Computer Architecture

Lec 1: Introduction

Slides developed by *Yusen Li* at Nankai University
with sources from many other universities

Course Logistics

Course Staff

- Instructor Prof. **Li Yusen**
  - @Nankai-Baidu Joint Lab
  - PhD, Nanyang Technological University, Singapore
  - Web: https://liyusen-nku.github.io/
  - Email: liyusen@nbjl.nankai.edu.cn
  - Interests : Parallel and Distributed Systems, Computer Architecture
- Lab Instructor
  - **Dr. Dong Qiankun**
  - **Dr. Yan Meng**
- TAs
  - Ling Feng (MS)
  - Cui Jiahe (BS)
  - Cui Lixiao (PhD)
  - Yang Fan (MS)

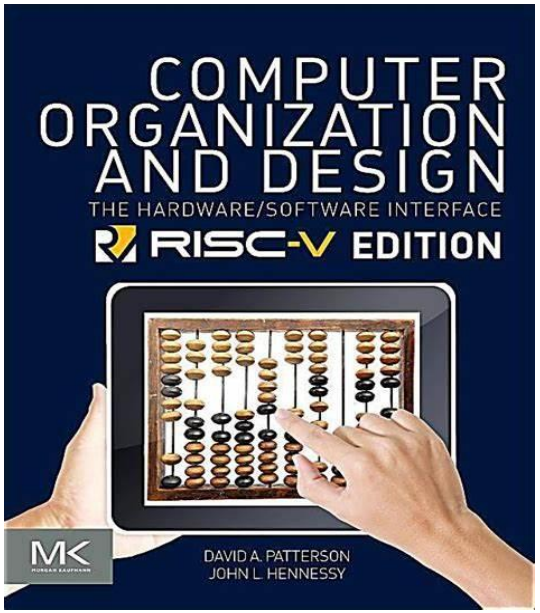
Course Resources

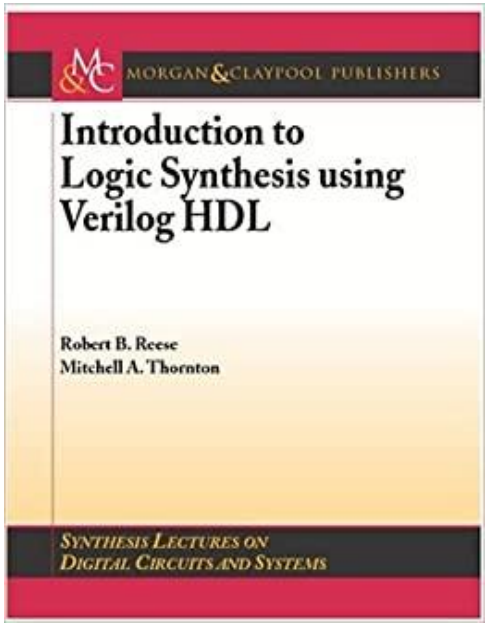
- Feishu Group
  - notice, slides, Q&A

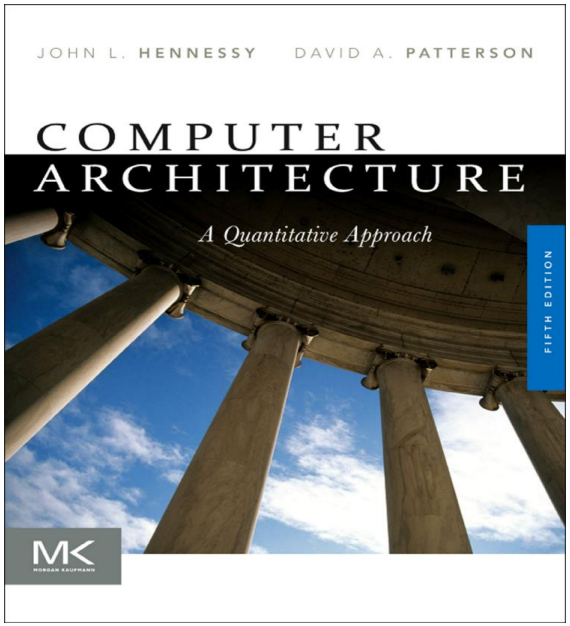
- 雨课堂
  - https://www.yuketang.cn/
  - notice, slides, quizzes

Course Resources

- Textbooks
  - Patterson & Hennessy, *Computer Organization and Design,* **4th or 5th edition**
  - Reese & Thornton, *Intro to Logic Synthesis using Verilog HDL*
  - *Digital Design: A Systems Approach* by Dally and Harting
  - Patterson & Hennessy, *Computer Architecture: A Quantitative Approach*, **5th or 6th edition**







Labs

- Lab descriptions
  - Lab 1: single-cycle processor (Review, Dong Qiankun)
  - Lab 2: multi-cycle processor (Dong Qiankun)
  - Lab 3: pipelined processor (Dong Qiankun)
  - Lab 4... to be determined (Li Yusen, Yan Meng)
- Labs 1-3 require **Verilog** implementation
- Lab 4... is based on **simulation**

