University Of Diyala
College Of Engineering
Computer Engineering Department

COMPUTER ARCHITECTURE II

PART 1: INTRODUCTION

Asst. Prof. Ahmed Salah Hameed

Second stage

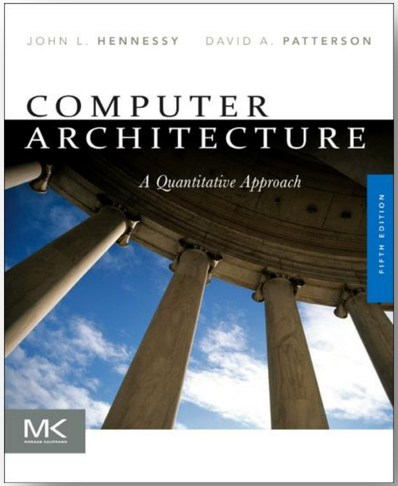
2022-2023

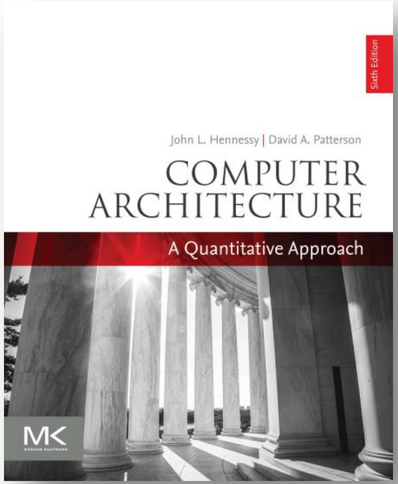
Pre-requisites:

- [CPE 201] Computer Architecture I & [CPE 105] Digital Logic Circuits I

Textbook:

John L. Hennessy and David A. Patterson, Morgan Kaufmann,
Computer Architecture: A Quantitative Approach, 5th Edition/6th Edition





- **Chapter 1:** Fundamentals of Quantitative Design and Analysis
- **Chapter 2:** Memory Hierarchy Design
- **Chapter 3:** Instruction-Level Parallelism and Its Exploitation
- **Chapter 4:** Data-Level Parallelism in Vector, SIMD, and GPU Architectures
- **Chapter 5:** Thread-Level Parallelism

INTRODUCTION:

- **Computer architecture overview**

Von Neumann Architecture	**Harvard Architecture**
➢ Same memory used for both data and instruction (code)	➢ Different memory blocks for data and instruction(code)
➢ One bus for the single memory (CPU has only one operation at a time)	➢ Separate buses for two memories (data memory and code memory)
➢ Two different set of clock cycles	➢ One set of clock cycle
➢ Pipelining is not better	➢ Pipelining is work better
➢ Simple	➢ Complex

DEFINING COMPUTER ARCHITECTURE

Computer architecture: referred to only instruction set design. Other aspects of computer design were called implementation.

↑ **OLD**

NEW ↓

Organization (microarchitecture) includes the high-level aspects of a computer's design, such as the memory system, the memory interconnect, and the design of the internal processor or CPU .

AMD Opteron	**80x86 instruction set**	**different pipeline and cache organizations**
Intel Core i7	**80x86 instruction set**	

DEFINING COMPUTER ARCHITECTURE

Hardware refers to the specifics of a computer, including the detailed logic design and the packaging technology of the computer.

Intel Core i7	**Nearly identical in instruction set and organization**	**offer different clock rates and different memory systems, making the Xeon E7 more effective for server computers.**
Intel Xeon E7		

THE NEW DEFINITION

Architecture covers all three aspects of computer design—instruction set architecture, organization or microarchitecture, and hardware.
Computer architects must design a computer to meet functional requirements as well as price, power, performance, and availability goals.

