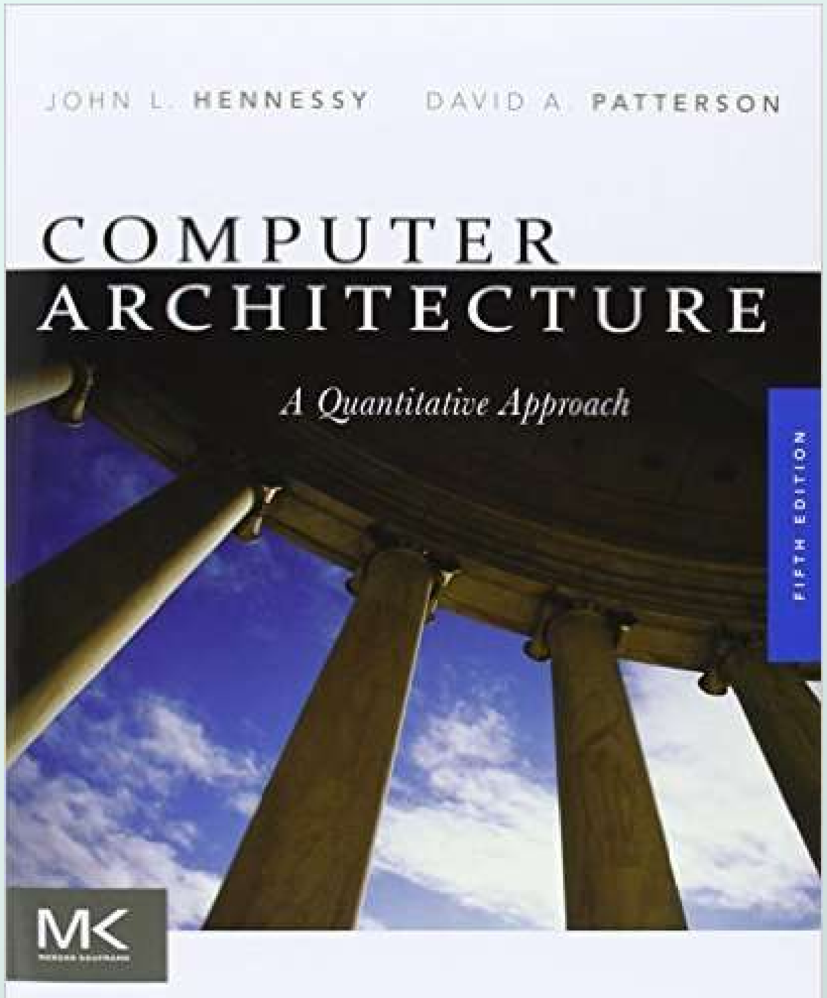


Technology Trends

Original slides from:

Computer Architecture

A Quantitative Approach

Hennessy, Patterson

Modified slides by
Yashwant Malaiya
Phil Sharp
Colorado State University

Exponential Growth

- Grows by a factor of (1+x) per year.
- By a factor of $(1+x)^n$ for n years.
- Example: An investment of $1000
  - 100% return in one year (i.e. doubles)
  - When will it become a million dollars?
  - Answer: 2^y=1000, y = ?

The computer industry has experienced exponential growth for decades: memory density, processor performance, circuit density, communications bandwidth, …

