CENG 3420
Computer Organization & Design
Lecture 01: Introduction

Textbook: Chapter 1.1, 1.5-1.8

Zhengrong Wang

CSE Department, CUHK

zhengrongwang@cuhk.edu.hk

Course Information

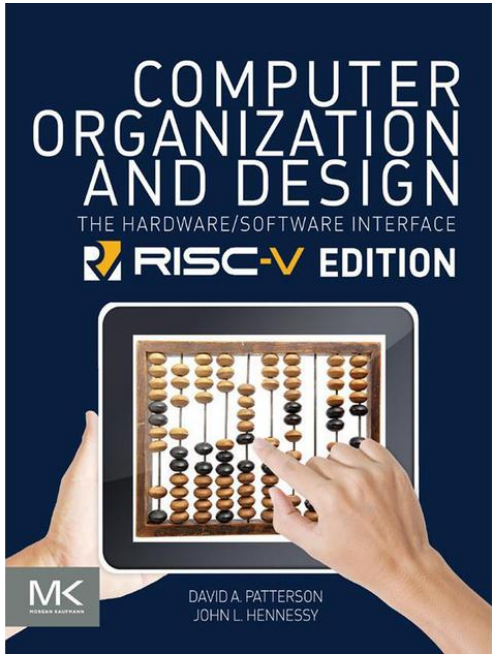
- Instructor:
  - Zhengrong Wang (zhengrongwang@cuhk.edu.hk)
  - SHB 931
  - Office Hour: W15:00-17:00
- Tutors:
  - Fangzhou Liu fzliu23@cse.cuhk.edu.hk
  - Jiahao Xu jhxu24@cse.cuhk.edu.hk
  - Yifan Shi yfshi24@cse.cuhk.edu.hk
  - Libo Shen lbshen24@cse.cuhk.edu.hk
  - Xiaoman Yang xmyang25@cse.cuhk.edu.hk
- Piazza for Q&A
  - https://piazza.com/cuhk.edu.hk/spring2026/ceng3420

Grading Information

- Grade determined by:
  - 20% Homework
  - 20% Midterm
  - 20% Three labs (individual work)
  - 40% Final exam
  - Late submission **per natural day** is subject to **10%** of penalty, maximum **5** days.
    - E.g., two days' late submission with original score 90 → 90 – 10 * 2 = 70.
  - You must gain at least **50%** of the full marks to pass.

General References

- Textbook:
  - Computer Organization and Design, RISC-V Edition
- Manuals:
  - RV32 Reference Card
  - RV32-I Manual (on course webpage)
  - Lab tutorials (slides)
- Slides:
  - On the course webpage before lecture
  - I may tweak the slides before the lecture



Course Content

- Introduction to the major components of a computer system, how they function together in executing a program
- Introduction to CPU datapath and control unit design
- Introduction to techniques to improve performance and energy-efficiency of computer systems
- Introduction to multiprocessor architecture
- So that...
  - ***Future software designers*** (compiler writers, operating system designers, database programmers, application developers, ...) can achieve the best cost-performance trade-offs.
  - ***Future architects*** understand the effects of their design choices on software.

Why Learn This Stuff?

- You want to call yourself a "computer scientist/engineer"
- You want to build HW/SW people use (so need performance/power)
- You need to make a purchasing decision or offer "expert" advice

Both hardware and software affect performance/power:

- Algorithm determines number of source-level statements
- Language/compiler/architecture determine the number of machine-level instructions
- Process/memory determine how fast and how power-hungry machine-level instructions are executed

