High Performance Computing

I-Introduction

Mahendra Verma
IIT Kanpur

Why Computation?

Science problems are complex.

Presently available analytical tools don't work.

Solved problems:

H atom, simple harmonic oscillator

So, for complex problems

Either perturbative methods

Or numerical methods

Computer simulations: Numerical experiments

Just like experiments

We make models based on numerical results.

Often complements experiments

Excellent visualisation

COMPUTATION is the third pillar of science along side
THEORY & EXPERIMENT

Major HPC Applications

Flows

- Weather and climate forecast
- Flows around automobiles, aeroplanes, space vehicles, and rockets
- Oil exploration
- Physics of turbulence
- Flows in and around stars, blackholes, galaxies, and planets
- Quantum turbulence

Materials and Quantum Systems

- Simulations of quantum matter
- Quantum Monte Carlo
- Density functional theory (DFT)
- Particle simulation

Nonlinear Physics, Health

- Epidemic evolution
- Understanding brain
- Network
- Human body
- Plate tectonics and earthquakes

Complex Physics

- Study dark matter & dark energy
- Evolution of the universe, galaxy, stars
- Fusion research and Tokomak
- Lattice quantum chromodynamics
- Accelerator simulations and data analysis

Complex Engineering

- Structures: Buildings, water network
- Chip design
- Complex machines: Rocket, Engines, Airplanes, Cars, ..
- Transport, Uber and Ola

Machine learning, Defence, Economics

- Machine learning
- Defence
- Economics: Stock market, banking, etc.

Example of a complex calculation

Weather forecasting

Estimates

- Typical laptop/desktop has 4 to 8 Gigabytes of RAM.
- To save a matrix A(N,N) with $N=10^3$, we need
- $8x10^6$ bytes for double, and $4x10^6$ bytes for float
- We deal with 4096^3 array. Memory requirement is $8x64x10^9$ = (1/4) Terabytes.

$10^5 \times 10^5$ $10^{10} \times 8$

$1024 \neq 10^3$

Space complexity

wikipedia

- Make a grid of Earth's atmosphere.
- Grids for the best simulations: 3 km x3 km ⟹ 12000x12000 horizontal g...
- Vertical direction: 1000 spheres
- Total grid points: 144×10^9.
- With 7 variables, memory required = $8 \times 7 \times 144 \times 10^9$ = 8.064 TB (single copy)

Time complexity

- Let us take 1 minute time resolution.
- For 3 days forecast, no of steps = 3*24*60 = 4320
- Grid points = 144×10^9
- Assuming 10000 FPs at each grid point, no of FP operations required
  - $144 \times 10^9 \times 4320 \times 10^4 \approx 6.2 \times 10^{18}$.
- Petaflop machine with 100% computational efficiency will require 6200 sec ≈100 minutes
- With 10% efficiency: 1000 minutes

Increase in Processor power

1972: Intel 4004 had 2300 transistors and performs 60K ops/sec.

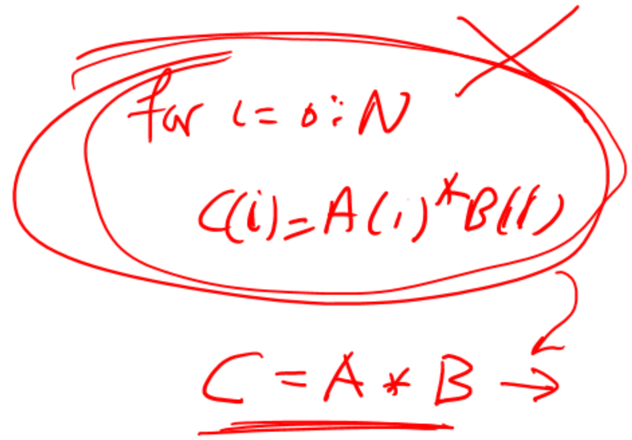
2023: EPYC 9754 (Genoa) has 82 billion transistors and can perform ~ 4×10^{12} ops/sec.

2023: GPU cards are even faster.

