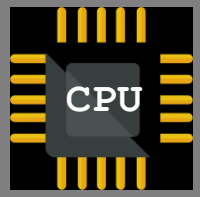


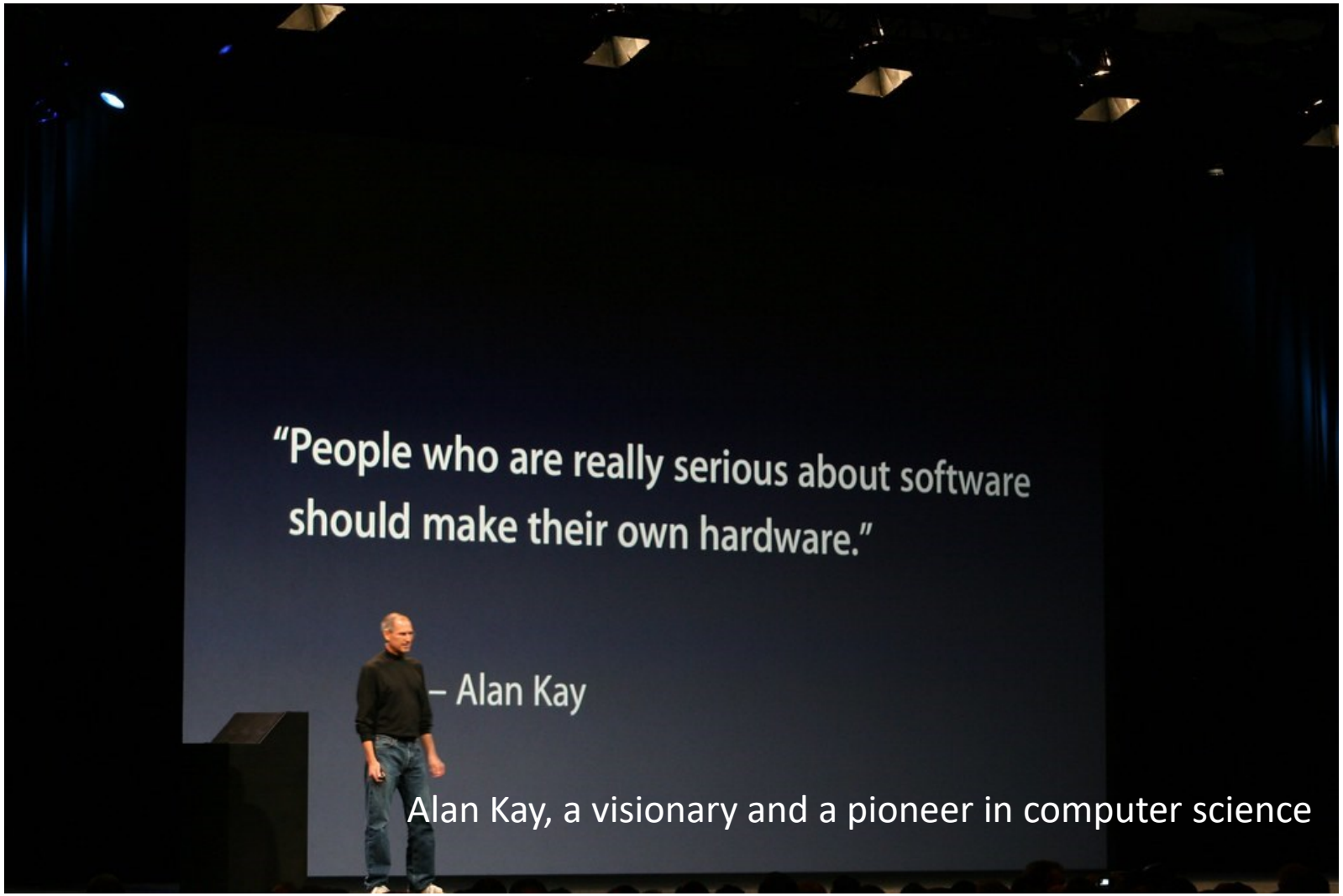
Topic 1

Introduction to Computer Architecture

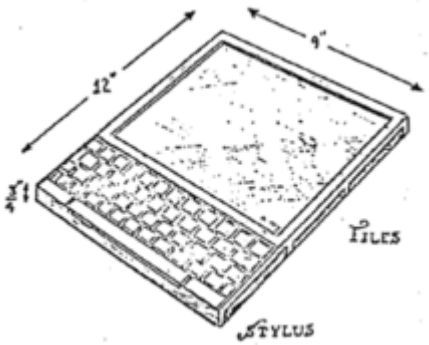
Xinfei Guo
xinfei.guo@sjtu.edu.cn

May 9th, 2022

Hardware is essential!



Alan Kay, a visionary and a pioneer in computer science



Dynabook 1972

https://en.wikipedia.org/wiki/Dynabook

Image: https://www.flickr.com/photos/nobihaya/5151743358

Everyone is building chips…


网易首页 > 网易号 > 正文

不造芯，不配做互联网巨头

2022-04-30 08:16:36 来源: 与非网eefocus 江苏

谷歌：互联网造芯“开山鼻祖”

Sources:

- https://www.163.com/dy/article/H66HJHOK0552YI1V.html
- https://36kr.com/p/1659735378901125
- https://www.cnbc.com/2021/09/06/why-tesla-apple-google-and-facebook-are-designing-their-own-chips.html
- https://www.thestreet.com/technology/tech-firms-design-custom-chips-how-to-invest-14866420
- https://www.yahoo.com/now/big-tech-companies-shift-making-030000821.html


Tech Giants Are Increasingly Designing Their Own Custom Chips - How to Play It

The next great advance in computing will be the product of software optimization.

JON MARKMAN • FEB 14, 2019 9:41 AM EST


36Kr 科技巨头们为什么要“跨界造芯”?

首页

科技云报道 · 2022-03-18 13:18 关注


TECH

Tech giants are rushing to develop their own chips — here’s why

ON, SEP 6 2021·7:13 AM EDT | UPDATED TUE, SEP 7 2021·1:03 AM EDT

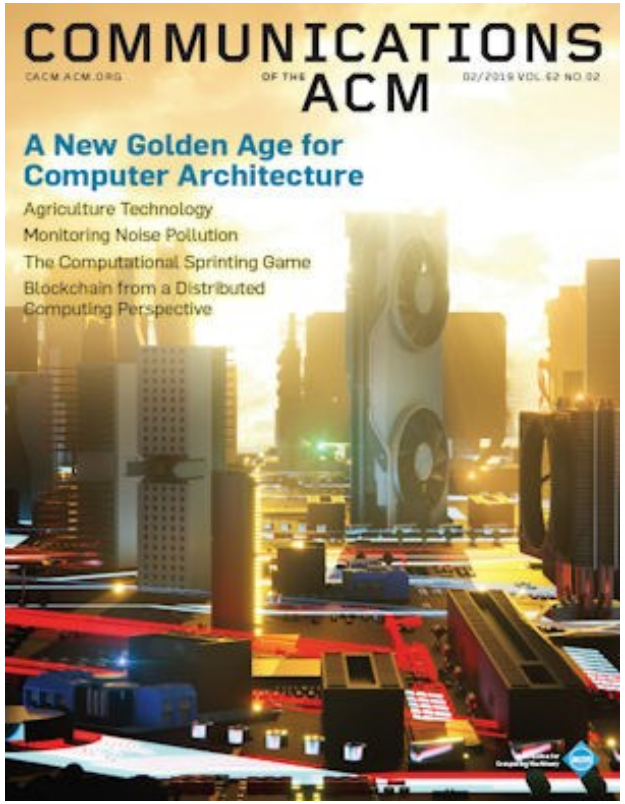

Big Tech companies shift to making semiconductors in-house

Christopher Hutton

September 9, 2021 · 3 min read

TRENDING

Golden Age for Computer Architecture



A New Golden Age for Computer Architecture:

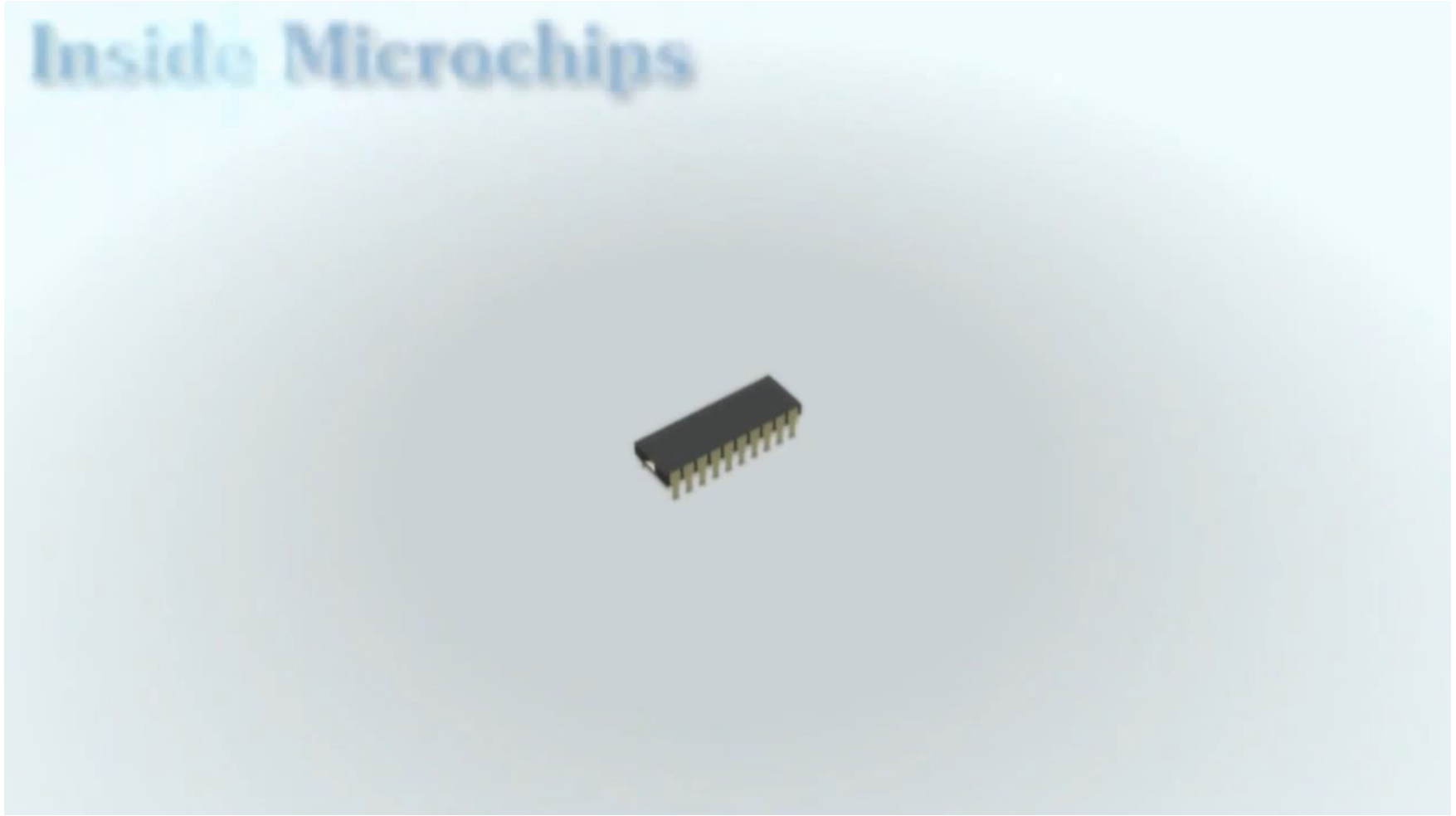
History, Challenges, and Opportunities

David Patterson
UC Berkeley and Google

December 5, 2018

Full Turing Lecture:
https://www.acm.org/hennessy-patterson-turing-lecture

Inside a Processor



video: https://www.youtube.com/watch?v=voRQSDyPIaM&ab_channel=Markus

What Is Computer Architecture?

