Lecture 3

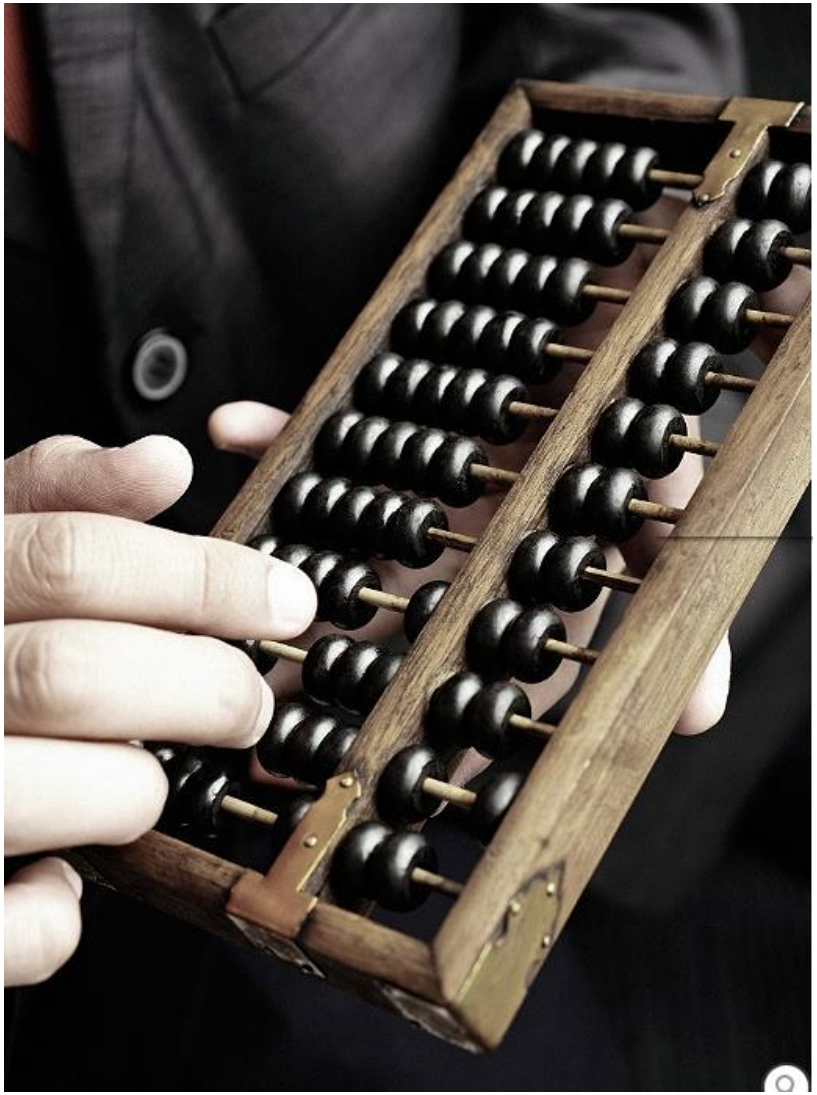
Computer Architecture
(Spring 2020)

Introduction

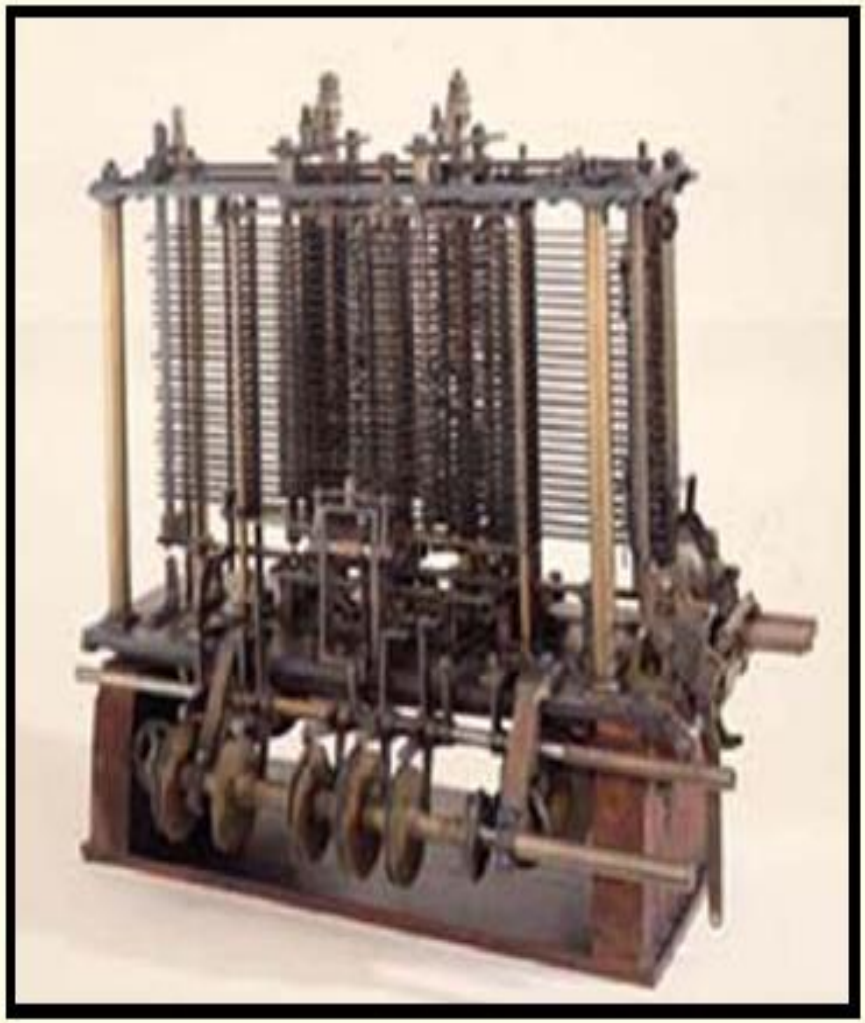
Dr. Duo Liu (刘铎)
Office: Main Building 0626
Email: liuduo@cqu.edu.cn

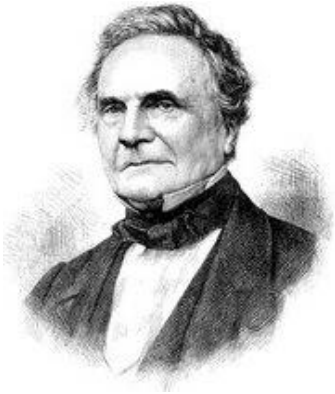
Mathematical Tools

Mathematical Tools
Throughout the history of calculating we've devised ways to add speed and accuracy while subtracting the drudgery.
Many solutions used body parts, notably fingers. A 19th century Chinese technique can count to 10 billion using just two hands! Other solutions were mechanical — both general-purpose tools for everyday calculations and specialized instruments for engineering, navigational, or other scientific and technical problems.



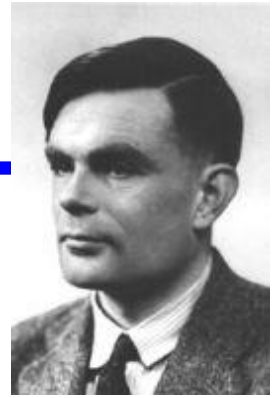
History: Analytical Engine (1883)