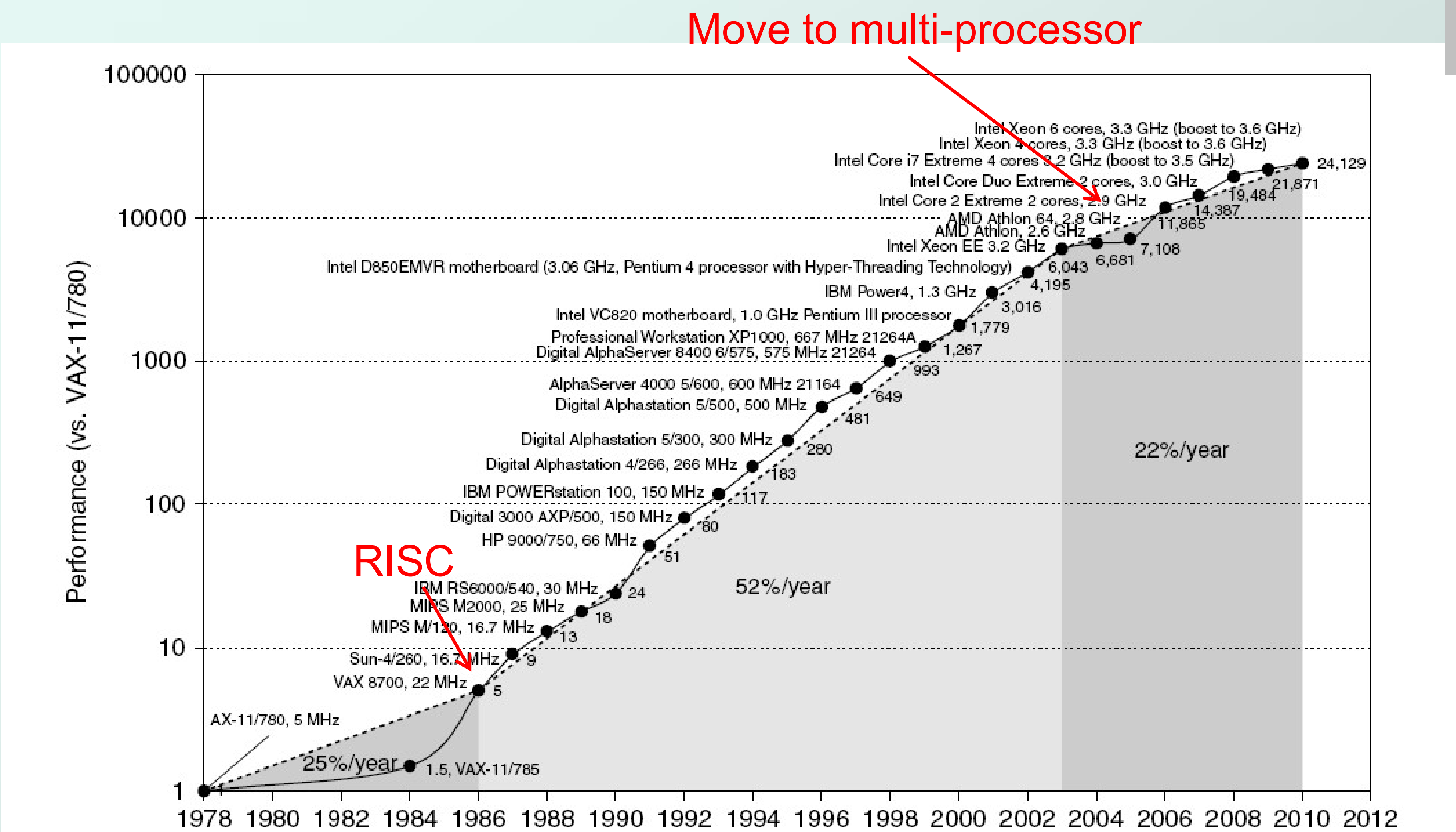
Order of Magnitude

- An **order of magnitude** is an approximate measure of the number of digits that a number has in the commonly-used base-ten number system. – Wikipedia
  - Similar to scientific notation
- Used for approximations when dealing with data that contains large differences in value
- X is one order of magnitude larger than Y means X is ~10 times larger than Y
  - X is three orders of magnitude larger that Y: ~1000 times
- Useful when dealing with exponential growth

Computer Technology

- Performance improvements:
  - Improvements in semiconductor technology
    - Reduced feature (circuit) size
    - Higher clock speeds
  - Improvements in computer architectures
    - Enabled by HLL compilers, UNIX
    - RISC architectures
  - **Together have enabled:**
    - Lightweight, portable, cheap, fast computers
    - Productivity-based programming languages
    - Advanced development environments and tools

Single Processor Performance



Defining Computer Architecture

- "Classical" computer architecture:
  - Instruction Set Architecture (ISA) design
  - i.e. decisions regarding:
    - registers, memory addressing, addressing modes, instruction operands, available operations, control flow instructions, instruction encoding

- "New" computer architecture:
  - Specific requirements of the target machine
  - Design to maximize performance within constraints:
    - **cost, power, and availability**
  - Includes ISA, microarchitecture, hardware

Trends in Technology

- Integrated circuit technology
  - Transistor density: 35%/year
  - Die size: 10-20%/year
  - Integration overall: 40-55%/year
- DRAM capacity: 25-40%/year (slowing)
- Flash capacity: 50-60%/year
  - 15-20X cheaper/bit than DRAM
- Magnetic disk technology: 40%/year
  - 10-20X cheaper/bit then Flash
  - 200-400X cheaper/bit than DRAM