Course contents

- Hardware & HPC systems (processors, memory, interconnect, shared & distributed memory)
- Software Aspects: Languages & Oop
- Simple tricks to speed up codes (e.g., loop optimisation, vectorisation).
- MPI programming: Python & C++
- Multiprocessing and multithreading: Python
- OpenMP
- GPU Programming:
  - GPU Hardware
  - CUDA, Cupy, Cu-numeric
  - OpenACC
- Handling large data
- Visualization

Q Compute



References

- P. S. Pacheco, An Introduction to Parallel Programming, Elsevier (2011)
- M. Quinn, Parallel Programming in C and OpenMP, McCraw Hill Education (India) (2003)
- A. Grama, A. Gupta, G. Karypis, and V. Kumar, Introduction to Parallel Computing, Pearson (2007)
- G. Zaccone. Python Parallel Programming Cookbook, Packt Publ. (2015)
- https://cupy.dev (for CuPy)
- R. Farber, Parallel Programming with OpenACC, Morgan Kaufmann (2016)

Bird’s-eye view of Computer Systems

Mahendra K. Verma

Ref: Hennessy & Patterson: Computer Architecture
& various websites

1. Basic design

- A driver of a luxury car does not need to know the intricacies of a car.
- But, Formula One car driver needs to know quite a lot about the car.
- For web-browsing, writing small programs, a user does not need know much about a computer.
- But, for HPC, a user would need to know about fast-changing computer hardware/software.

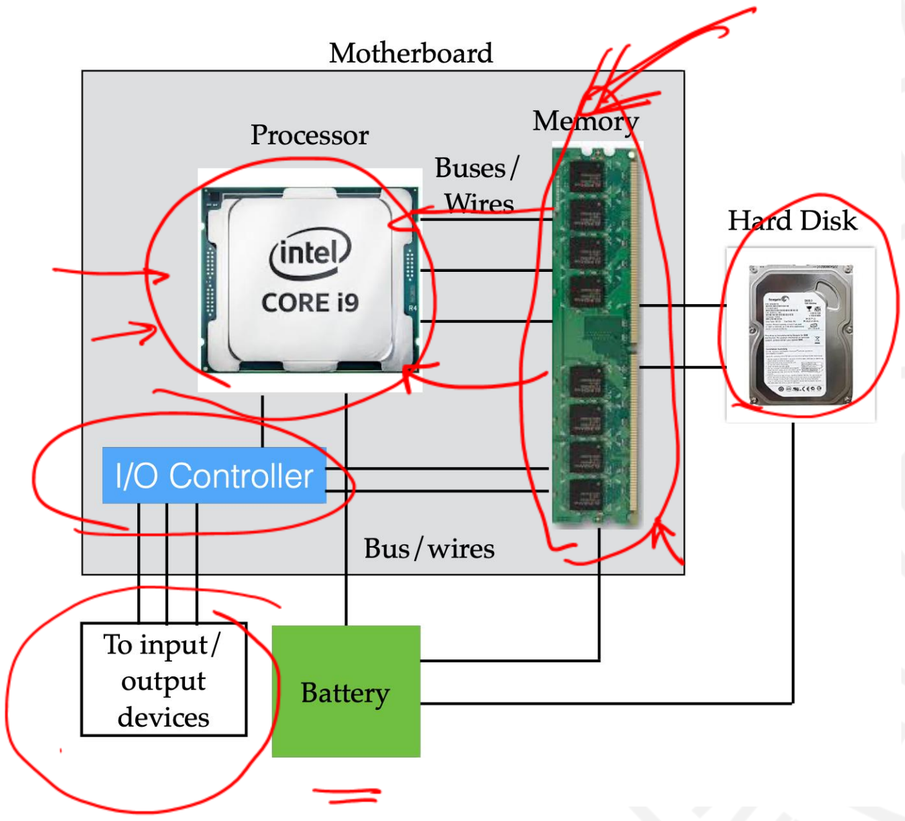
Brief history

The Idea Factory

- First computer was mechanical, made by Babbage. Later with valve -tubes.
- Then came electronics, ICs, sophisticated fabs.
- Contributors: Intel, Motorola, IBM, Apple, Microsoft, Cray.
- USA dominated the computer revolution, from large scale machines to personal computers

https://www.youtube.com/watch?v=-M6lANfzFsM&list=PLzzEvSGvmTX8Pnr5pDMUbPH3SSYMFYV1w