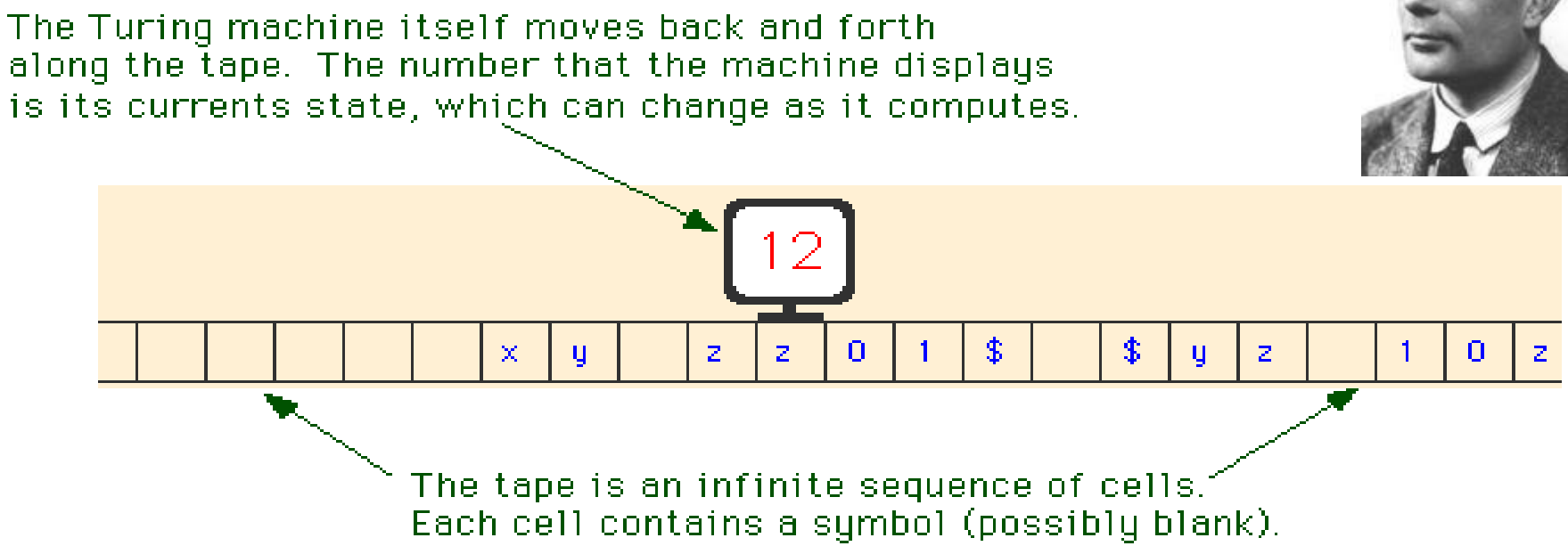




- Charles Babbage (1791-1871, UK)
- Made by mechanical parts.
- Started in 1833; never finished.

Turing Machine (1937)

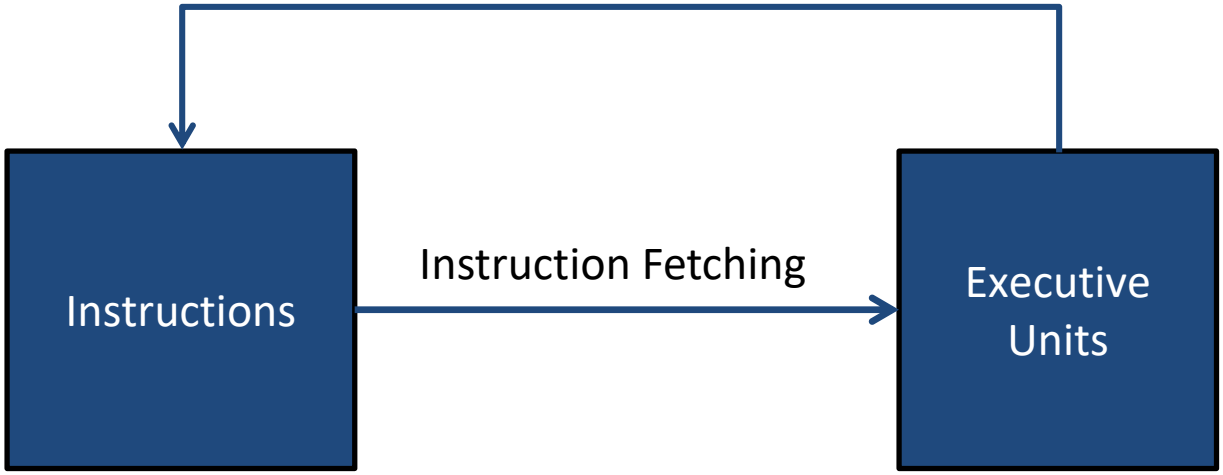




- **A theoretical computing machine, the basis of modern computing.**
  - Theoretically, as powerful as any other computer
  - Conceptually, a finite set of states, a finite alphabet and a finite set of instructions.
  - Physically, it has a head (read, write), and move along an infinitely long tape that is divided into cells storing a letter.

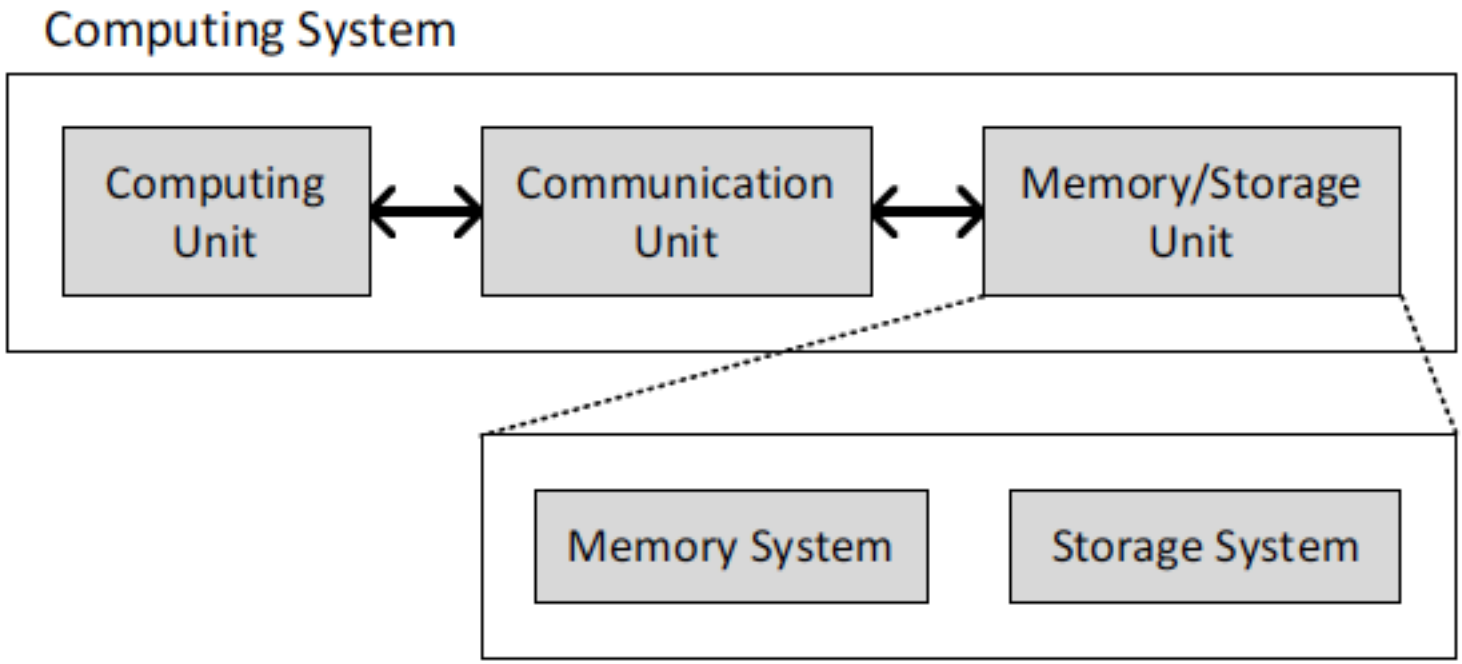
What is a Computer?