- "The term **architecture** is used here to describe the attributes of a system as seen by the programmer, i.e., the conceptual structure and functional behavior as distinct from the organization of the dataflow and controls, the logic design, and the physical implementation."

  - Gene Amdahl, IBM Journal of R&D, April 1964

From VE370 to VE470

- Architecture, Organization, Implementation
  - Computer architecture: SW/HW interface
    - instruction set
    - memory management and protection
    - interrupts and traps
    - floating-point standard (IEEE)
  - Organization: also called microarchitecture
    - number/location of functional units
    - pipeline/cache configuration
    - datapath connections
  - Implementation
    - low-level circuits

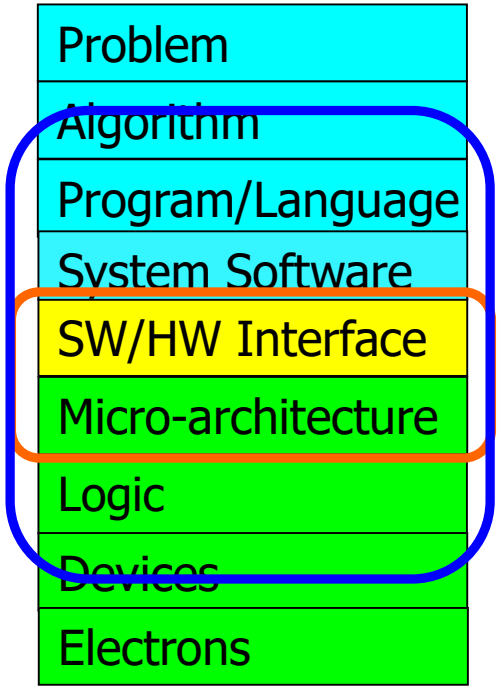
Another View: Levels of Abstraction

- Architecture
  - A set of specifications that allows developers to write software and firmware
  - These include the instruction set.
- Microarchitecture
  - The logical organization of the inner structure of the computer
- Hardware or Implementation
  - The realization or the physical structure, i.e., logic design and chip packaging

Levels of Abstraction

Computer Architecture (expanded view)

Computer Architecture (narrow view)



- Each level of design imposes different requirements and constraints, which change over time.
- History and economics: there is commercial pressure to evolve in a way that minimizes disruption and possible costs to the ecosystem (e.g., software).
- There is also a need to look forward and not design for yesterday's technology and workloads!
- Design decisions should be carefully justified through experimentation.

Figure: Prof. Onur Mutlu @ ETH

What is an architect usually doing?

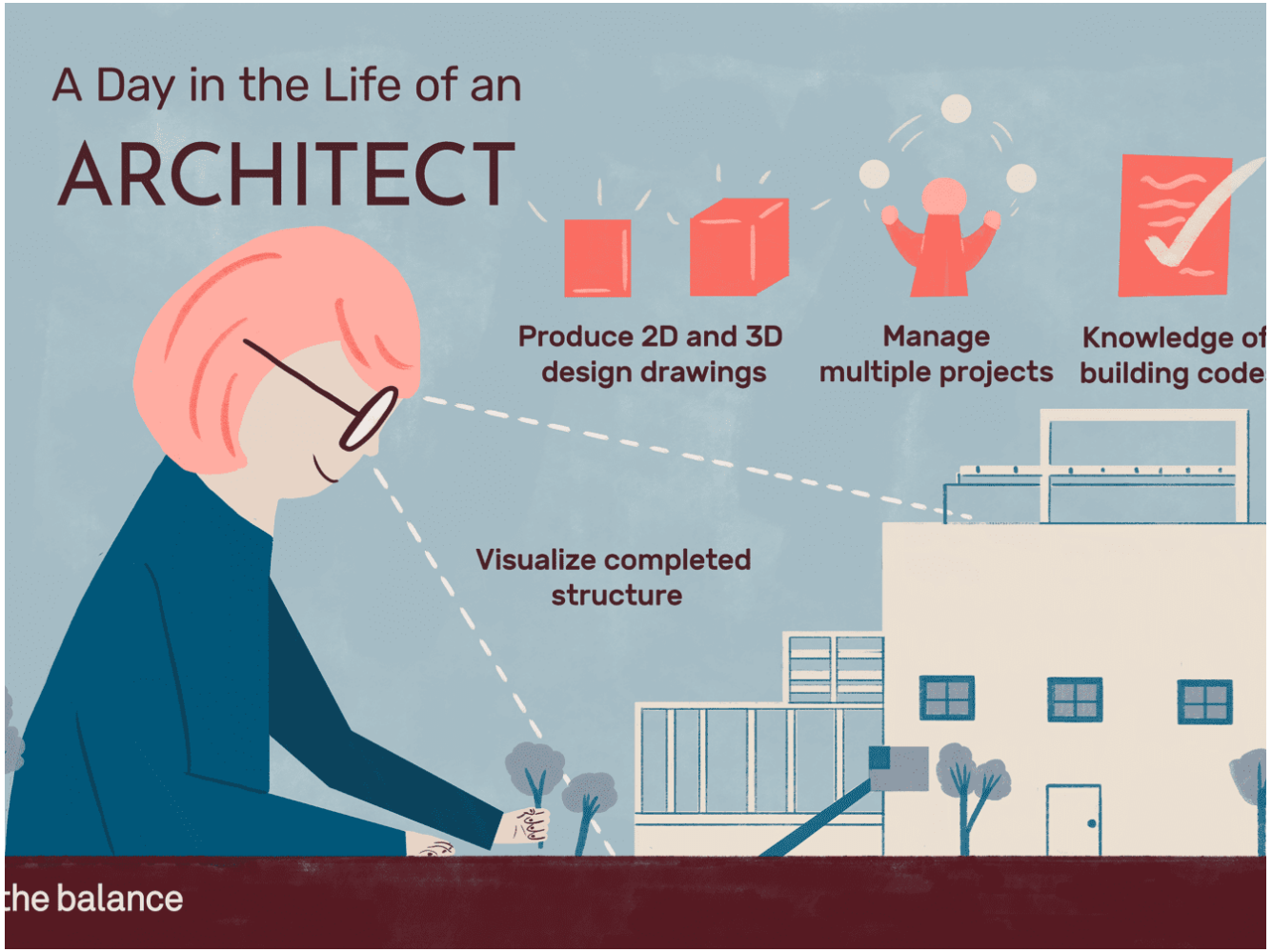


Image: https://darnelltechnical.com/4-crucial-skills-architects-need-to-develop-to-achieve-success/

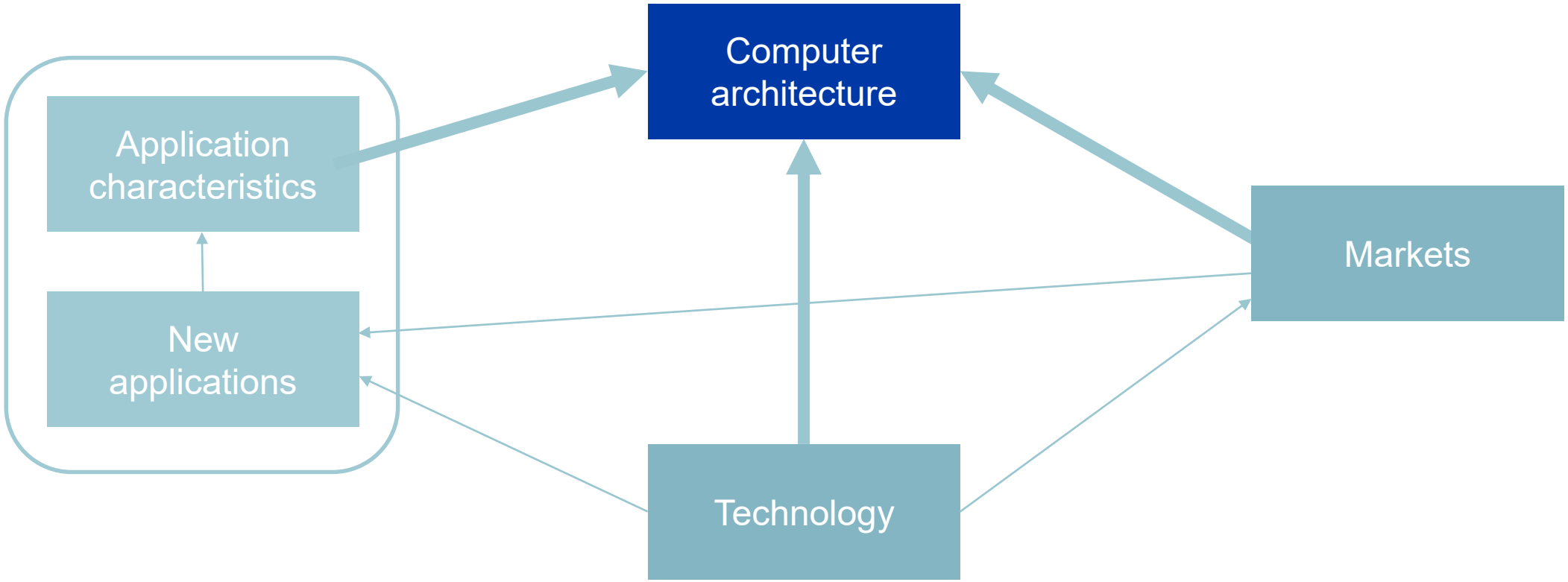
Computer Architecture

- Computer architecture is much more than the task of defining the instruction-set or high-level architecture.
- The computer architect must contribute to, and understand, **all levels of design** in order to deliver the most appropriate design for a particular application and target market.

Computer Architecture

- Computer architecture is concerned with how best to exploit fabrication technology to meet marketplace demands.
  - *e.g., how best might we use five billion transistors and a power budget of two watts to design the chip at the heart of a mobile phone?*
- Computer architecture builds on a few simple concepts, but is challenging as we must constantly seek new solutions.
- What constitutes the “best” design changes over time and depending on our use-case. It involves considering many different trade-offs.

The Computer Architecture Arena



Source: "Early 21st Century Processors," S. Vajapeyam and M. Valero, IEEE Computer, April 2004

Design Goals I

- **Functional** – hard to correct (unlike software). **Verification** is perhaps the highest single cost in the design process. We also need to **test** our chips once they have been manufactured, again this can be a costly process and requires careful thought at the design stage
- **Performance** – what does this mean? No single best answer, e.g., sports car vs. off-road 4x4 vehicle – performance will always depend on the "workload"
- **Power** – a first-order design constraint for most designs today. Power limits the performance of most systems.

Source: ARM

Design Goals II

- **Security** – e.g., the ability to control access to sensitive data or prevent carefully crafted malicious inputs from hijacking control of the processor
- **Cost** – design cost (complexity), die costs (i.e., the size or area of our chip), packaging, etc.
- **Reliability** – do we need to try to detect and/or tolerate faults during operation?