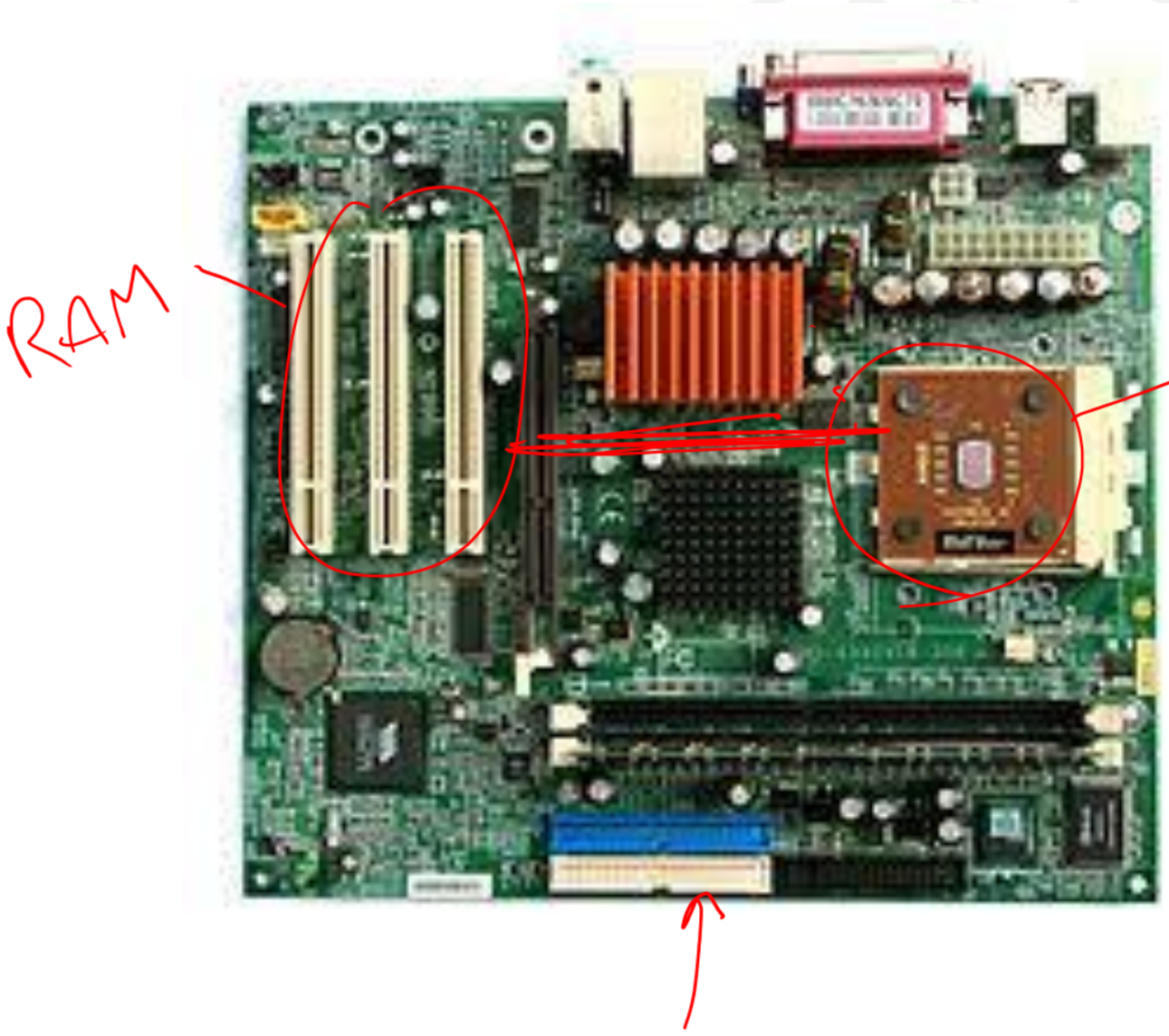
von Neumann architecture



- Computational unit: Processor
- Memory
- Input/Output
- A processor has several compute units called cores.

