Advanced Computer Architecture

Fall 2020

Hamed Farbeh

farbeh@aut.ac.ir

Department of Computer Engineering

Amirkabir University of Technology

Grading

- Final Exam: 50%
- Quiz: 30%
- Assignments: 20%
- Projects: 10%

Course Outline

- Introduction
- Memory Hierarchy
- Prefetching
- Instruction-level parallelism
  - Pipeline, branch prediction, out-of-order execution, …
- Data-level parallelism
- Vector processors and SIMD
- Thread-level parallelism
  - Multicores and coherency protocols

Textbook

- *Computer Architecture: A Quantitative Approach*, 6th edition, John L. Hennessy, David A. Patterson, MK pub., 2019

- *Processor Architecture From Dataflow to Superscalar and Beyond*, Jurij Silc, Borut Robic, Theo Ungerer, Springer, 1999.

- *High Performance Computer Architecture*, 3rd Edition Harold S. Stone, 1987

Copyright Notice

Lectures adopted from

- Computer Architecture: A Quantitative Approach, 6th edition, John L. Hennessy, David A. Patterson, MK pub., 2019

- Graduate Computer Architecture, handouts, by Prof. Asanovic, University of California at Berkeley, Spring 2018.

What is Computer Architecture?

Application

Gap too large to bridge in one step

(but there are exceptions, e.g. magnetic compass)

Physics

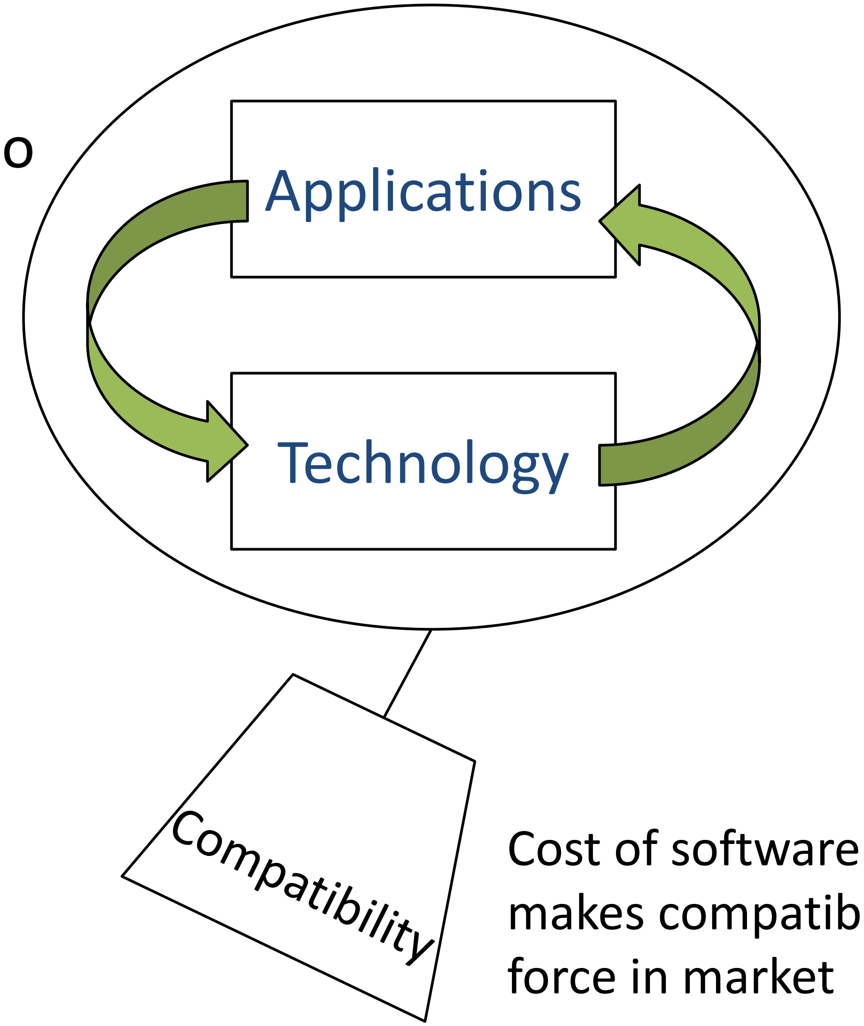
In its broadest definition, computer architecture is the *design of the abstraction layers* that allow us to implement information processing applications efficiently using available manufacturing technologies.