Corollary of exponential growth

- When two quantities grow exponentially, but at different rates, their ratio also grows exponentially.
- 1.1n ≠ O(2n)
  - or 2n grows a lot faster than (1.1)
- Consequence for computer architecture: growth rate for e.g. memory is not as high as for processors, therefore, memory gets slower and slower (in terms of clock cycles) as compared to processors.
- This gives rise to so called gaps or walls

Bandwidth and Latency

- Bandwidth or throughput
  - Total work done in a given time
  - 10,000-25,000X improvement for processors
  - 300-1200X improvement for memory and disks
- Latency or response time
  - Time between start and completion of an event
  - 30-80X improvement for processors
  - 6-8X improvement for memory and disks

Memory Gap

- Memory bandwidth and latency improve much slower than processor speeds
  - Especially latency
- Reading from memory takes roughly the same number of clock cycles today as reading from disc in the 70s and 80s
- Addressed by something known as a cache
- Caches discussed in a future lecture

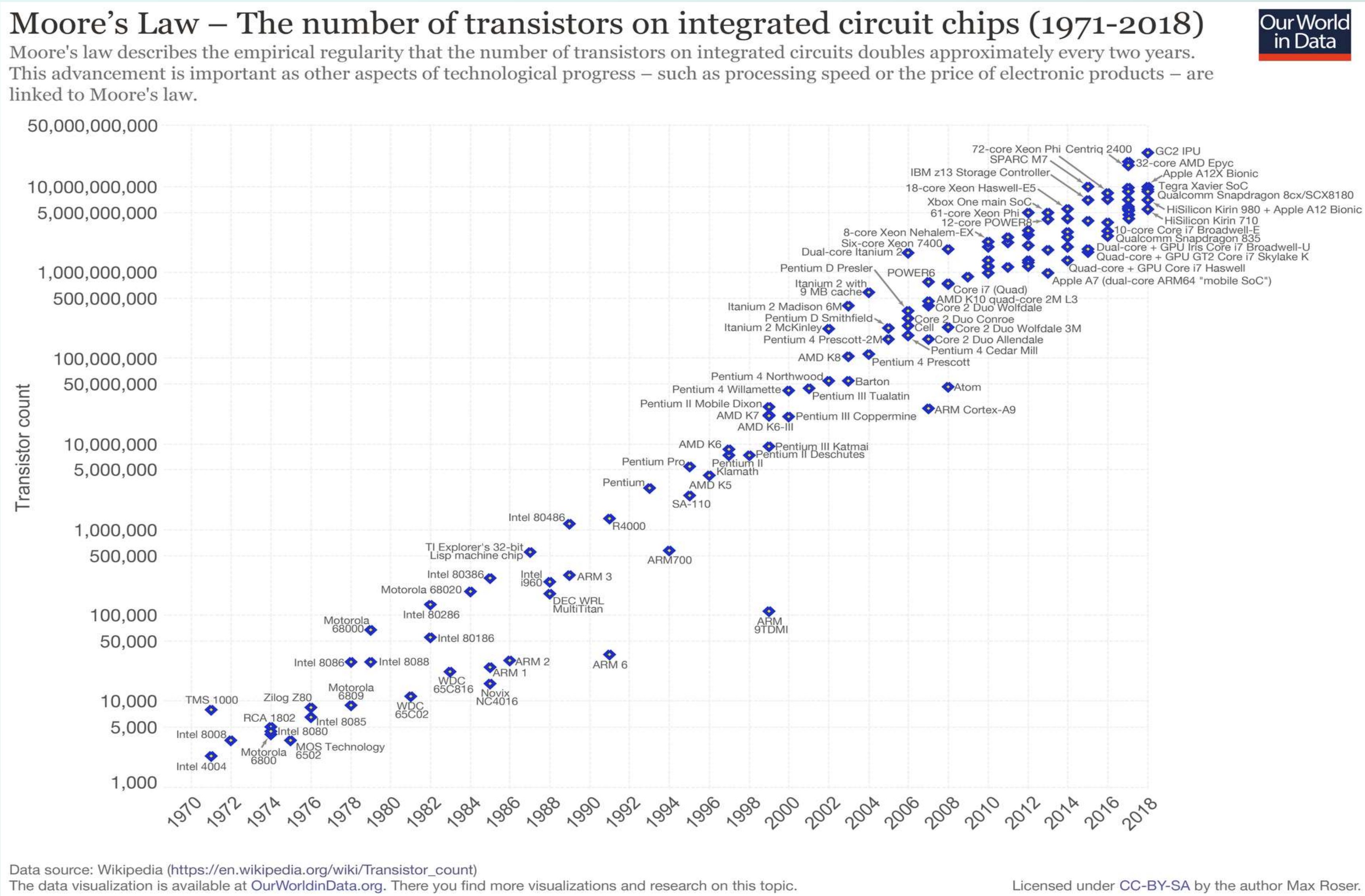
CS475 intro lecture slide

Technology Laws

- **Moore's Law:** formulated by Gordon Moore of Intel in the early 70's - **the number of transistors on a chip doubles every 18 months**; corollary, computers become faster and the price of a given level of computing power halves every 18 months.
  - Slowed around 2012
- **Dennard Scaling:** as transistors get smaller their power density stays constant, this means **power use is proportional with chip area not number of transistors.**
  - Ended around 2005
  - Smaller transistors have higher leakage power
  - Limits to lowering voltage

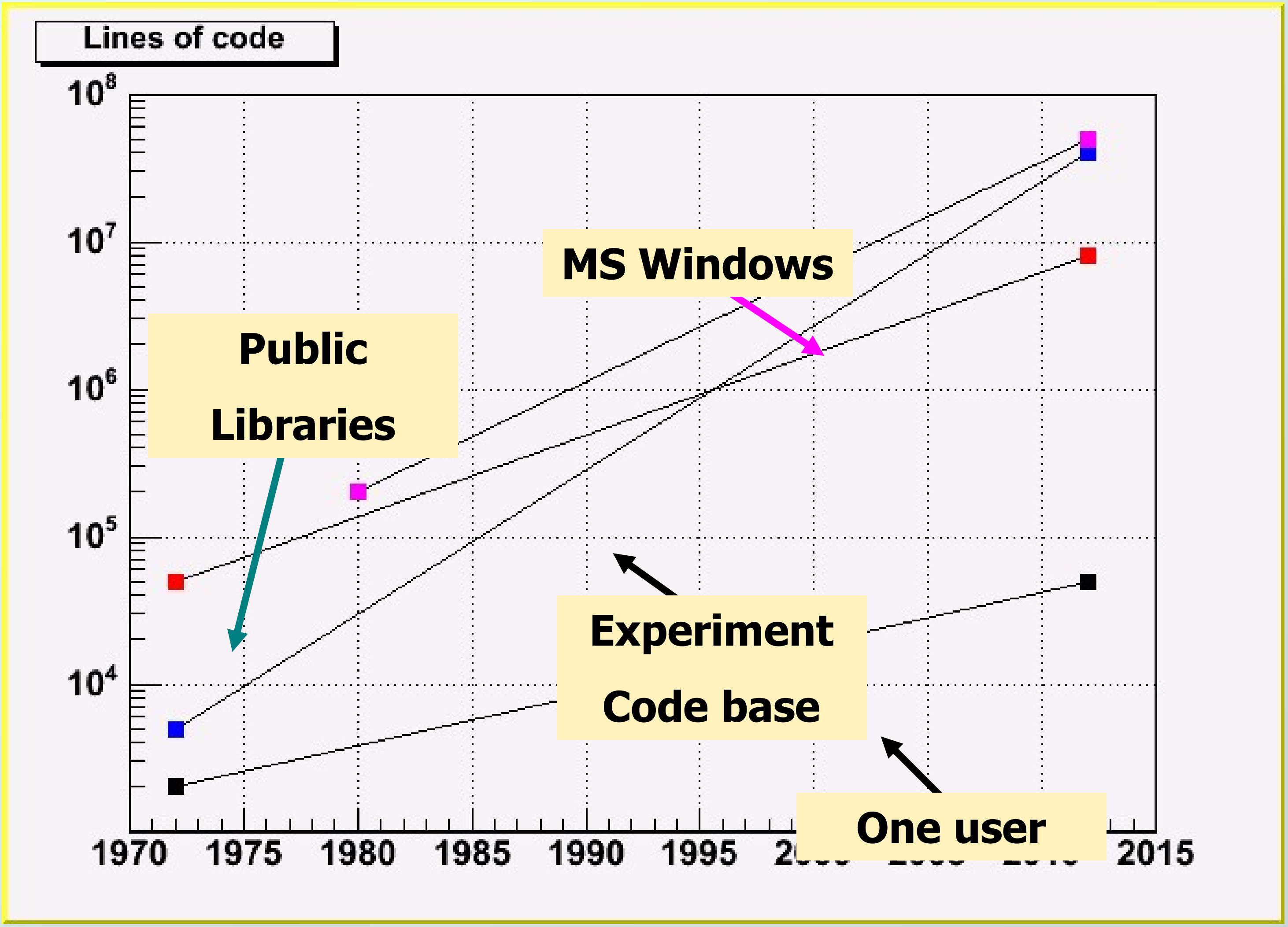
https://en.wikipedia.org/wiki/Moore%27s_law
https://www.micron.com/about/blog/2018/october/metamorphosis-of-an-industry-part-two-moores-law