Grading

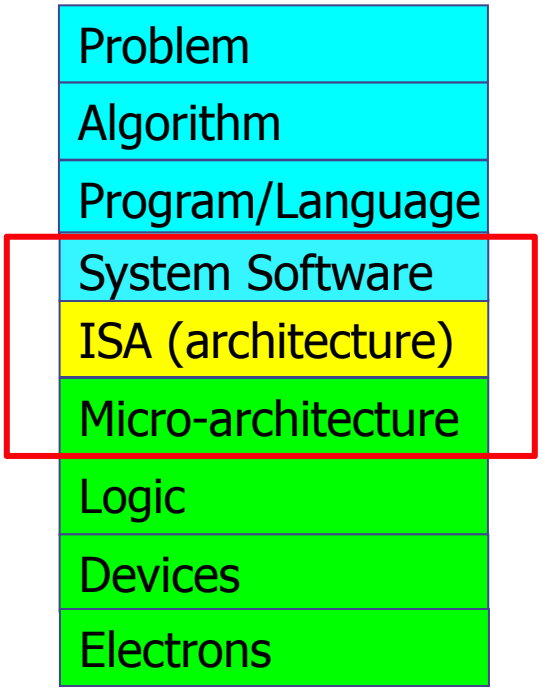
- Tentative grade contributions:
  - Quizzes: 10%
  - Assignments: 30%
  - Labs: 30%
  - Final Exam: 30%

Any assignment can be submitted **up to 24 hours** late, for **80% credit**

Cheating will **not** be tolerated

Why Study Computer Architecture?

Computer Architecture



- **Computer architecture**
  - Definition of ISA to facilitate implementation of software layers
  - The hardware/software interface
- **Computer micro-architecture**
  - Design processor, memory, I/O to implement ISA
  - Efficiently implementing the interface
- This course is mostly about processor **micro-architecture**

Computer Architecture

- In *Compuer Organization and Design* you learned how a processor worked, in this course we will tell you how to make it work **well**.

Computer Organization and Design

Computer Architecture

Computer Architecture

- Computer Organization and Design
  - Focus on one toy ISA
  - Focus on functionality: "just get something that works"
  - Instructive, learn to crawl before you can walk

- Computer Architecture
  - Less emphasis on any particular ISA during lectures
  - Focus on quantitative aspects: **performance**, cost, power, etc.
  - Representative of ~1980s hardware

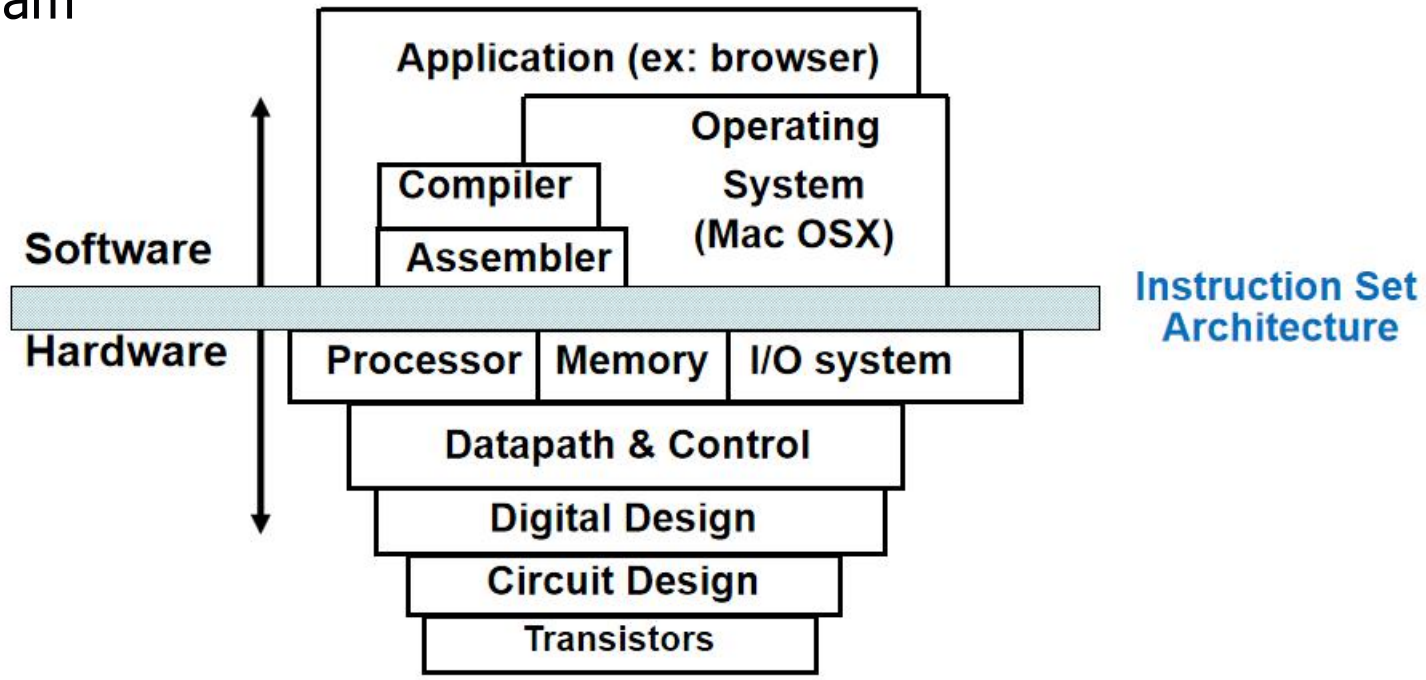
Why Study Computer Architecture

- **Understand where computers are going**
  - Future capabilities drive the (computing) world
  - Real world-impact: no computer architecture → no computers!
- **Understand high-level design concepts**
  - The best system designers understand all the levels
    - Hardware, compiler, operating system, applications
- **Understand computer performance**
  - Writing well-tuned (fast) software requires knowledge of hardware
- **Understand cutting-edge research in computer systems**
  - High quality research works can open your eyes

