COMPUTER ABSTRACTION AND TECHNOLOGY

Debiprasanna Sahoo

Assistant Professor

Department of Computer Science and Engineering

Indian Institute of Technology Roorkee

Content

Book

Computer Organization and Design: The Hardware/Software Interface- RISC-V Edition, 5th Edition, 2017

Chapter-2

David A. Patterson and John L. Henessey

Reference Books

Computer Architecture: A Quantitative Approach, 6th Edition, 2017

Chapter-1

David A. Patterson and John L. Henessey

Classes of Computers

Personal Computers: A computer designed for use by an individual, usually incorporating a graphics display, a keyboard, and a mouse.

Servers: A computer used for running larger programs for multiple users, often simultaneously, and typically accessed only via a network.

Supercomputers: A class of computers with the highest performance and cost; they are configured as servers and typically cost tens to hundreds of millions of dollars.

Classes of Computers

Embedded Computers: A computer inside another device used for running one predetermined application or collection of software.

Personal Mobile Devices (PMD): PMDs are small wireless devices to connect to the Internet; they rely on batteries for power, and software is installed by downloading apps. Conventional examples are smartphones and tablets.

Cloud Computing: It refers to large collections of servers that provide services over the Internet; some providers rent dynamically varying numbers of servers as a utility.

Commonly Used Acronyms

- CPU: Central Processing Unit/Core/Compute Units/Processing Elements (PE)
- GPU: Graphics Processing Unit
- GPGPU: General Purpose Graphics Processing Unit
- DRAM: Dynamic Random Access Memory
- SRAM: Static Random Access Memory
- SSD: Solid State Drives
- BIOS: Basic Input Output Systems
- DIMM: Dual Inline Memory Modules
- SATA/PATA: Serial/Parallel Advanced Technology Attachments
- PCIe: Peripheral Connecting Interface Express
- HDD: Hard Disk Drives

Great Ideas for Designing Computers

Enhancing Process Technology

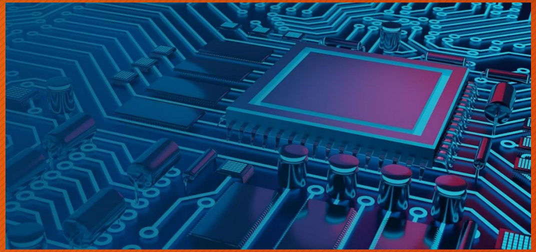
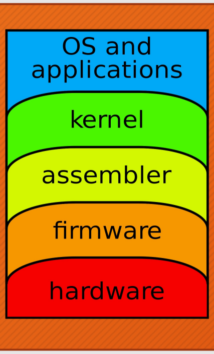


Image from Dell

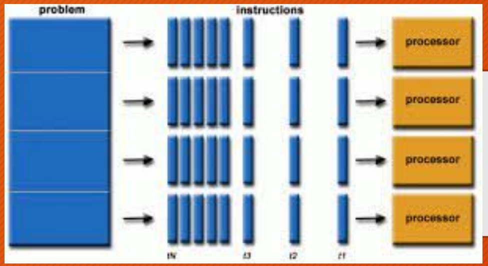


Abstraction

Image from Wikipedia

Make Common Cases Faster

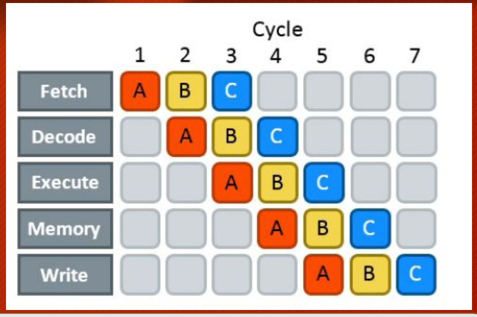
Image from Vecteezy



Performance via Parallelism

Image from HPC@LLNL

Performance via Pipelining



	Cycle 1	2	3	4	5	6	7
Fetch	A	B	C				
Decode		A	B	C			
Execute			A	B	C		
Memory				A	B	C	
Write					A	B	C

Image from Algorithmica

Performance via Prediction

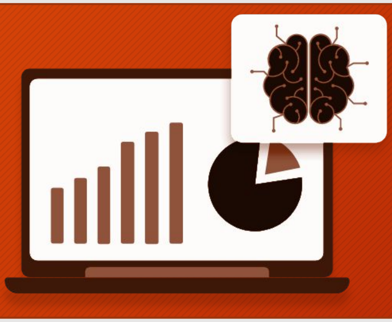


Image from Coditation

Hierarchy of Memory

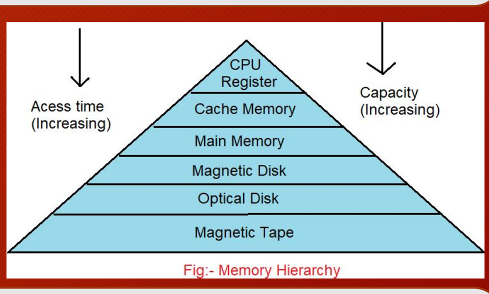


Fig:- Memory Hierarchy

Image from CodeNinja

Dependability via Redundancy

Image from Mahinda

What’s below the program?

Operating System: Supervising program that manages the resources of a computer for the benefit of the programs that run on that computer. Kernel is the heart of the Operating System.

Compiler: A program that translates high-level language statements into assembly language statements or set of instructions.

Assembler: A program that translates a symbolic version of instructions into the binary version. Assembly Language to Machine Language.

Assembly Language: A symbolic representation of machine instructions.

Machine Language: A binary representation of machine instructions.

