Syllabus

Kun-Chih (Jimmy) Chen 陳坤志

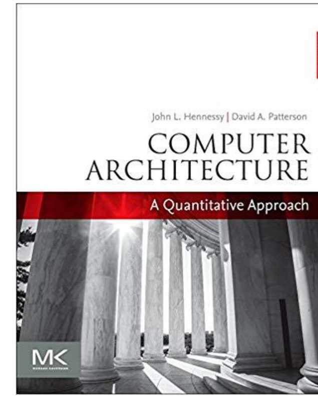
Department of Computer Science and Engineering
National Sun Yat-sen University

Fact Sheet

- **Lecture**
  - EC Rm.9032-1 Fri. 9:10am – 12:00pm
- **Instructor**
  - Kun-Chih Chen (EC Rm.9033) kcchen@mail.cse.nsysu.edu.tw
- **Office Hour**
  - Mon. 10am – 12pm, EC Rm.9033
- **TA Information**
  - TA: Pavan Kumar MP pavankumarmp@cereal.cse.nsysu.edu.tw
    李政融 (Andy) andylee@cereal.cse.nsysu.edu.tw
  - TA Lab: EC Rm. 5037
- **Class web page (video, handout, announcement)**
  - 中山大學網路大學 (cyber university)
- **Prerequisites (but not necessary)**
  - Computer Organization
  - Computer programming (C or C++)
  - Logic design
- **Grading**
  - Homework 20% (HW#1:10%, HW#2: 10%)
  - Final project 40% (Proj#1:10%, Proj#2:10%, Proj#3:10%, Final report:10%, Bonus: 20%)
  - Exam 40% (including midterm and final exam)

Textbook and References

- **Textbook**
  - John L. Hennessy and David A. Patterson, ***Computer Architecture : A Quantitative Approach* 5th ed.**, Morgan Kaufmann Publishers, 2017. (ISBN: 978-0-12-811905-1)



- **Reference (Not required to purchase)**
  - John L. Hennessy and David A. Patterson, ***Computer Organization and Design: The Hardware/Software Interface 5th ed.***, Elsevier, 2014. (ISBN: 978-986-6052-67-5)
  - John L. Hennessy and David A. Patterson, ***Computer Organization and Design: The Hardware/Software Interface RISV-V ed.***, Morgan Kaufmann Publishers, 2018. (ISBN: 978-0-12-812275-4)

Schedule (1/2)

Week	Date	Lecture	Handout	Submit
1	9/9	Mid-autumn Festival		
2	9/16 (virtual)	Syllabus		
3	9/23 (virtual)	Fundamentals of Computer Architecture: cost and performance measurement		
4	9/30 (virtual)	Memory Hierarchy	Project#1	
5	10/7 (virtual)	Memory Hierarchy		
6	10/14 (virtual)	Virtual Memory		
7	10/21 (virtual)	Virtual Memory	HW#1	Project#1
8	10/28	**Midterm**	Project#2, Final Project	
9	11/4	Baseline MIPS Architecture		HW#1

Schedule (2/2)

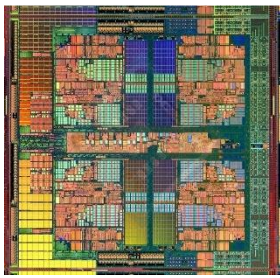
Week	Date	Lecture	Handout	Submit
10	11/11	Instruction-Level Parallelism (ILP): Pipeline Architecture		Project#2
11	11/18	Instruction-Level Parallelism (ILP): Pipeline Architecture		
12	11/25	Data-level Parallelism (DLP): Vector, SIMD, and GPU Architecture	Project#3	
13	12/2	Thread-level Parallelism (TLP): Distributed Memory Communication		
14	12/9	Interconnection in Multicore System		Project#3
15	12/16	Reliable System Design	HW#2	
16	12/23	**Final Exam**		
17	12/30	TBD		HW#2
18	1/6	TBD		