What You Should Already Know

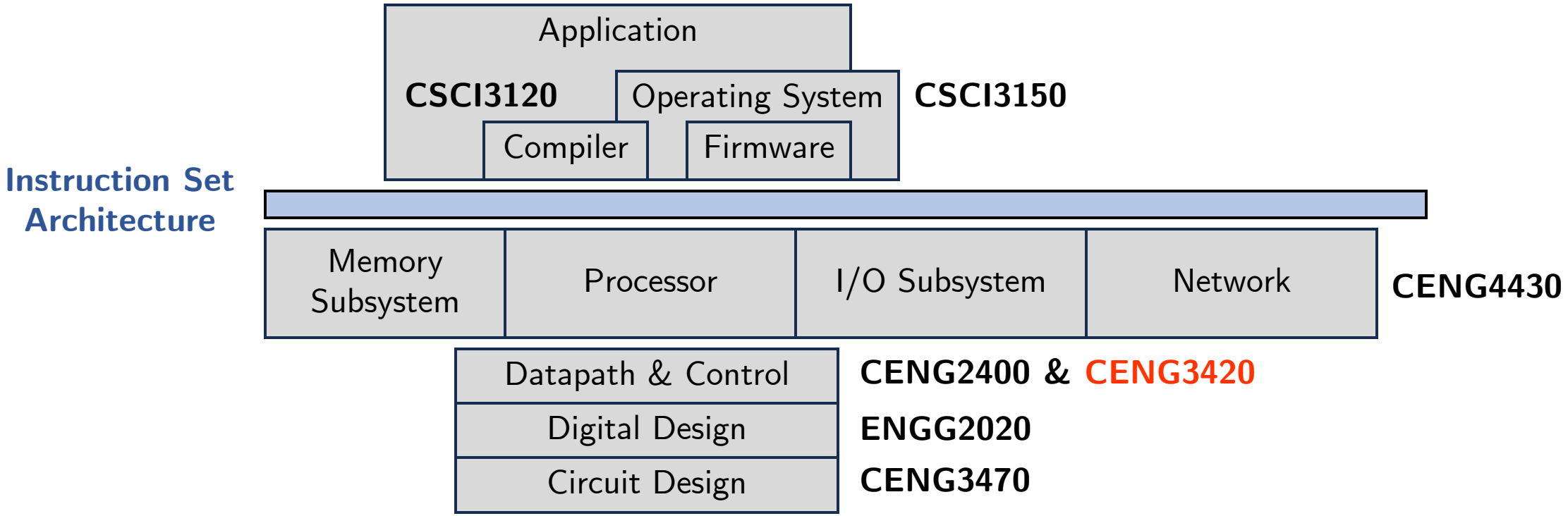
- Basic logic design & machine organization
  - Logical minimization, FSMs, component design
  - Processor, memory, I/O
- Create, run, debug programs in assembly language
  - Will be introduced in tutorial
- Create, compile, and run C/C++ programs
- Create, organize, and edit files and run programs on Unix/Linux

One example here to compile and run simple program in WSL.

Computer Organization and Design

- This course is all about how computers work
- But what do we mean by a computer?
  - Different types: embedded, laptop, desktop, server, datacenter, ...
  - Different users: automobiles, graphics, finance, genomics, ...
  - Different manufacturers: Intel, Apple, AMD, NVIDIA, IBM, Oracle, ...
  - Different underlying technologies and different costs
- Analogy: Consider a course on "automotive vehicles"
  - Many similarities among vehicles (e.g., wheels)
  - Huge differences among vehicles (e.g., gas vs. electric)
- Best way to learn:
  - Focus on a specific instance and learn how it works
  - While learning general principles and historical perspectives

How Do the Pieces Fit Together?



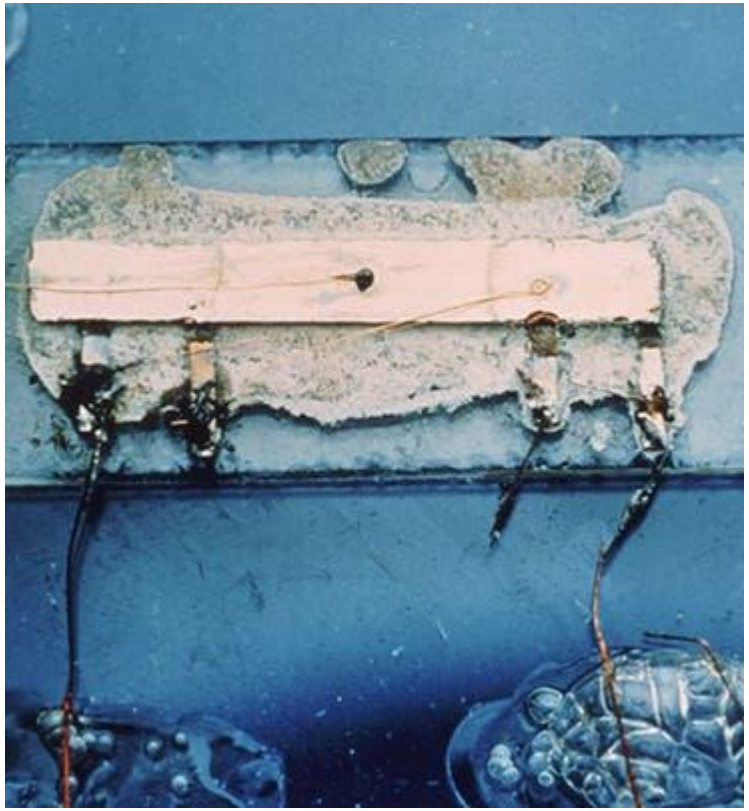
- Coordination of many levels of abstraction
- Under a rapidly changing set of forces
- Design, measurement, and evaluation

The Evolution of Computer Hardware

- When was the first transistor invented?
  - 1947, bi-polar transistor, by John Bardeen, Walter Brattain, and William Shockley at Bell Labs
  - Nobel Prize in Physics in 1956
  - Laid the foundation for integrated circuits and modern computing
  - Smaller, more reliable, more energy efficient, faster than vacuum tubes