Source: ARM

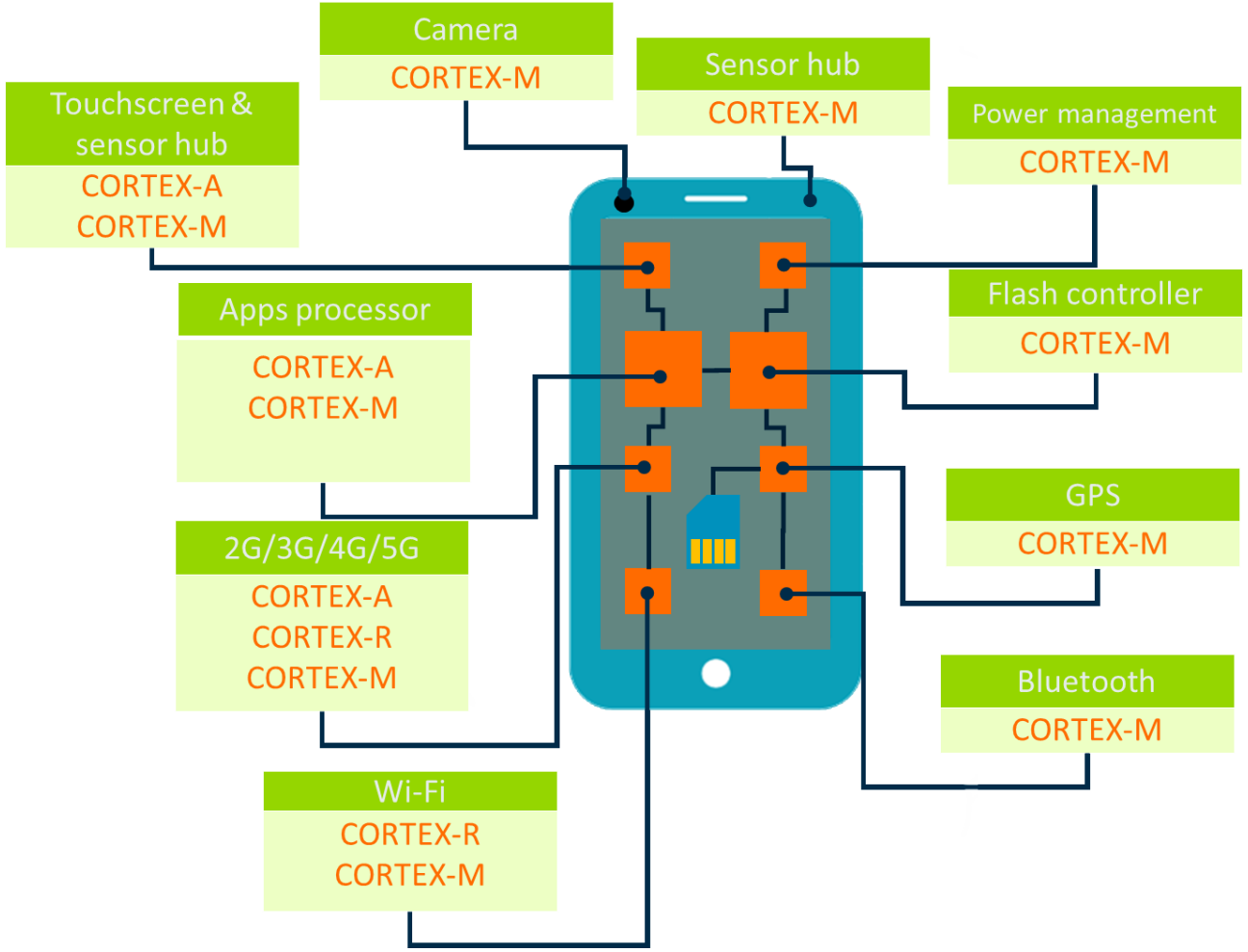
Markets and Features

- Each target market will require a different trade-off in terms of *power consumption, cost, area, performance, security, reliability,* etc.
- Here are some example processor classes from Arm:
  - **Cortex-A:** high-performance application processors, e.g., for mobile phones
  - **Cortex-R**: deterministic real-time performance, fault detection, and tolerance.
  - **Cortex-M**: energy-efficient embedded devices (“microcontroller” class cores)
  - **Neoverse**: scalable networks of processors on a single chip
    - e.g., 8, 16, 64, or 128 cores. Used in datacenters, edge servers, and storage

Source: ARM

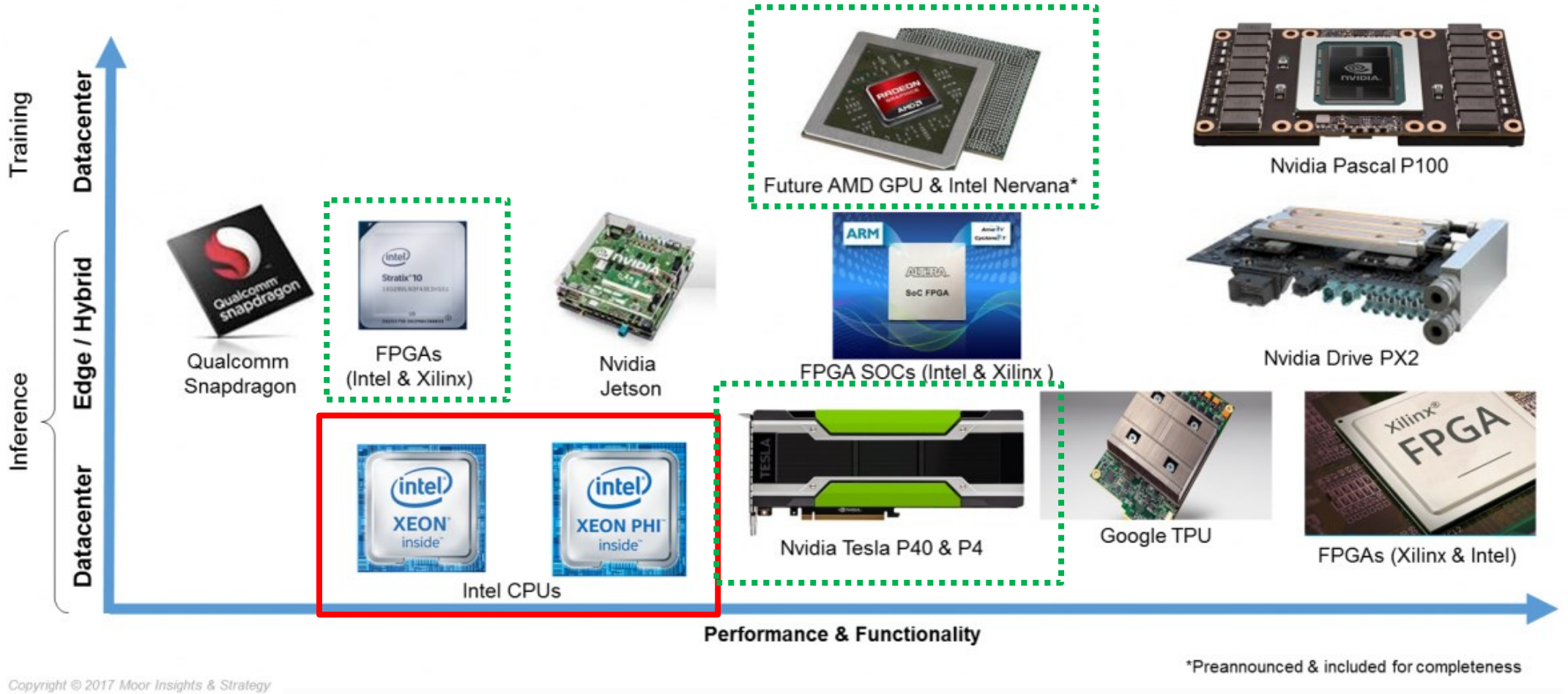
The Smartphone

- A single smartphone will contain many different processor cores.
- *Why not use a single processor?*



Source: ARM

Architecture landscape



Hardware across the Machine Learning landscape (Source: Moor Insights & Strategy)

CPU Vendors (only Intel and AMD?)

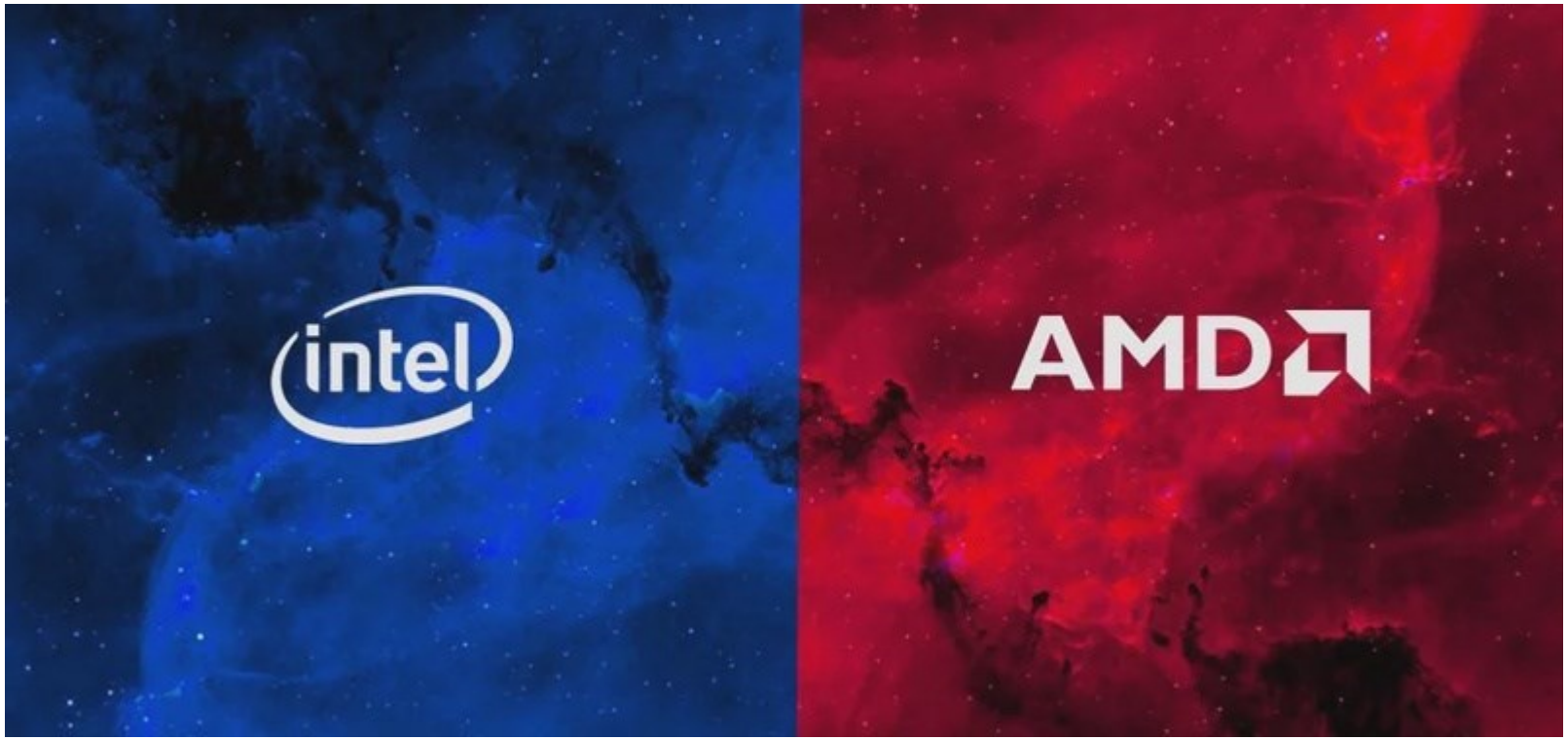


Image: https://techplanet.today/post/why-are-there-only-two-cpu-companies

CPU Vendors (much more!)