- **A simplest computer model**



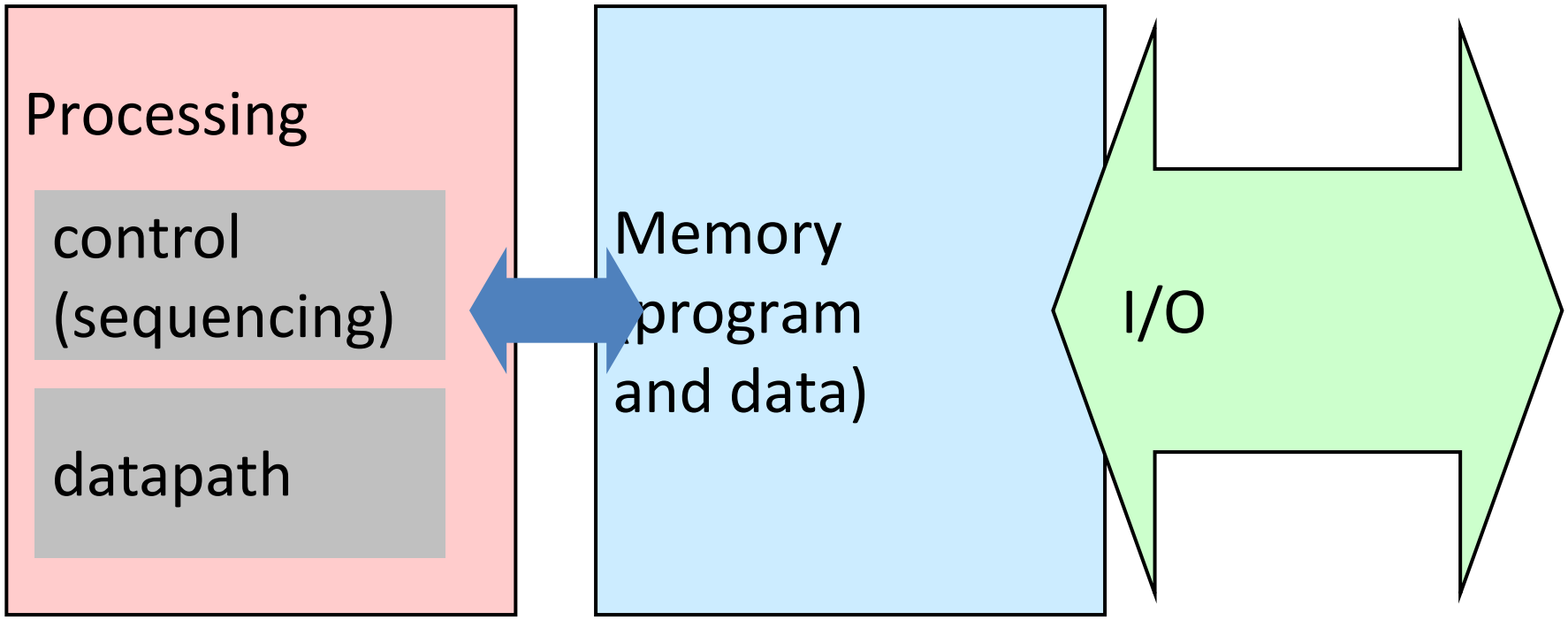
What is a Computer?

- Three key components
  - Computation
  - Communication
  - Storage (memory)



What is a Computer?

- We will cover all three components



Stored-program Concept (1945)

John von Neumann and Alan Turing, proposed the stored program concept in the 1940s.

Stored-program Concept

- Store program and data in memory
- A computer can read them from memory.
- Program can be altered.

The Von Neumann Model/Architecture

- **Also called *stored program computer* (instructions in memory). Two key properties:**

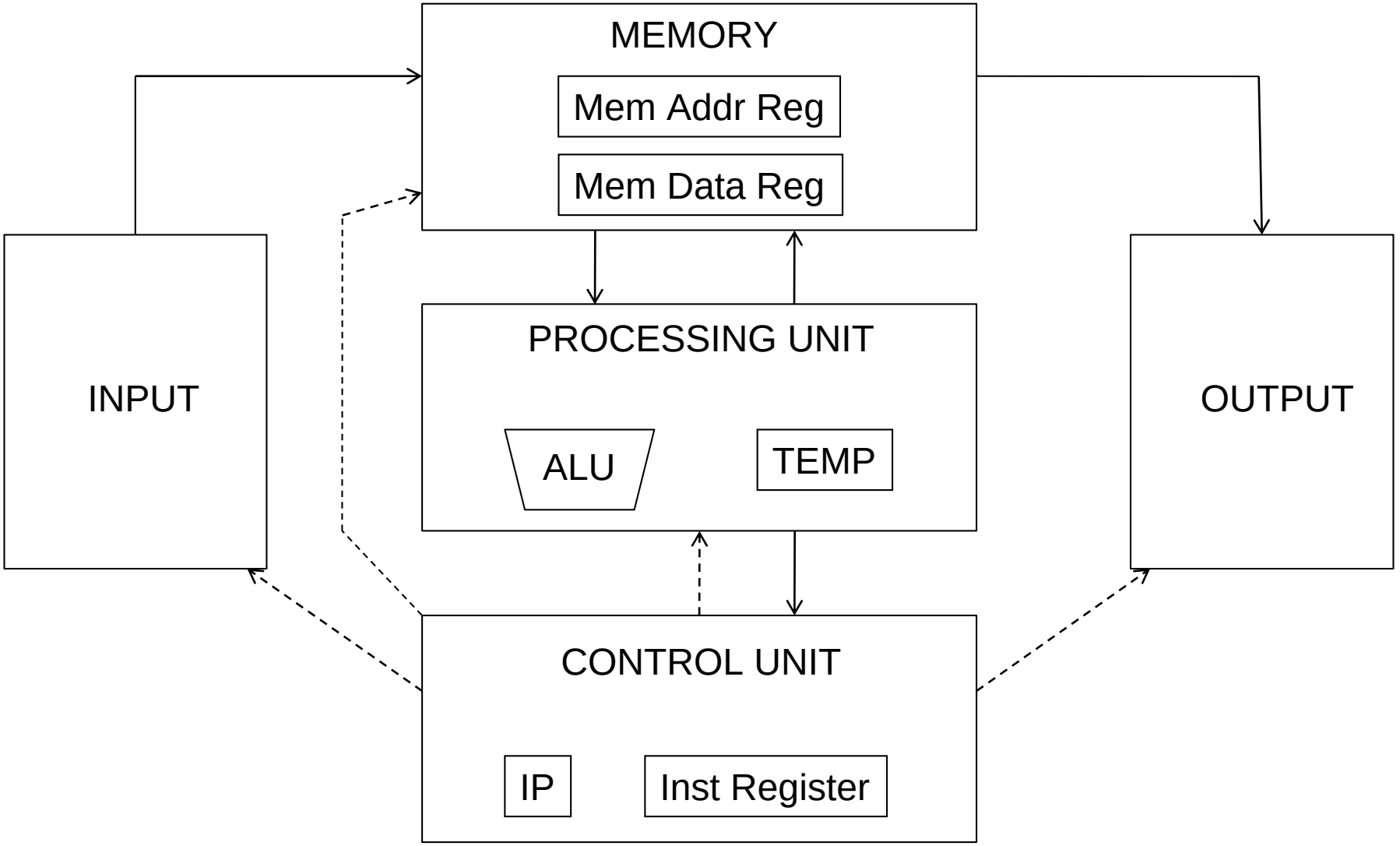
- **Stored program**
  - Instructions stored in a linear memory array
  - Memory is unified between instructions and data
    - The interpretation of a stored value depends on the control signals

- **Sequential instruction processing**
  - One instruction processed (fetched, executed, and completed) at a time
  - Program counter (instruction pointer) identifies the current instr.
  - Program counter is advanced sequentially except for control transfer instructions

The Von Neumann Model/Architecture

- **Recommended reading**
  - Burks, Goldstein, von Neumann, "Preliminary discussion of the logical design of an electronic computing instrument," 1946.
  - Patt and Patel book, Chapter 4, "The von Neumann Model"
- **Stored program**
- **Sequential instruction processing**

The Von Neumann Model (of a Computer)

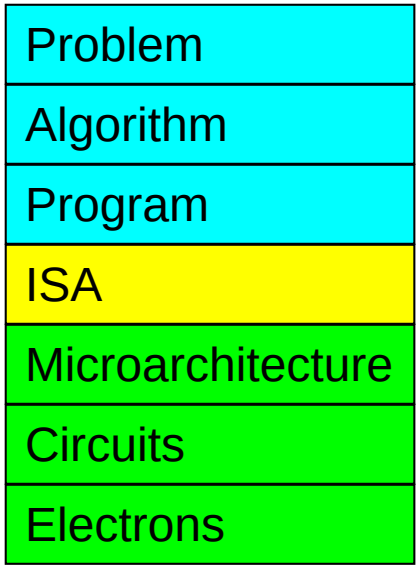


The Von-Neumann Model

- **All major *instruction set architectures* today use this model**
  - x86, ARM, MIPS, SPARC, Alpha, POWER
- **Underneath (at the microarchitecture level), the execution model of almost all *implementations (or, microarchitectures)* is very different**
  - Pipelined instruction execution: *Intel 80486 uarch*
  - Multiple instructions at a time: *Intel Pentium uarch*
  - Out-of-order execution: *Intel Pentium Pro uarch*
  - Separate instruction and data caches
- **But, what happens underneath that is *not* consistent with the von Neumann model is *not* exposed to software**
  - Difference between ISA and microarchitecture