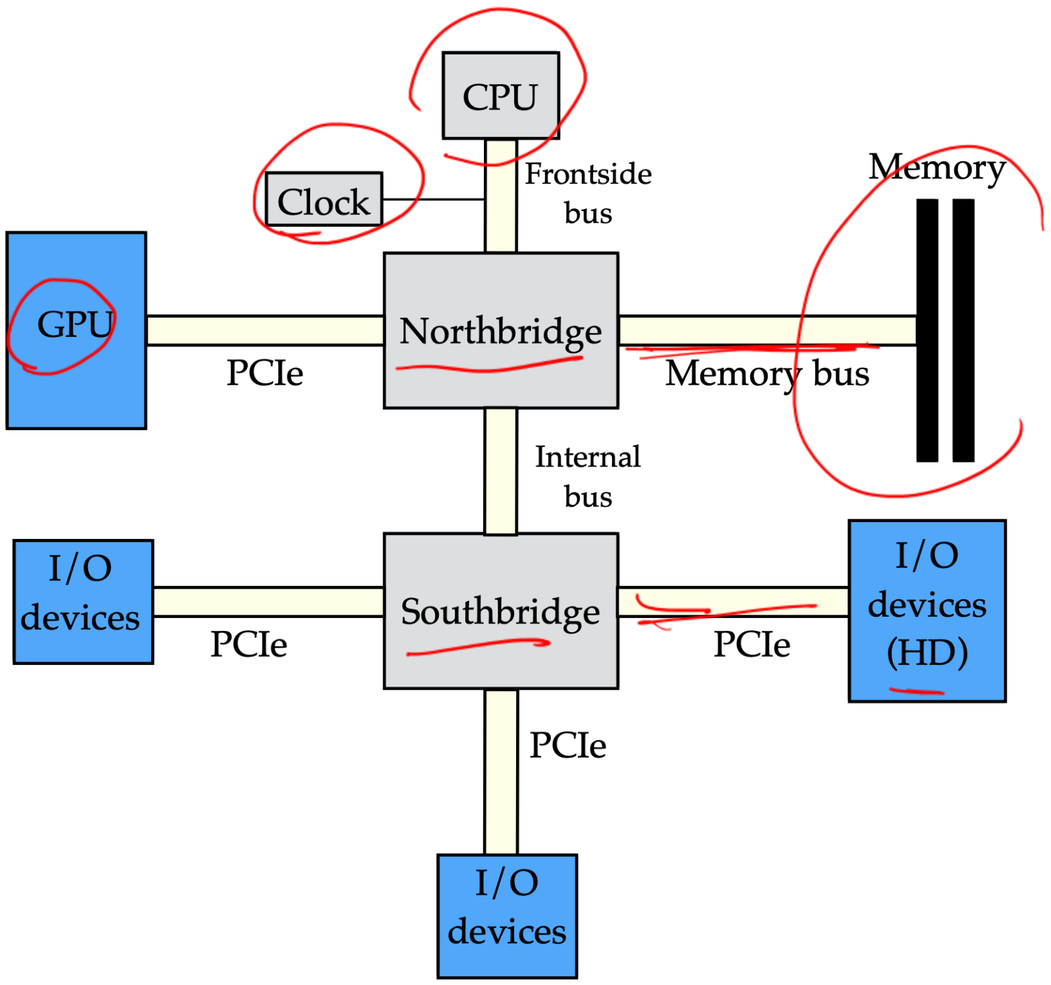
Need to speed up each component

Motherboard



wikipedia

Inside Motherboard



- CPU connected to Northbridge and Southbridge
- Northbridge connects to memory and GPU
- Southbridge connects to I/O devices
- Clock: 1-4 GHz
- PCIe 4.0: Peripheral component interface express (8000 MBps)
- PCIe 5.0: 14000 MBps
- Memory bus

Some terminologies

- Memory:
  - Bit: single unit of information: 0 or 1
  - Byte: 8 bits
  - KiB = $2^{10} \approx 10^3$
  - MiB = $(1024)^2 \approx 10^9$
  - GiB = $(1024)^3 \approx 10^{12}$
  - TiB (Tera) = $(1024)^4 \approx 10^{12}$
  - PiB (Peta) = $(1024)^5 \approx 10^{15}$
  - EiB = $(1024)^5 \approx 10^{15}$
- Similar notation for clock speeds and floating-point operations.
- 64-bit processors (integers and float are by default 64 bits)
- Clock frequency: 1 to 4 GHz

Table 1: Ranges and precision for various floating point representations.