Class Policy

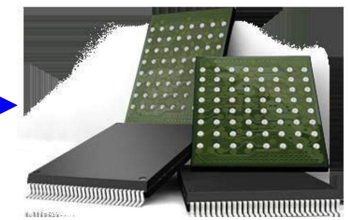
- **Lecture**
  - Do not hesitate to ask questions in office hour
  - The videos are only allowed to <u>keep it to yourself</u>
- **Homework/Project**
  - Homework: Submit the hardcopy report in class
  - Project: Submit the softcopy report and your final program to TA (zip file)
    - File Naming Rule: (Student ID #)_(Student Name)_(HW#) ex. D12345_王小明_HW1
  - Late homework/project 1/3 off each week, no late homework after 3 weeks
  - Discussion with classmates is encouraged!
  - **Cheating = zero grade for <u>both</u> students!**
- **Midterm/Final Exam**
  - Close book
  - Bag isolation
  - Seat assignment
  - **Cheating = zero grade for <u>both</u> students!**

Current Computer Architecture

Computer architecture, HW/SW, systems, bioinformatics, security

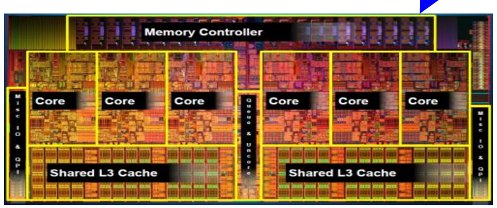


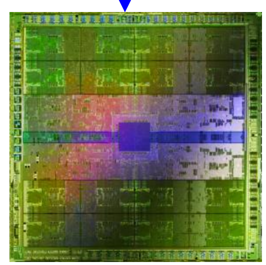
Hybrid Main Memory



Persistent Memory/Storage

Heterogeneous Processors and Accelerators





Graphics and Vision Processing

Build fundamentally better architectures

Four Key Current Directions

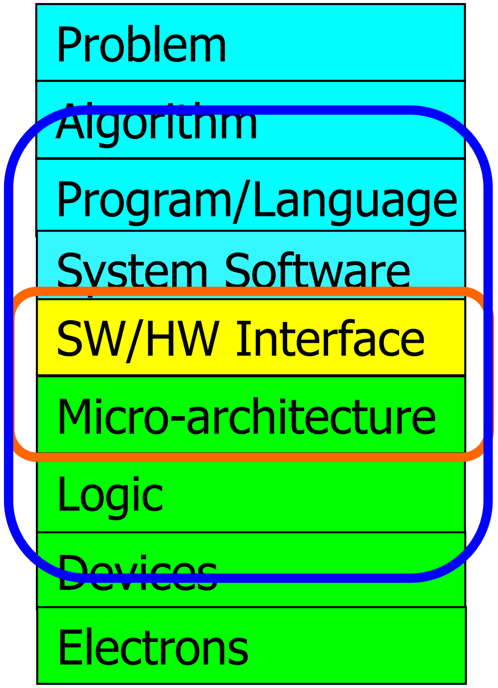
- Fundamentally Secure/Reliable/Safe Architectures

- Fundamentally Energy-Efficient Architectures
  - Memory-centric (Data-centric) Architectures

- Fundamentally Low-Latency and Predictable Architectures

- Architectures for AI/ML, Genomics, Medicine, Health

The Transformation Hierarchy



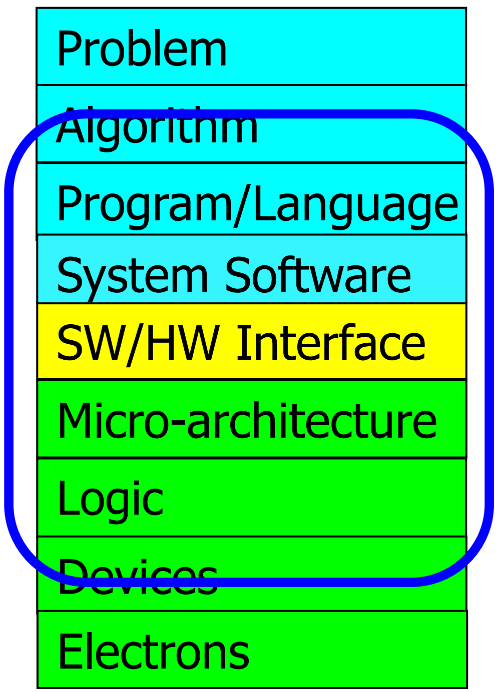
Computer Architecture (expanded view)