Firmware: A software stack that computer hardware uses for basic operations and to run applications

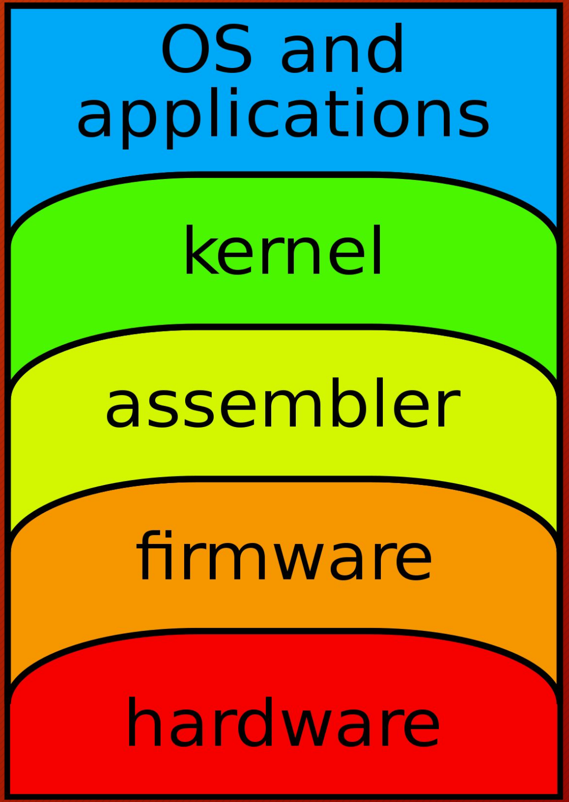


Image from Wikipedia

Technology of Building Processor and Memory

Transistors: An on/off switch controlled by an electric signal.

Very Large Scale Integration (VLSI) Circuits: A device containing hundreds of thousands to millions of transistors.

Silicon: A natural element that is a semiconductor.

Semiconductor: A substance that does not conduct electricity well. Specific areas of the silicon is transformed to conduct or insulate electricity using chemical processing.

Year	Technology	Performance/Unit Cost
1951	Vacuum Tube	1
1965	Transistor	35
1975	Integrated Circuits (IC)	900
1995	Very Large Scale Integrated (VLSI) Circuits	2,400,000
2013	Ultra Large Scale Integrated Circuits	250,000,000,000

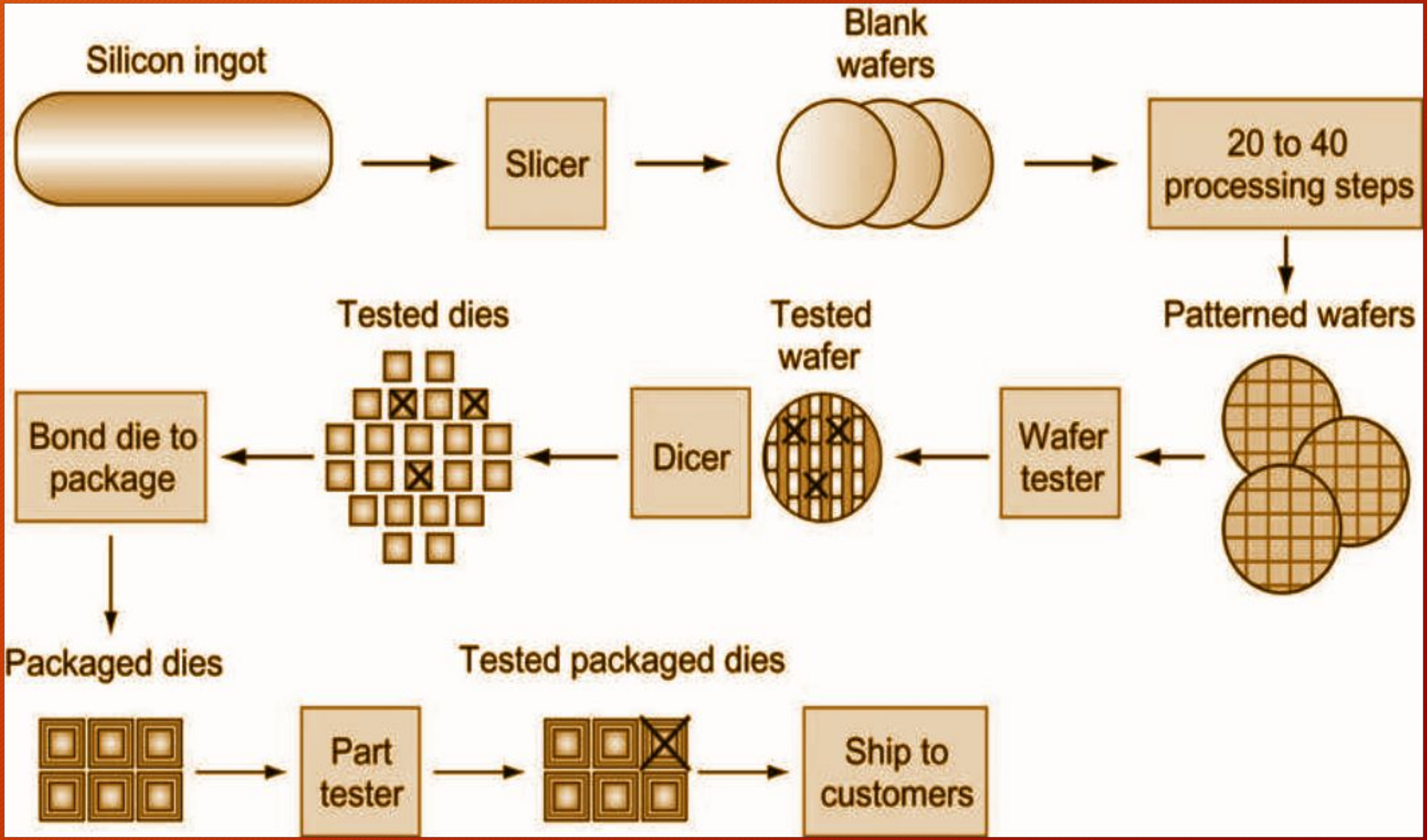
Semiconductor Manufacturing Process

Silicon Ingot: A rod composed of a silicon crystal sliced into wafers.