Top x86 CPU manufacturers:

- AMD
- Intel

List of CPU manufacturers (most of the companies sell ARM-based CPUs, assumed if nothing else stated):

- Arm Ltd. (sells designs only)
- Apple Inc. (ARM-based CPUs)
- Broadcom Inc. (ARM-based, e.g. for Raspberry Pi)
- Fujitsu (its ARM-based CPU used in top supercomputer, still also sells its SPARC-based servers)
- Hitachi (its own designs and ARM)
- Hygon (x86-based)
- HiSilicon (acquired by Huawei), stopped making its ARM-based design
- IBM (now only designs two architectures)
- Ingenic Semiconductor (MIPS-based)
- Marvell (its ThunderX3 ARM-based)
- MCST (its own designs and SPARC)
- MediaTek (ARM chips, and MIPS chips)[2]
- Nvidia (sells ARM-based, and unsuccessfully attempted to buy the ARM company)
- Qualcomm (ARM-based)
- Rockchip (ARM-based)
- Amlogic (ARM-based)
- Allwinner (ARM-based)
- Samsung (ARM-based)
- SiFive (RISC-V-based, e.g. HiFive Unleashed)
- Texas Instruments (its own designs and ARM)
- Via (formerly Centaur Technology division), its own x86-based design
- Wave Computing (previously MIPS Technologies), licenses MIPS CPU design
- Zhaoxin (its own x86 design based on Via's)

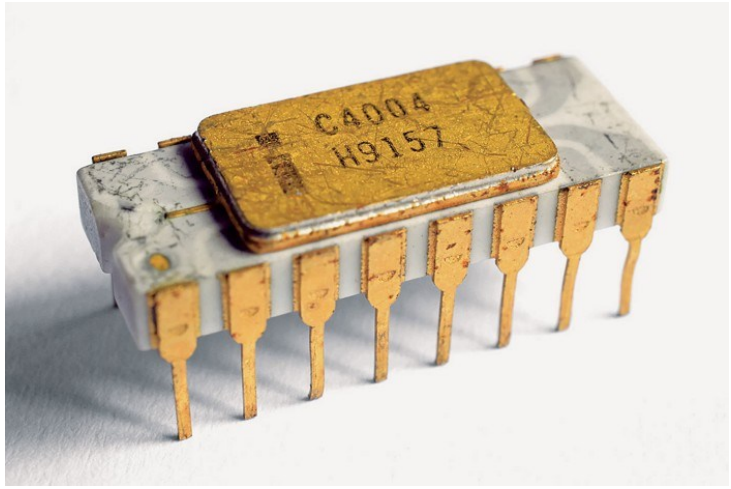
source: https://en.wikipedia.org/wiki/List_of_computer_hardware_manufacturers

Great Reading* (highly recommended)

*Available on Canvas > Files > Reading Materials

The first commercial (micro)processor

Year 1971: the unveiling of the first commercially produced microprocessor.



- 2,300 transistors
- 12 mm^2 chip for a printing calculator made by the Japanese company Busico
- Packing the computing power of the room-sized, vacuum tube–based first computers into a chip with the size of a fingernail!

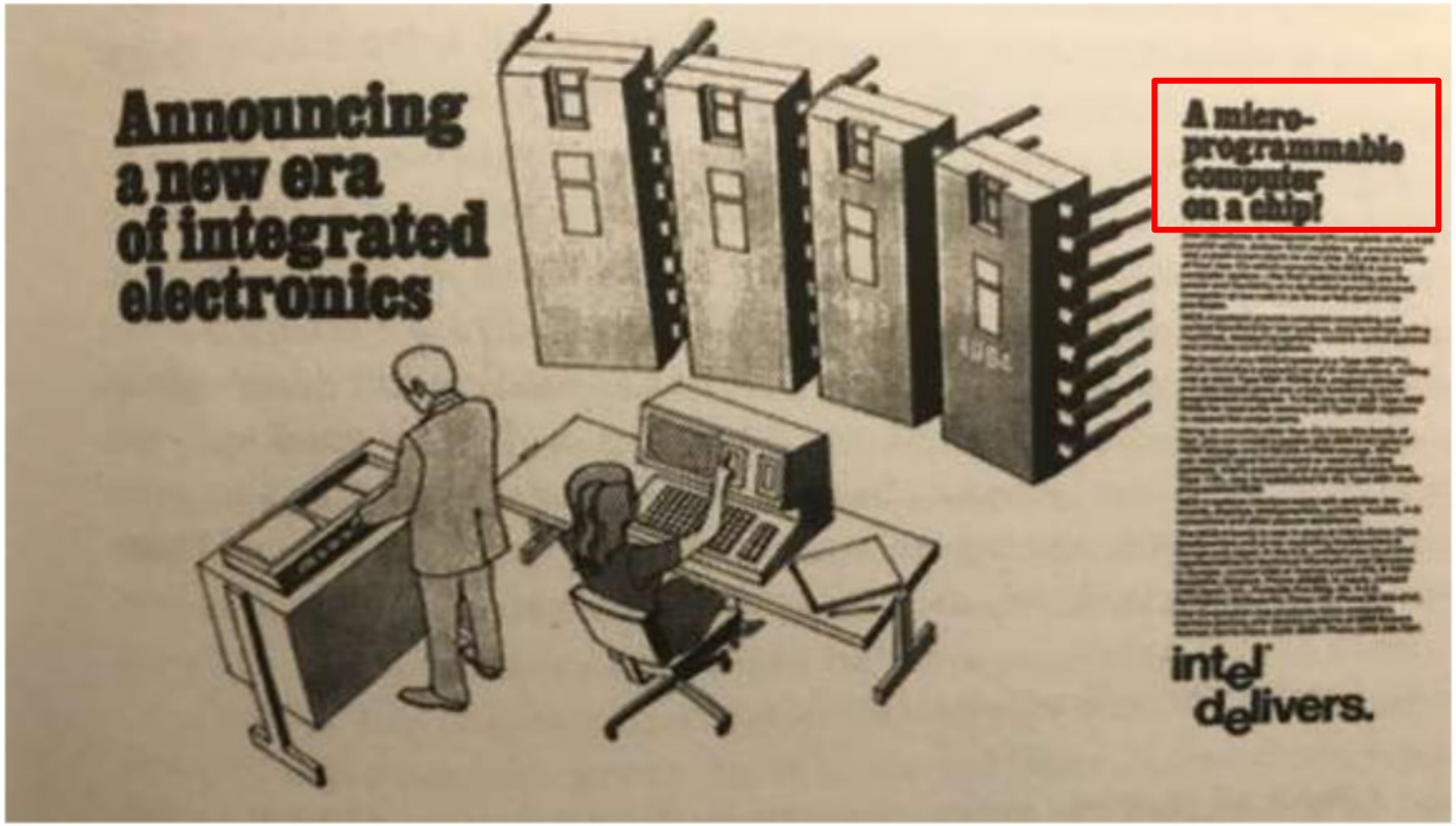
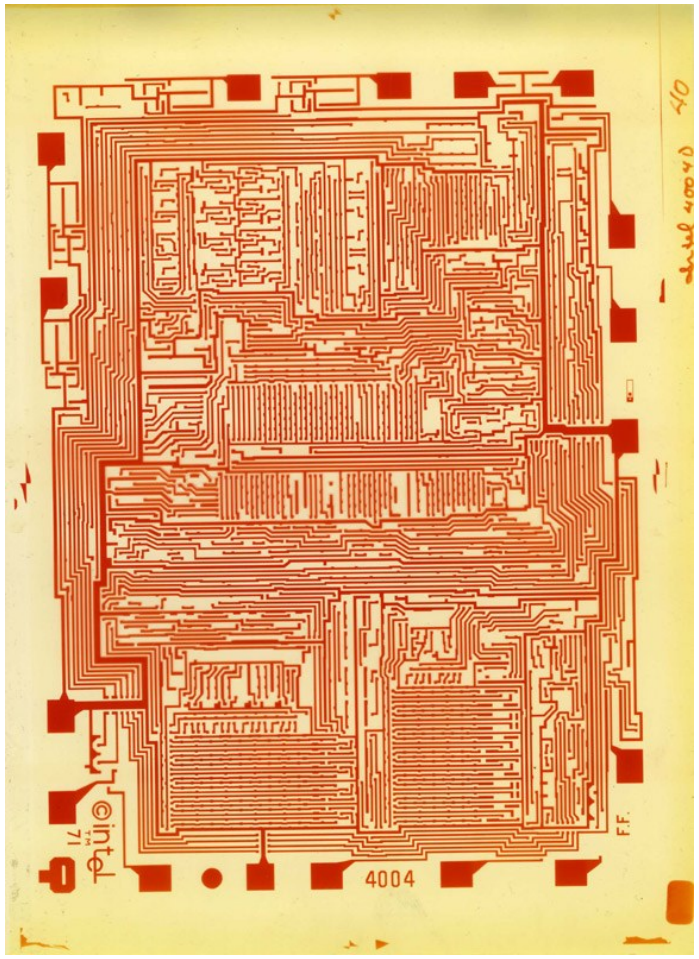


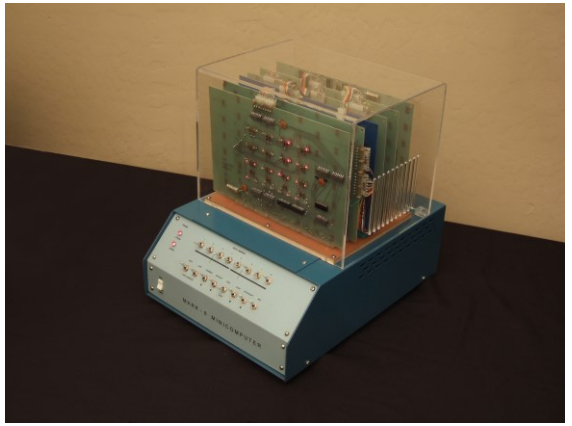
FIGURE 1. Advertisement in the *Electronics News Weekly* in November 1971 announcing the Intel 4004.

Intel 4004 Fun Facts



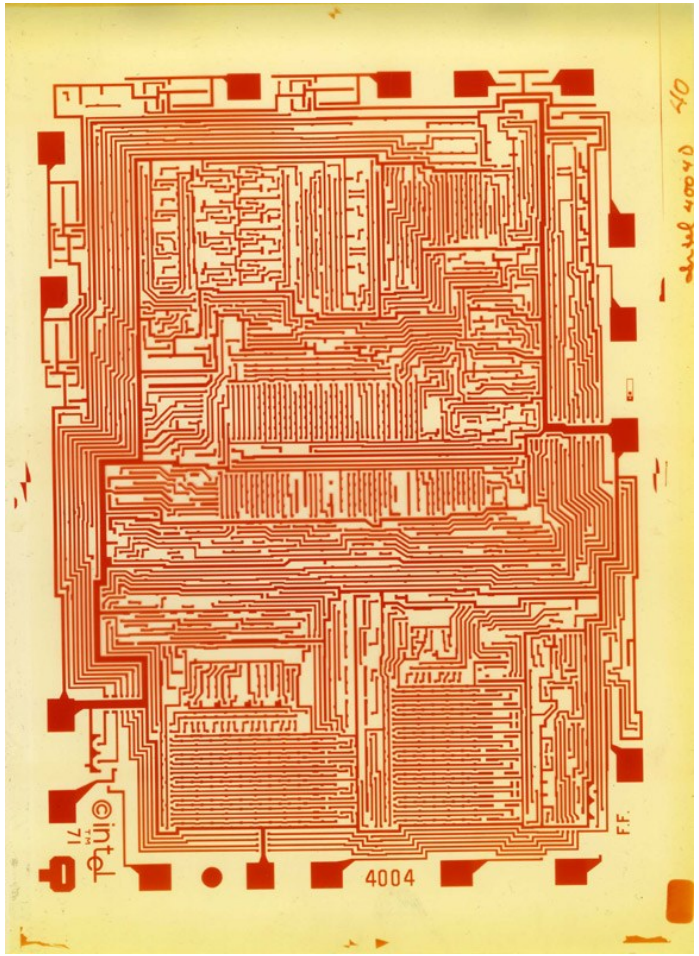
Intel 4004 metal layout.

- "Prior to the Intel 4004, computer central processing units ("CPUs") were usually one or several circuit boards packed with ICs and discrete electronic components."
- "The 4004 didn't see widespread use relative to the 8-bit Intel chips that followed."
- "Only five months after Intel announced the 4004 in Electronic News, the firm shipped the Intel 8008, the first 8-bit microprocessor"
- The 8008 launched an era of at-home hobby computers like the Mark-8 in 1974 and spurred the personal computer industry.