Moore’s law

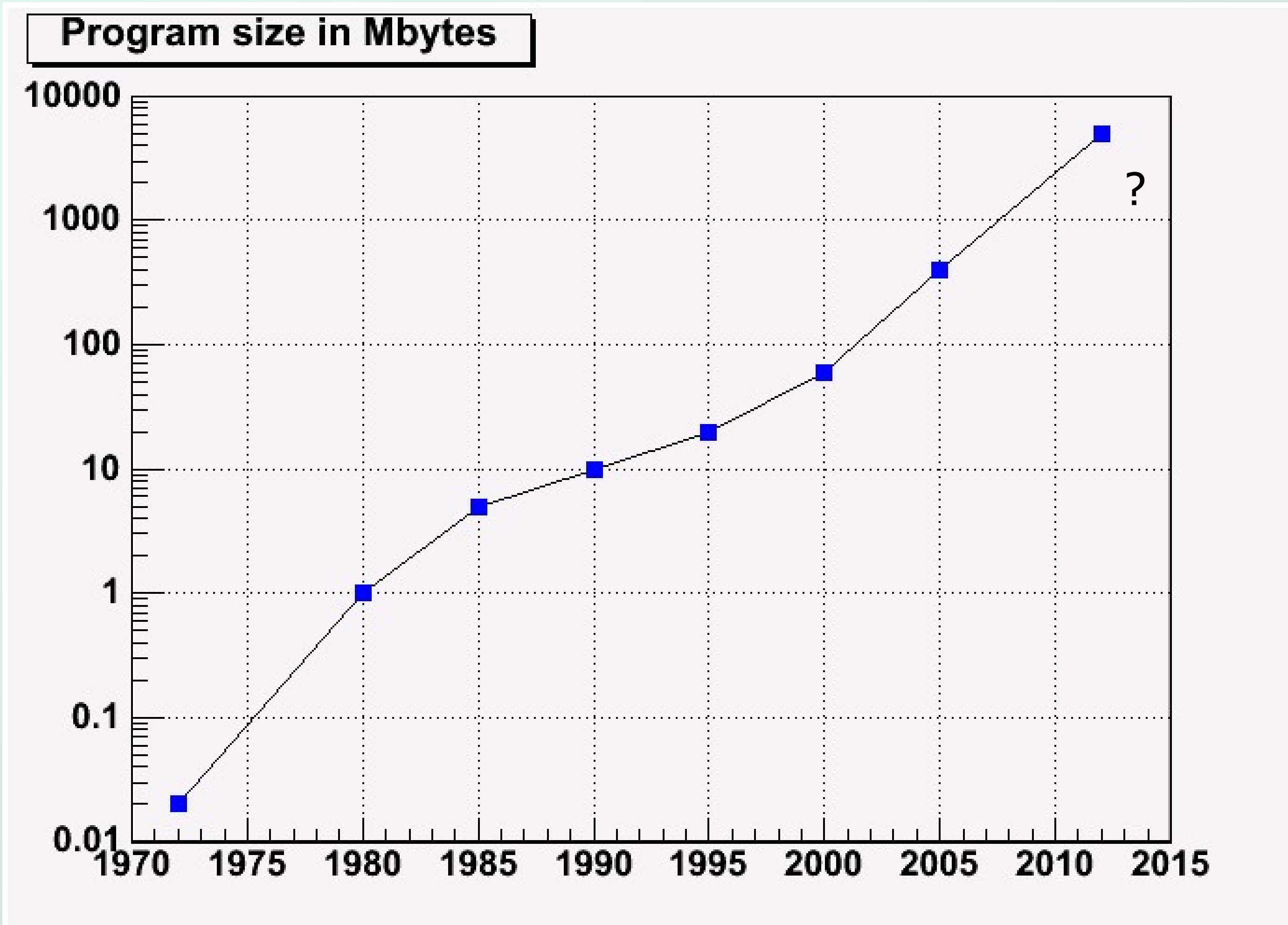


By Max Roser - https://ourworldindata.org/uploads/2019/05/Transistor-Count-over-time-to-2018.png, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=79751151