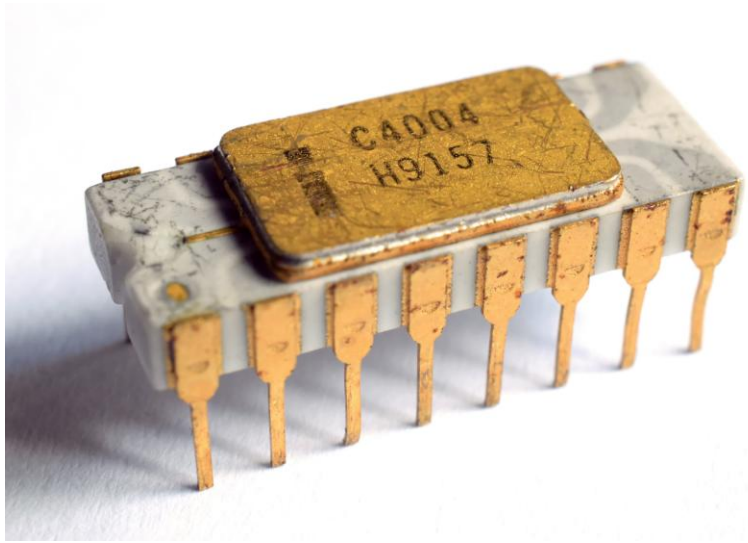
The Evolution of Computer Hardware

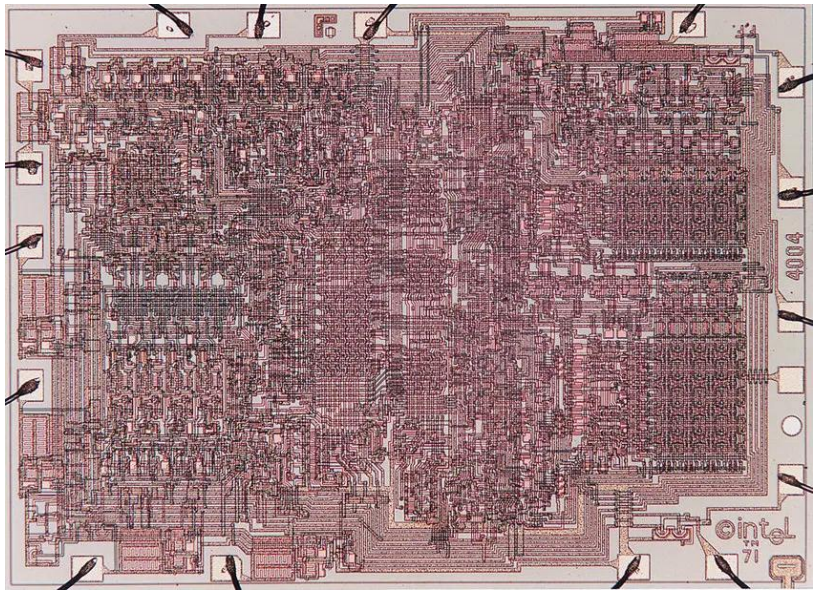
- When was the first integrated circuit (IC) invented?
  - 1958, by Jack Kilby at Texas Instruments
  - Nobel Prize in Physics in 2000
  - Laid the foundation for microprocessor and modern computing
  - Integrates components on single substrate