Computer Architecture (narrow view)

Axiom

- To achieve the highest energy efficiency and performance:

we must take the expanded view
of computer architecture



Co-design across the hierarchy:
Algorithms to devices

Specialize as much as possible
within the design goals

Current Research Mission & Major Topics

Build fundamentally better architectures

- Problem
- Algorithm
- Program/Language
- System Software
- SW/HW Interface
- Micro-architecture
- Logic
- Devices
- Electrons

Broad research spanning apps, systems, logic with architecture at the center

- Data-centric arch. for low energy & high perf.
  - Proc. in Mem/DRAM, NVM, unified mem/storage
- Low-latency & predictable architectures
  - Low-latency, low-energy yet low-cost memory
  - QoS-aware and predictable memory systems
- Fundamentally secure/reliable/safe arch.
  - Tolerating all bit flips; patchable HW; secure mem
- Architectures for ML/AI/Genomics/Graph/Med
  - Algorithm/arch./logic co-design; full heterogeneity
- Data-driven and data-aware architectures
  - ML/AI-driven architectural controllers and design
  - Expressive memory and expressive systems

What is computer architecture?

- is the science and art of designing computing platforms (hardware, interface, system SW, and programming model)

- to achieve a set of design goals
  - E.g., highest performance on earth on workloads X, Y, Z
  - E.g., longest battery life at a form factor that fits in your pocket with cost < $$$ CHF
  - E.g., best average performance across all known workloads at the best performance/cost ratio
  - …

  - Designing a supercomputer is different from designing a smartphone → But, many fundamental principles are similar