Abstraction Layers in Modern Systems

- Application
- Algorithm
- Programming Language
- Operating System/Virtual Machines
- Instruction Set Architecture (ISA)
- Microarchitecture
- Gates/Register-Transfer Level (RTL)
- Circuits
- Devices
- Physics

Architecture continually changing

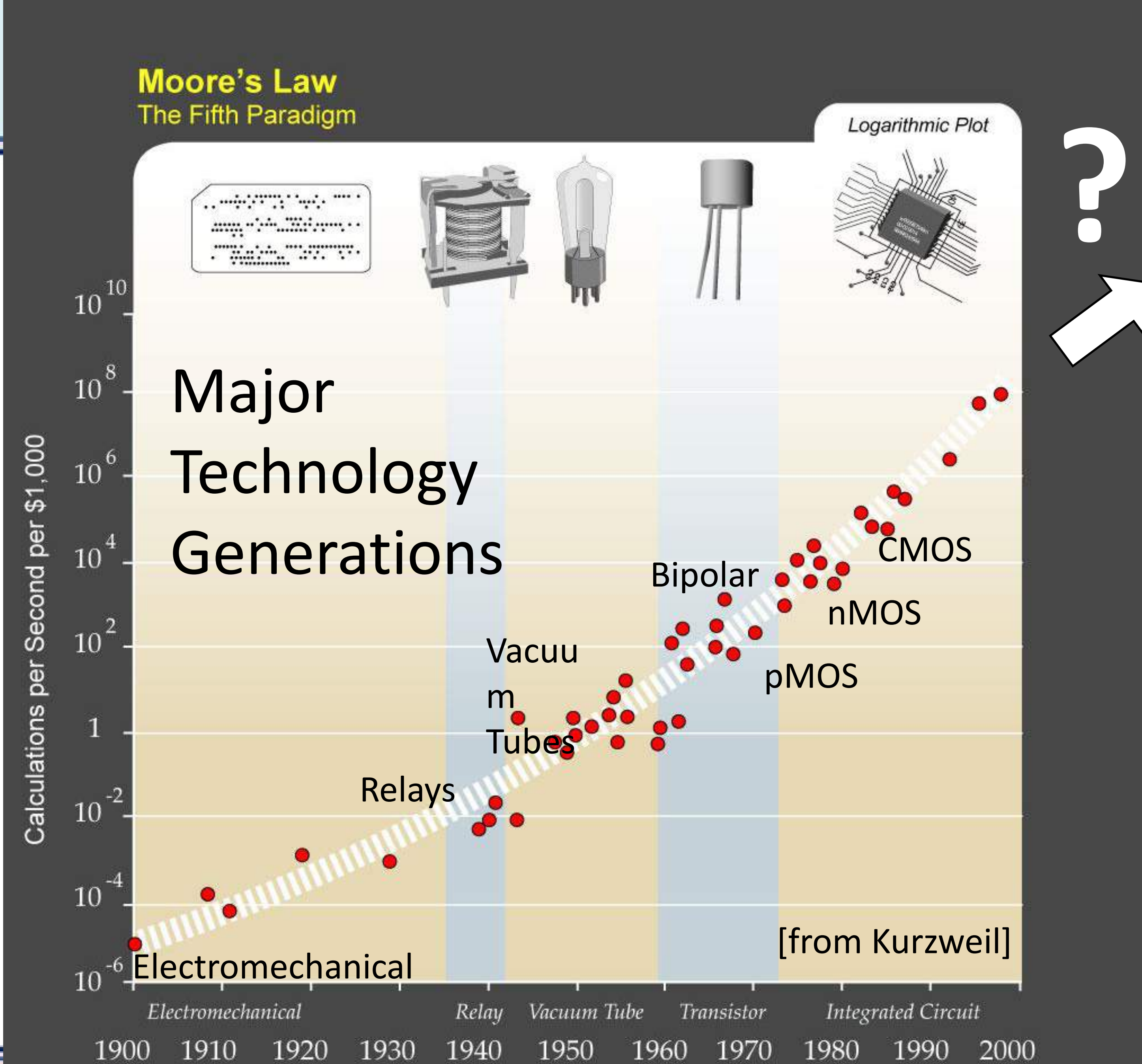
Applications suggest how to improve technology, provide revenue to fund development