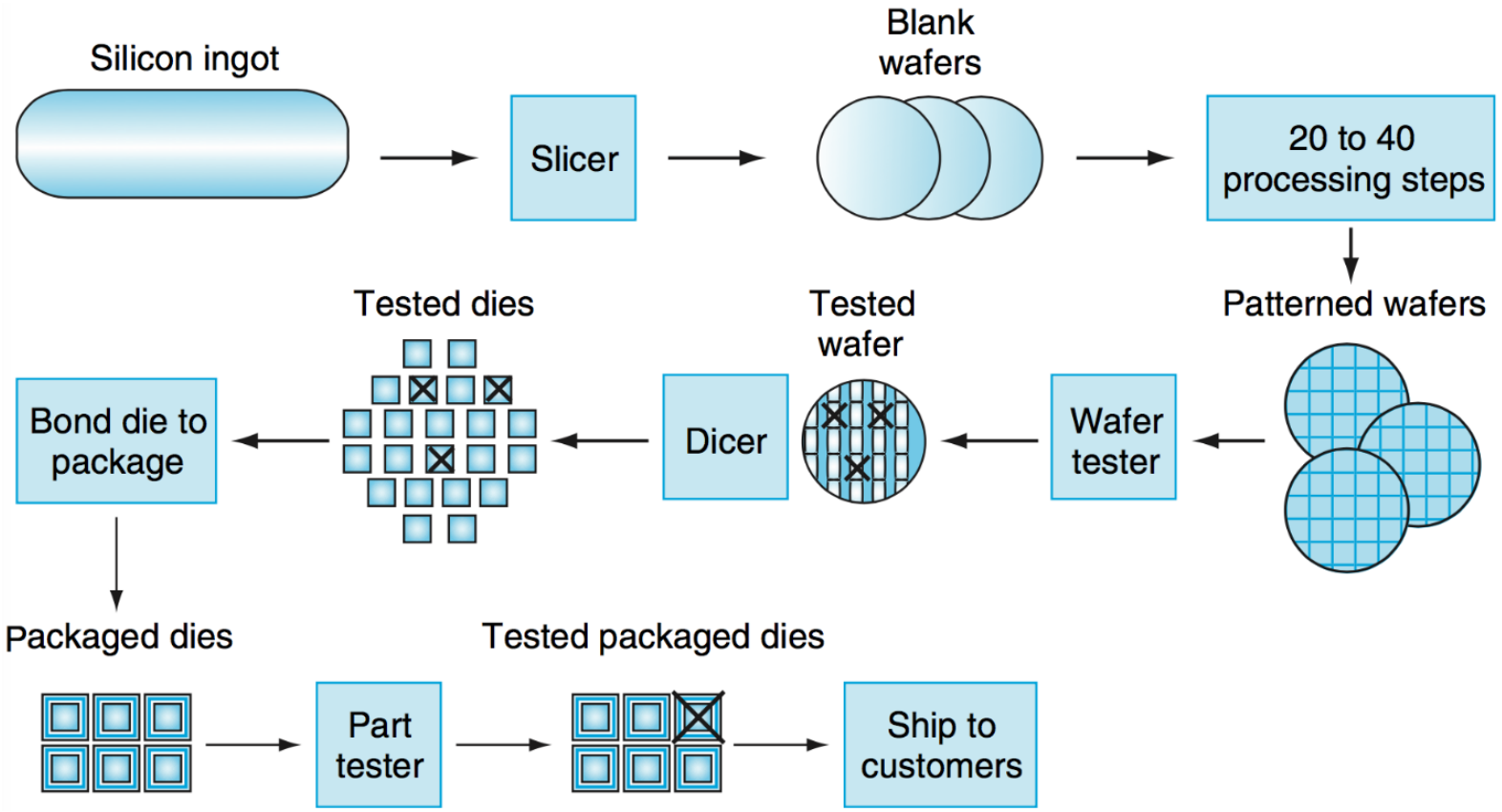
The Evolution of Computer Hardware

- When was the first microprocessor?
  - 1971, Intel 4004
  - 4-bit CPU with 2,300 transistors, 740 kHz, 640 bytes of addressable memory
  - First chip to integrate all functions of a CPU