Defect: A microscopic flaw in a wafer or in patterning steps that can result in the failure of the die containing that defect.

Die/Chips: The individual rectangular sections that are cut from a wafer.

Yield: The percentage of good dies from the total number of dies on the wafer.



Moore’s Law

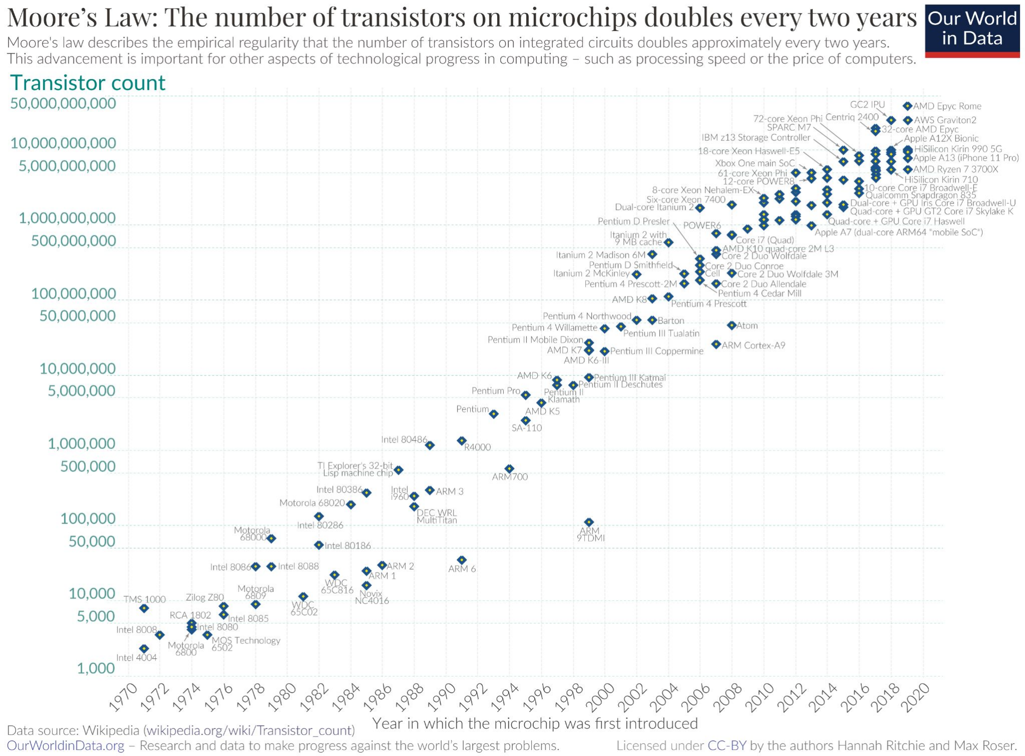
From Wikipedia

Gordon Earle Moore

Jan-3rd 1929 to Mar-24th 2023

Ph.D. California Institute of Technology

Founder of Intel Corporation



Moore’s Law: The number of transistors on microchips doubles every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

OurWorldinData.org – Research and data to make progress against the world’s largest problems.

Moore's law is *the observation that the number of transistors in an integrated circuit (IC) doubles about every two years.*

https://en.wikipedia.org/wiki/Moore%27s_law

The Performance Wall!

Response/Execution Time: The total time required for the computer to complete a task, including disk accesses, memory accesses, I/O activities, operating system overhead, CPU execution time, and so on.

Throughput/Bandwidth: The number of tasks completed per unit time.

Clock Cycle/Tick/Clock/Cycle/Processor Clock: The time for one clock which runs at a constant rate.

Performance = 1/Execution Time

$$\text{Performance}_X > \text{Performance}_Y$$

$$\Downarrow$$

$$\text{ExecutionTime}_X < \text{ExecutionTime}_Y$$

Frequency: 1/Length of Clock Cycle

The Performance Wall!

CPU Execution Time = #CPU Clocks for a Program * Clock Cycle Time

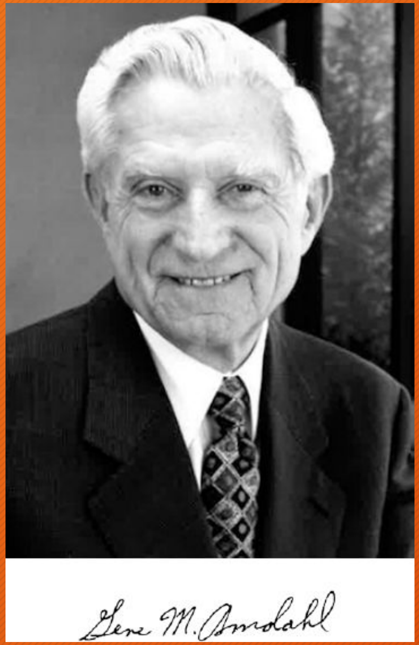
CPU Execution Time = #CPU Clocks for a Program / Clock Rate

#CPU Clocks for a Program = #Instructions * Average Clock Cycles/Instructions

CPU Execution Time = #Instructions * Average Clock Cycles/Instructions / Clock Rate

CPU Execution Time = #Instructions * Average Clock Cycles/Instructions * Clock Cycle Time

