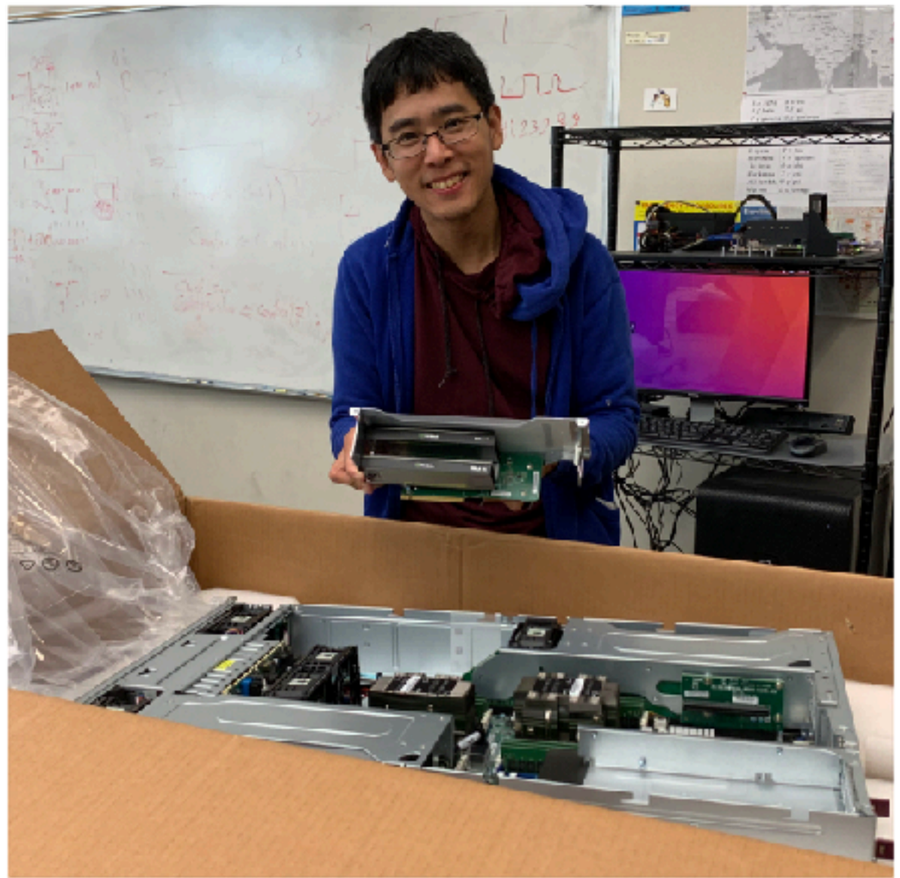
By loading different programs into memory, your computer can perform different functions