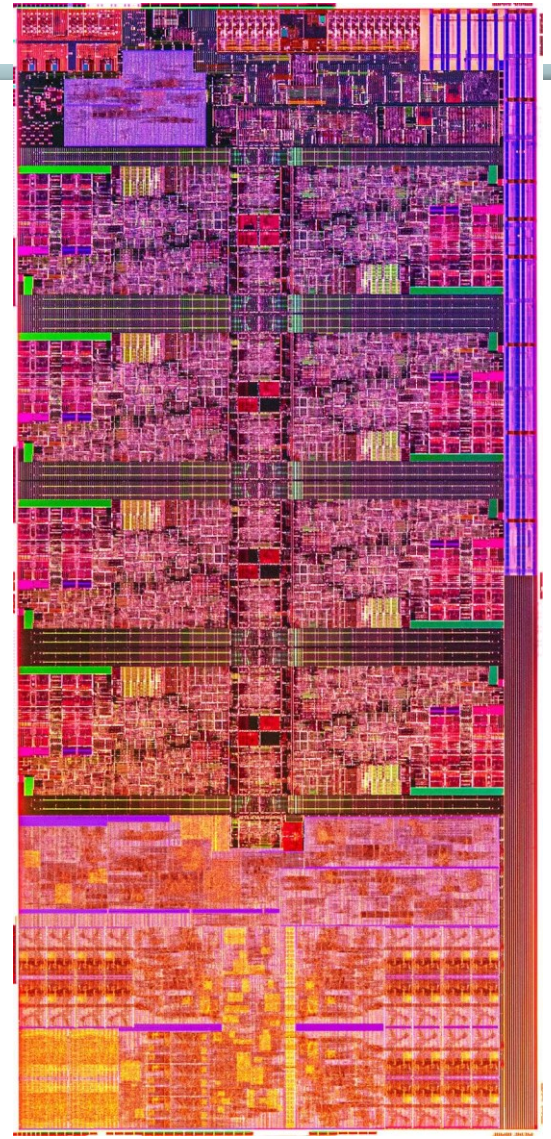
Images:
- https://arstechnica.com/information-technology/2011/11/the-40th-birthday-ofmaybethe-first-microprocessor/
- https://hackaday.com/2016/12/24/mark-8-2016-style/
- https://bytecollector.com/mark_8.htm

What has changed?



Intel 4004, Released in 1971

???



Intel i9, Released in 2021

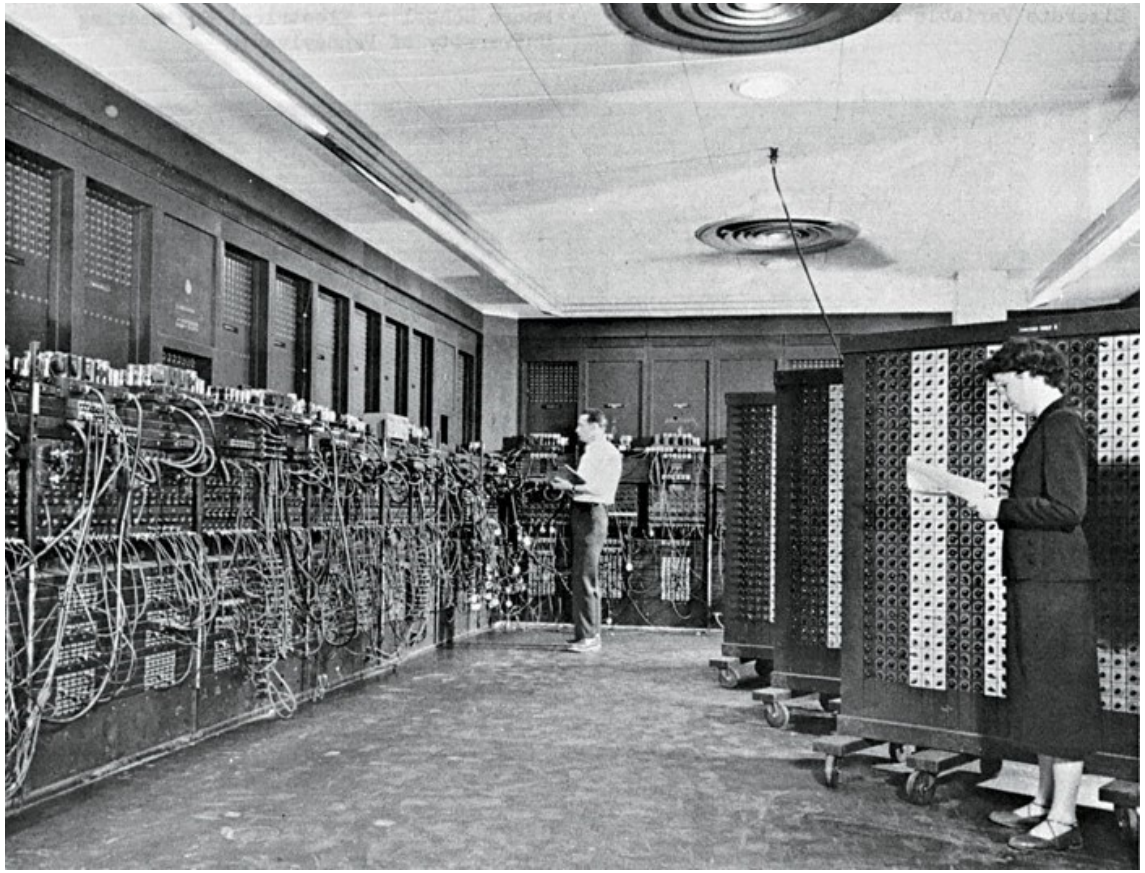
source: Intel

Then and Now: The Intel 4004 vs an Intel Core i9-12900K

CPU Model	Intel 4004 (1971)*	Intel Core i9-12900K (2021)*
Date Publicly Announced	November 15, 1971	October 27, 2021*
Price (2021 Dollars)	$401.41	$589.00
Price (1971 Dollars)	$60.00*	$87.82
Max Clock Speed	0.00074 GHz (740 kHz)	5.20 GHz (5,200,000 kHz)
Word Size	4-bit	64-bit
Cores	1	16
Threads	1	24
Memory Limit	0.000004 GB (4 KB)	128 GB (134,217,728 KB)
Power Usage	1 W	125-241 W
Process Size	10,000 nm (10 µm)	10 nm (0.010 µm)
Die Size	12 mm² (4 mm × 3 mm)	215.25 mm² (20.5 mm × 10.5 mm)*
Transistor Count	2,250	~21,700,000,000

source: https://www.howtogeek.com/766783/the-microprocessor-is-50-celebrating-the-intel-4004/

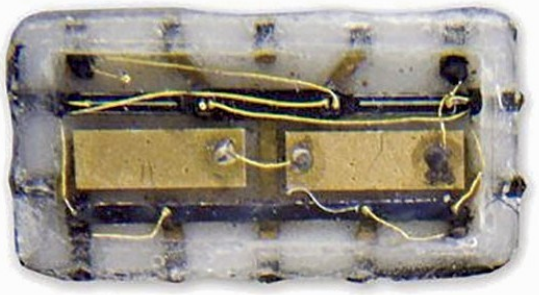
Then (Many “first”)



The first computer: vacuum tube–based ENIAC took up an entire room.

The first transistor (replica shown here) was made at AT&T’s Bell Labs in 1947.

The first commercial silicon transistor was the Texas Instruments Type 905.



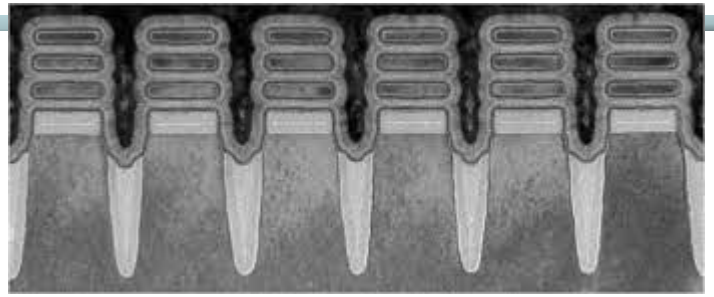
Texas Instruments’ first solid circuit chip used flying gold wires.

https://cen.acs.org/materials/electronic-materials/microprocessor-turns-50/99/i25

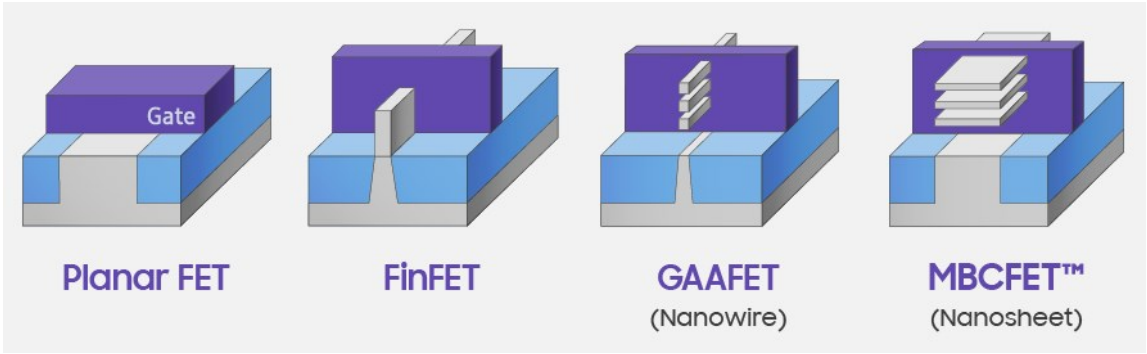
Now (Still many “first”)

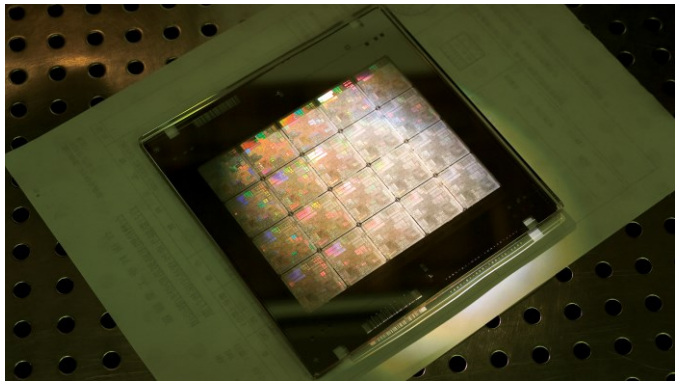
The Fugaku supercomputer, World's Fastest Supercomputer as of Nov. 2021.

https://www.japantimes.co.jp/



Nanosheet

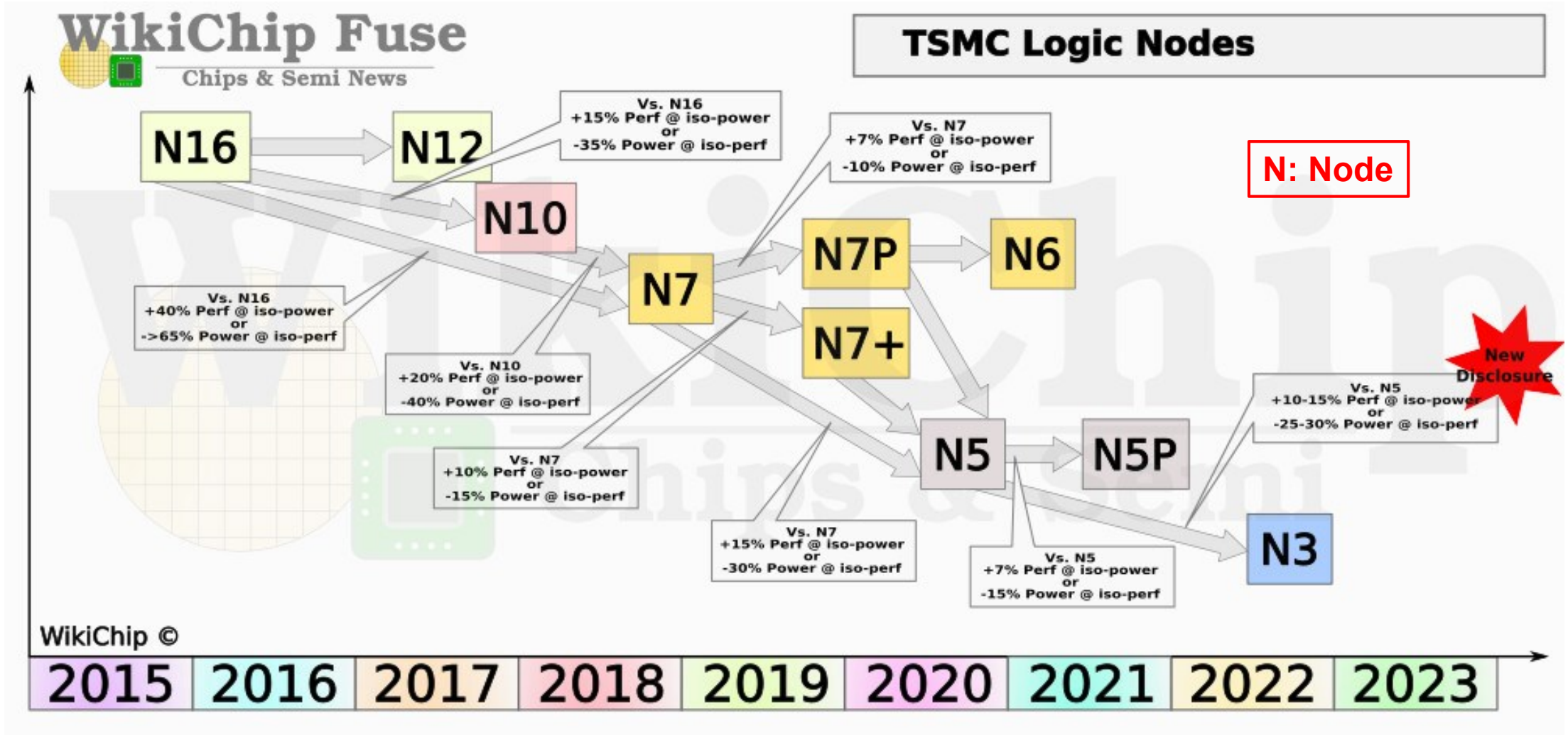




2nm Transistor Wafer, IBM

What is happening now in computer hardware world?