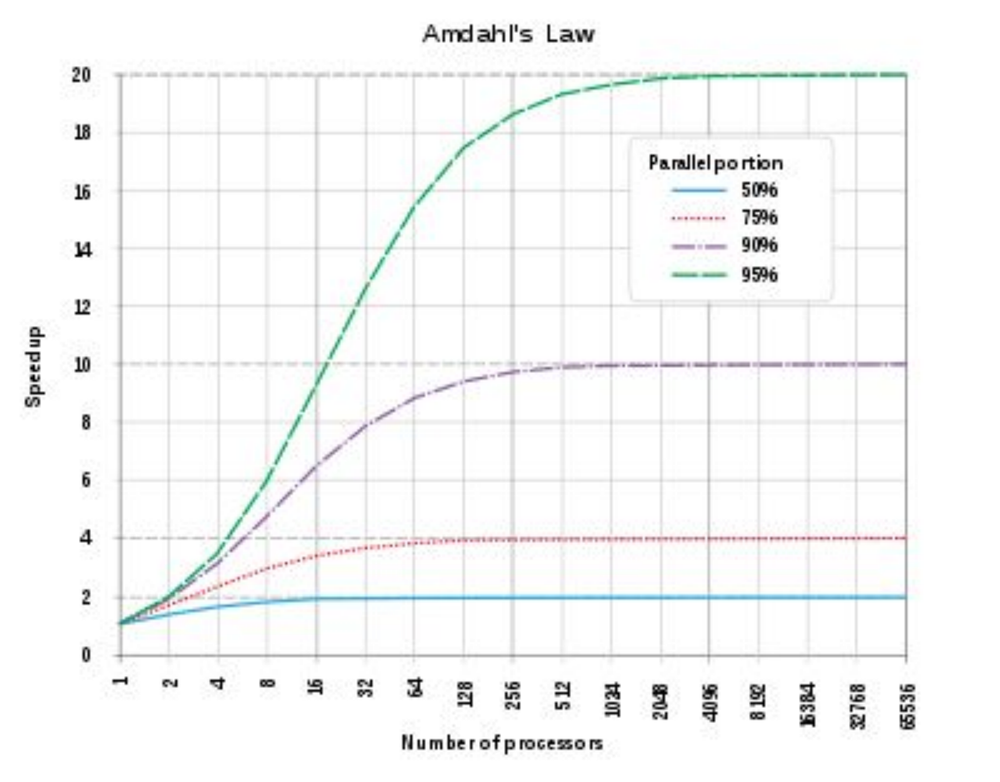
Amdahl’s Law



https://nap.nationalacademies.org/read/25543/chapter/3

Gene Myron Amdahl

November 16, 1922 – November 10, 2015
Founder of Amdahl Corporation
One of the Major Contributors of IBM 360, Main frame computers



Wikipedia

Amdahl’s law is a principle that states that the maximum potential improvement to the performance of a system is limited by the portion of the system that cannot be improved. In other words, the performance improvement of a system as a whole is limited by its bottlenecks.

Amdahl’s Law

Execution Time after improvement = Execution Time Unaffected +
Execution Time affected by improvement/Amount of Improvement

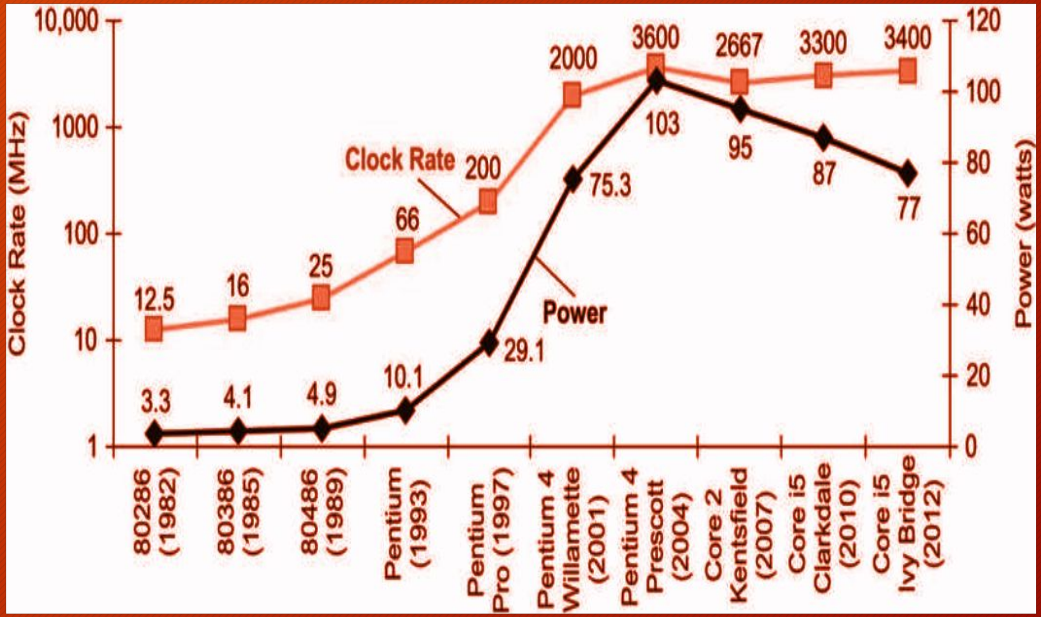
$\text{Speedup}_{\text{overall}}$ = $\text{Execution Time}_{\text{old}}/\text{Execution Time}_{\text{new}}$
= 1 / ((1 - $\text{Fraction}_{\text{enhanced}}$) +
$\text{Fraction}_{\text{enhanced}}/\text{Speedup}_{\text{enhanced}}$)