COMPUTER TECHNOLOGY

- **Computer technology:** (70 years)

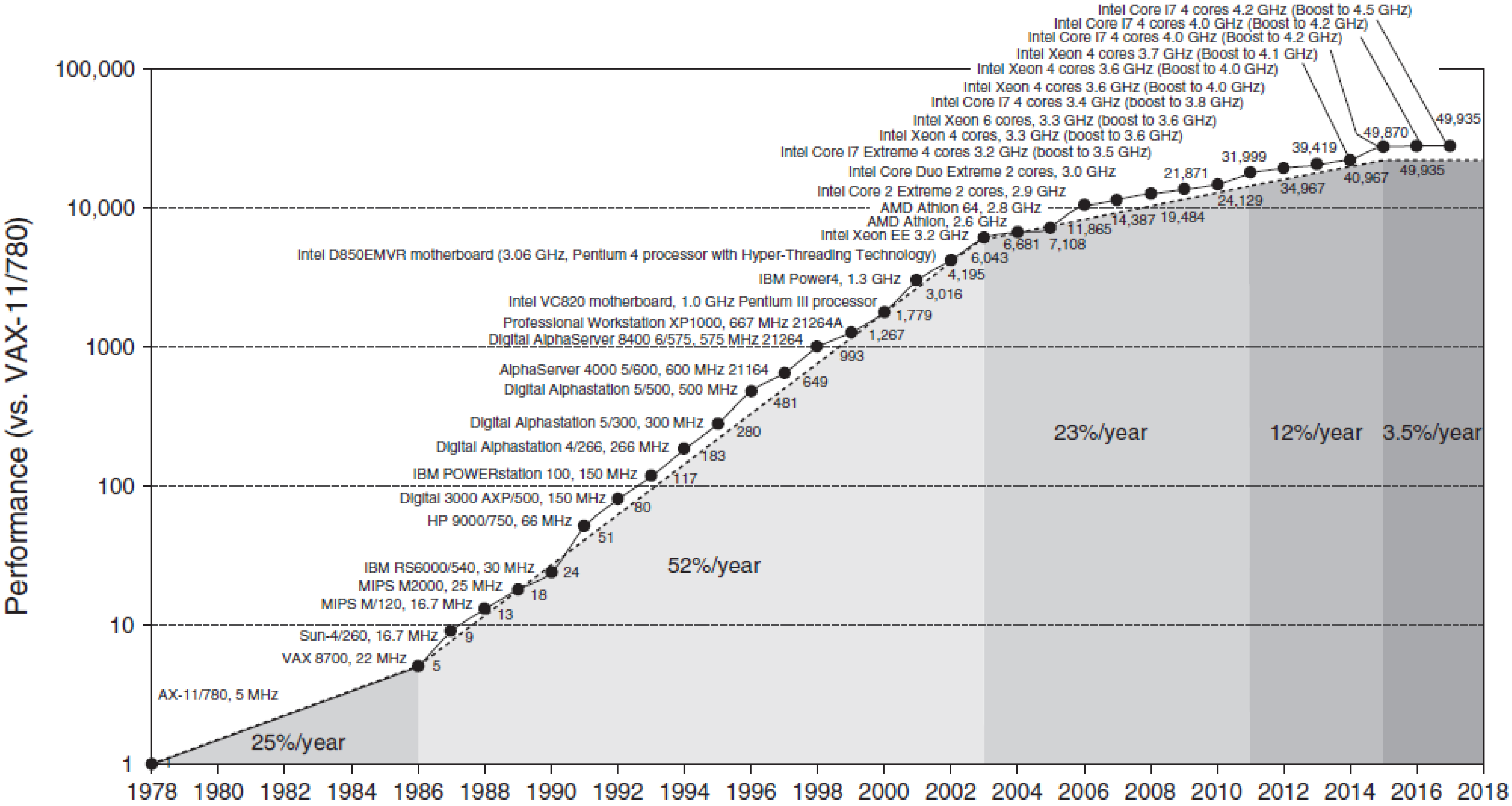
➢ Advances in the technology used to build computers

➢ Innovations in computer design.

1993	Today
the world's fastest computer in 1993 cost $50 million	A cell phone with performance better than as the world's fastest computer.