ISA vs. Microarchitecture

- **ISA**
  - Agreed upon interface between software and hardware
    - SW/compiler assumes, HW promises
  - What the software writer needs to know to write and debug system/user programs
- **Microarchitecture**
  - Specific implementation of an ISA
  - Not visible to the software
- **Microprocessor**
  - **ISA, uarch**, circuits
  - "Architecture" = ISA + microarchitecture



ISA vs. Microarchitecture

- **Implementation (uarch) can be various as long as it satisfies the specification (ISA)**
  - Add instruction vs. Adder implementation
    - Bit serial, ripple carry, carry lookahead adders are all part of microarchitecture
  - x86 ISA has many implementations: 286, 386, 486, Pentium, Pentium Pro, Pentium 4, Core, …

- **Microarchitecture usually changes faster than ISA**
  - Few ISAs (x86, ARM, SPARC, MIPS, Alpha) but many uarchs
  - *Why?*

ISA

- **Instructions**
  - Opcodes, Addressing Modes, Data Types
  - Instruction Types and Formats
  - Registers, Condition Codes
- **Memory**
  - Address space, Addressability, Alignment
  - Virtual memory management
- **Call, Interrupt/Exception Handling**
- **Access Control, Priority/Privilege**
- **I/O: memory-mapped vs. instr.**
- **Task/thread Management**
- **Power and Thermal Management**
- **Multi-threading support, Multiprocessor support**

Intel® 64 and IA-32 Architectures
Software Developer's Manual

Volume 1:
Basic Architecture

Microarchitecture

- **Implementation of the ISA under specific design constraints and goals**
- **Anything done in hardware without exposure to software**
  - Pipelining
  - In-order versus out-of-order instruction execution
  - Memory access scheduling policy
  - Speculative execution
  - Superscalar processing (multiple instruction issue?)
  - Clock gating
  - Caching? Levels, size, associativity, replacement policy
  - Prefetching?
  - Voltage/frequency scaling?
  - Error correction?

1st Electronic Computer (1946)

ENIAC

- J. Eckert and J. Mauchly at Univ. of Pennsylvania
- 18,000 vacuum tubes
- 80 feet long; 8.5 feet high and several feel wide.
- 1,900 additions/sec

1st Full-Scale Stored-program Machine (1949)

EDSAC 1 (1949)

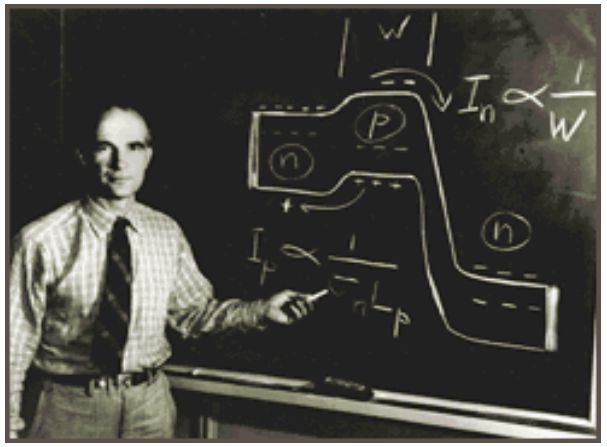
- **Maurice Wiles at Cambridge University**
- **1st stored-program machine.**
- **600 instructions/sec**
- **Occupied a room 5m by 4m**

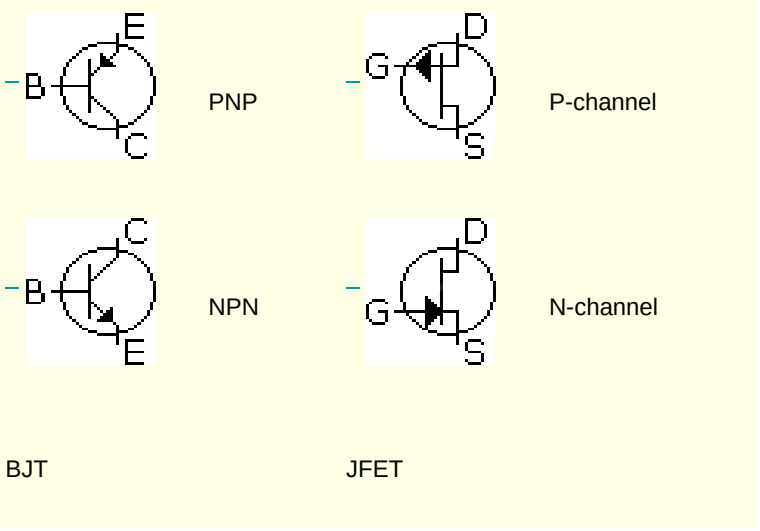
IBM 7094 (1962)

- Made by transistors (invented at Bell Lab in 1947).
- Manufacture components separately and then wire together onto circuit board.
- 10,000 transistors
- Clock Cycle = 2 us (500 KHz)
- 32 K Memory
- Price: $3M.

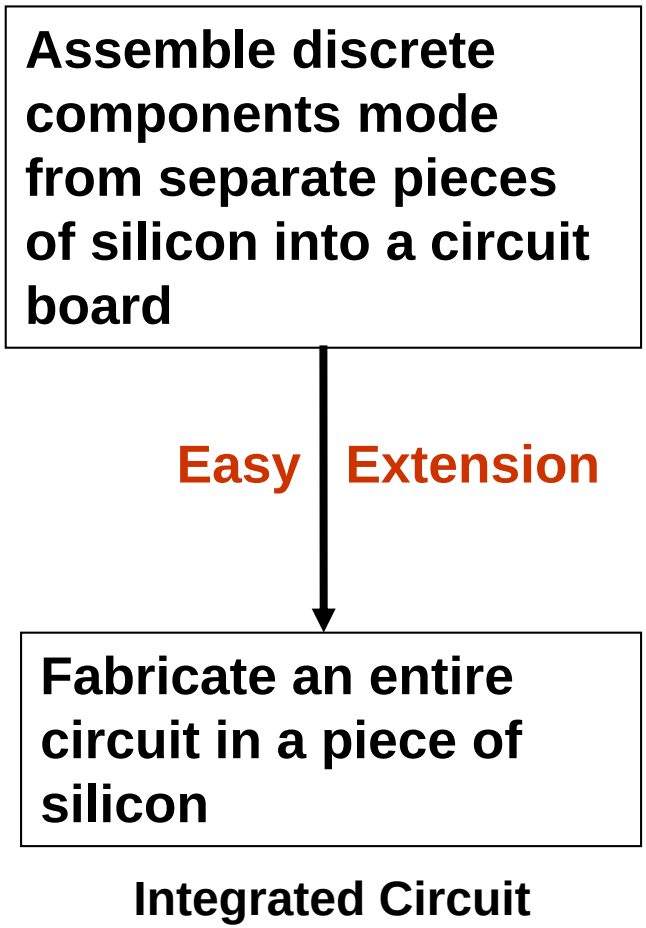
Transistor (1947)

- **Solid State device made from Silicon (Sand)**
- **Act as a variable valve: based on its input current (BJT) or input voltage (FET), *allows* a precise amount of current to flow through it from the circuit's voltage supply.**
- **Invented by John Bardeen, Walter H. Brattain, and William B. Shockley, awarded the Nobel Prize in physics in 1956.**
- **Smaller and cheaper with less heat dissipation**





Integrated Circuits (1958)



Integrated Circuit

- The cost of a chip virtually **unchanged** with the growth in density.
- Components are placed **closer**, the electrical path is shorten so the speed and capacity are **increased**.
- Computers become smaller
- Reduction in power and cooling requirement
- The interconnection is much more **reliable** than solder connections.