TSMC's Roadmap



source: wikichip

From Sensors and Smartphones to Servers

An area optimized microcontroller core (e.g., Arm Cortex-M0)

1X

1 square represents the **area** of this core

High-performance 32-bit core (e.g., Arm Cortex-M7) Used in automotive, sensor hub, and other embedded applications.

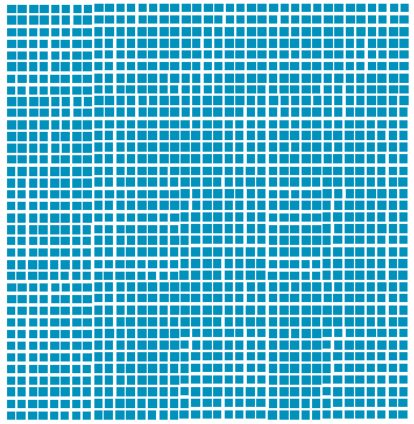
13X

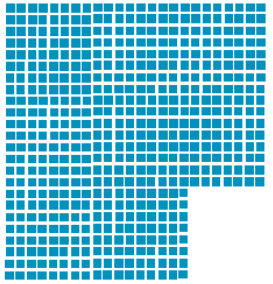
Mid-range 64-bit processor (e.g., Arm Cortex-A55). For smartphones, TVs, network infrastructure, …

130X

1 laptop or server class processor (e.g., A76 core with 512KB of L2 cache)

1380X



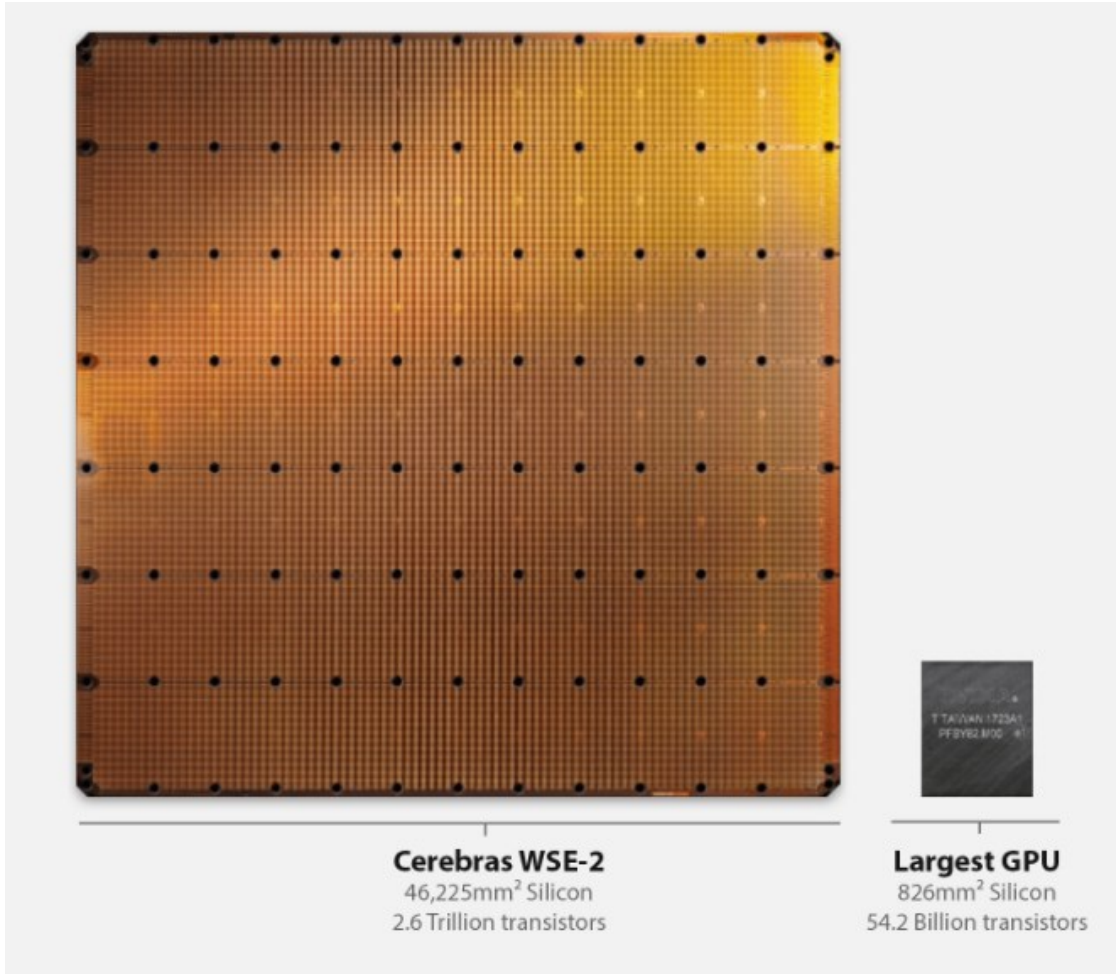


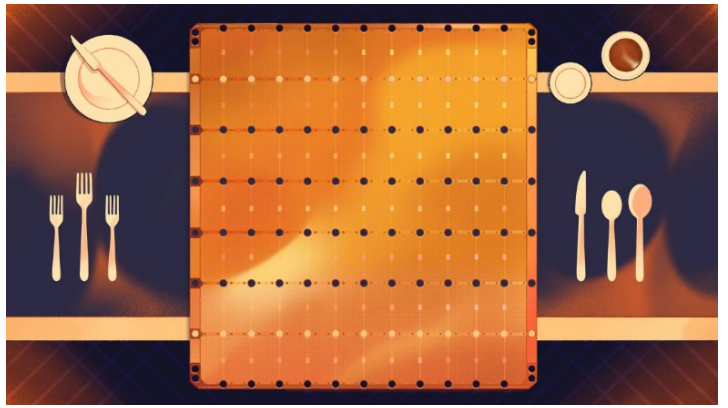
520X

High-performance processor (e.g., Arm Cortex-A73). For mobile and consumer devices.

source: ARM

The World's Largest Computer Chip

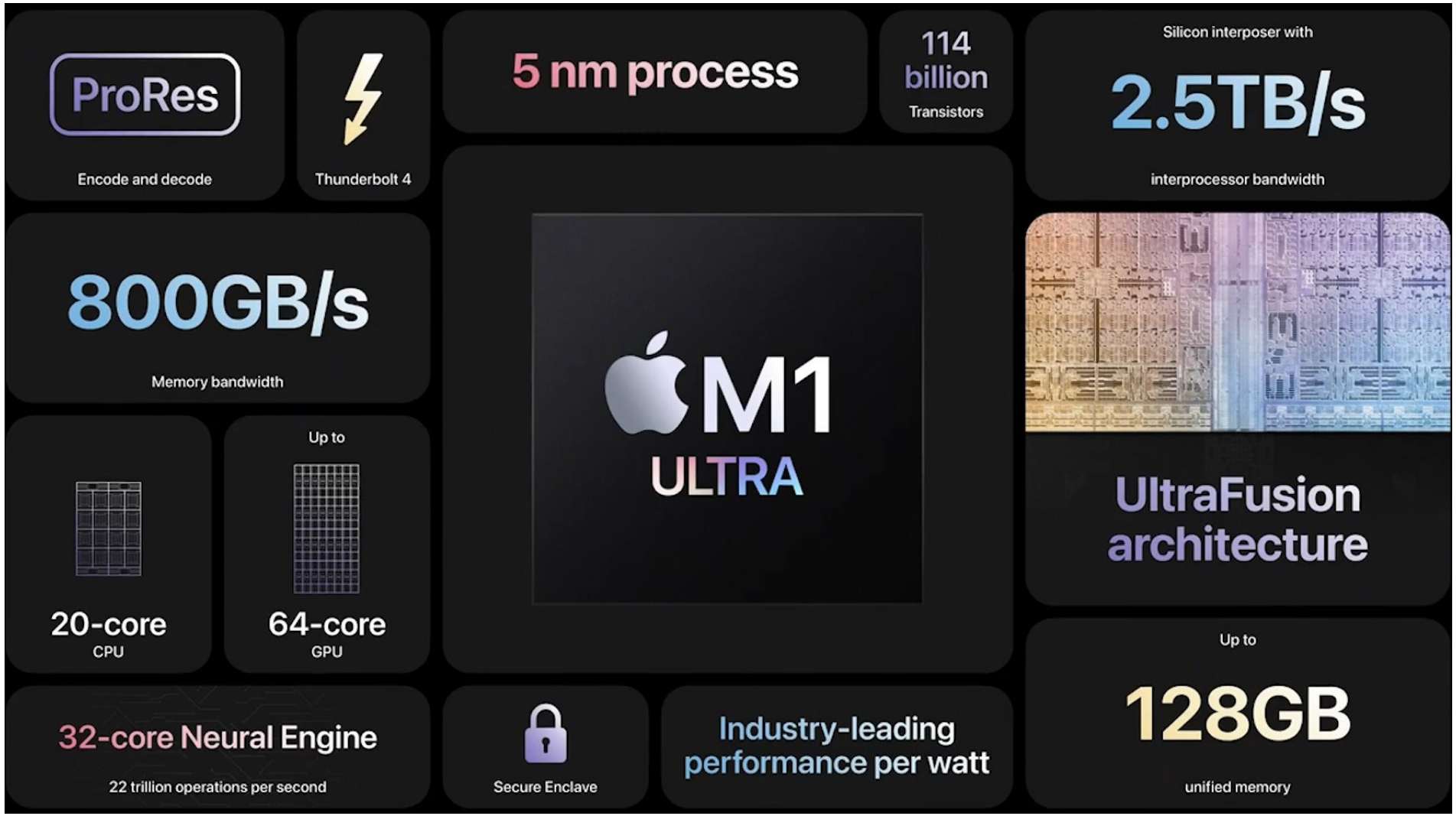




Images:

- https://www.newyorker.com/
- https://cerebras.net/chip/

Apple M1 Ultra (2022)