The Power Wall

Energy proportional to ½ * Capacitive Load * Voltage^2

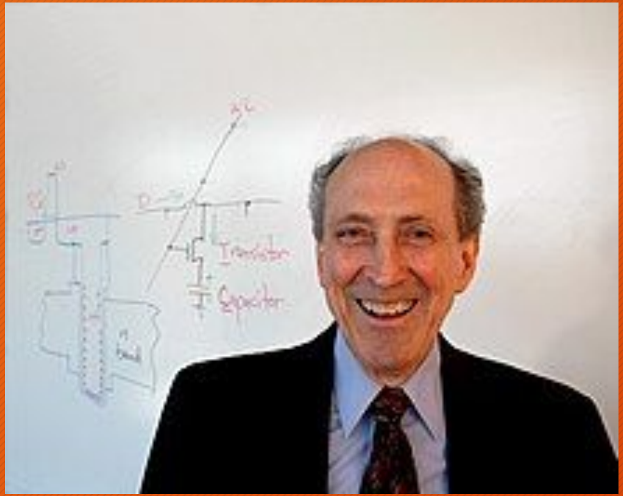
Dynamic Power proportional to ½ * Capacitive Load * Voltage^2 * Frequency

Voltage has come down from 5V to 1V in 20 Years (15% per generation)



The capacitive load per transistor is a function of both the number of transistors connected to an output (called the fanout) and the technology, which determines the capacitance of both wires and transistors.

Dennard’s Scaling

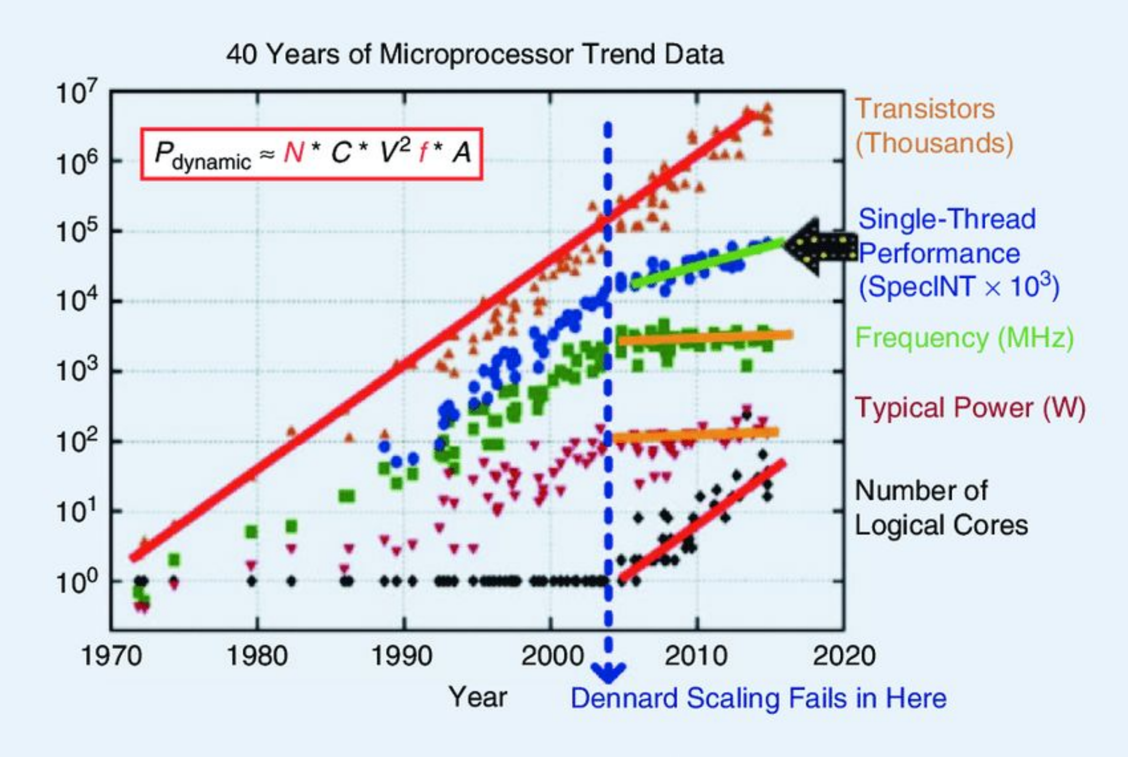


Robert Heath Dennard

Sep 5th 1932-

Inventor of DRAMs

From Wikipedia



Dennard Scaling states roughly that, as transistors get smaller, their power density stays constant, so that the power use stays in proportion with area.

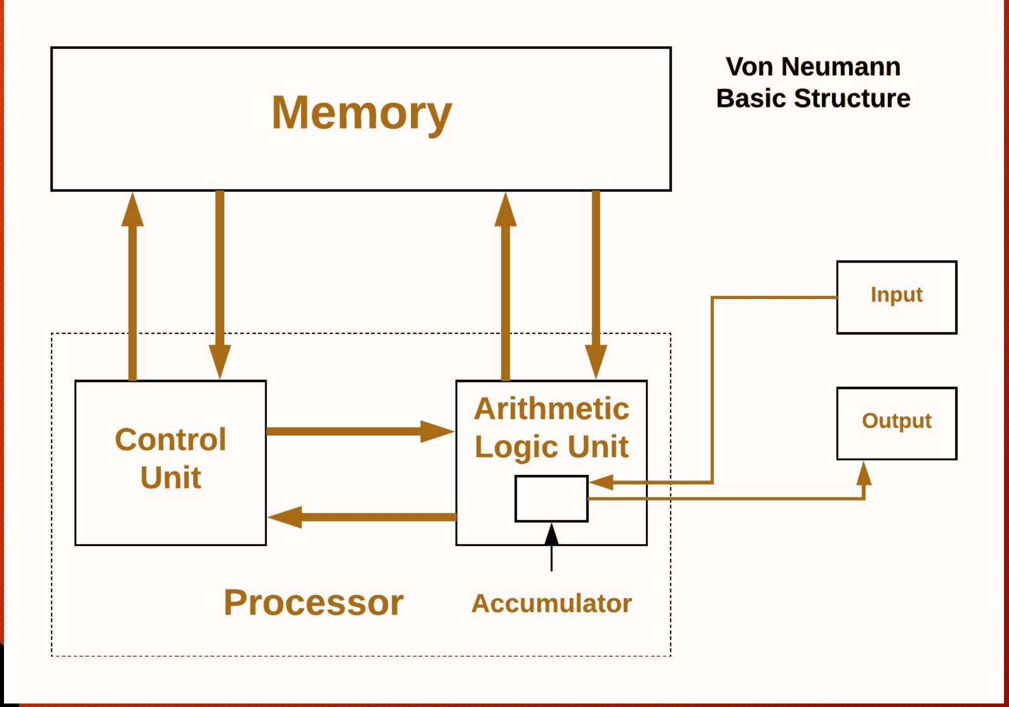
https://www.researchgate.net/figure/The-Dennard-scaling-failed-around-the-middle-of-the-2000s-24_fig1_330893452