IBM System/360 (1964)

Model No.	Clock Rate	Memory Size	Price
40	1.6 MHz	32-256KB	$225,000
50	2.0 MHz	128-256KB	$550,000
65	5.0 MHz	256KB-1MB	$1,200,000
75	5.1 MHz	256KB-1MB	$1,900,000

- **A line of computers: the first planned family of computers (mainframe).**
- **Made by Integrated Circuits.**
- **Four models varied in cost and performance.**

PDP-8 (1965): The 1st Minicomputer

- Made by Integrated Circuits.
- Price: $16,000, cheap enough for each lab technician to have one.
- 500K Instruction/s (0.5 MIPS)
- 8 Cubic Feet

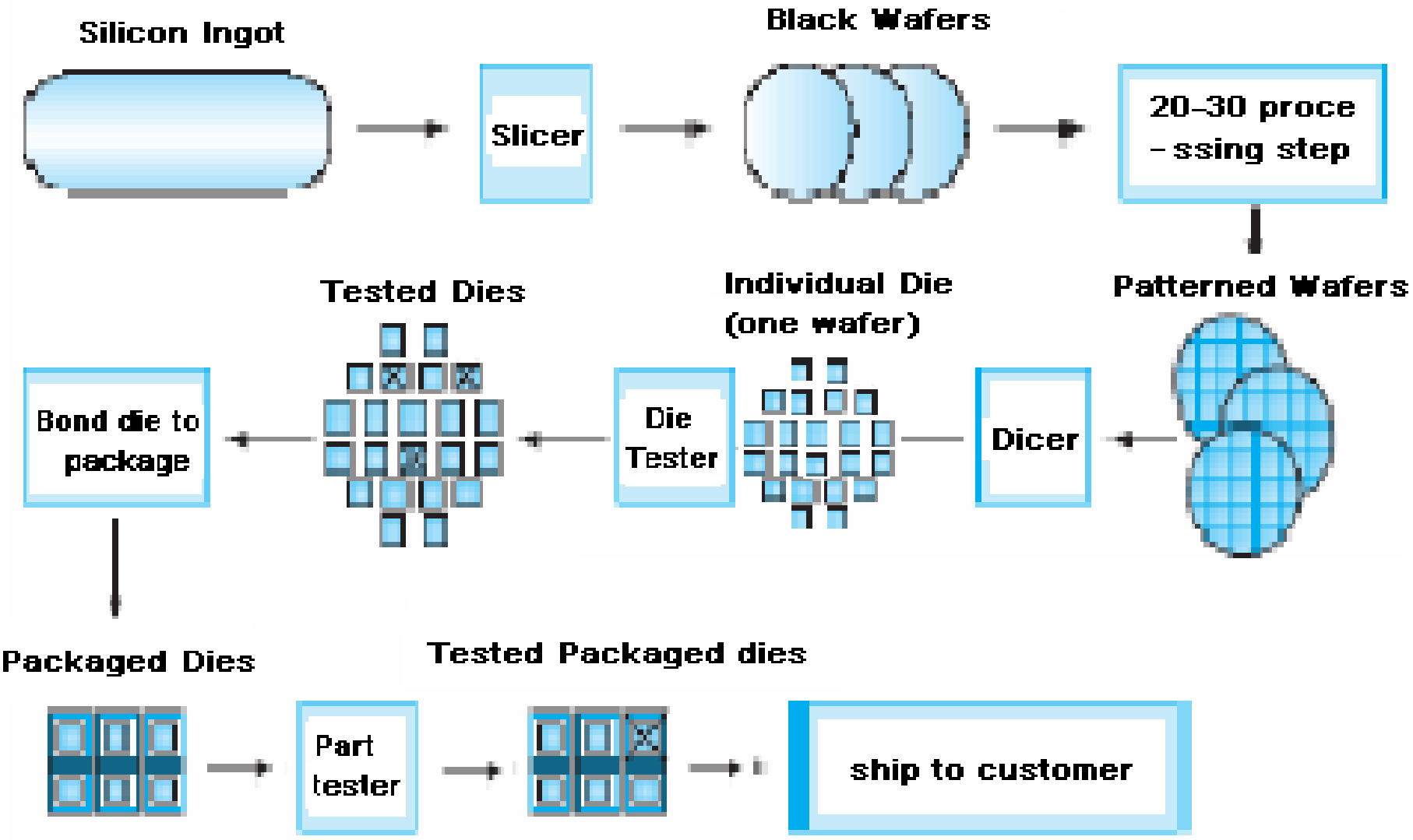
Moore's Law (1965)

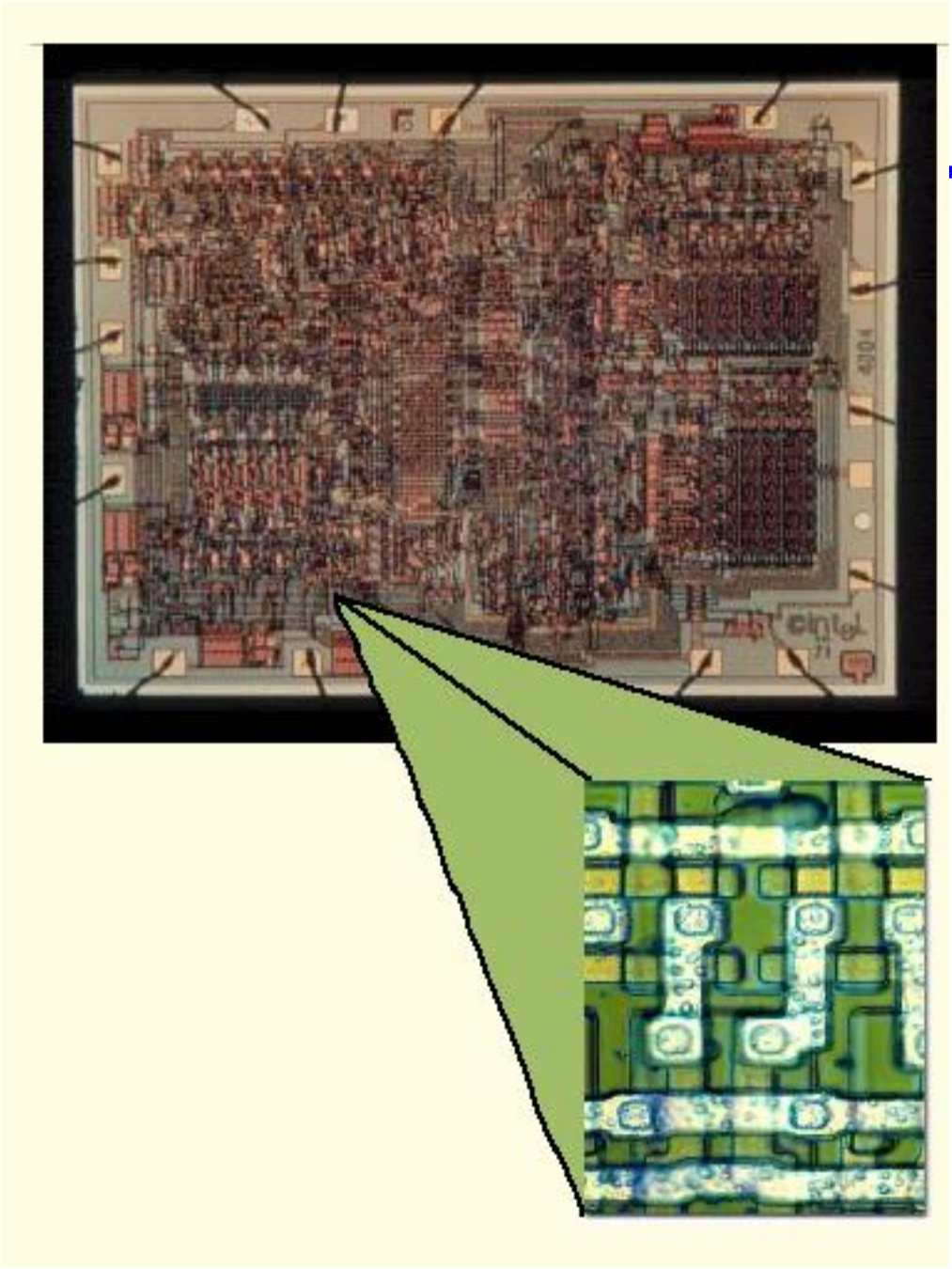
- **In 1965, Moore predicted that the number of transistors that could be put on a single chip will be doubling every year.**

- **The pace slowed to a doubling every 18 months in 70s.**

- **Moore's Law turns 40. What is its future?**
  - From a theoretical point of view, silicon transistors could continue to be shrunk until about the 4-nanometer manufacturing generation, which could appear about 2023.
  - At that point, the source and the drain, which are separated by the transistor gate and gate oxide, will be so close that electrons will drift over on their own.