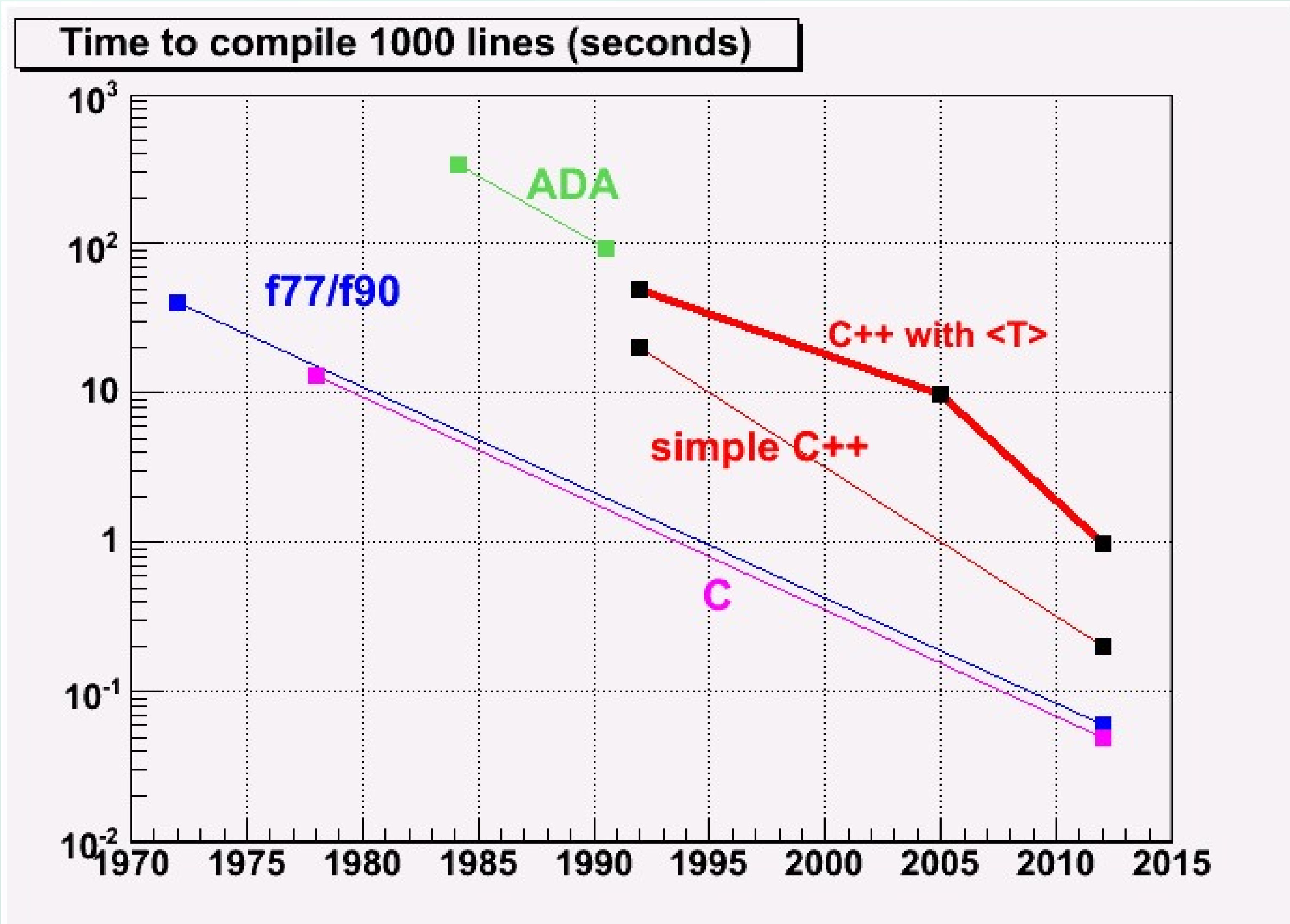
Program Size (lines of code)