Von Neumann Architecture

Wikipedia

John von Neumann

December 28, 1903 – February 8, 1957
Mathematician, Physicist, Computer Scientist, known for many theorems



Geeks for Geeks

Components of a Computer:

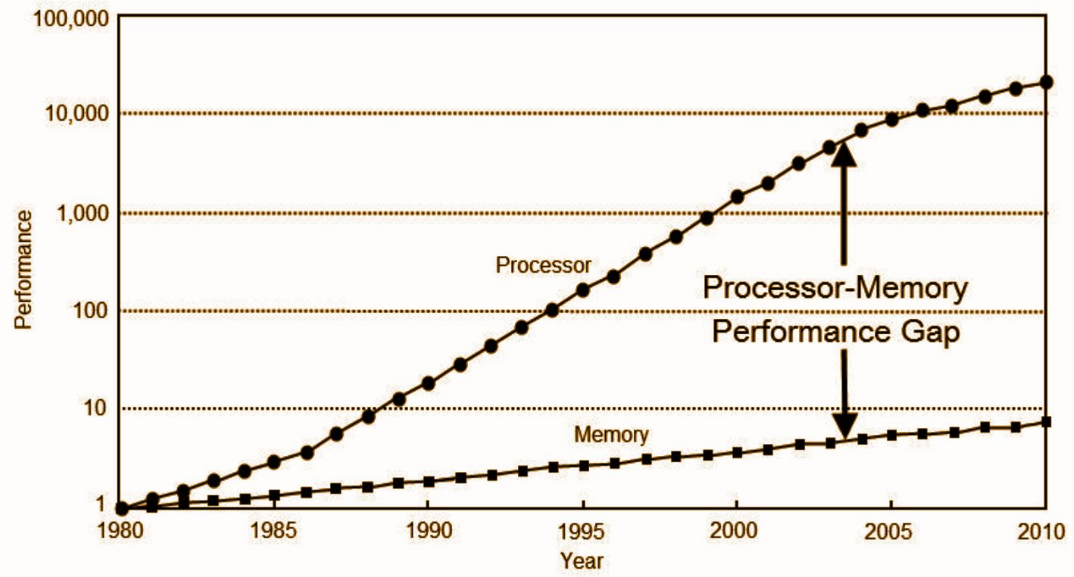
- Processor which does the computation with the help of Arithmetic and Logic Unit (ALU) and Control Unit (CU).
- Memory that stores data on which computation can happen.
- Input and Output Devices that supplies or uses data item.

Von Neumann Bottleneck

The fundamental idea of the **von Neumann Bottleneck** is most commonly described as the system slowdown due to the separation of the CPU and Main Memory.

Main memory is designed using DRAMs. The process technology used to design DRAMs are not scaling at the same rate.

The gap between the processor and memory is increasing with time. Computes are moving faster than memory.

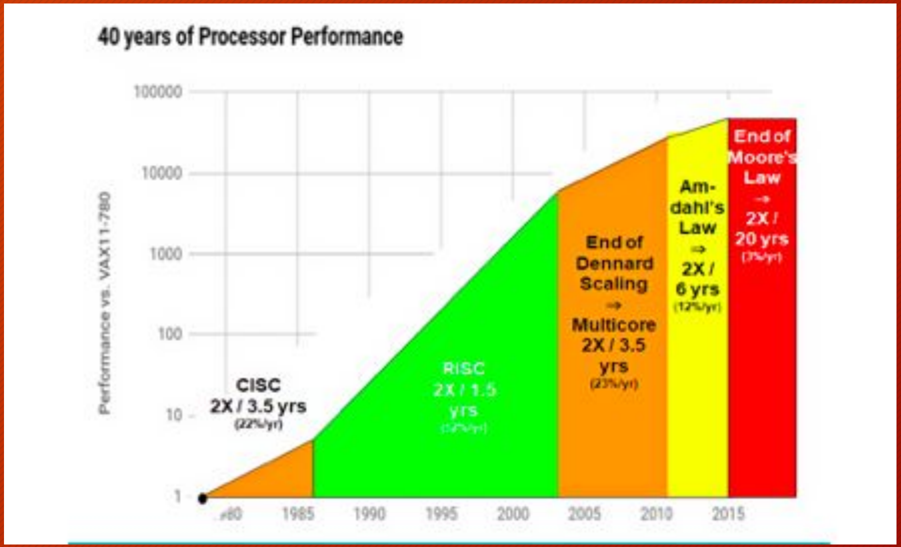


Computer Architecture: A Quantitative Approach by John L. Hennessy, David A. Patterson, Andrea C. Arpaci-Dusseau