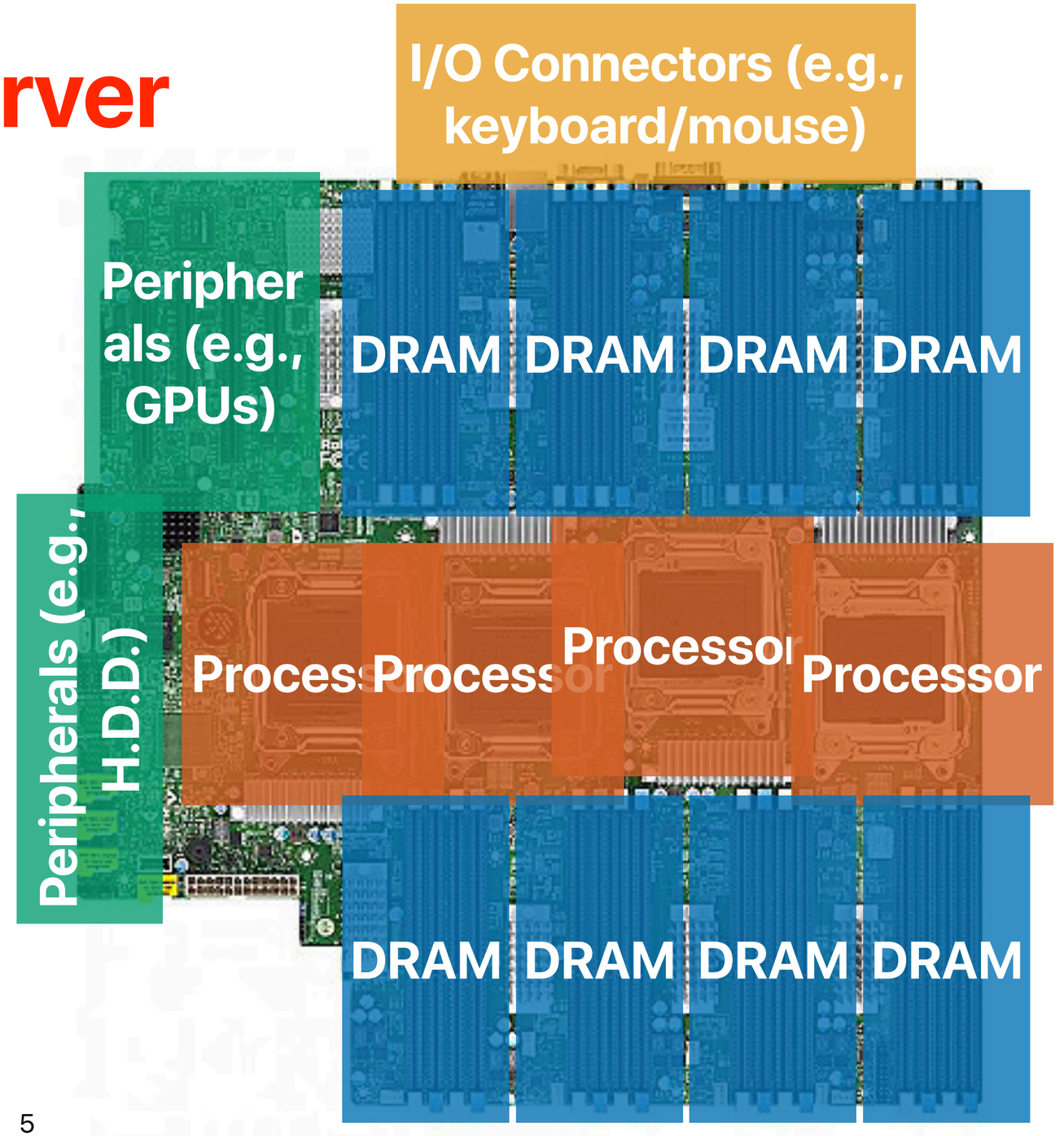
Desktop Computer



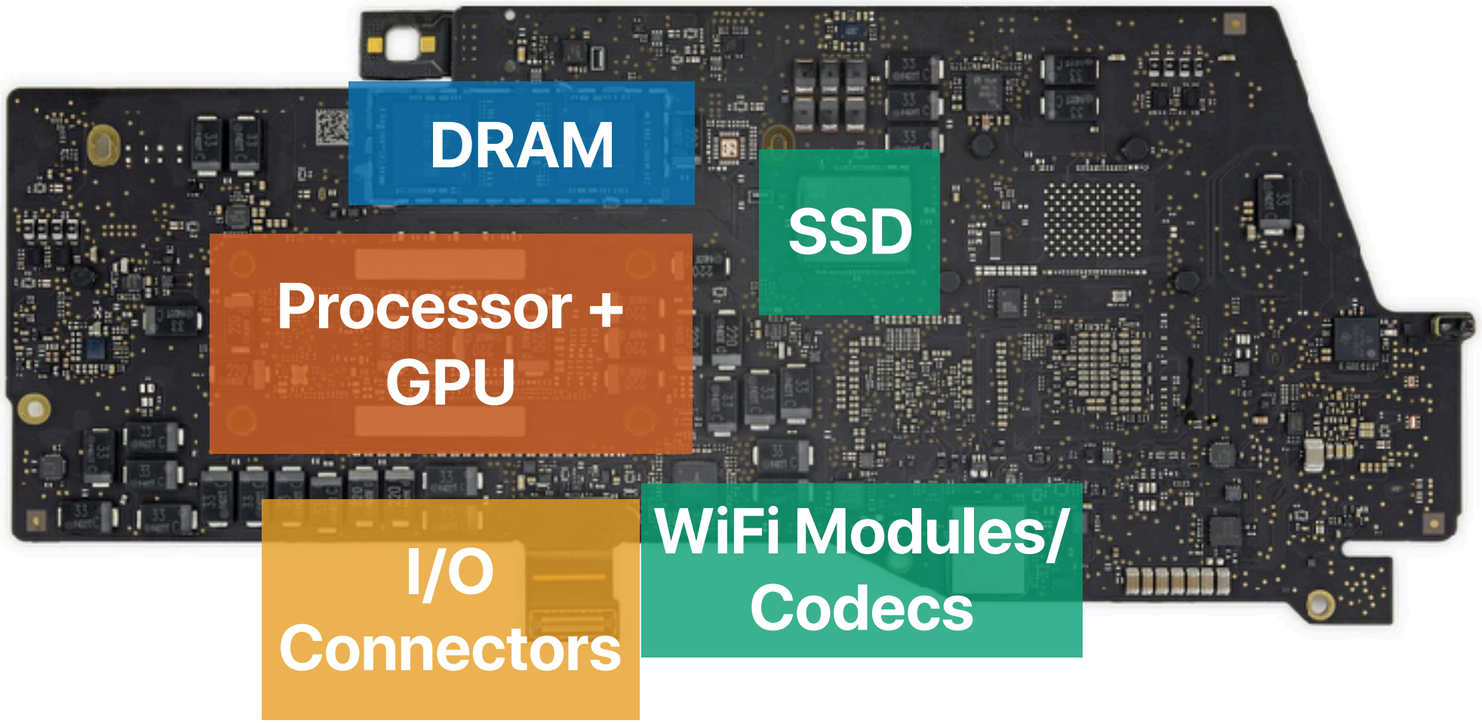
Server