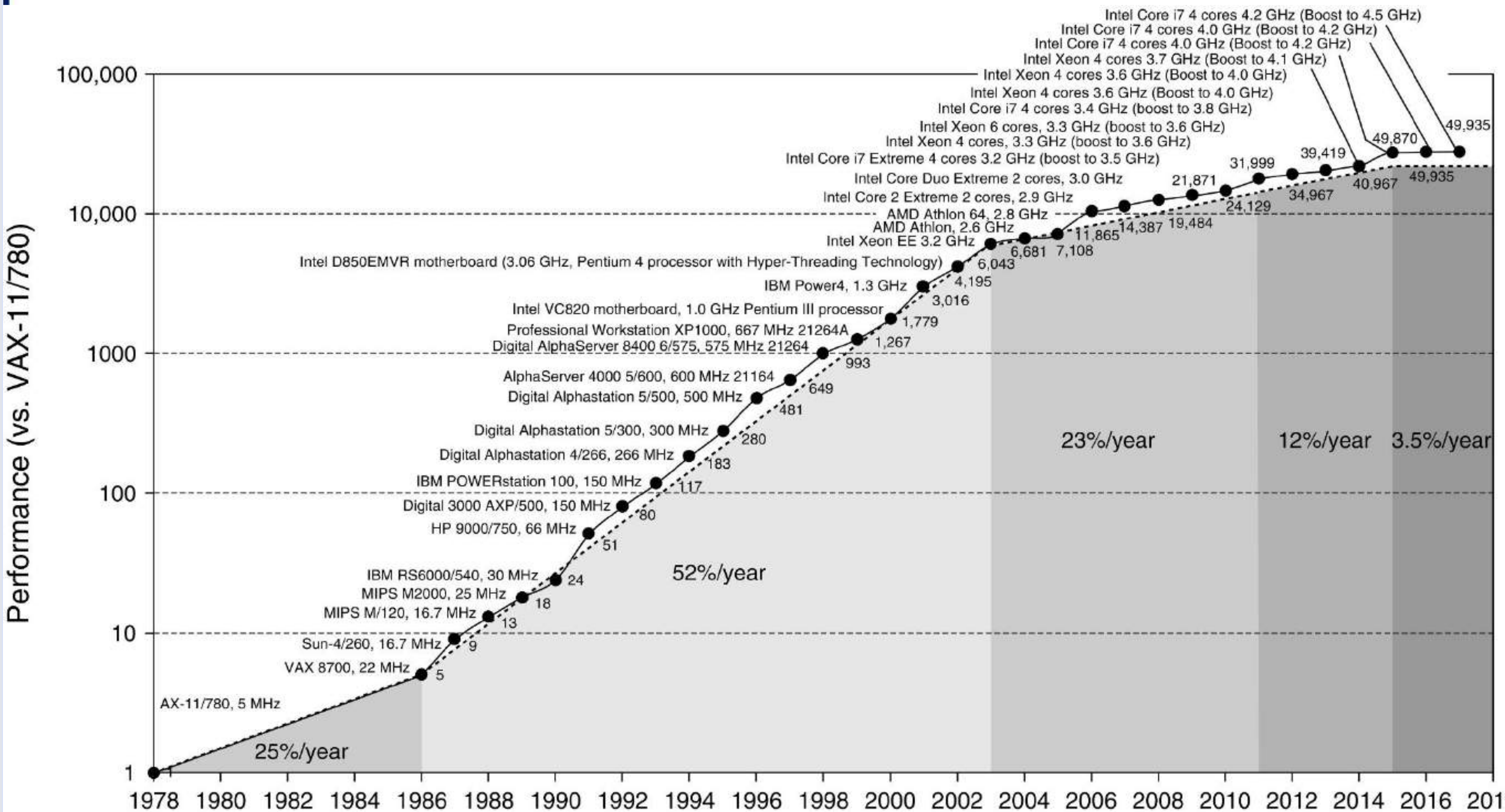
Improved technologies make new applications possible