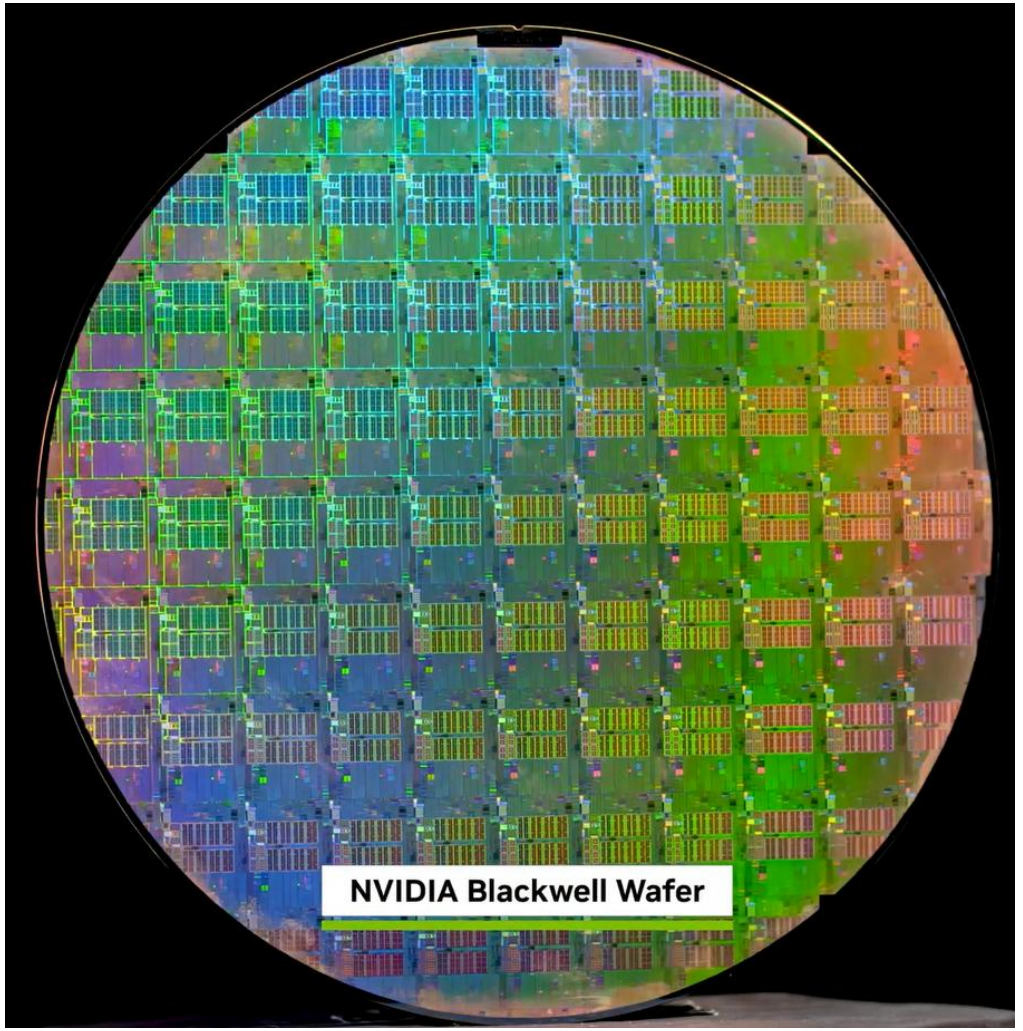
The IC Manufacturing Process



- Yield: Proportion of working dies per wafer
- Check this: https://www.youtube.com/watch?v=d9SWNLZvA8g&list=FLELqiXCJQW-jcijW8ZAbA8w
- Longer version: https://www.youtube.com/watch?v=dX9CGRZwD-w

NVIDIA Blackwell NVL72

- A mega system for AI
  - 104 billion transistors per GB200 GPU, TSMC 4NP process
  - 72 GPUs in single rack
  - 1.4 exaFLOPS of compute
  - 130 trillion transistors in total
  - Check this: https://www.youtube.com/watch?v=YsiccpVKUIs
- One of humanity's most refined achievements



Integrated Circuit Cost

$$Cost\ per\ die = \frac{Cost\ per\ wafer}{Dies\ per\ wafer \times Yield}$$

$$Dies\ per\ wafer = \frac{Wafer\ area}{Die\ area}$$

$$Yield = \frac{1}{[1 + (Defects\ per\ area \times Die\ area)]^N}$$

Nonlinear relation to area and defect rate

- Wafer cost and area are fixed
- Defect rate determined by manufacturing process
- Die area determined by architecture and circuit design