Source: Apple

Exascale supercomputer

- Count how many zeros here

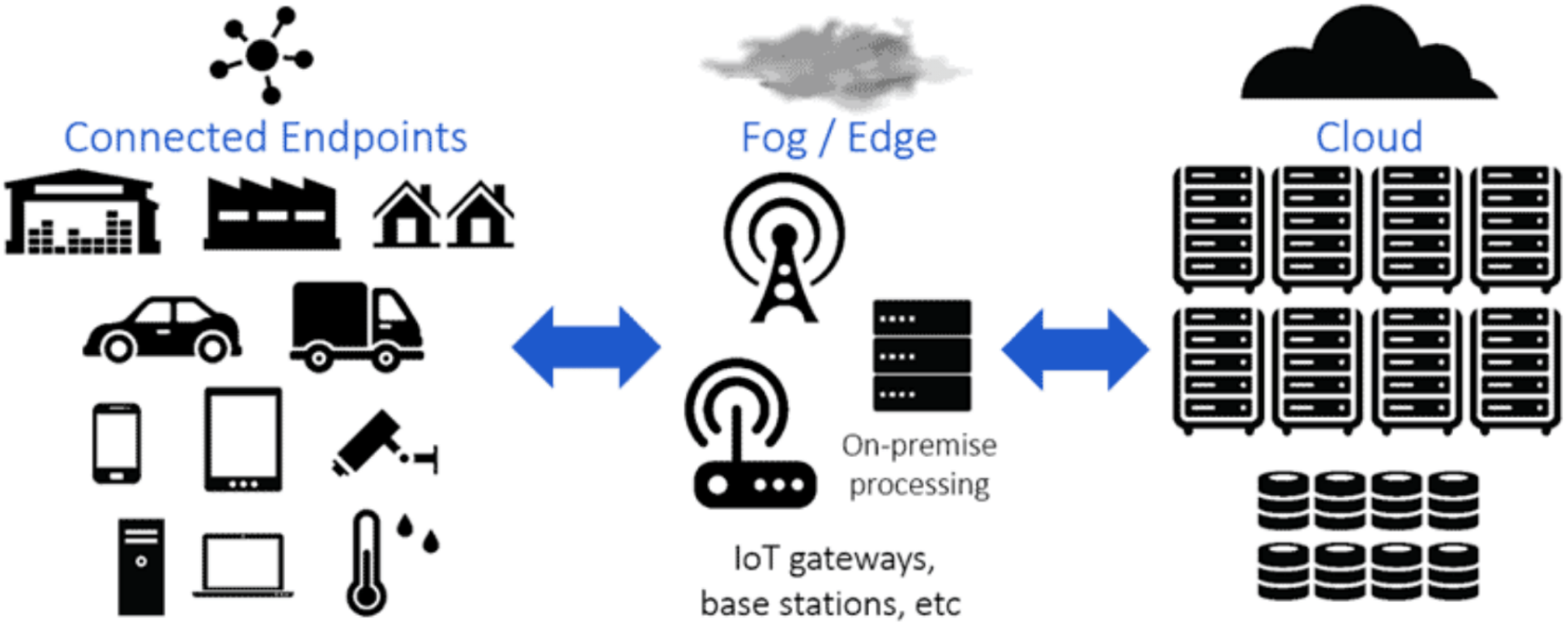
10000000000000000

- What is this?

One quintillion operations (per second)

- Frontier: the newest supercomputer underway at ORNL

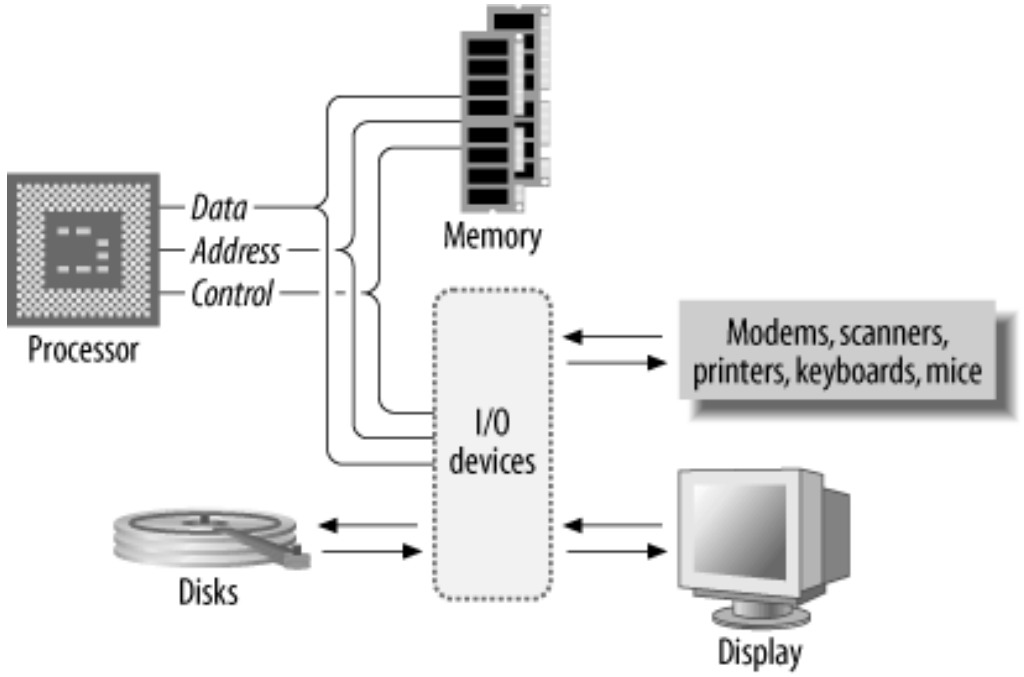
Trillions of edge devices



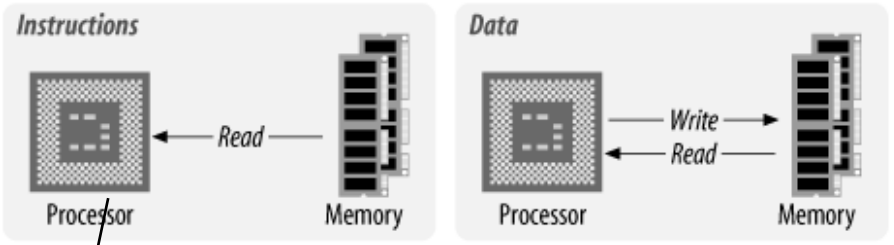
source: https://semiengineering.com/winners-and-losers-at-the-edge/

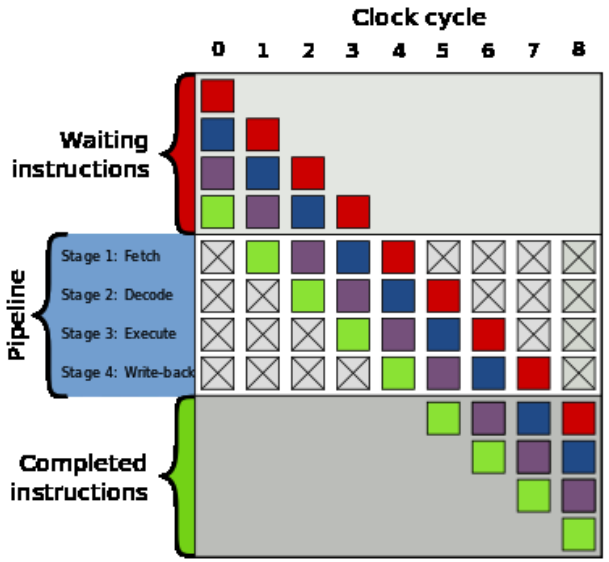
Back to this course: Computer Architecture (Evolving through ages…)

A basic computer system

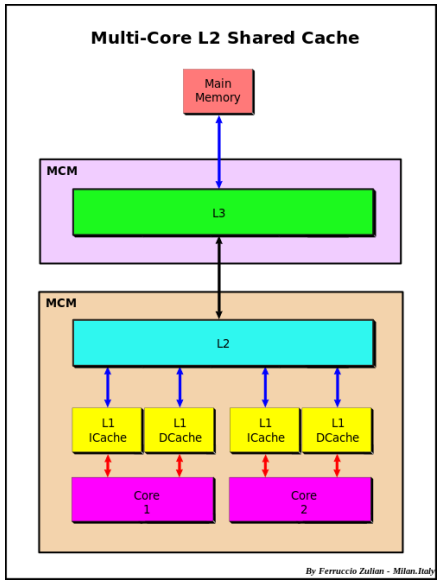


Dataflow





Pipeline



Cache

https://hackernoon.com/

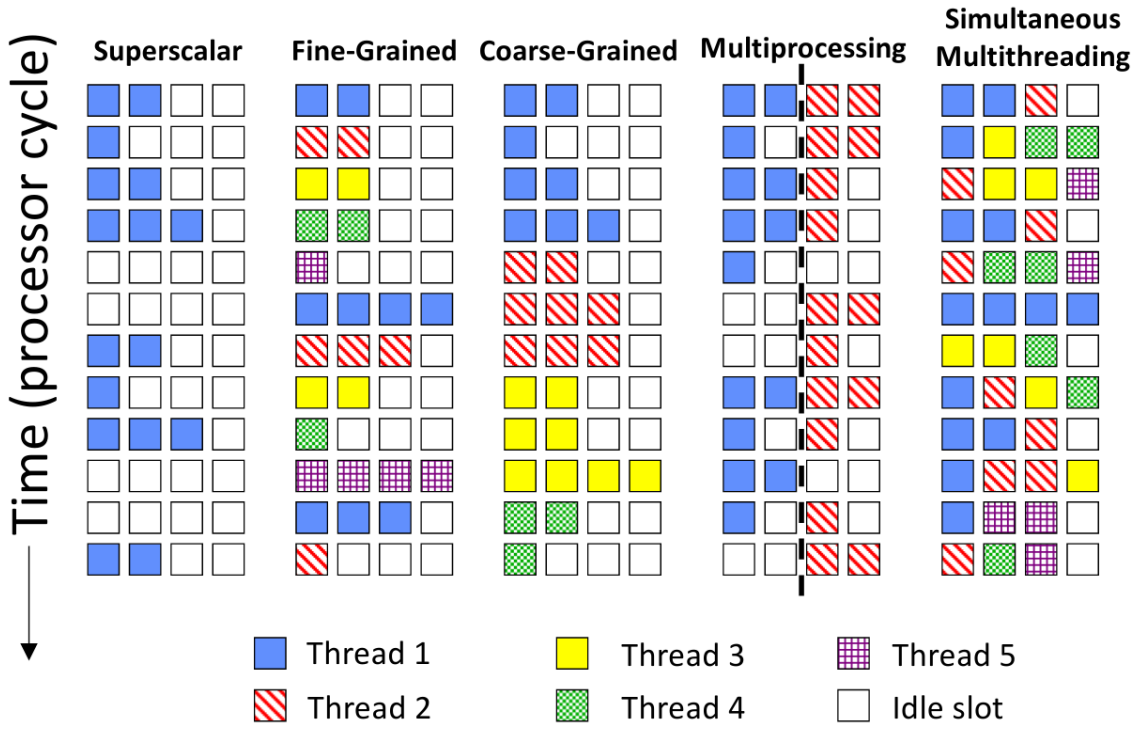
source: https://www.oreilly.com/library/view/designing-embedded-hardware/0596007558/ch01.html

The trending architectures

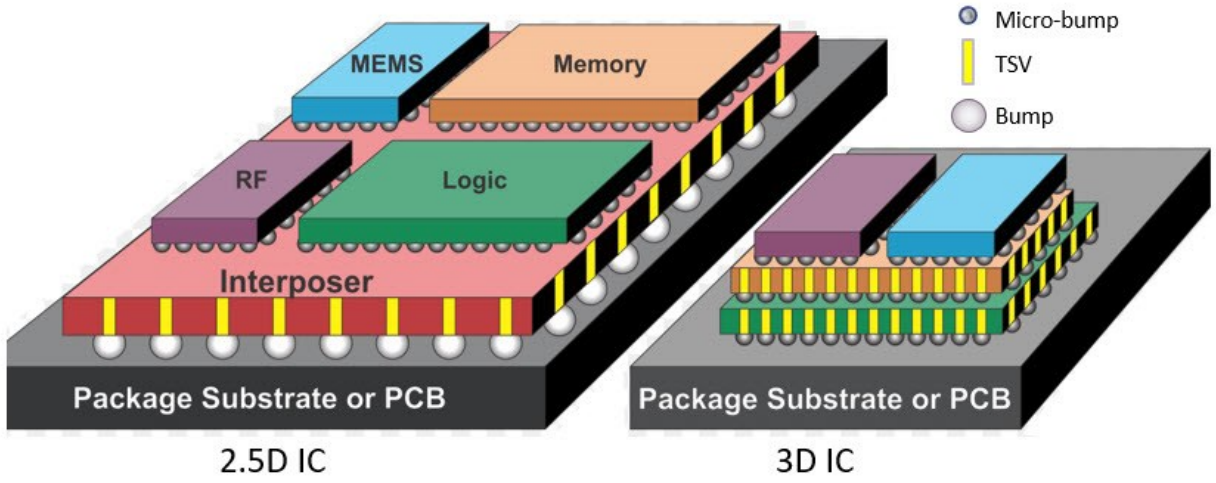
- From sequential to parallel
- From horizontal to vertical
- From homogeneous to heterogenous
- From single die to multidie