Cost of software development makes compatibility a major force in market

Major Technology Generations



Single-Thread Processor Performance



Upheaval in Computer Design

- Most of last 50 years, Moore's Law ruled
  - Technology scaling allowed continual performance/energy improvements without changing software model
- Last decade, technology scaling slowed/stopped
  - Dennard (voltage) scaling over (supply voltage ~fixed)
  - Moore's Law (cost/transistor) over?
  - No competitive replacement for CMOS anytime soon
  - Energy efficiency constrains everything
- No "free lunch" for software developers, must consider:
  - Parallel systems
  - Heterogeneous systems

Today’s Dominant Target Systems

- Mobile (smartphone/tablet)
  - >1 billion sold/year
  - Market dominated by ARM-ISA-compatible general-purpose processor in system-on-a-chip (SoC)
  - Plus sea of custom accelerators (radio, image, video, graphics, audio, motion, location, security, etc.)
- Warehouse-Scale Computers (WSCs)
  - 100,000’s cores per warehouse
  - Market dominated by x86-compatible server chips
  - Dedicated apps, plus cloud hosting of virtual machines
  - Now seeing increasing use of GPUs, FPGAs, custom hardware to accelerate workloads
- Embedded computing
  - Wired/wireless network infrastructure, printers
  - Consumer TV/Music/Games/Automotive/Camera/MP3
  - Internet of Things!

Computer Architecture: A Little History

Why worry about old ideas?

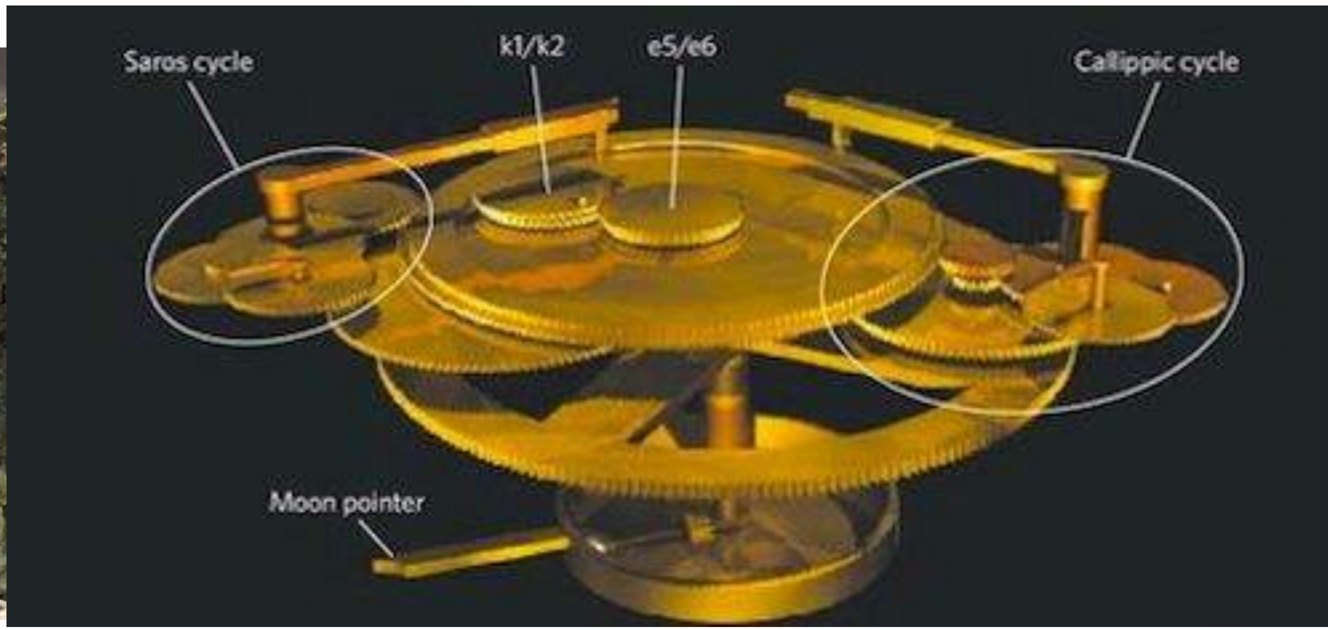
- Those who ignore history are doomed to repeat it
- Helps to illustrate the design process, and explains why certain decisions were taken
- Because future technologies might be as constrained as older ones