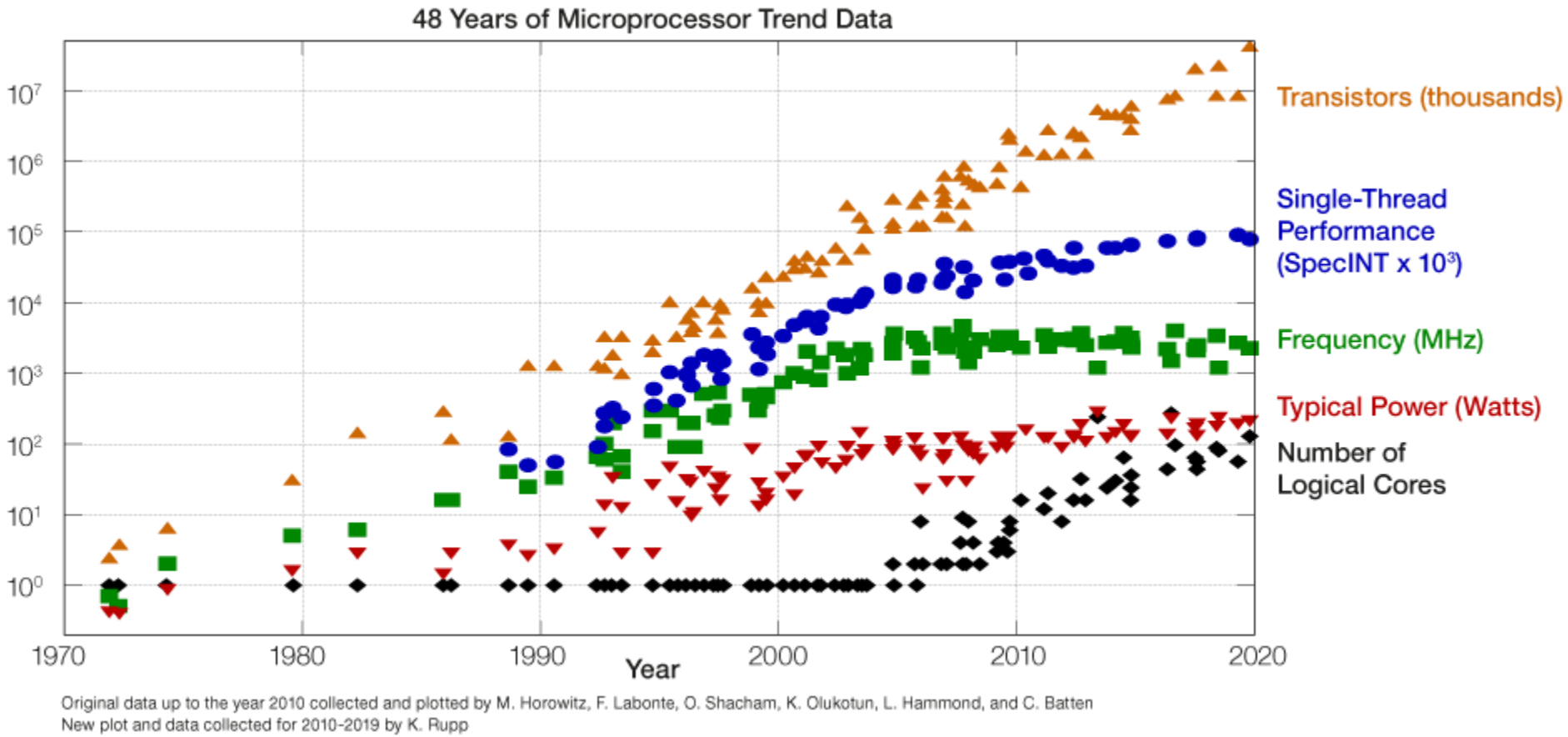
Impacts of Advancing Technology

- Processor
  - Logic capacity: increases about 30% per year; now no longer true
  - Performance: 2x every 1.5 years; now 2x per 2.2-2.4 years[1]
- Memory
  - DRAM capacity: 4x every 3 years, ~60% per year; now about 2x every 3 years
  - Memory speed: 1.5x every 10 years
  - Cost per bit: decreases about 25% per year; now about 15% per year
- Disk
  - Capacity: increases about 60% per year;
  - Now 10-20% per year (HDD), 30%-40% per year (SSD)

We had a good run, but now technology advancement is slowing down...

[1] https://techovedas.com/is-moores-law-dead-assessing-moores-law-over-the-past-decade-with-amd-ceo/