COMPUTER TECHNOLOGY

Growth in processor performance over 40 years



PERFORMANCE IMPROVEMENTS

- Improvements in semiconductor technology
  - Feature size, clock speed
- Improvements in computer architectures
  - Enabled by HLL compilers, UNIX
  - Lead to RISC architectures
- Together have enabled:
  - Lightweight computers
  - Productivity-based managed/interpreted programming languages
  - SaaS, Virtualization, Cloud
- Applications evolution:
  - Speech, sound, images, video, “augmented/extended reality”, “big data”

CLASSES OF COMPUTERS

- Internet of Things/Embedded Computers
  e.g. microwaves, washing machines, most printers, networking switches, and all automobiles. (19 billion sold 2015)
- Personal Mobile Device (PMD)
  e.g. smart phones, tablet computers (1.6 billion sold 2015)
  Emphasis on energy efficiency and real-time
- Desktop Computing
  Emphasis on price-performance (275 million desktop PCs)
- Servers
  Emphasis on availability (very costly downtime!), scalability, throughput (15 million servers)
- Clusters / Warehouse Scale Computers
  Used for "Software as a Service (SaaS)", PaaS, IaaS, etc.
  Emphasis on availability ($6M/hour-downtime at Amazon.com!) and price-performance (power=80% of TCO!)
  Sub-class: Supercomputers, emphasis: floating-point performance and fast internal networks, and big data analytics

CLASSES OF COMPUTERS

Feature	Personal mobile device (PMD)	Desktop	Server	Clusters/warehouse-scale computer	Internet of things/ embedded
Price of system	\$100–\$1000	\$300–\$2500	\$5000–\$10,000,000	\$100,000–\$200,000,000	\$10–\$100,000
Price of microprocessor	\$10–\$100	\$50–\$500	\$200–\$2000	\$50–\$250	\$0.01–\$100
Critical system design issues	Cost, energy, media performance, responsiveness	Price-performance, energy, graphics performance	Throughput, availability, scalability, energy	Price-performance, throughput, energy proportionality	Price, energy, application-specific performance

CLASSES OF PARALLELISM AND PARALLEL ARCHITECTURES

There are basically two kinds of parallelism in applications:

1. Data-level parallelism (DLP)

arises because there are many data items that can be operated on at the same time.

2. Task-level parallelism (TLP)

arises because tasks of work are created that can operate independently and largely in parallel.

CLASSES OF PARALLELISM AND PARALLEL ARCHITECTURES

Computer hardware in turn can exploit the two kinds of application parallelism in four major ways:

1. **Instruction-level parallelism**

exploits data-level parallelism at modest levels with compiler help using ideas like pipelining and at medium levels using ideas like speculative execution.

2. Vector architectures, graphic processor units (GPUs), and multimedia instruction sets exploit data-level parallelism by applying a single instruction to a collection of data in parallel.

3. Thread-level parallelism

exploits either data-level parallelism or task-level parallelism in a tightly coupled hardware model that allows for interaction between parallel threads.

4. Request-level parallelism

exploits parallelism among largely decoupled tasks specified by the programmer or the operating system.

TRENDS IN TECHNOLOGY

1. Integrated circuit logic technology

The number of devices per chip is still increasing, but at a decelerating rate. Unlike in the Moore's Law era, it is expected that the doubling time to be stretched with each new technology generation.

2. Semiconductor DRAM

The growth of DRAM has slowed dramatically, from quadrupling every three years as in the past. The 8-gigabit DRAM was shipping in 2014, but the 16-gigabit DRAM won't reach that state until 2019, and it looks like there will be no 32-gigabit DRAM

3. Semiconductor Flash (electrically erasable programmable read-only memory) In recent years, the capacity per Flash chip increased by about 50%–60% per year, doubling roughly every 2 years. Currently, Flash memory is 8–10 times cheaper per bit than DRAM.

4. Magnetic disk technology

Between 2004 and 2011, it dropped back to about 40% per year, or doubled every two years. Recently, disk improvement has slowed to less than 5% per year.