Computer architecture is perhaps the most fundamental subject in computer science

Relations with OS, Complier

- **Operating system** (OS) is system software that manages computer hardware, software resources, and provides common services for computer programs
- **Compiler** is a program that translates source code from a high-level programming language to a lower level language (e.g. assembly language, object code, or machine code) to create an executable program



Computer System in China

top-heavy, weak foundation

- In top 2000 listed company, 14 + 14 chip and software companies in US, but 0 in China
- Proportions of people in **AI systems**, **AI algorithms** and **AI applications**:
  - **China**: **3.3%**, 34.9%, 61.8%
  - **US**: **22.7%**, 37.4%, 39.4%

from Prof. Jin Hai at HUST

Computer System in China

dilemma : assessment criteria

- Points-based PhD thesis assessment
  - e.g., : TsingHua (6 points：CCF A 5p，B 3p, C 1.5p)
  - NIPS 23：**2000+** accepted
  - AAAI 21：**1692** accetepd
  - ASPLOS 22：**80** accepted
  - SOSP 19：**38** accepted

bad situation for computer system researchers

from Prof. Jin Hai at HUST

CSCC全国大学生系统能力大赛

CSCC 全国大学生计算机系统能力大赛

编译系统赛　操作系统赛　CPU设计赛

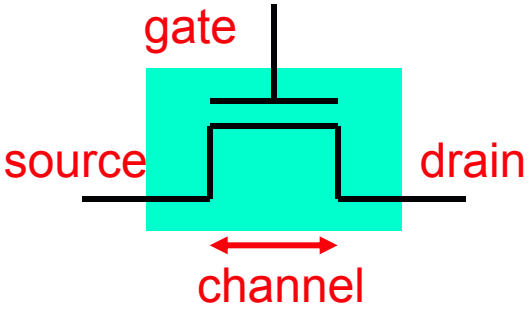
二等奖		
华南理工大学	TINBAC Is Not Building A Compiler	郭传鉉 刘泽森 王延葵 姚文浩
西北工业大学	胡编乱造不队	王炳杰 张少腾 毛心东 李国旭
北京航空航天大学	真实匿名队	陈思言 胡琎 乔盛业 田昶尧
北京大学	全场景分布式优化队	郭资政 顾宇晨 潘樾阳 肖元安
中国科学技术大学	Maho_shojo	陈金宝 黄庄湫 王原龙 吴毓辰
南开大学	天津泰达	杨科迪 费迪 孙一丁 沈哲

二等奖		
HexonZ	北京理工大学	朱桂潮 陈新 林恺
邯郸路企鹅编译器	复旦大学	谭一凡 陈立达 杜雨轩 聂绍珩
嘉然今天偷着乐	国防科技大学	黄子潇 熊思民 黎梓浩 贺子然
萝杨空队	哈尔滨工业大学（深圳）	梁韬 杨博康 苏亦凡
NKUER4	南开大学	时浩铭 严诗慧 林坤 梅骏逸
从容应队	西北工业大学	王翰墨 王玉佳 郑世杰 乔袁飞龙