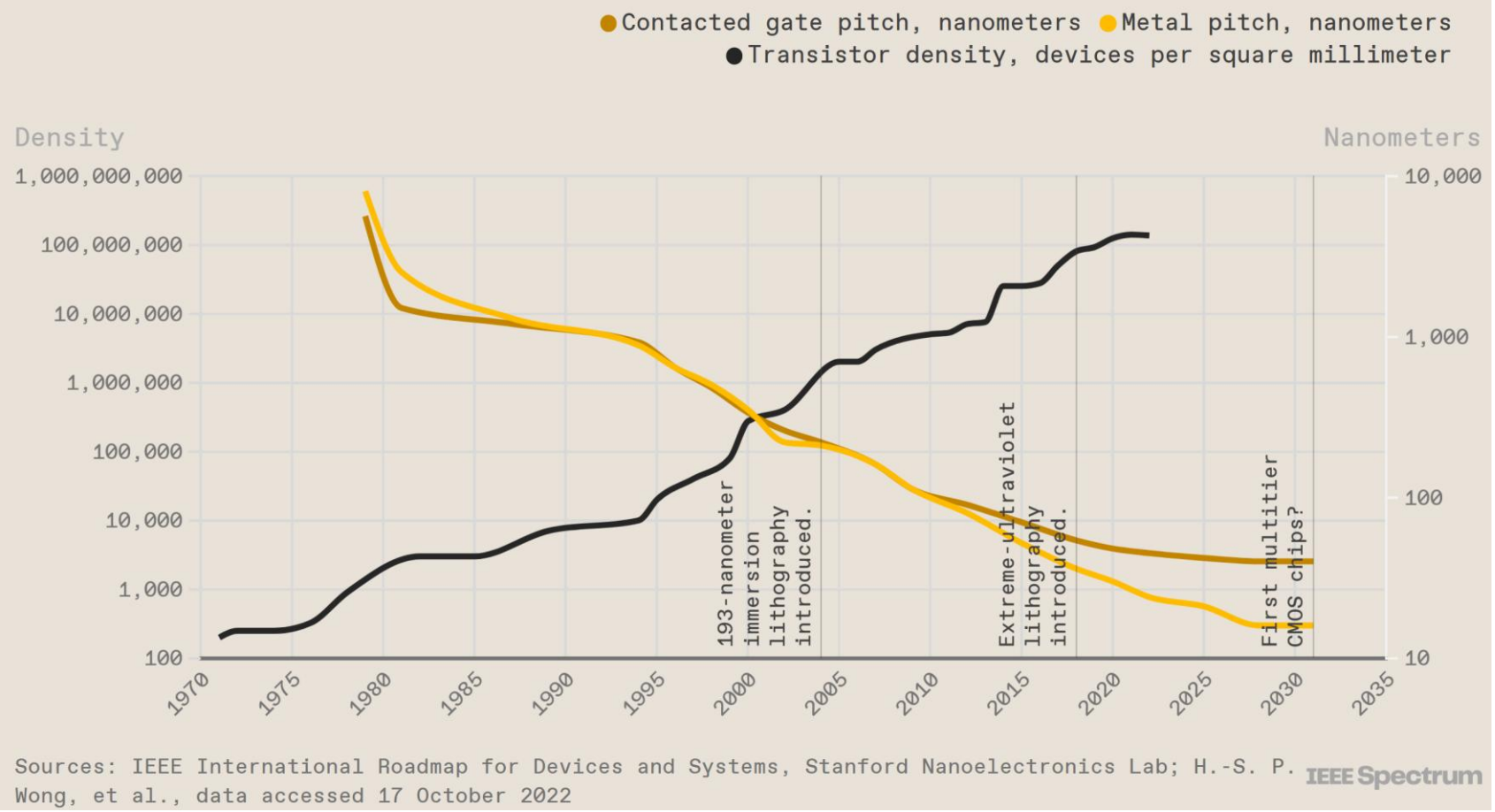
Moore's Law



- Number of transistors doubles every 2 years, driven by device scaling

Main Driver: Device Scaling

- Transistor density in logic circuits has increased more than **600,000-fold** since 1971.
  - https://spectrum.ieee.org/transistor-density
  - https://en.wikipedia.org/wiki/List_of_semiconductor_scale_examples



How to Measure Performance

- Latency vs. Throughput
  - Latency: the time taking to complete one job, e.g., 10ms.
  - Throughput: number of jobs completed per unit time, e.g., 10 jobs/hour.
  - Example: plane (fast but less capacity) vs. ship (slow but massive).
- Execution time = Instruction count x CPI x Clock cycle time
  - Instruction: basic "job" for CPU, e.g., add, multiply, read from memory, ...
  - Instruction count: Number of executed instructions in the program.
  - CPI: cycle per instruction.
  - Clock cycle time: determined by hardware, i.e., 1 / Frequency. 1GHz $\rightarrow$ 1ns = 10^{-9}s.
- Determined by various aspects:
  - Software: Algorithm, programming language, compiler.
  - Architecture: Instruction set architecture.
  - Hardware: technology scaling.

Intel's Frequency Scaling

- Increasing frequency is often the most effective strategy.
  - Great for branding too, but does it always translate to better performance?