Changing Face of Computing

- 1960s: mainframes
- 1970s: mini-computers
- 1980s: Desktop computers
  - Desktop
  - Workstation
  - → Servers
    - Reliable
    - Long term file storage and access
    - Larger memory
    - More computing power
- 1990s: Embedded, Internet
  - PDA
  - Set top boxes, Game consoles
- 2000s:
  - ? (multi-core, networked, ..)
  - Personal Mobile Devices
  - Clusters / Cloud

Analog Computers

- Analog computer represents problem variables as some physical quantity (e.g., mechanical displacement, voltage on a capacitor) and uses scaled physical behavior to calculate results



tive Commons BY-SA 3.0]

Antikythera mechanism c.100BC

Wingtip vortices of Cessna tail in wind tunnel

Digital Computers

- Represent problem variables as numbers encoded using discrete steps
  - Discrete steps provide noise immunity
- Enables accurate and deterministic calculations
  - Same inputs give same outputs exactly
- Not constrained by physically realizable functions

IBM 701 (1952)

- IBM’s first commercial scientific computer
- Main memory was 72 William’s Tubes, each 1Kib, for total of 2048 words of 36 bits each
  - Memory cycle time of 12μs
- Accumulator ISA with multipler/quotient register
- 18-bit/36-bit numbers in sign-magnitude fixed-point
- Misquote from Thomas Watson Sr/Jr:

  “I think there is a world market for maybe five computers”
- Actually TWJr said at shareholder meeting:

“as a result of our trip [selling the 701]***, on which we expected to get orders for five machines, we came home with orders for 18.”***

IBM 650 (1953)

- The first mass-produced computer
- Low-end system with drum-based storage and digit serial ALU
- Almost 2,000 produced

IBM 360 : A General-Purpose Register (GPR) Machine

- Processor State
  - 16 General-Purpose 32-bit Registers
    - *may be used as index and base register*
    - *Register 0 has some special properties*
  - 4 Floating Point 64-bit Registers
  - A Program Status Word (PSW)
    - *PC, Condition codes, Control flags*
- A 32-bit machine with 24-bit addresses
  - But no instruction contains a 24-bit address!
- Data Formats
  - 8-bit bytes, 16-bit half-words, 32-bit words, 64-bit double-words

The IBM 360 is why bytes are 8-bits long today!

IBM Mainframes survive until today

z14 processor design summary

Micro-Architecture

- 10 cores per CP-chip
- 5.2GHz

- Cache Improvements:
    - 128KB I$ + 128KB D$
    - 2x larger L2 D$ (4MB)
    - 2x larger L3 Cache
    - symbol ECC

- New translation & TLB design
    - Logical-tagged L1 directory
    - Pipelined 2nd level TLB
    - Multiple translation engines

- Pipeline Optimizations
    - Improved instruction delivery
    - Faster branch wakeup
    - Improved store hazard avoidance
    - 2x double-precision FPU bandwidth
    - Optimized 2nd generation SMT2

- Better Branch Prediction
    - 33% Larger BTB1 & BTB2
    - New Perceptron & Simple Call/Return Predictor

Architecture

- PauseLess Garbage Collection
- Vector Single & Quad precision
- Long-multiply support (RSA, ECC)
- Register-to-register BCD arithmetic

Accelerators

- Redesigned in-core crypto-accelerator
    - Improved performance
    - New functions (GCM, TRNG, SHA3)

- Optimized in-core compression accelerator
    - Improved start/stop latency
    - Huffman encoding for better compression ratio
    - Order-preserving compression