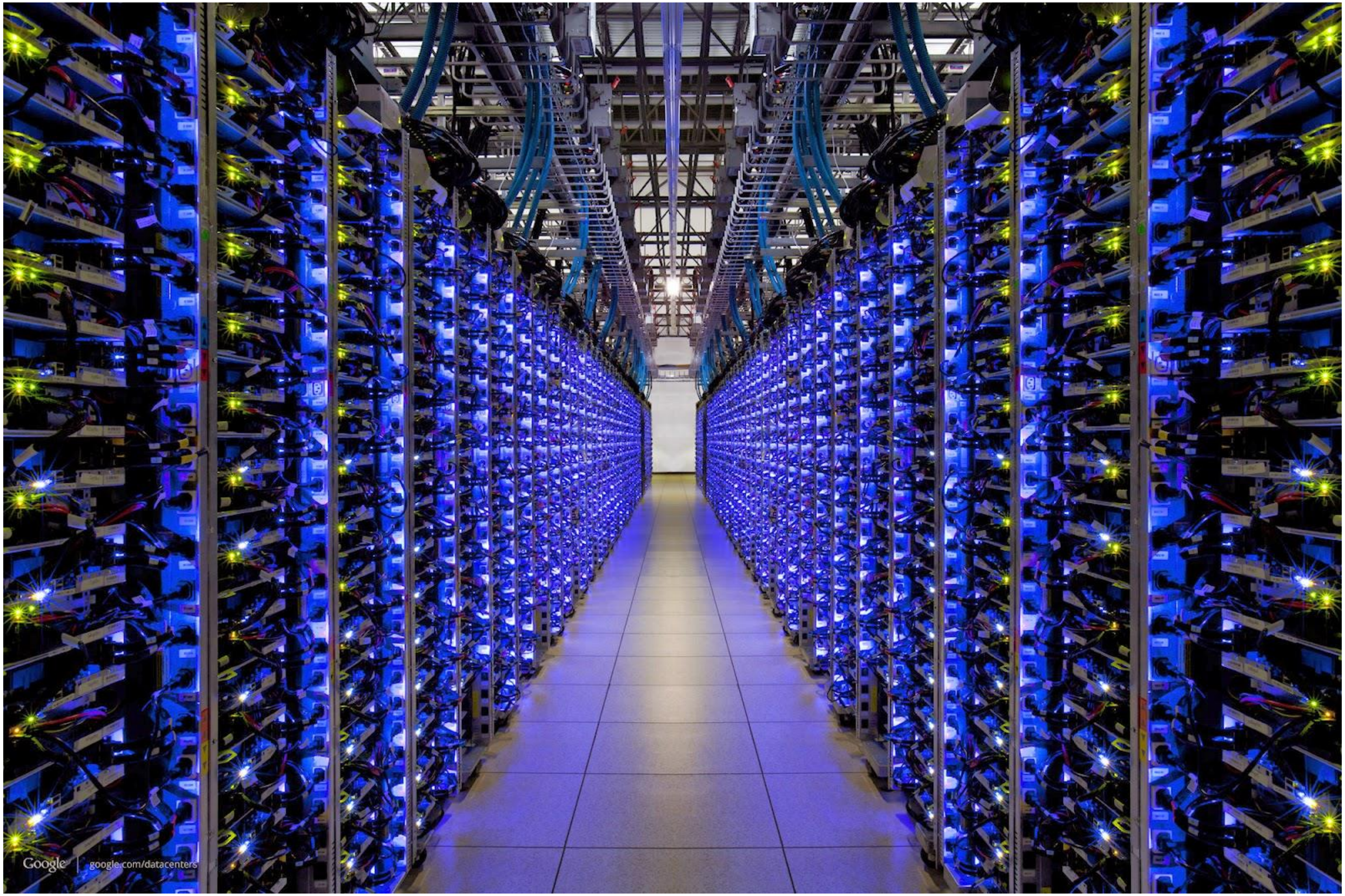
Different Platforms, Different Goals

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

Different Platforms, Different Goals

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

Different Platforms, Different Goals