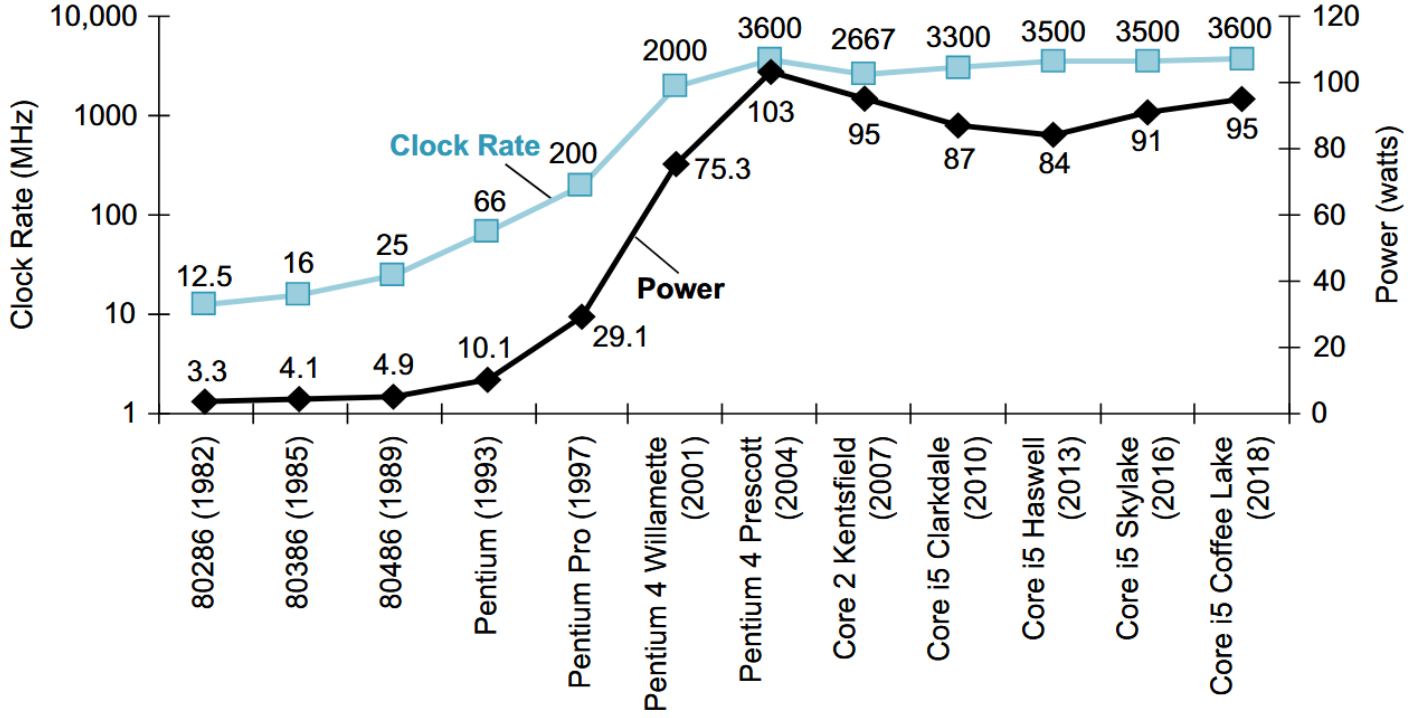


FIGURE 1.16 Clock rate and power for Intel x86 microprocessors over nine generations and 36 years. The Pentium 4 made a dramatic jump in clock rate and power but less so in performance. The Prescott thermal problems led to the abandonment of the Pentium 4 line. The Core 2 line reverts to a simpler pipeline with lower clock rates and multiple processors per chip. The Core i5 pipelines follow in its footsteps.

Hitting the Power Wall

$$\boldsymbol{Power} \propto \frac{\mathbf{1}}{\mathbf{2}} \boldsymbol{Capacitive\ load} \times \boldsymbol{Voltage}^{\mathbf{2}} \times \boldsymbol{frequency}^{\mathbf{1}}$$

- For a simple processor, if capacitive load is reduced by 15%, voltage is reduced by 15%, maintain the same frequency, how much power consumption can be reduced?
- A: 27.8%
- B: 38.6%
- C: 85.0%

- However, increasing $\boldsymbol{f}$ demands higher $\boldsymbol{V}$ → $\boldsymbol{P}$ **scales faster than** $\boldsymbol{f}$**.**
- **A new scaling strategy is essential.**

[1]Here we only consider dynamic power, but not static power.

Massive Multi-Core System

- Scale in the number of cores → Improves throughput, not latency.
- Very aggressive (and successful) strategy by AMD.

Year	Generation/Codename	Max Core Count	Notes
2003	Opteron (SledgeHammer)	2	AMD's first 64-bit server CPU
2005-2010	Opteron Dual/Quad Core	8	Gradual scaling with K8 architecture
2011	Opteron (Interlagos)	16	Bulldozer architecture debut
2012	Opteron (Abu Dhabi)	16	Piledriver refinement
2017	EPYC 7001 (Naples)	32	Zen architecture launch
2019	EPYC 7002 (Rome)	64	Zen 2
2021	EPYC 7003 (Milan)	64	Zen 3
2022	EPYC 7003X (Milan-X)	64	Added 3D V-Cache
2022	EPYC 9004 (Genoa)	96	Zen 4
2023	EPYC 9004X (Genoa-X)	96	3D V-Cache variant
2023	EPYC 8004 (Siena)	64	Optimized for edge and telco
2023	EPYC 9004 (Bergamo)	128	Cloud-native
2025*	EPYC 9005 (Turin Dense)	192	Expected Zen 5c architecture

An Exciting Era of Computer Systems

- There is too much going on:
  - **Massive Multi-Core Servers**: Powering global-scale internet services.
  - **Multi-GPU Clusters**: Fueling generative AI models like ChatGPT with millions of GPUs.
  - **Supercomputers**: Delivering real-time climate modeling.
  - **Gaming PCs and Laptops**: Pushing the limits of graphics and immersive experiences.
  - **Mobile SoCs**: Integrating CPU, GPU, NPU, and baseband in your pocket.
  - **Autonomous Vehicles**: Combining real-time sensing, decision-making, and edge.
  - **Embodied AI & Robotics**: Enabling intelligent agents to interact with physical world.
  - **Edge Devices & IoT**: Bringing computation closer to data sources for speed and privacy.
  - **Quantum Computing**: Exploring radically new paradigms beyond classical computing.
  - ...