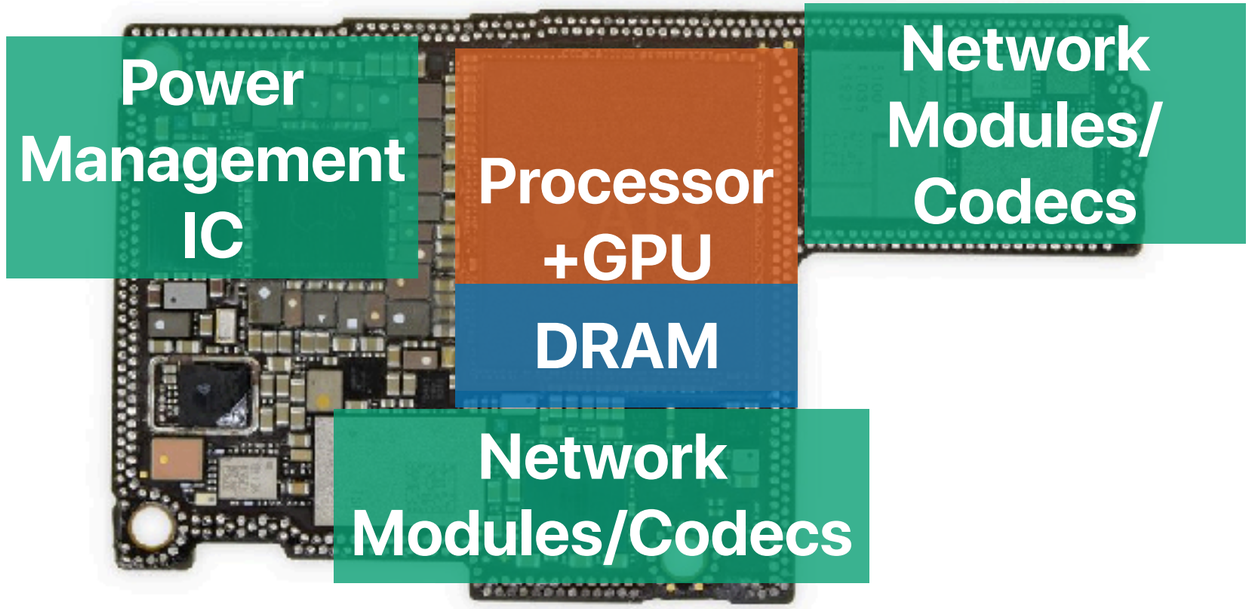


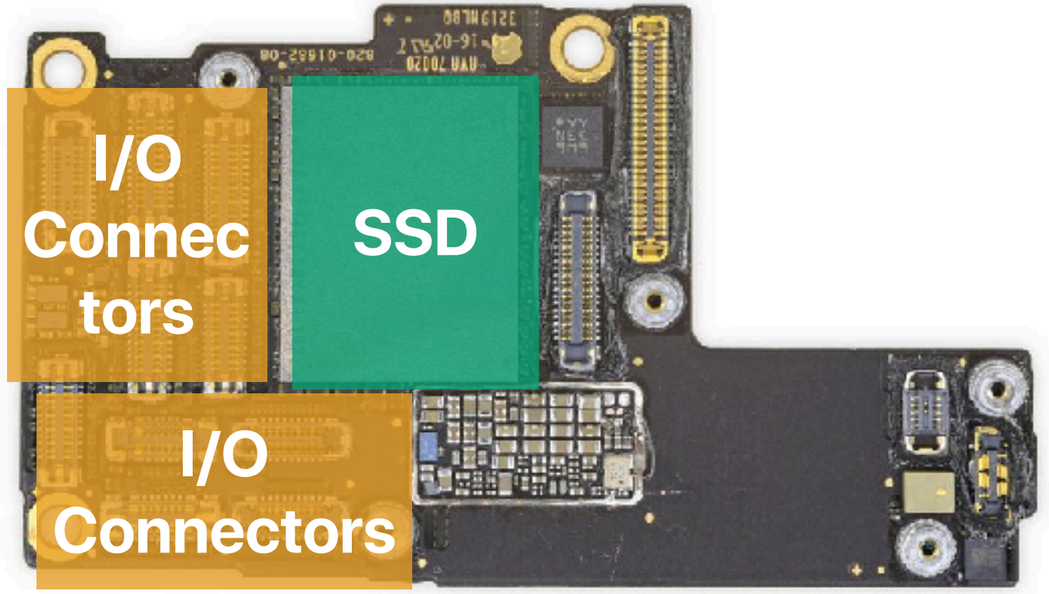

I/O Connectors (e.g., keyboard/mouse)
Peripherals (e.g., GPUs)
DRAM
DRAM
DRAM
DRAM
Peripherals (e.g., H.D.D.)
Processor
Processor
Processor
Processor
DRAM
DRAM
DRAM
DRAM