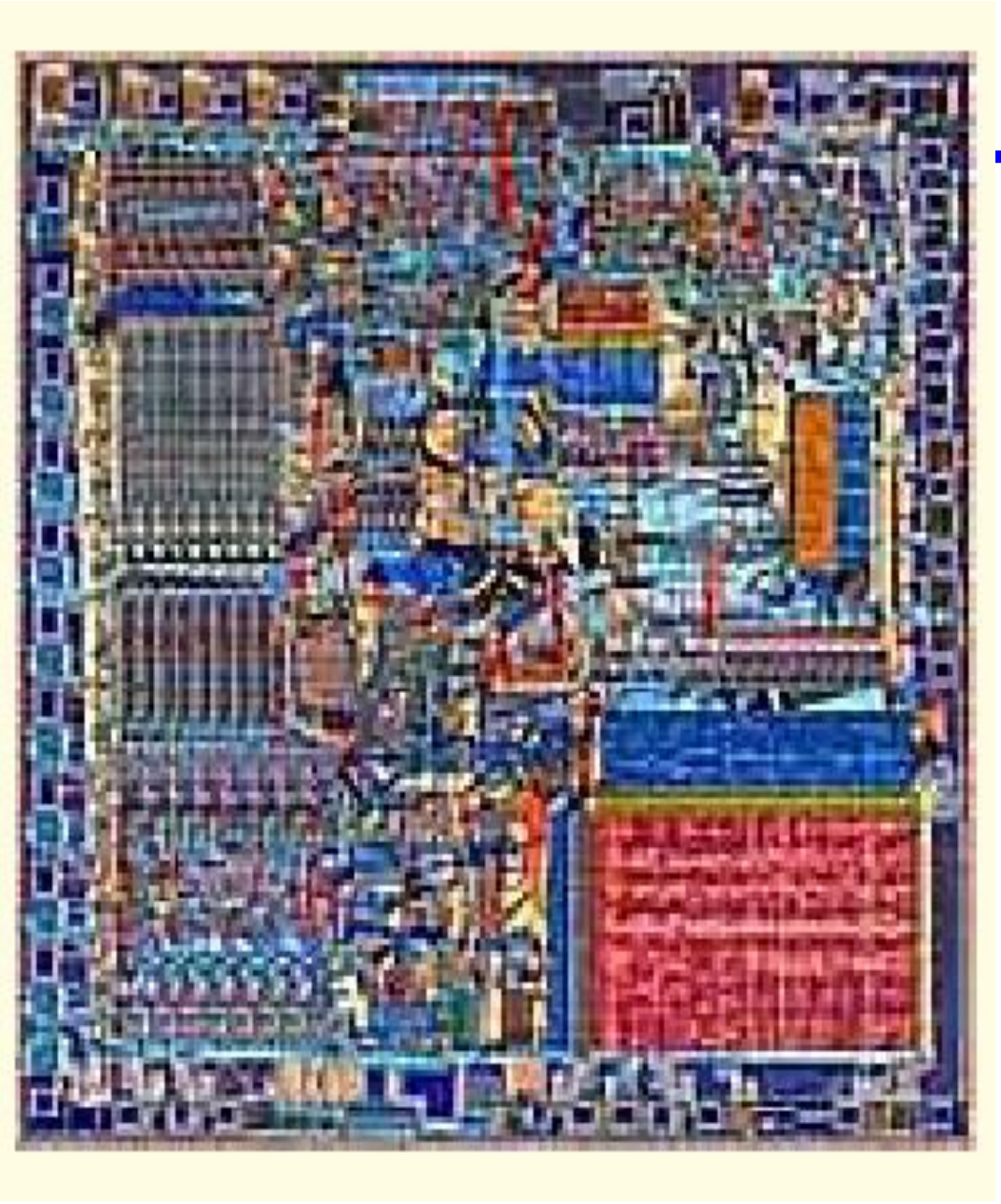
The Chip Manufacturing Process





Intel 4004 (1970)

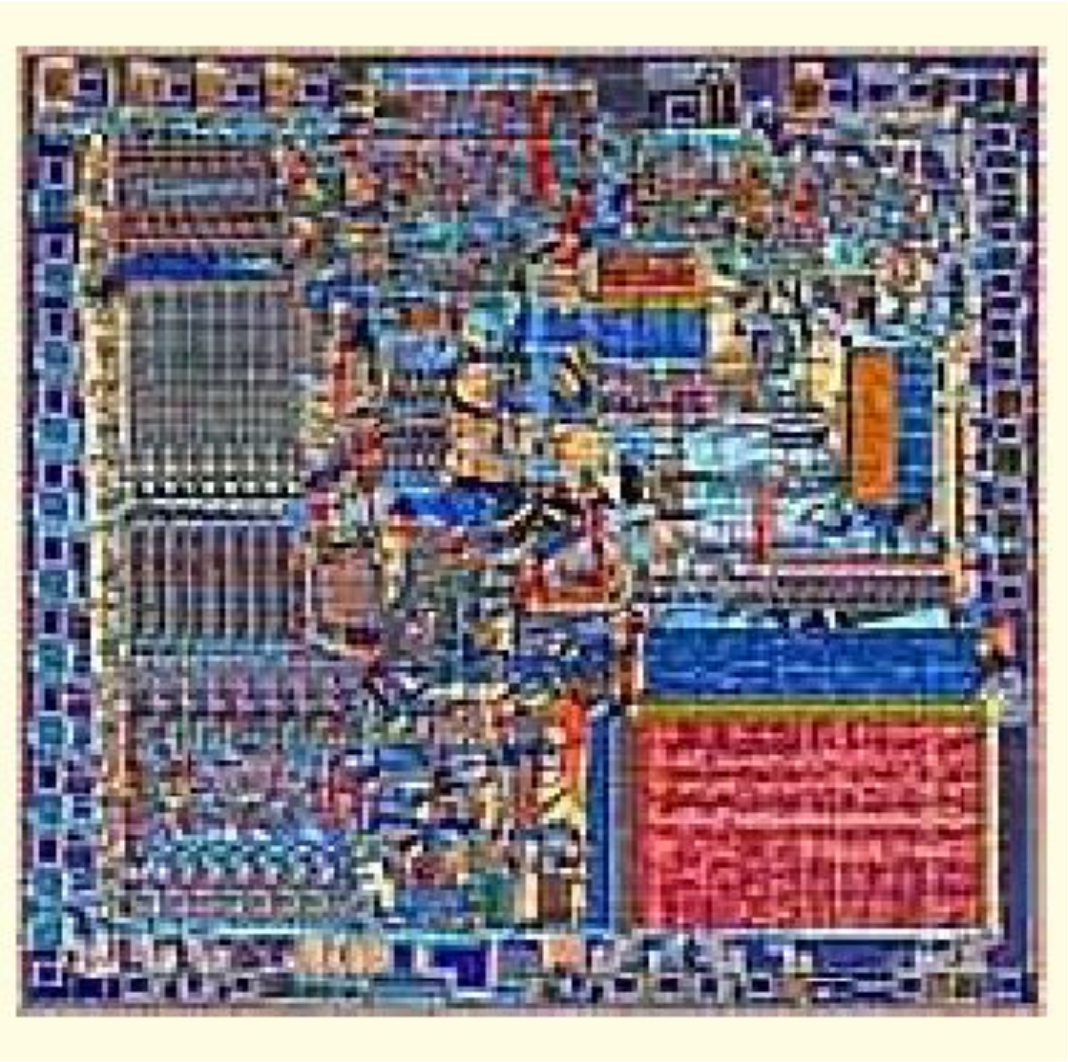
- The first microprocessor (all of the components of a CPU on a single chip).
- 2,300 Transistors
- 108 KHz
- Addressable memory: 640 bytes