Technology Trends

Technology that Drives

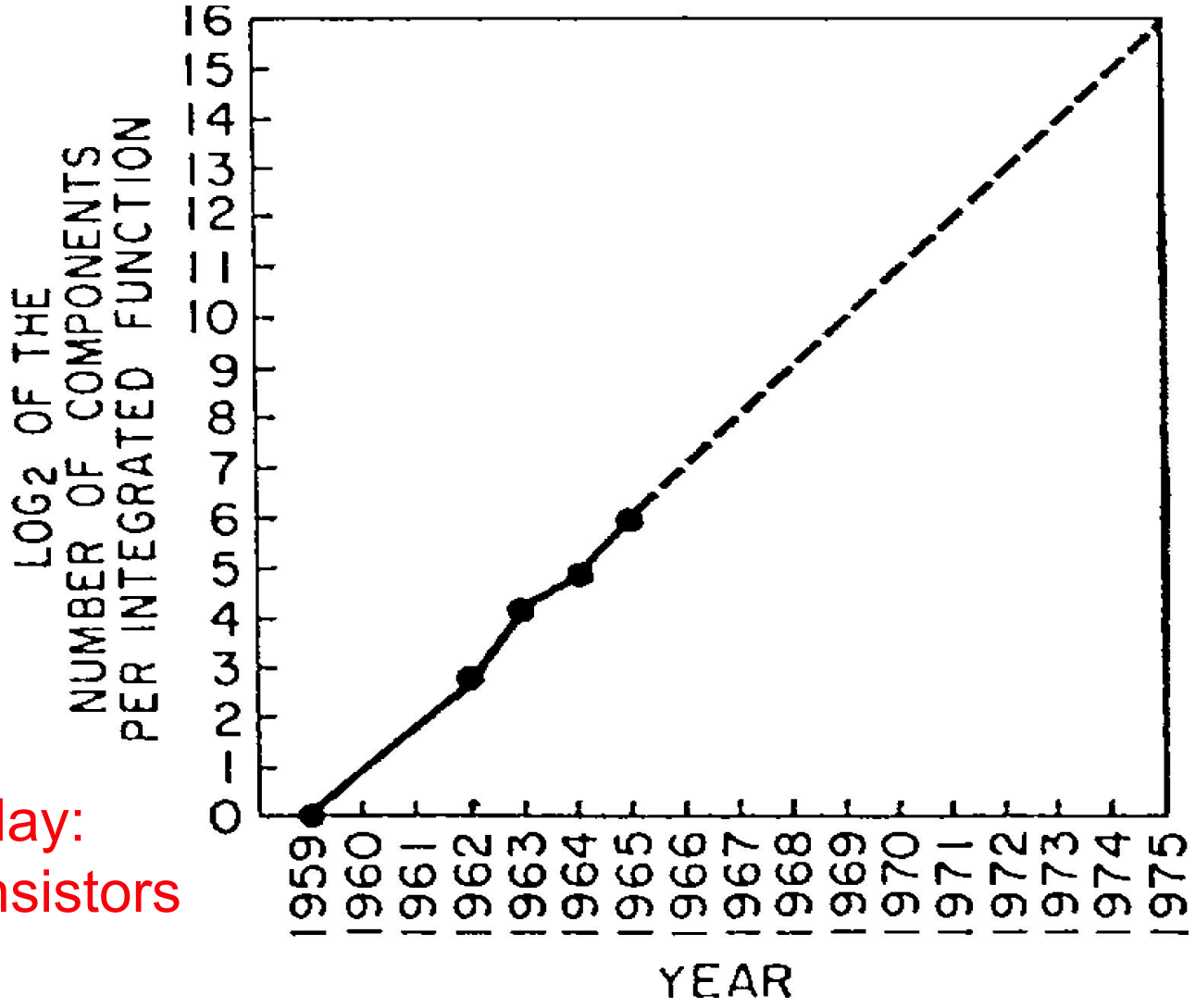
- Basic element
  - Solid-state **transistor** (i.e., electrical switch)
  - Building block of **integrated circuits (ICs)**



- What's so great about ICs? Everything
  + High performance, high reliability, low cost, low power
  + Lever of mass production

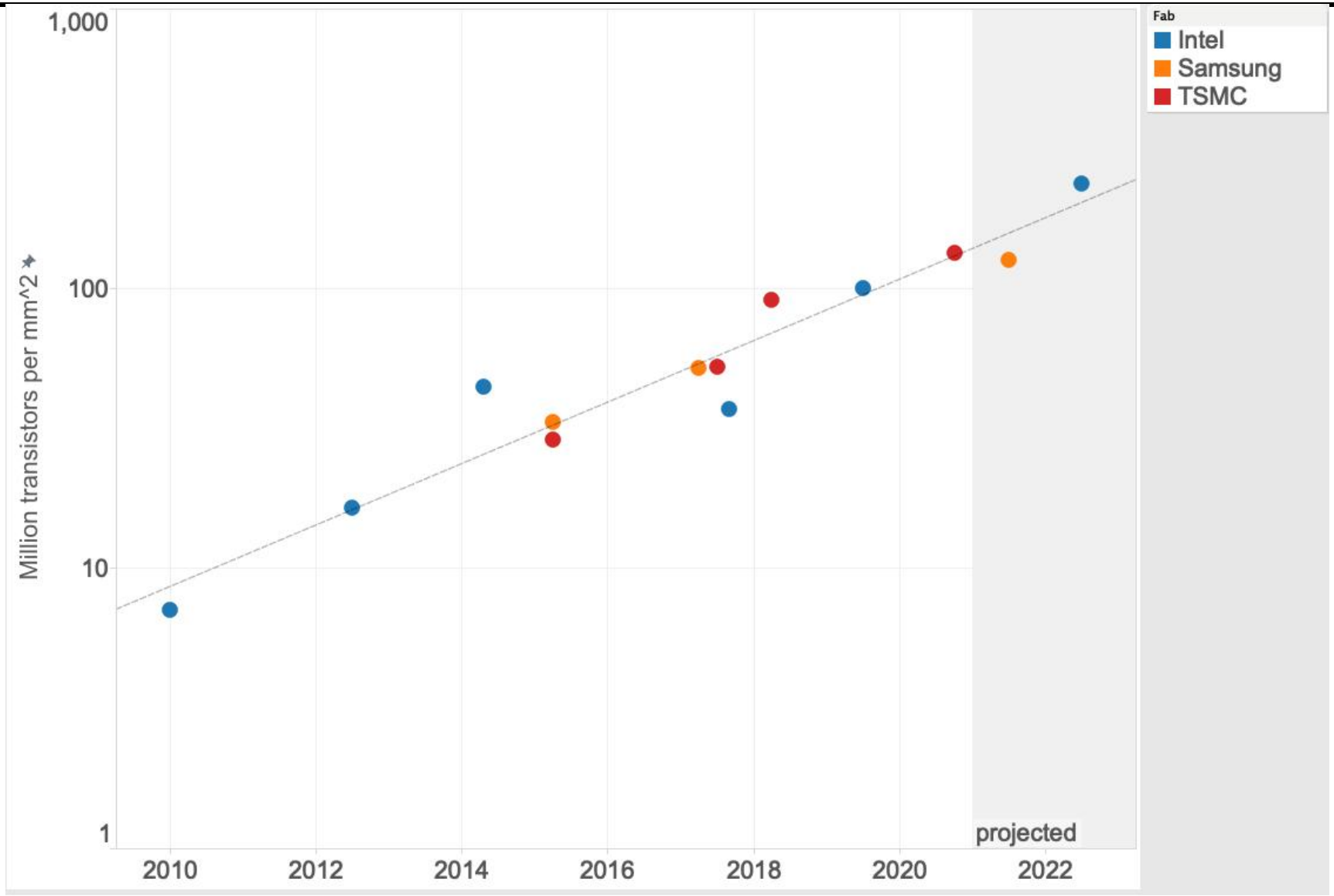
- Several kinds of integrated circuit families
  - **SRAM/logic**: optimized for speed (used for processors)
  - **DRAM**: optimized for density, cost, power (used for memory)
  - **Flash**: optimized for density, cost (used for storage)
  - Increasing opportunities for integrating multiple technologies