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

Different Platforms, Different Goals

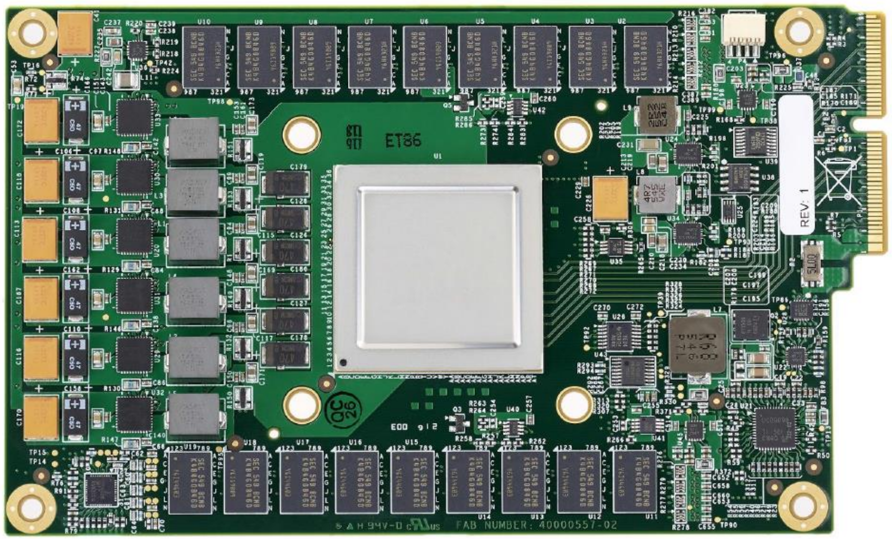


Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

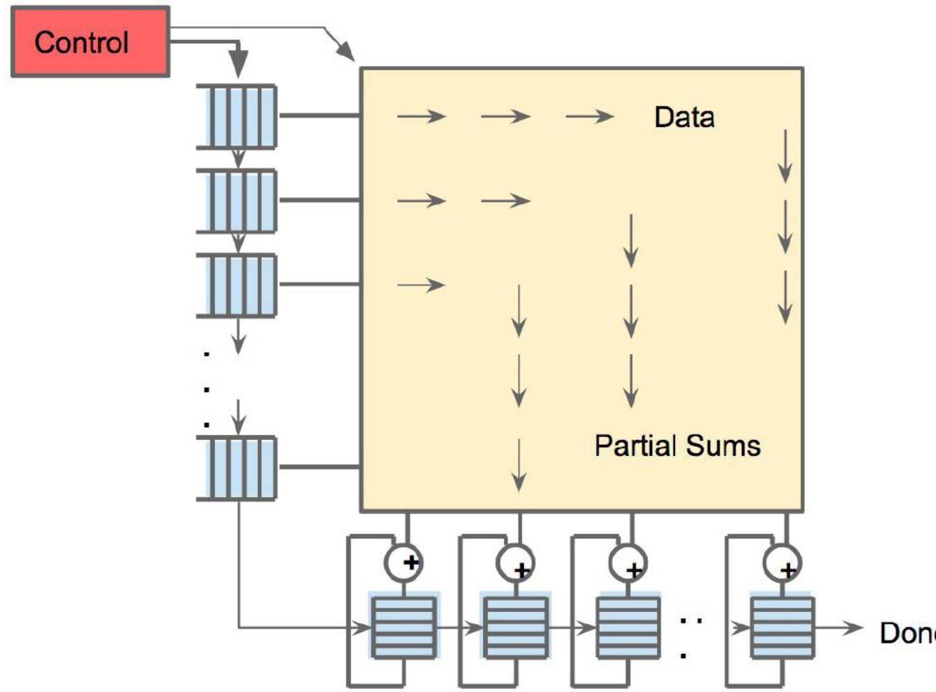
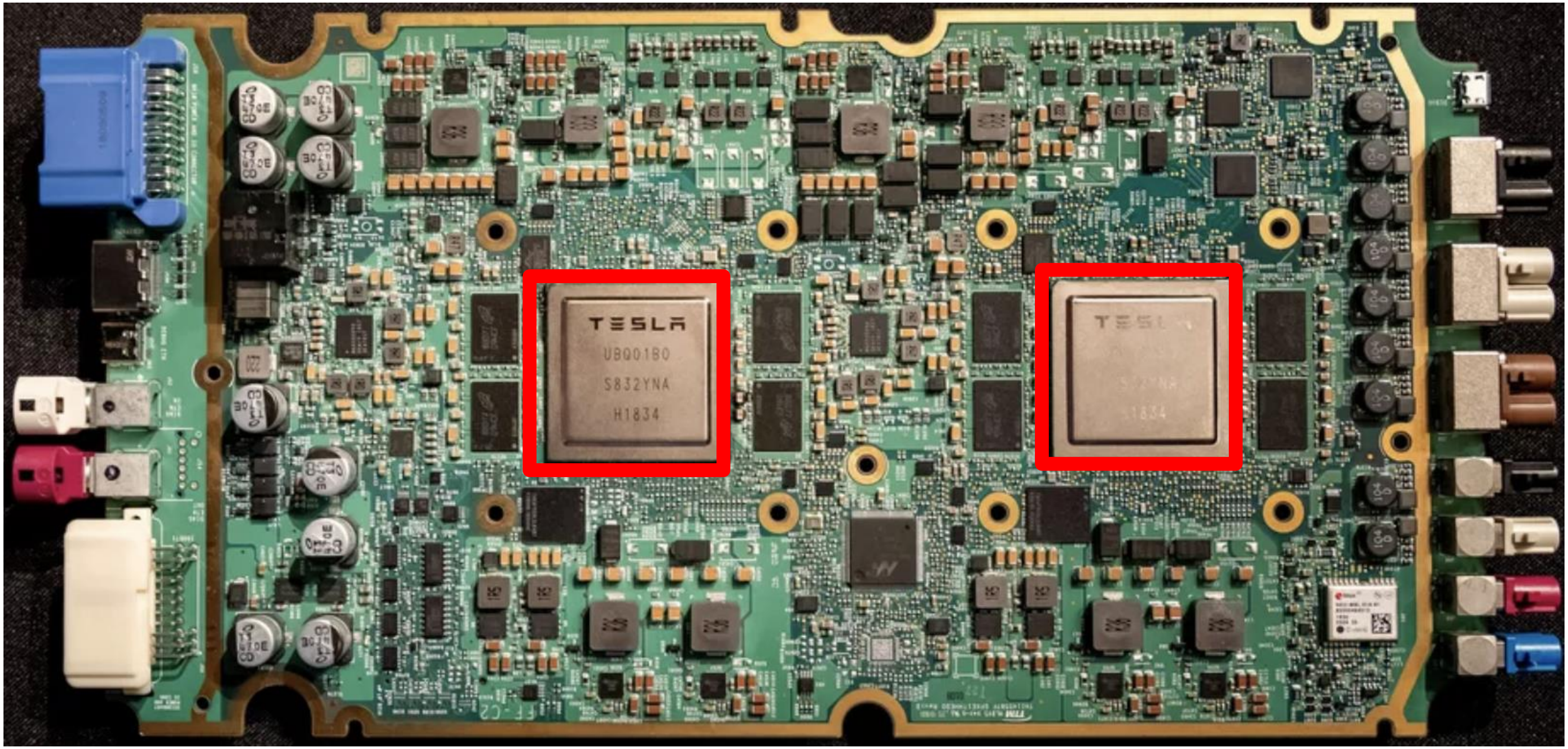


Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

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

Different Platforms, Different Goals

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



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

Why Study Computer Architecture?

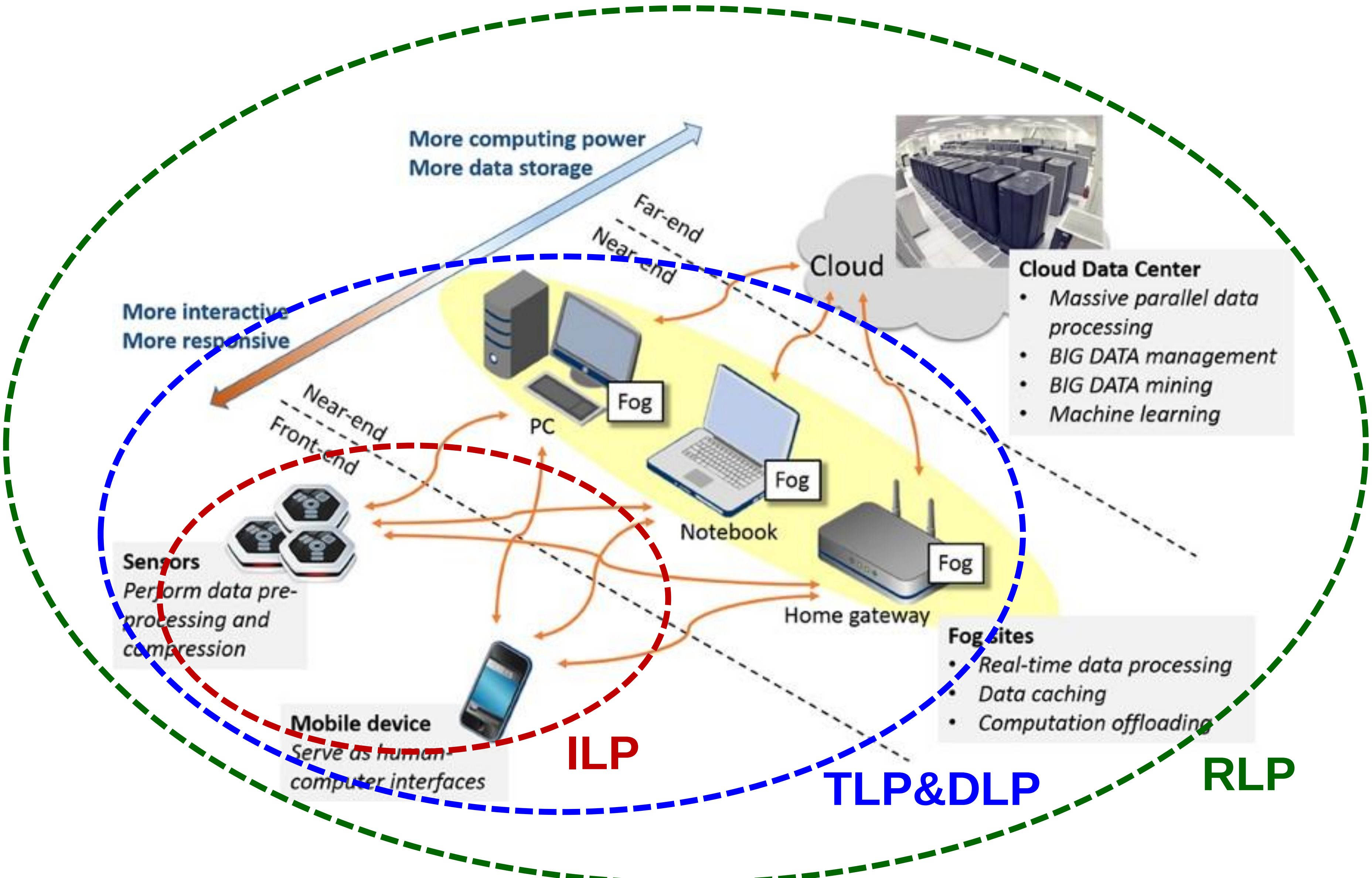
- Enable better systems: make computers faster, cheaper, smaller, more reliable, …
  - By exploiting advances and changes in underlying technology/circuits