Hitting the Limit! Golden Era of Computer Architecture Research and Development

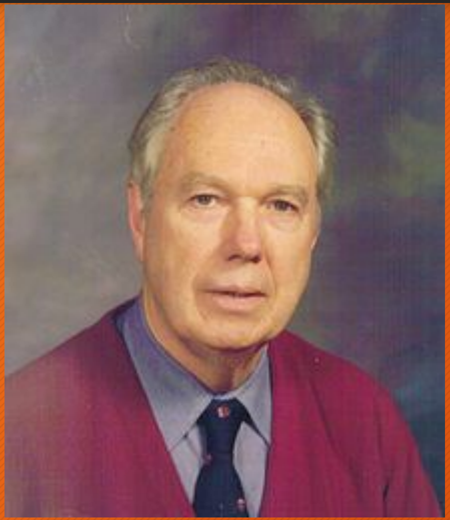
Dying era of Moore’s Law, Amdahl’s Law, and Dennard’s Scaling

- Semiconductor Manufacturers are hitting 2 nm process technology limits! -> Moore’s Law
- Hitting Frequency Limits! What about Intel Itanium Processor? -> Dennard’s Scaling
- Hitting limits on Instruction-level Parallelism (ILP) -> Amdahl’s Law
- Memory is still an Issue!
- What is the catastrophic effect of AI/Graphics application?



https://www.rcrwireless.com/20191001/opinion/as-moores-law-ends-hardware-acceleration-takes-center-stage-part-1-reality-check

Flynn’s Classification of Computers

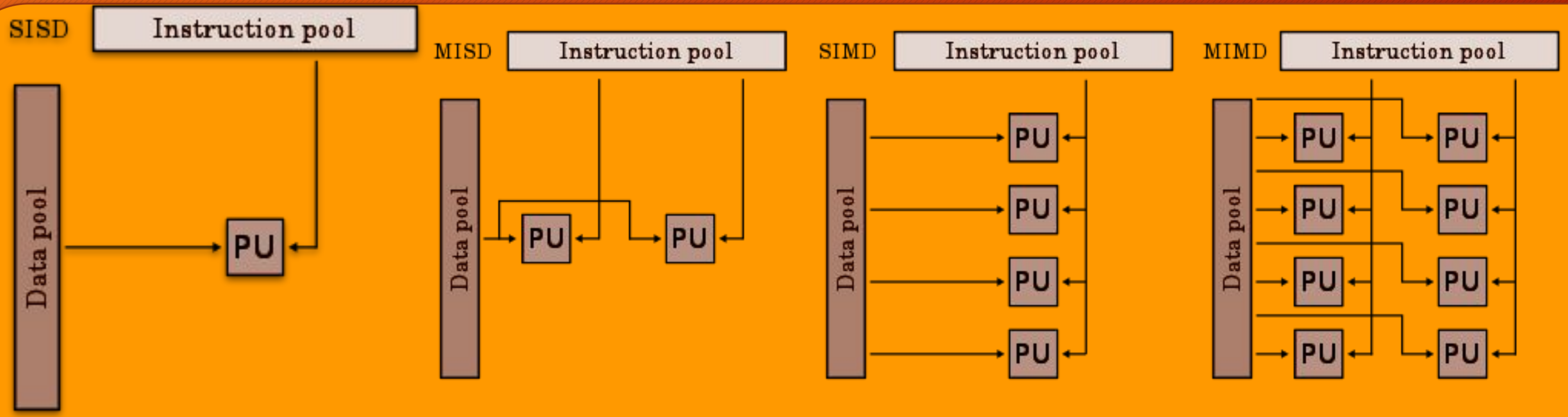


Wikipedia

Michael J. Flynn

May 20, 1934 -

- Single Instruction Single Data (Von-Neuman-Single Core)
- Single Instruction Multiple Data (GPU and Vector Processor)
- Multiple Instruction on Single Data (Systolic Arrays and TPUs)
- Multiple Instruction Multiple Data (Multicore CPUs)