Moore's Law - 1965



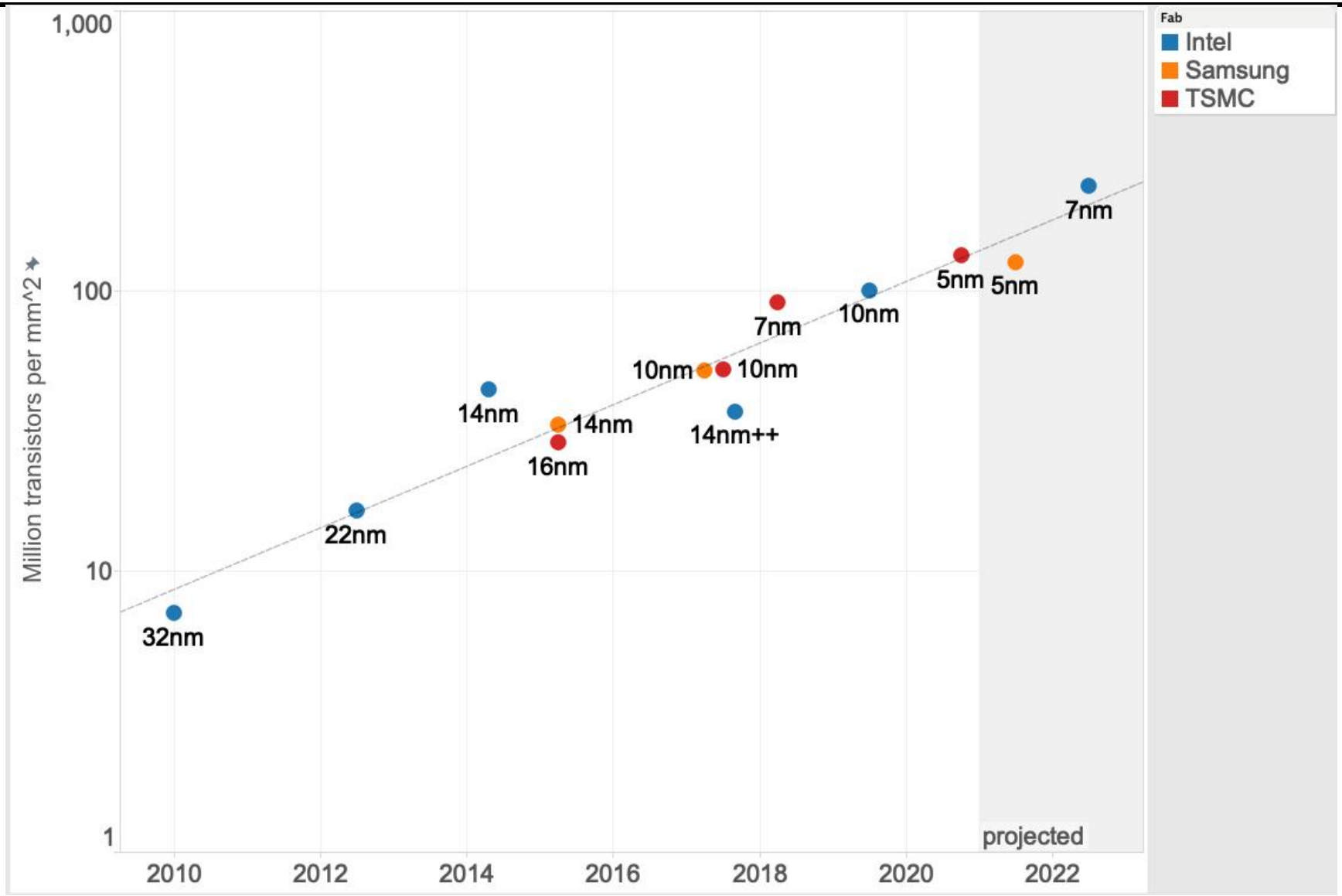
Today:
2^{35} transistors

Moore's Law today



data c/o WikiChip, Wikipedia

Moore's Law today



data c/o WikiChip, Wikipedia

Revolution I: The Microprocessor

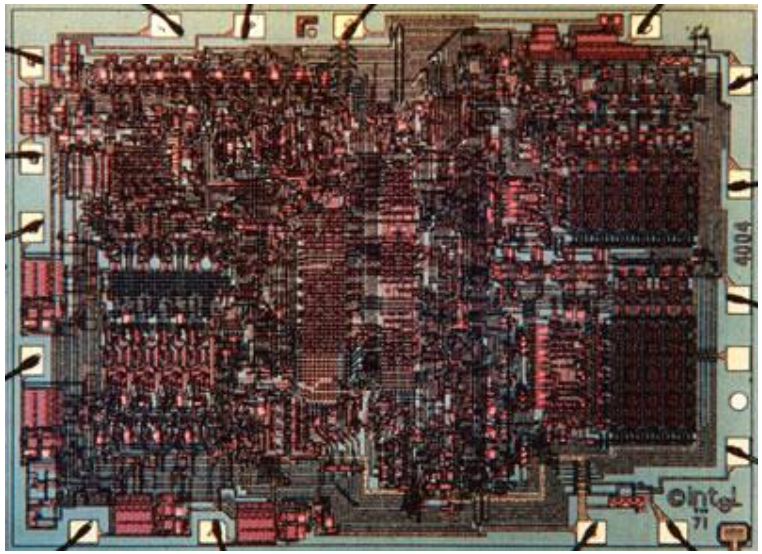
- **Microprocessor revolution**
  - One significant technology threshold was crossed in 1970s
  - Enough transistors (~25K) to put a 16-bit processor on one chip
  - Huge performance advantages: fewer slow chip-crossings
  - Even bigger cost advantages

- Microprocessors have allowed new market segments
  - Desktops, CD/DVD players, laptops, game consoles, set-top boxes, mobile phones, digital camera, mp3 players, GPS, automotive

- And replaced incumbents in existing segments
  - Microprocessor-based system replaced supercomputers