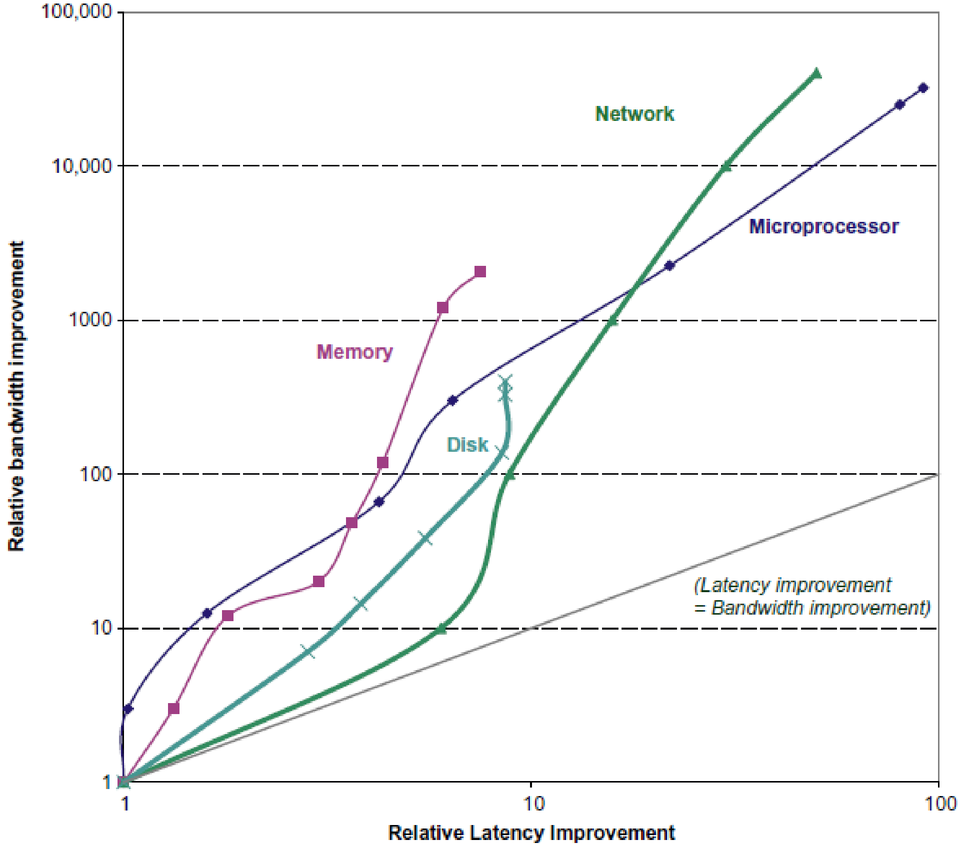
5. Network technology

Network performance depends both on the performance of switches and on the performance of the transmission system.

PERFORMANCE TRENDS: BANDWIDTH OVER LATENCY

Bandwidth or throughput is the total amount of work done in a given time, such as megabytes per second for a disk transfer.

Latency or response time is the time between the start and the completion of an event, such as milliseconds for a disk access.