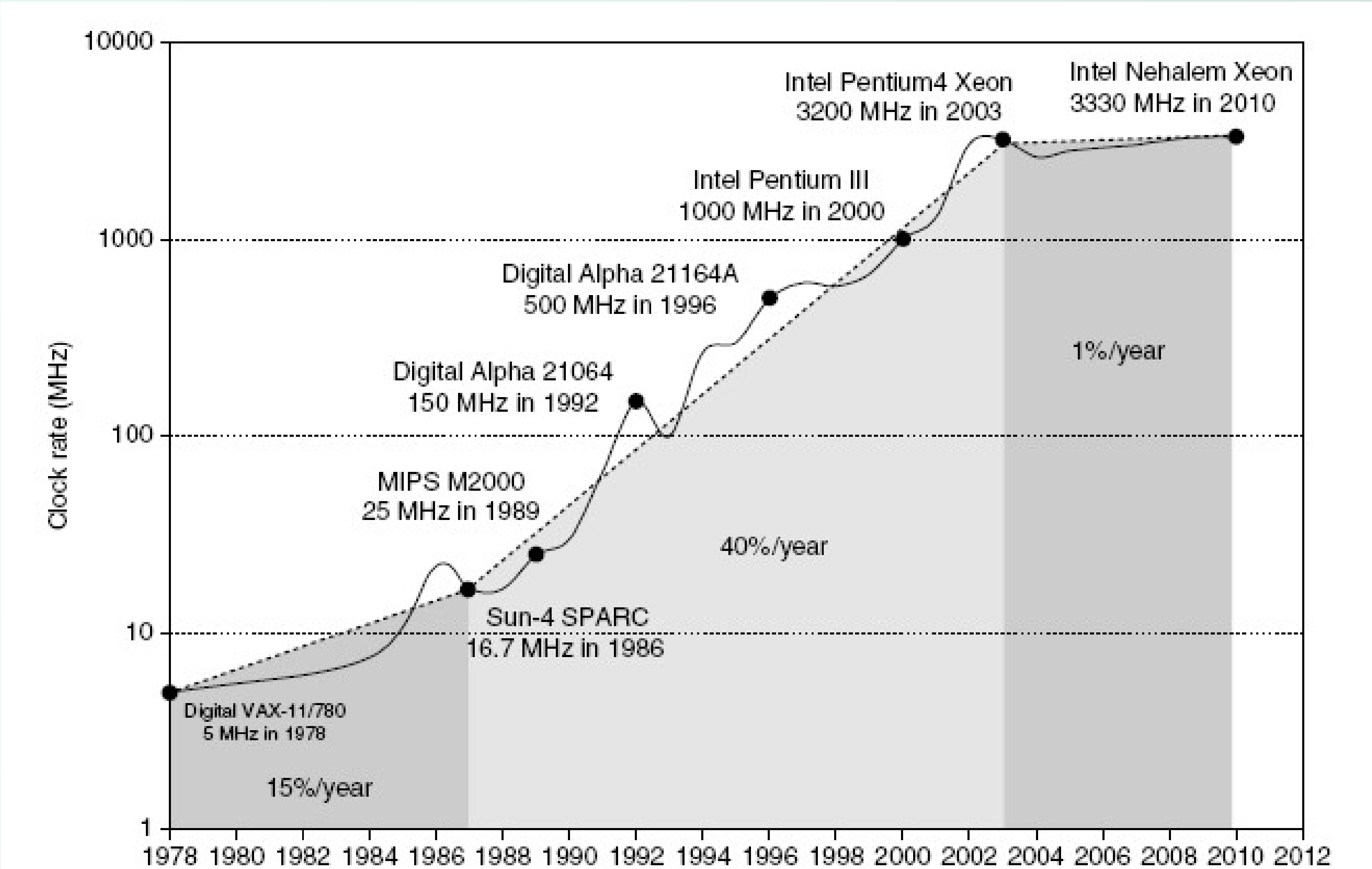
Program Size (RAM)



Compile Time (seconds)



Processor Speed



Power Scaling

- Intel 80386 running at 16 Mhz consumed around 2 Watts, less than a LED light bulb.
  - Qualcomm 855 SOC in cell phone ~5 Watts
- Intel Core i7 running at 3.3 GHz consumes 130 Watts, still less than a television.
  - Xeon 9200 400Watts
- However, heat must be dissipated from 1.5 x 1.5 cm chip in a closed case.
- Even with aluminum cooling fins and a fan, this is close the limit of what can be cooled.
  - Water cooling useful but still only ~2x cooling vs air
- Furthermore, the power consumption (based on CMOS technology) scales faster than clock speed.