First Microprocessor

- Intel 4004 (1971)
  - Application: calculators
  - Technology: 10,000 nm
  - 2300 transistors
  - 13 mm^2
  - 108 KHz
  - 12 Volts
  - 4-bit data
  - Single-cycle datapath



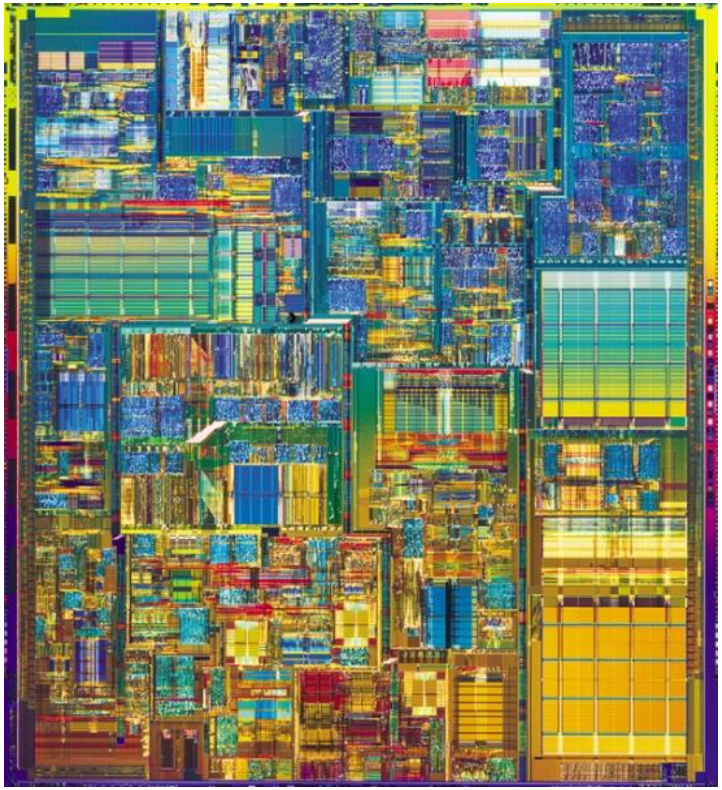
Revolution II: Implicit Parallelism

- **Implicit instruction-level parallelism**
  - Hardware provides parallel resources, figures out how to use them
  - Software is oblivious

- Initially using pipelining …
  - Which also enabled increased clock frequency
- … caches …
  - Which became necessary as processor clock frequency increased
- … and integrated floating-point
- Then deeper pipelines and branch speculation
- Then multiple instructions per cycle (superscalar)
- Then dynamic scheduling (out-of-order execution)