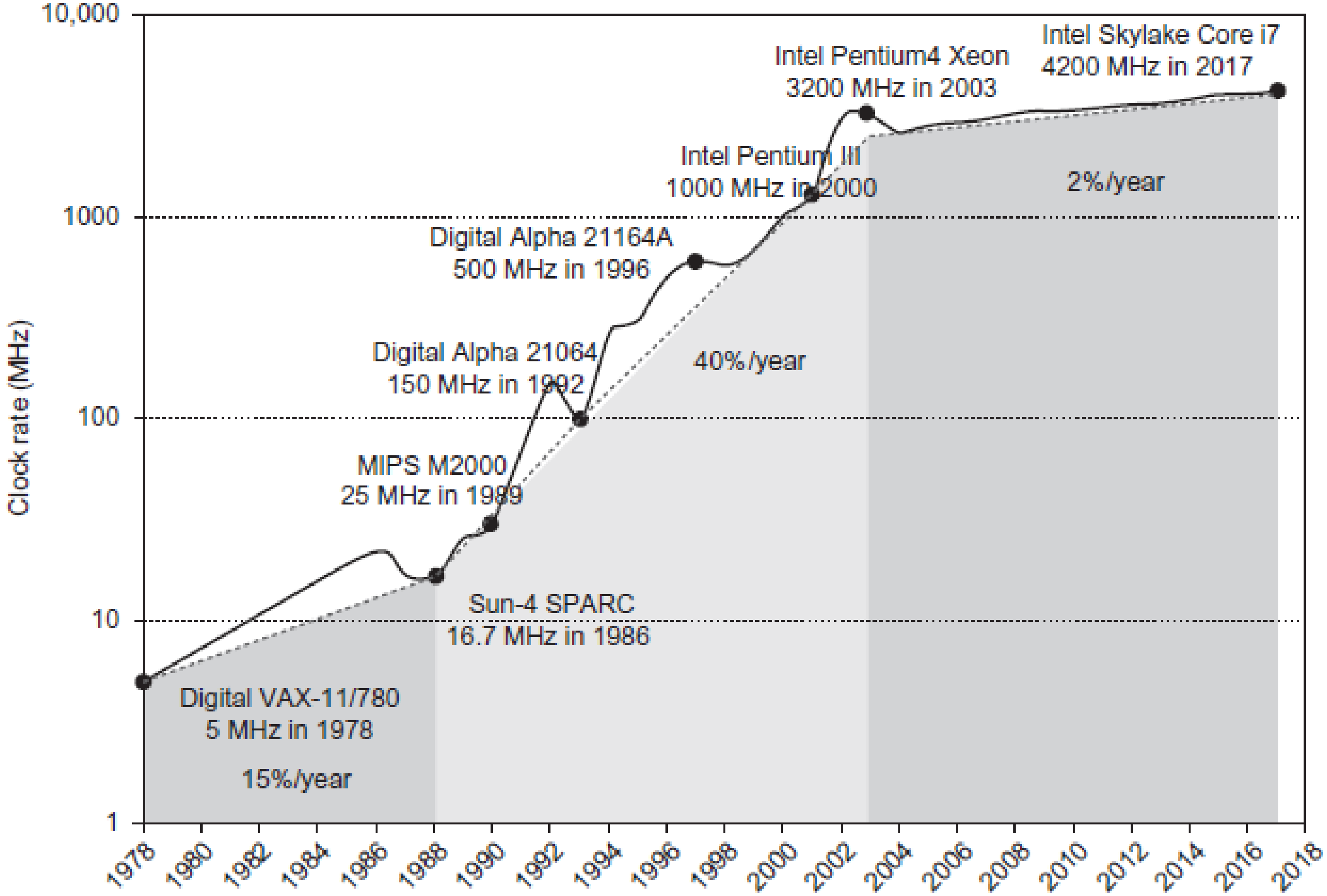


TRENDS IN POWER AND ENERGY IN INTEGRATED CIRCUITS

Energy is the biggest challenge facing the computer designer for nearly every class of computer.

Intel 80386
consumed ~ 2 W

3.3 GHz Intel Core i7
consumes 130 W



TECHNIQUES TO IMPROVE ENERGY EFFICIENCY DESPITE FLAT CLOCK RATES AND CONSTANT SUPPLY VOLTAGES:

1. Do nothing well

Most microprocessors today turn off the clock of inactive modules to save energy and dynamic power. For example, if no floating-point instructions are executing, the clock of the floating-point unit is disabled. If some cores are idle, their clocks are stopped.

2. Dynamic voltage-frequency scaling (DVFS).

Modern microprocessors typically offer a few clock frequencies and voltages in which to operate that use lower power and energy.

3. Design for the typical case.

Given that PMDs and laptops are often idle, memory and storage offer low power modes to save energy. For example, DRAMs have a series of increasingly lower power modes to extend battery life in PMDs and laptops, and there have been proposals for disks that have a mode that spins more slowly when unused to save power.

4. Overclocking.

Intel started offering Turbo mode in 2008, where the chip decides that it is safe to run at a higher clock rate for a short time, possibly on just a few cores, until temperature starts to rise. For example, the 3.3 GHz Core i7 can run in short bursts for 3.6 GHz.

SECOND SEMESTER REPORT

1) **Von Neumann Architecture Vs. Harvard Architecture**
2) **Classes of Computers**