- Enable new applications
  - Life-like 3D visualization 20 years ago? Virtual reality?
  - Self-driving cars?
  - Personalized genomics? Personalized medicine?

- Enable better solutions to problems
  - Software innovation is built on trends and changes in computer architecture
    - > 50% performance improvement per year has enabled this innovation

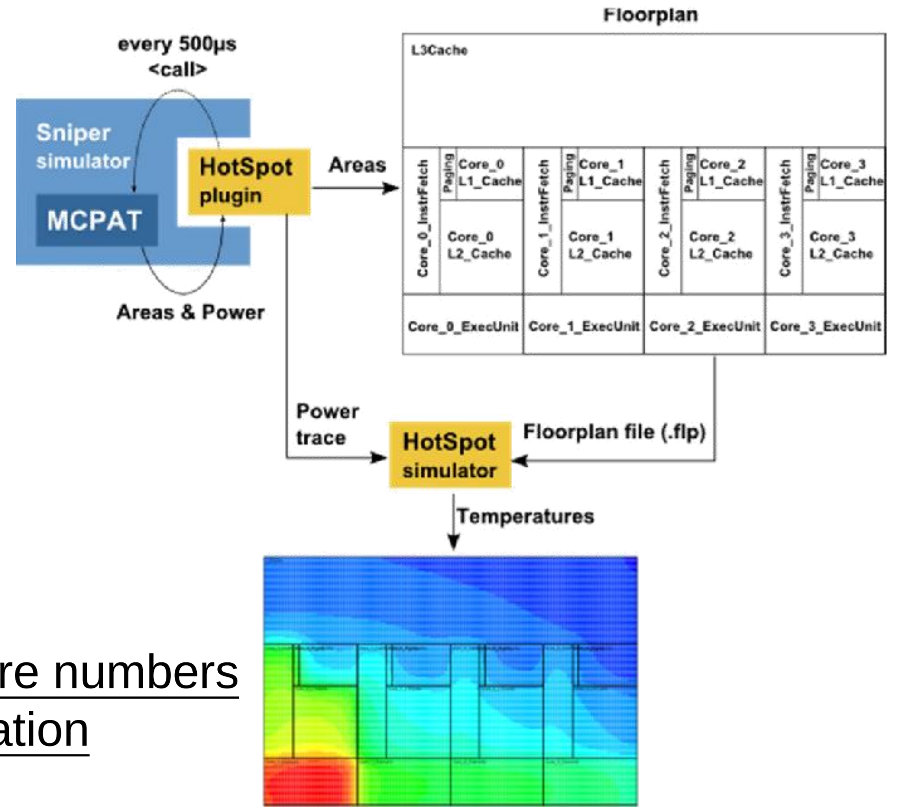
- Understand why computers work the way they do

Cloud, Fog, and Edge Computing



Introduction of Projects

- **Sniper Multi-Core Simulator**
  - Cycle-accurate multi-core x86 simulator with max. 100 cores
- **Hotspot Temperature Simulator**
  - Accurate and fast thermal model suitable for use in architectural studies



- **Project #1**
  - Sniper evaluation
  - Performance analysis under <u>different core numbers and memory levels with different association</u>
- **Project #2**
  - Performance and temperature analysis under <u>different core numbers and memory levels</u>
- **Project #3**
  - Performance analysis under TDP consideration
    (Bonus 20%: Find the optimal design parameter)