Intel 8086 (1978)

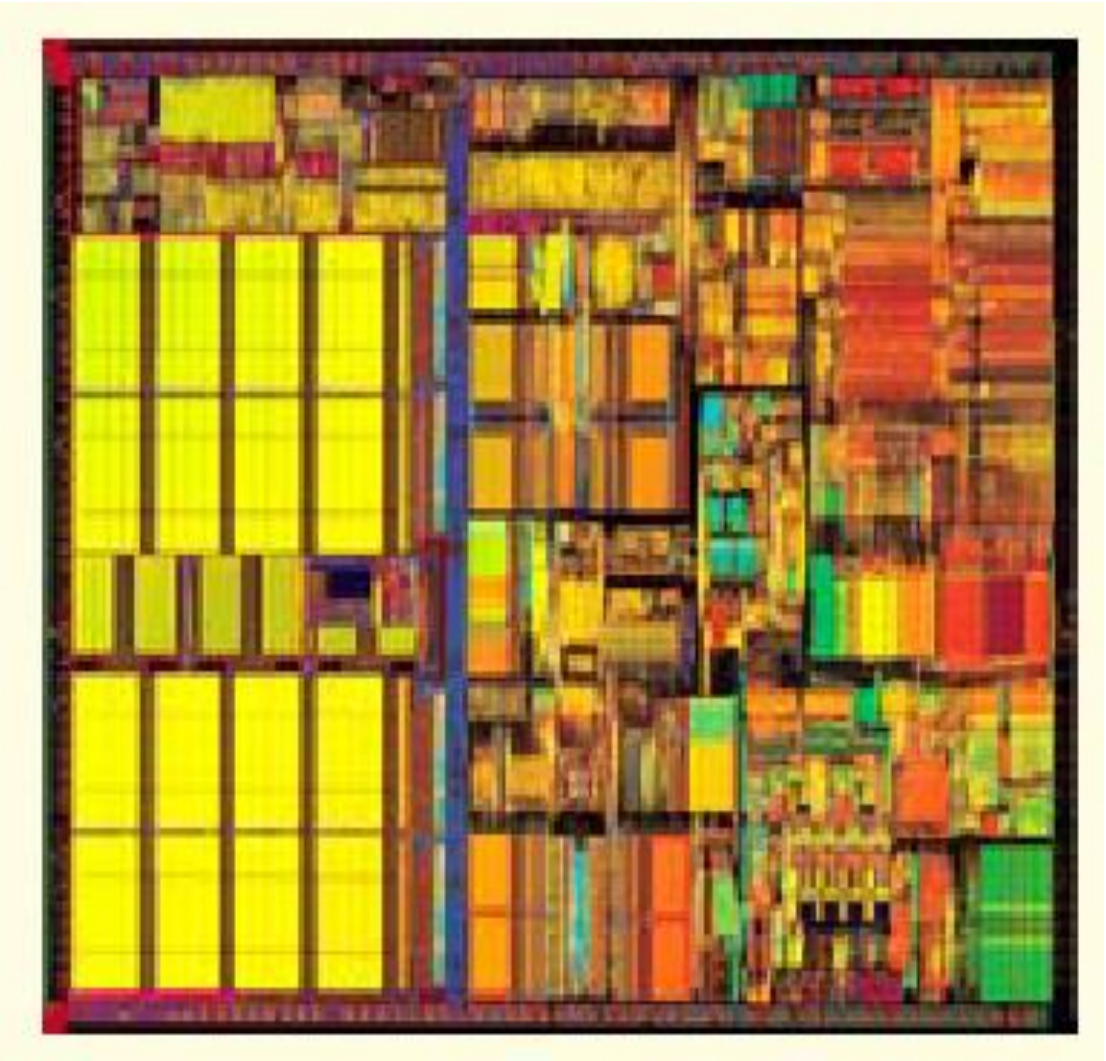
- 16-bit general-purpose processor
- 29,000 Transistors
- 5MHz, 8MHz, 10MHz
- 33 square mm

Pentium (1993)



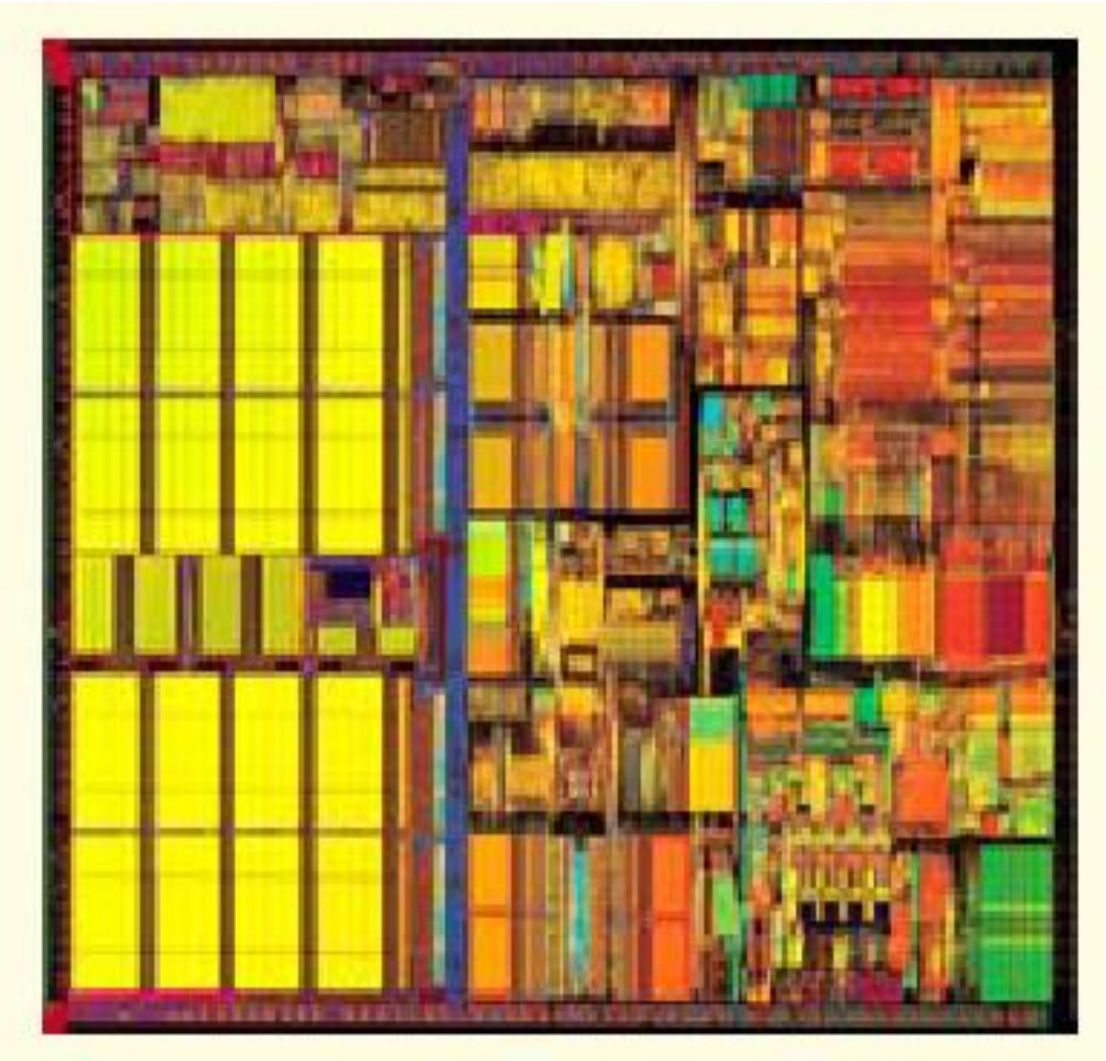
- 32-bit general-purpose processor
- 3.1M Transistors
- 60MHz-166MHz
- 296 Square mm
- The 1st Superscalar Implementation of IA 32

Pentium III (1999)



- 9.5 M Transistors
- 455 MHz
- 125 Square mm

Pentium 4 (2000)