Multithreading and Multicore

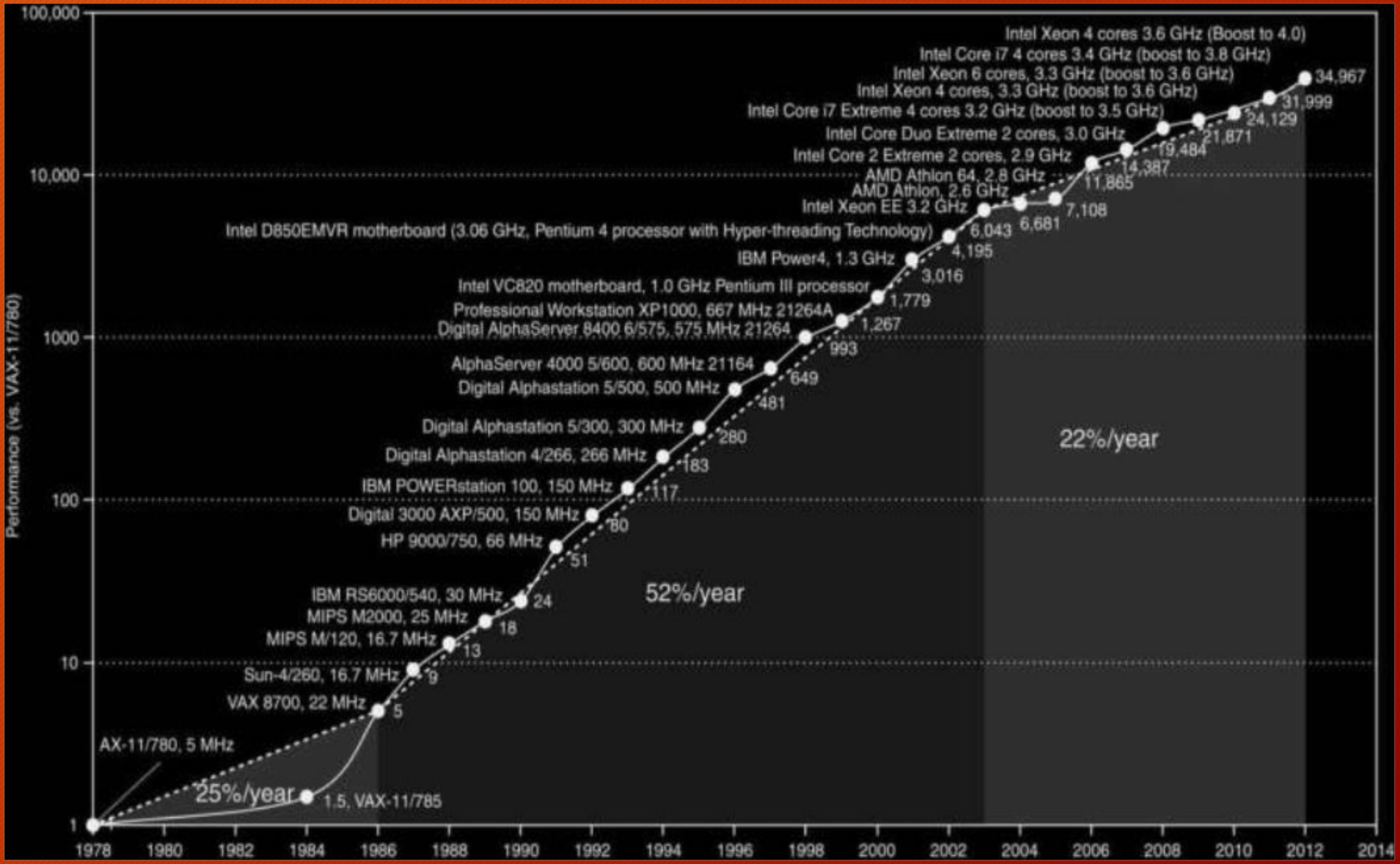
Thread: Independent Flow of Control.

Multithreading: Executing multiple threads on one core.

Software Multithreading: Supported by Operating System.

Hardware Multithreading: Supported by Hardware.

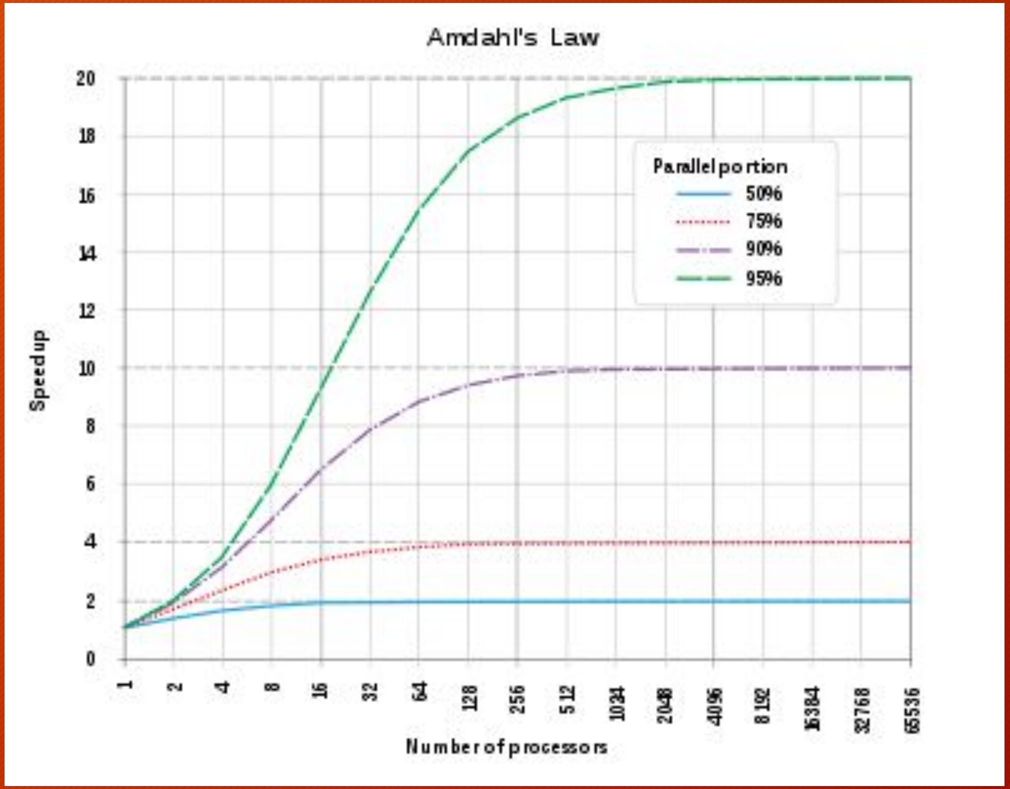
Multicore: Enables running threads on two or more independent cores.



Multithreading and Multicore

What about Amdahl’s Law for Multi-core and Multi-thread?

What are Shared Resources?



Workloads and Benchmarks

Workload: A set of programs run on a computer that is either the actual collection of applications run by a user or constructed from real programs to approximate such a mix.

Benchmark: A program selected for use in comparing computer performance

Benchmark Suites: SPEC CPU 1992, SPEC CPU 2000, SPEC CPU 2006, SPEC CPU 2017, PARSEC, SPLASH, Rodinia, Parboil, SPEC JBB, SPEC Web, Cassandra, MLPerf, Custom Benchmarks

“

No matter how many times computing power doubles, the time needed to train deep models never decreases.

”

Let’s have fun!