Power

- Increasingly important to design chips with power consumption in mind
  - Mobile
  - Warehouse scale – cloud
- Dynamic Power α .5 x C_L x V^2 x F
- Static Power α C_S x V
  - C_L - Capacitive Load
  - C_S - Static Current

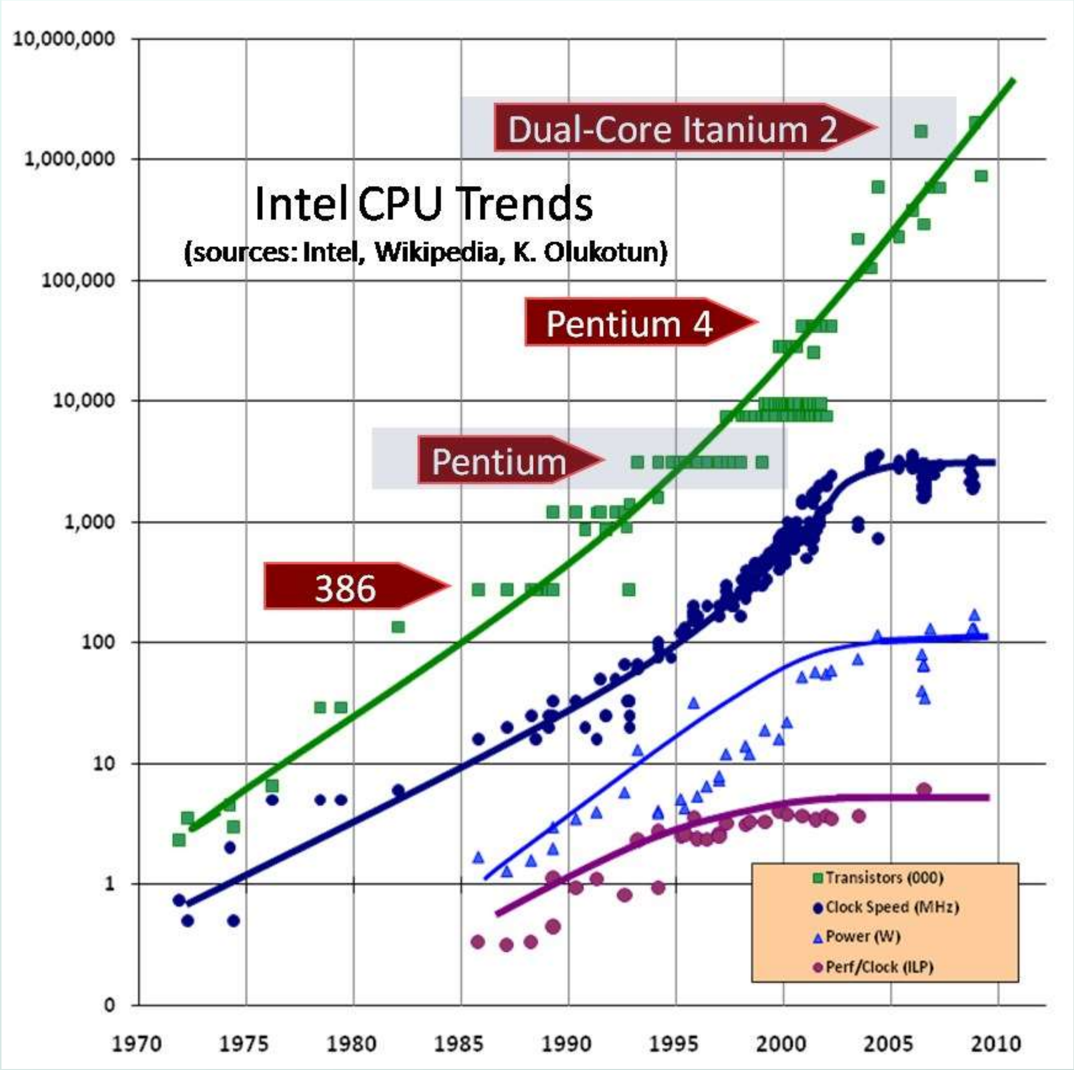
    C_L and C_S related to specific process technology
  - V - Voltage
  - F - Frequency

Power

- Dynamic power dominated by voltage
  - Double the voltage, quadruple the power
  - Higher frequency requires higher voltage for a particular process
    - DVFS
  - Power consumed only when transistors switch states

- Static power
  - Larger proportion of total power in smaller process nodes, i.e., sub 90 nm
    - Leakage
    - 25% to 50% of total power in modern chips
    - Power consumed even when clock is off / no switching
    - Dark Silicone

Power Wall



What would happen if clock speed and power consumption scaled as it did from the 1980's to 2000's?

Dr. Dobbs Journal, March 2003

Ways forward

How do engineers continue to increase performance in the face of current challenges

- Multicore
  - More operations at same frequency
  - Management and programming issues
  - Dark Silicone?
- Special IP
  - IP: block of logic on a chip
  - Do one job very fast
  - Extra room on chip from process shrink
  - Must be able to use functionality
- Accelerators
  - GPU / TPU
  - Many small simple cores
  - More work for programmers