MacBook Pro 13″

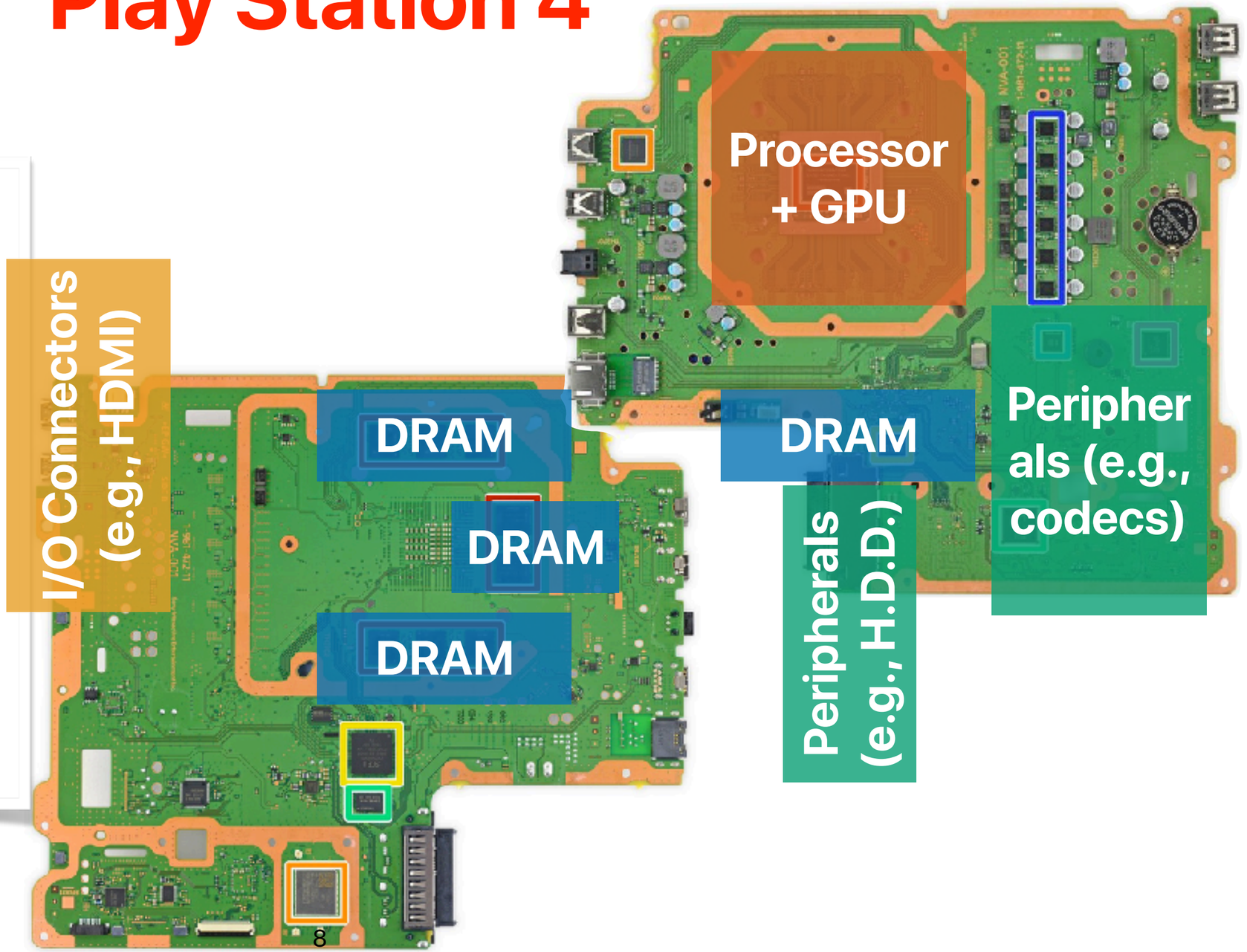


iPhone 11 Pro

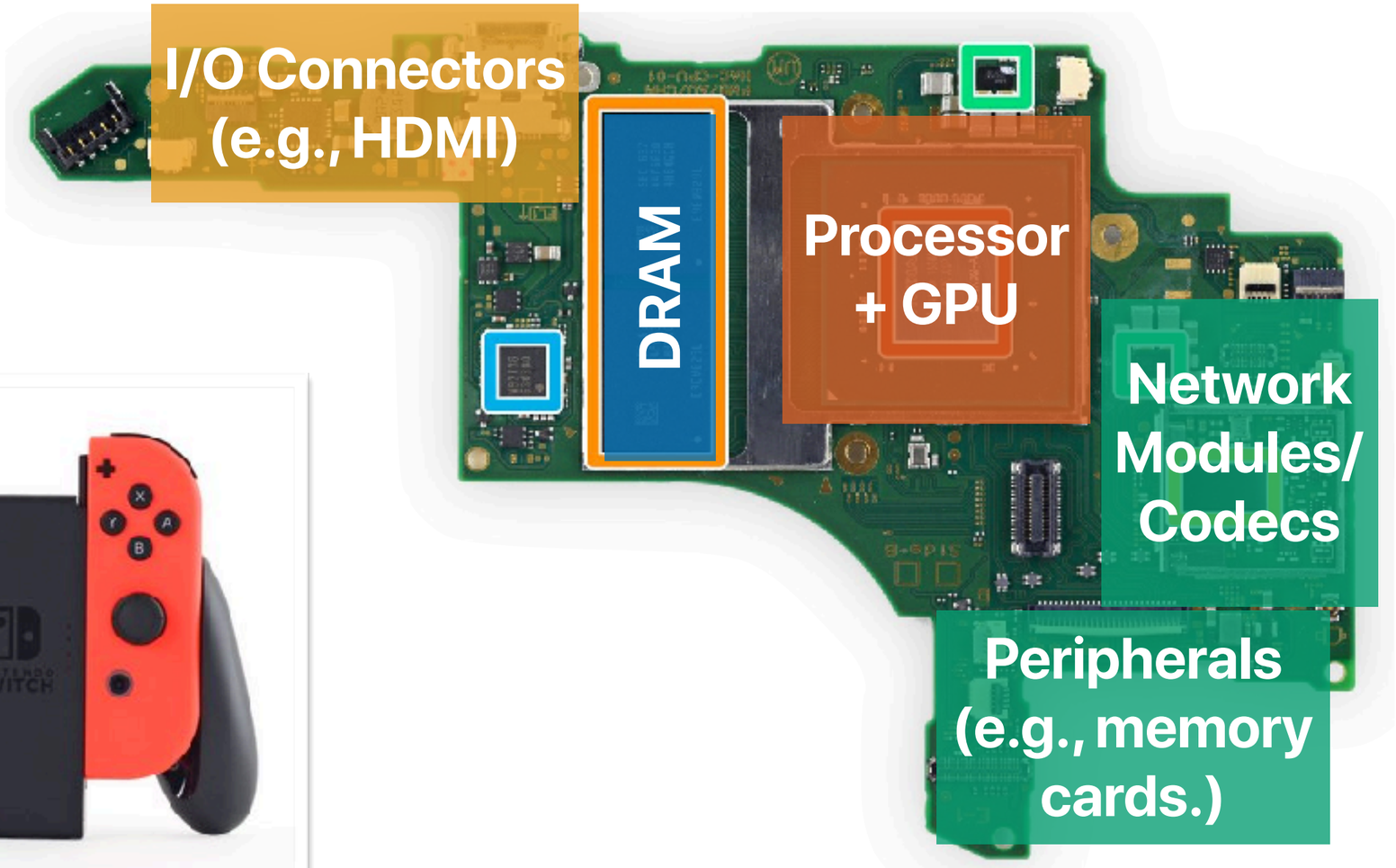




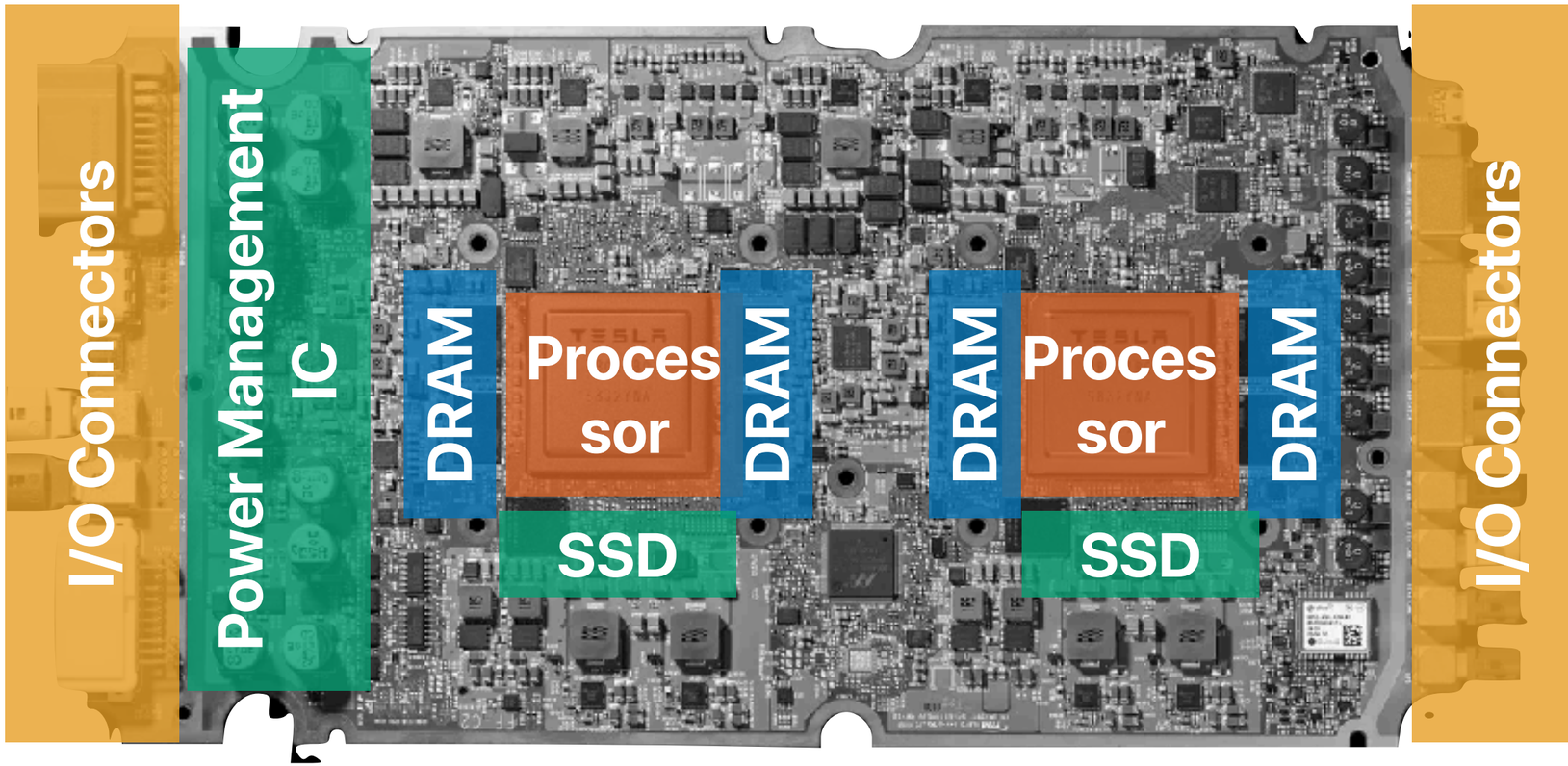
Play Station 4



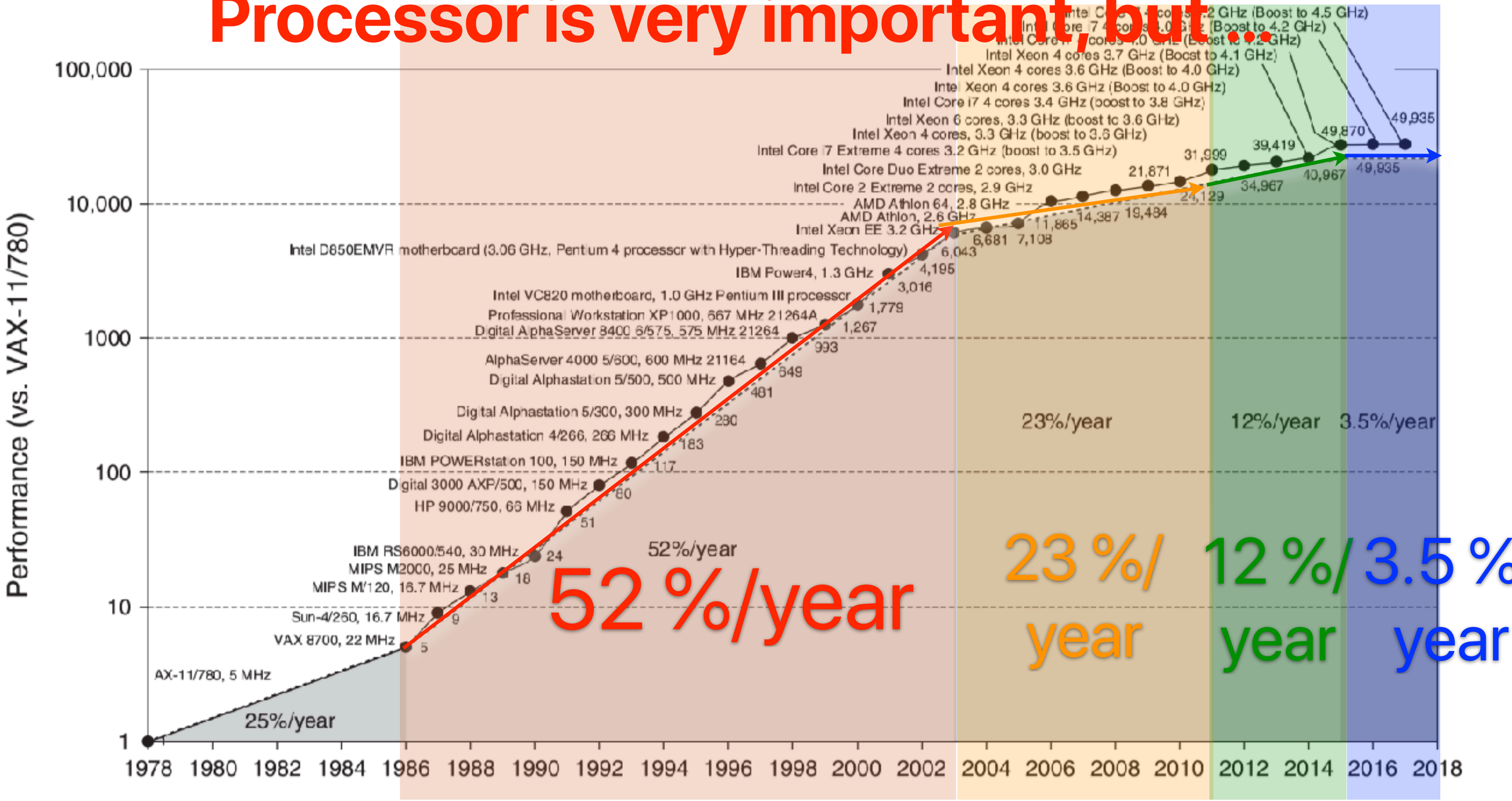
Nintendo Switch



Tesla Model 3



Processor is very important, but…



What can we do?

- Holistic/cross-layered system design instead of single-point/ local optimizations.
- Distributed computing instead of centralized computing.
- Massive, wimpy processing units instead of single, powerful processing unit.
- Domain-specific design instead of general-purpose architectures.

Holistic System Designs

- Key-value SSDs — SSD as a key-value store
  - Samsung Key Value SSD enables High Performance Scaling https://www.samsung.com/semiconductor/global.semi.static/Samsung_Key_Value_SSD_enables_High_Performance_Scaling-0.pdf
  - KAML: A Flexible, High-Performance Key-Value SSD
- P2P Data Exchange
  - SPIN: Seamless Operating System Integration of Peer-to-Peer DMA Between SSDs and GPUs

Distributed computing instead of centralized computing

- Edge computing
- IoT
- Data center v.s. Supercomputer

Massive, wimpy processing units instead of single, powerful processing unit.