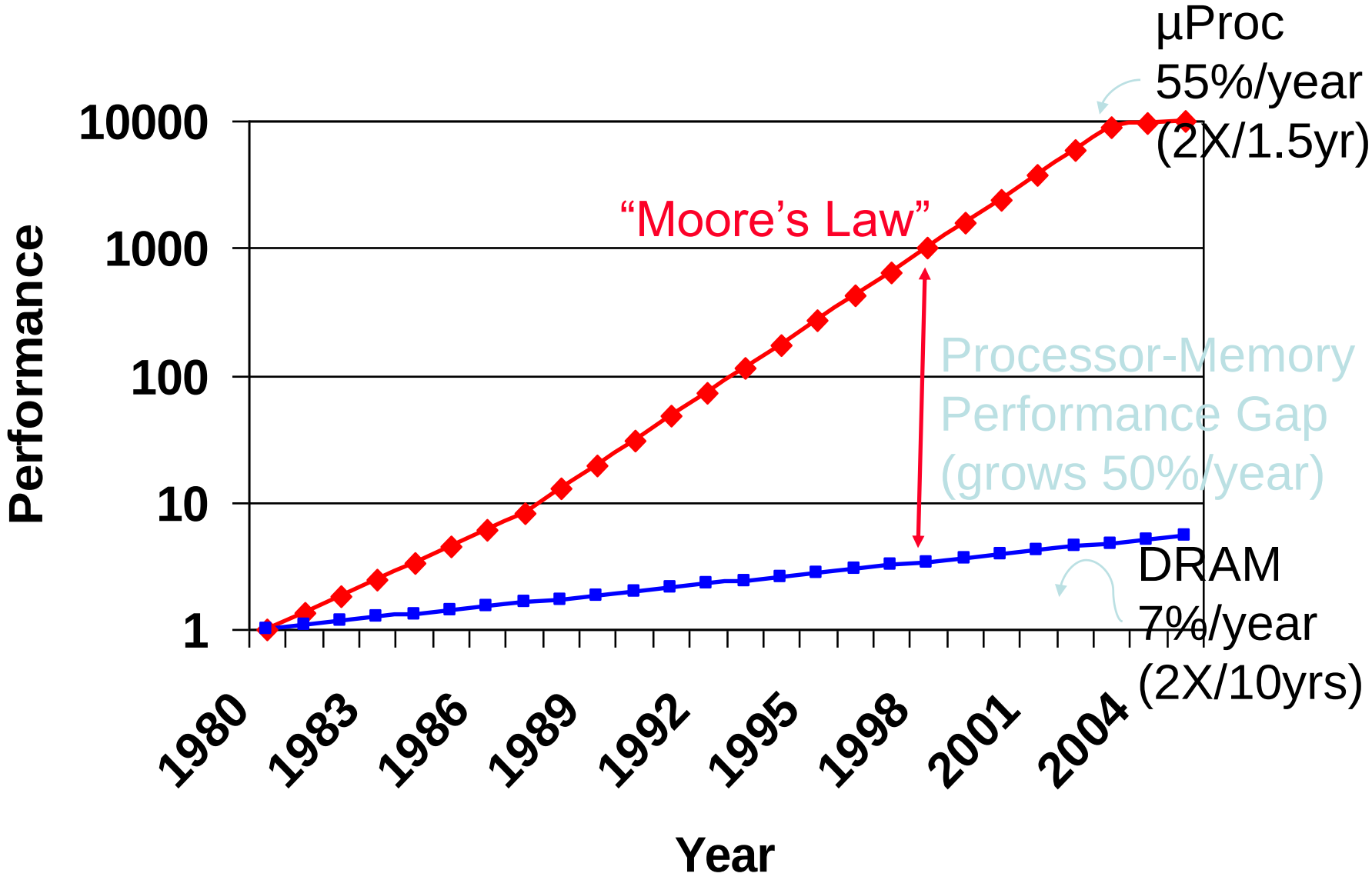


- 42 M Transistors
- 1.3-1.9 GHz
- 146 Square mm

Generations of Computer

Gen.	Dates	Technology	Typical Speed	Typical Products
1	1946-1957	Vacuum tube	40,000	ENIAC
2	1958-1964	Transistor	200,000	IBM 7000
3	1965-1971	Small/medium Scale Integrated Circuit	1,000,000	IBM System/360, DEC PDP-11
4	1971-1977	Large Scale Integrated Circuit	10,000,000	Intel 4004
5	1978-	Very-Large-Scale Integrated Circuit	100,000,000	Intel Pentium

Processor-Memory Performance Gap



Better Computer-Architectures are NEEDED !!!!

- **We need better computer architectures to fill the gaps**
  - Architecture
    - von Neumann Architecture & Harvard Architecture
    - VLIW, SuperScalar
    - ?
  - Instruction Set
    - CISC, RISC
    - ?
  - CPU & Memory
    - Hierarchy of Memory: Cache, Memory, Disk, …
    - ?
  - Advanced Peripheral Devices

Past Architecture Techniques

- **Pipeline**
  - Separate the execution of an instruction into several stages using separate resources
  - Overlap the execution of multiple instructions
  - e.g., we can have 5 stages: IF → ID → EX → MEM → WB

- **Branch Prediction**
- **Data Flow Analysis: optimal instruction schedule**
- **Speculative Execution**
- **Performance Balance: Adjust organization and architecture to compensate for the mismatch among various components.**

Solve the Mismatch between Processor Speed and Dynamic RAM Speed

- **Use wide data buses so we can retrieve more bits at the same time.**
- **Include a cache /other buffer scheme to make memory chip work more efficiently.**
- **Put cache into processors**
- **Use high-speed buses to interconnect processor and memory.**

Summary

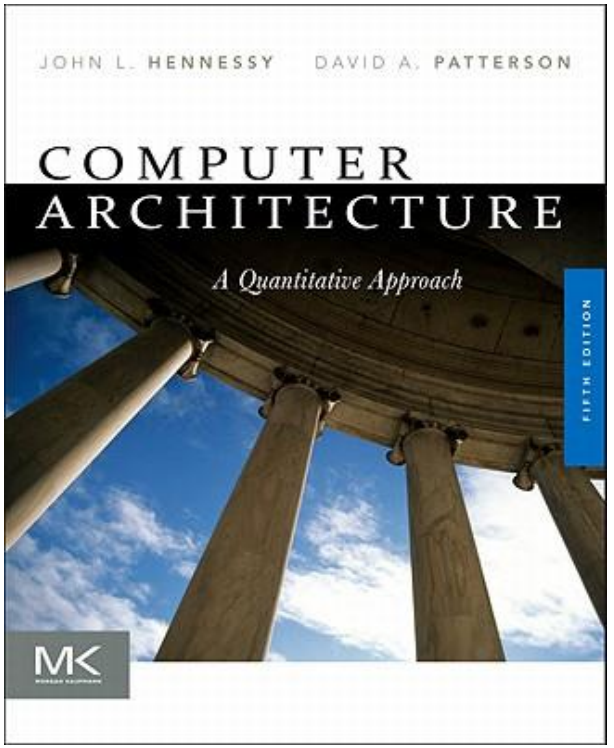
- **Technique advance made dramatic performance improvement in the past 40 years**
  - Processor
    - logic capacity: increases about 30% per year
    - performance: 2x every 1.5 years
  - Memory
    - DRAM capacity: 4x every 3 years, now 2x every 2 years
    - memory speed: 1.5x every 10 years
    - cost per bit: decreases about 25% per year
  - Disk
    - capacity: increases about 60% per year

  Moore's Law will continue work in the near future?
- **Better architectures are needed for better usage of IC capacity**

Assigned Readings

- Computer Architecture: A Quantitative Approach, 5th ed.," John L. Hennessey and David A. Patterson, Morgan Kaufman, 2011

- Sections: 1.1-1.6, Appendix L



Further Readings

- Martin Davis. “Engine of Logic”. Norton, 2000
- Hermann H. Goldstine. “The Computer: From Pascal to Von Neumann”. Princeton University Press, 1972
- Andrew Hodges. “Alan Turing: The Enigma”. Walker & Company, 2000
- Eloina Pelaez. “The Stored-Program Computer: Two Conceptions”. In Social Studies of Science 29(3), June 1999.
- Computer History Museum. http://computerhistory.org