From sequential to parallel

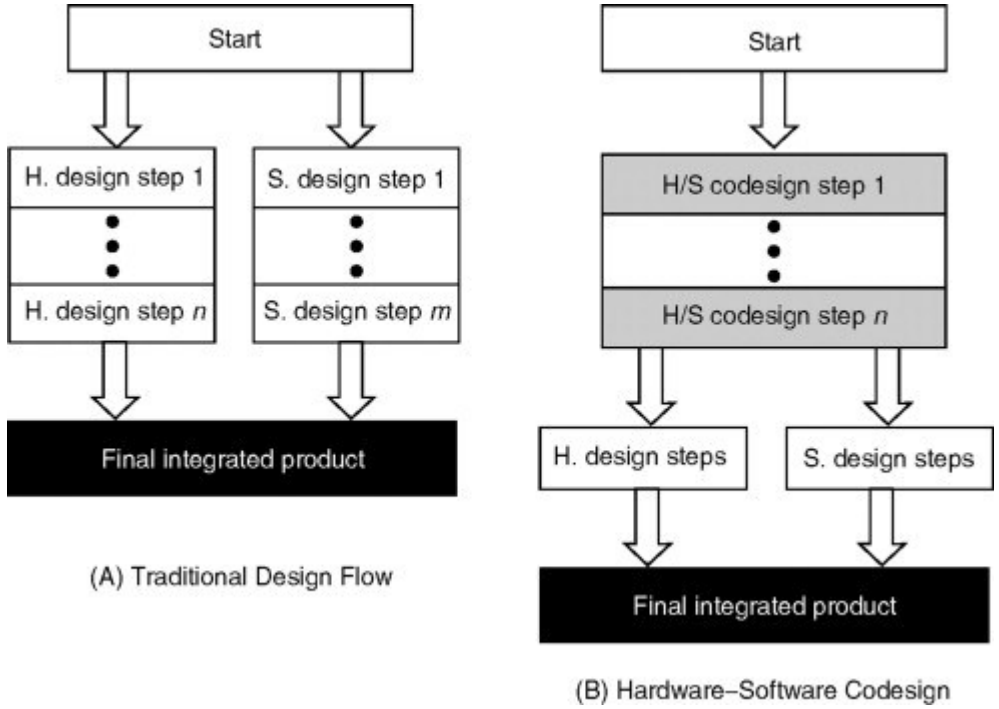
- Instruction count
  - Increased datapath width (e.g., 16-bit to 32-bit to 64-bit)
  - Larger register files (fewer load/store instructions)
  - More complex instructions?
  - SIMD instructions
- Parallel
  - Instruction-level parallelism
  - Multithreading
  - Multicore
  - Multidie



From horizontal to vertical



From 2D to 3D Integration

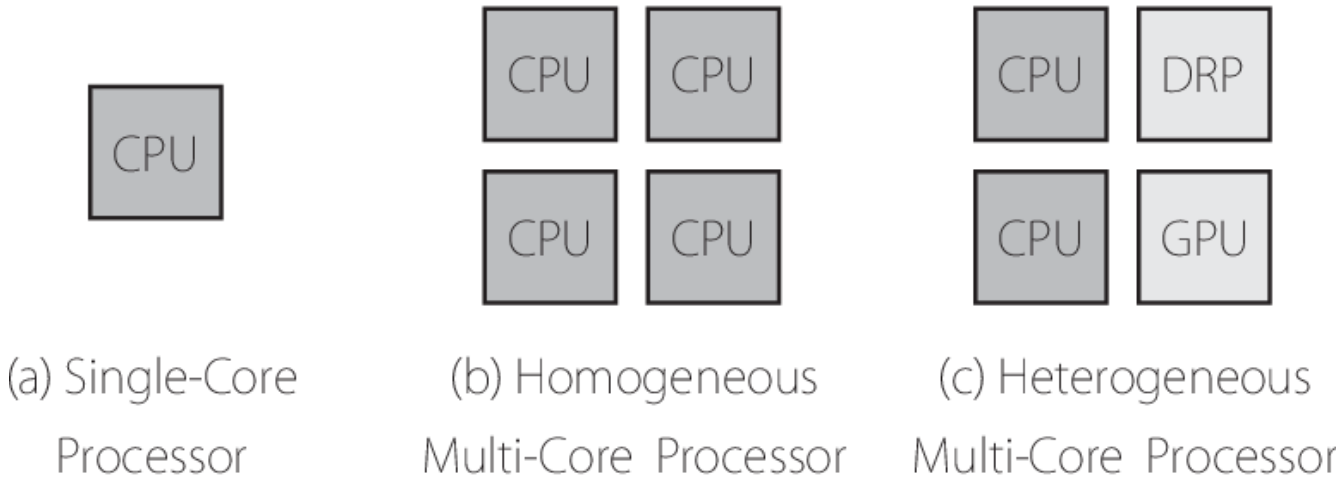


HW/SW Codesign

Sources:

- https://semiengineering.com/now-you-can-automate-latch-up-verification-for-2-5-3d-technologies/
- https://www.sciencedirect.com/topics/engineering/hardware-software-codesign

From homogeneous to heterogenous



source: Hayashi A. Studies on automatic parallelization for heterogeneous and homogeneous multicore processors[D]. Waseda University, 2012.

- System on Chip
- Specialization
  - Domain-specific Architecture
  - AI Accelerators
  - DSP
  - FPGAs

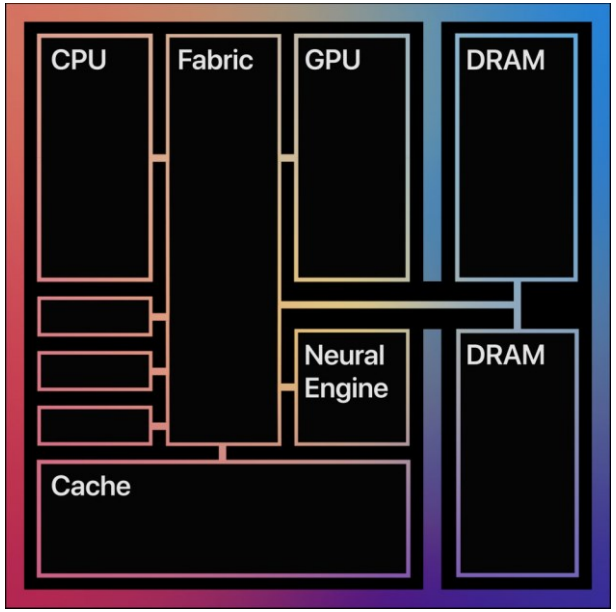
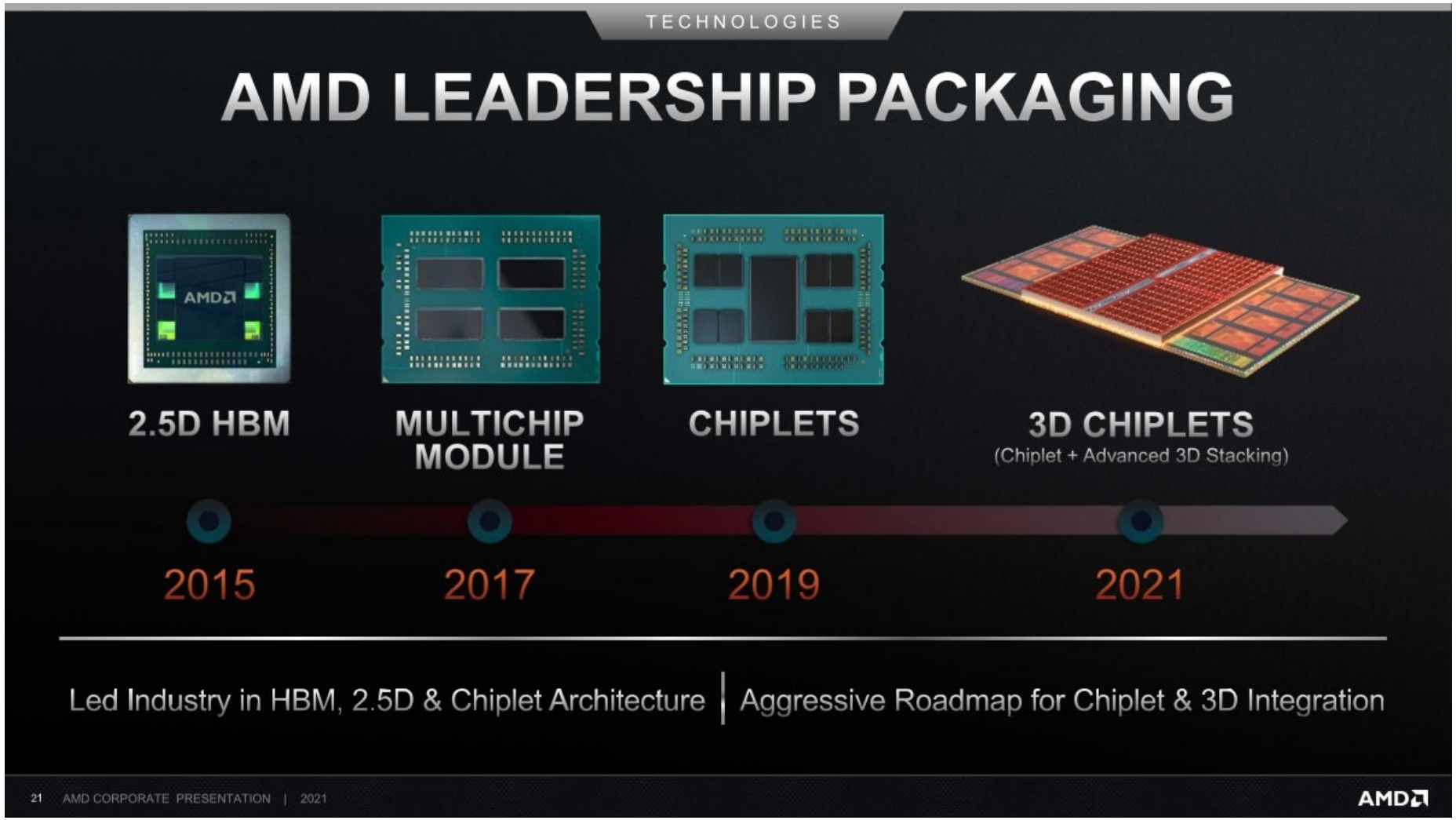


Image: Apple

From single die to multidie



source: AMD

Why did I show those?

- It is very relevant to this course!
- But we will not cover them all, will go deep on certain aspects, will cover basics of some of others…
  - Understand the motivation of these developments
  - Develop deep understandings of instruction level parallelism, out of order, and more
  - Being able to stay refreshed on top of these trends
  - …

Where are we Heading?

- T2: Fundamentals of Quantitative Design and Analysis
  - ISAs, power, performance, other measures

Acknowledgement

Slides in this topic are inspired in part by material developed and copyright by:

- ARM Courseware
- NVIDIA Courseware
- Prof. Ron Dreslinski @ UMich, EECS 470
- Prof. Onur Mutlu @ ETH

Action Items

- Read syllabus
- Join Feishu group
- Join piazza group
- Finish the course survey
- Reading Materials
  - Syllabus, Lecture slides, Reading materials on canvas