- GPU
- Intel's Xeon Phi

Table 1. Comparison of NVIDIA Pascal GP102 and Turing TU102

GPU Features	GTX 1080Ti	RTX 2080 Ti	Quadro P6000	Quadro RTX 6000
Architecture	Pascal	Turing	Pascal	Turing
GPCs	6	6	6	6
TPCs	28	34	30	36
SMs	28	68	30	72
CUDA Cores / SM	128	64	128	64
CUDA Cores / GPU	3584	4352	3840	4608
Tensor Cores / SM	NA	8	NA	8
Tensor Cores / GPU	NA	544	NA	576
RT Cores	NA	68	NA	72
GPU Base Clock MHz (Reference / Founders Edition)	1480 / 1480	1350 / 1350	1506	1455

2 Products	Intel® Xeon® Processor E7-8893 v4	Intel® Xeon Phi™ Processor 7295
Product Collection	Intel® Xeon® Processor E7 v4 Family	Intel® Xeon Phi™ 72x5 Processor Family
Vertical Segment	Server	Server
Processor Number	E7-8893V4	7295
Off Roadmap	No	No
Status	Launched	Launched
Launch Date	Q2'16	Q4'17
Lithography	14 nm	14 nm
Recommended Customer Price	$6,841.00	

Performance

# of Cores	4	72
# of Threads	8	72
Processor Base Frequency	3.20 GHz	1.50 GHz
Max Turbo Frequency	3.50 GHz	1.60 GHz
Cache	60 MB	36 MB L2 Cache
Bus Speed	9.6 GT/s	
# of QPI Links	3	

Domain-specific design instead of general-purpose architectures.

- Tensor Processing Units
  - In-Datacenter Performance Analysis of a Tensor Processing Unit
- NN accelerators
  - SCNN: An Accelerator for Compressed-sparse Convolutional Neural Networks

How’re we going to learn?

Logistics

- Instructor — Hung-Wei Tseng
- Website: https://www.escalab.org/classes/ee260-2020wi/
- Time/Location: Mostly Tuesdays 12:30p-1:50p @ **WCH 415**

Seminar-style

- We will meet every Tuesday @ WCH 415
- We will NOT meet on Thursdays unless it’s necessary
- We will present TWO papers per week

How to read research papers

- For each paper, you should identify the followings:
  - Why?
    - Why should we care about this paper?
    - What's the problem that this paper is trying to address?

  The most important thing when you're reading/writing a paper

  - What?
    - What has been proposed?
    - Contributions of the paper

  The second most important thing when you're reading/writing a paper

  - How?
    - How does the paper accomplish the proposed idea?
    - How does the result perform?

  They are important only if you want to implement the proposed idea

Recap & Brainstorm

- What are those related papers that you read before?
- Compare with those related papers and re-exam their **whys**, **whats** and hows
- **What will you propose** if you're solving the same "why"?

Why is reading papers important

- As a researcher
  - You want to identify important problems
  - You want to know what has been accomplished
- As an engineer
  - You want to know if there is a solution of the design problems of your systems, applications
  - You want to know if you can apply the proposed mechanism
  - You want to know how to do it

Your responsibility

- Being a presenter — once a quarter (50%)
- Being an audience/reader — every week (20%)
- Being a criticizer — once a quarter (30%)

Being a speaker

- Sign up for presentation
- Select one paper from a top-tier conference since 2016 to present
  - Architecture: ISCA, MICRO, ASPLOS, HPCA
  - System: OSDI, SOSP, USENIX ATC, FAST
  - Programming languages: PLDI, OOPSLA, POPL
  - Networks: SIGCOMM, NDSI, Mobisys, Mobicom, Sensys
  - Embedded systems: RTSS, RTAS, EMSOFT, DATE

Being a speaker

- Announce the paper a week before your presentation
- Each talk should be 20-minutes only
  - 8 minutes of the why
  - 8 minutes of the what
  - 4 minutes of the how
- 20 minutes for questions
  - You need to prepare three questions to discuss

Being an audience/reader

- Submit a “preview” before the lecture
  - At least four sentences about the why, what, how, and which trend(s) does the paper fulfill
- Ask a question during the lecture — you need to at least one question to pass the class
- Eat

Being a criticizer

- You need to write a “survey”-style paper for any topic you select.
  - You need to compare a few papers regarding a single topic
  - You need to write a least 6 pages in a conference format
  - It’s helpful for you to prepare your research exam

What do you need to do after today's lecture?

- Check your e-mail/piazza for signup instructions and upcoming papers
- Sign up!
- Read the paper for the next week and prepare a review!
- See you next Tuesday!