Pinnacle of Single-Core Microprocessors

- Intel Pentium4 (2003)
  - Application: desktop/server
  - Technology: 90nm



  - 55M transistors
  - 101 mm^2
  - 3.4 GHz
  - 1.2 Volts

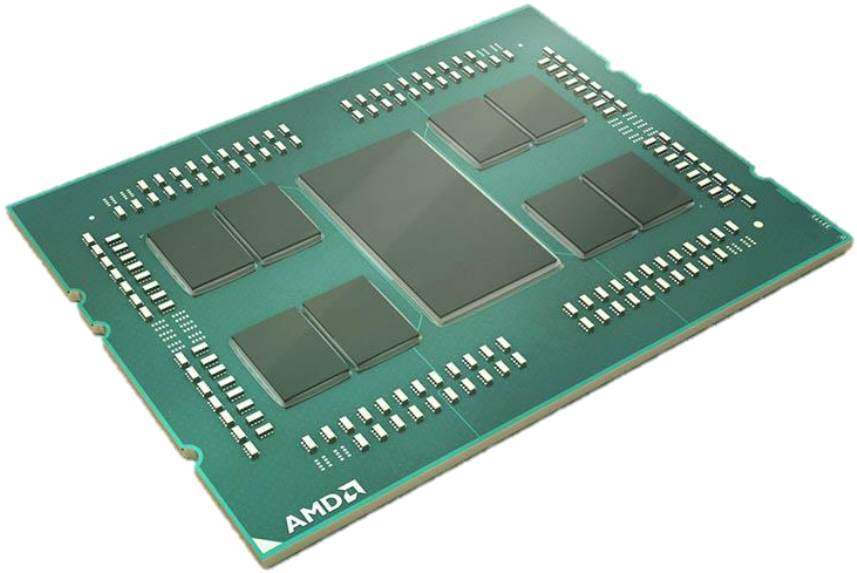
  - 32/64-bit data (16x)
  - 22-stage pipelined datapath
  - 3 instructions per cycle (superscalar)
  - Two levels of on-chip cache
  - data-parallel vector (SIMD) instructions, hyperthreading

Revolution III: Explicit Parallelism

- **Explicit data & thread level parallelism**
  - Hardware provides parallel resources, software specifies usage
  - Why? diminishing returns on instruction-level-parallelism
- First using (subword) vector instructions…, Intel's SSE
  - One instruction does four parallel multiplies
- … and general support for multi-threaded programs
  - Coherent caches, hardware synchronization primitives
- Then using support for multiple concurrent threads on chip
  - First with single-core multi-threading, now with multi-core
- Graphics processing units (GPUs) are highly parallel

Modern Multicore Processor

- AMD EPYC 7H12
  - Application: server
  - Technology: 7nm



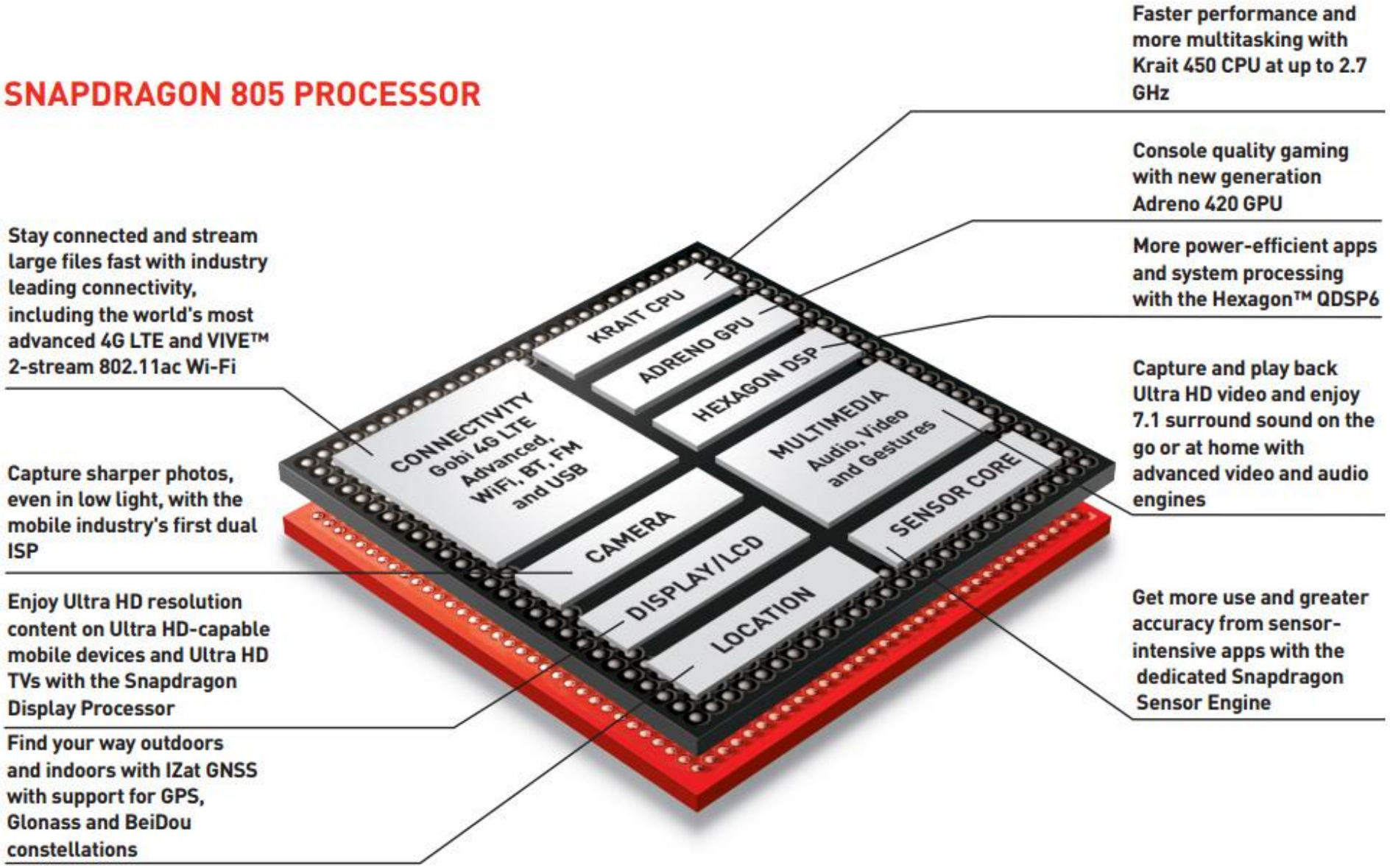
  - 39.5B transistors
  - 1008 mm^2
  - 2.6 to 3.3 Ghz
  - 256-bit data (2x)
  - 19-stage pipelined datapath
  - 4 instructions per cycle
  - 292MB of on-chip cache
  - data-parallel vector (SIMD) instructions, hyperthreading
  - **64-core multicore**

Historical Microprocessor Evolution

Feature	Intel 4004	Intel Pentium 4 Prescott	AMD EPYC Rome
release date	1971	2004	2019
transistor size	10,000 nm	90 nm	7 nm, 14 nm
transistor count	2,300	125M	39.5B
area	13 mm^2	112 mm^2	1008 mm^2
frequency	740 KHz	3.8 GHz	2.6-3.3 GHz
data width	4-bit	64-bit	256-bit
pipeline stages	n/a	31	19
pipeline width	n/a	3	4
core count	1	1	64
on-chip cache	n/a	1MB	292MB

Revolution IV: Accelerators

- Combining **multiple** kinds of compute engines in one die
  - not just homogenous collection of cores
  - System-on-Chip (SoC) is one common example in mobile space
- Lots of stuff on the chip beyond just CPUs
  - Graphics Processing Units (GPUs)
    - throughput-oriented specialized multicore processors
    - good for gaming, machine learning, computer vision, …
  - Special-purpose logic
    - media codecs, radios, encryption, compression, machine learning
- Excellent energy efficiency and performance
  - extremely complicated to program!


SNAPDRAGON 805 PROCESSOR
Faster performance and more multitasking with Krait 450 CPU at up to 2.7 GHz
Console quality gaming with new generation Adreno 420 GPU
More power-efficient apps and system processing with the Hexagon™ QDSP6
Capture and play back Ultra HD video and enjoy 7.1 surround sound on the go or at home with advanced video and audio engines
Get more use and greater accuracy from sensor-intensive apps with the dedicated Snapdragon Sensor Engine
Stay connected and stream large files fast with industry leading connectivity, including the world's most advanced 4G LTE and VIVE™ 2-stream 802.11ac Wi-Fi
Capture sharper photos, even in low light, with the mobile industry's first dual ISP
Enjoy Ultra HD resolution content on Ultra HD-capable mobile devices and Ultra HD TVs with the Snapdragon Display Processor
Find your way outdoors and indoors with IZat GNSS with support for GPS, Glonass and BeiDou constellations
KRAIT CPU
ADRENO GPU
HEXAGON DSP
CONNECTIVITY Gobi 4G LTE Advanced, WiFi, BT, FM and USB
MULTIMEDIA Audio, Video and Gestures
SENSOR CORE
CAMERA
DISPLAY/LCD
LOCATION

c/o Qualcomm

Technology Disruptions

- Classic examples:
  - transistor
  - microprocessor
- More recent examples:
  - flash-based solid-state storage
  - shift to accelerators (e.g., AI chip)
- Nascent disruptive technologies:
  - non-volatile memory ("disks" as fast as DRAM)
  - Chip stacking (also called "3D die stacking")
  - Processing near/in memory

"Golden Age of Computer Architecture"

- Hennessy & Patterson, 2018 Turing Laureates
- the end of **Dennard Scaling & Moore's Law** means no more free performance
  - "The next decade will see a Cambrian explosion of novel computer architectures"

Themes

- Parallelism
  - enhance system performance by doing multiple things at once
  - instruction-level parallelism, multicore, GPUs, accelerators
- Caching
  - exploiting locality of reference: storage hierarchies
  - try to provide the illusion of a single large, fast memory