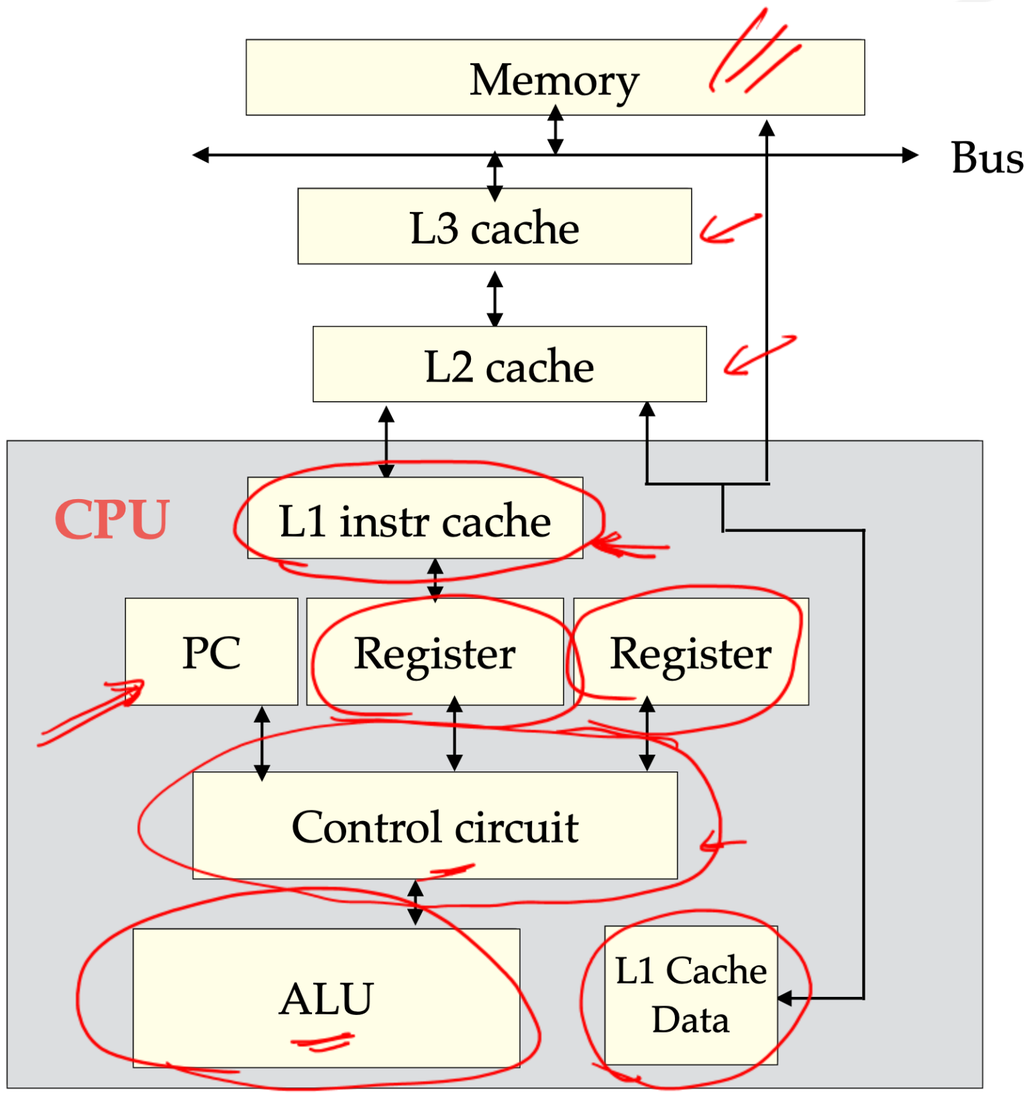
	Range	Precision
float64 (double)	-1.7×10^{308} to $+1.7\times10^{308}$	16 digits
float32 (float)	-3.4×10^{38} to $+3.4\times10^{38}$	7 digits
float8	-6.55×10^{4} to $+6.55\times10^{4}$	3 digits
float16	-6.5×10^{4} to $+6.5\times10^{4}$	2 digits

2. Processor

Functions of a processor

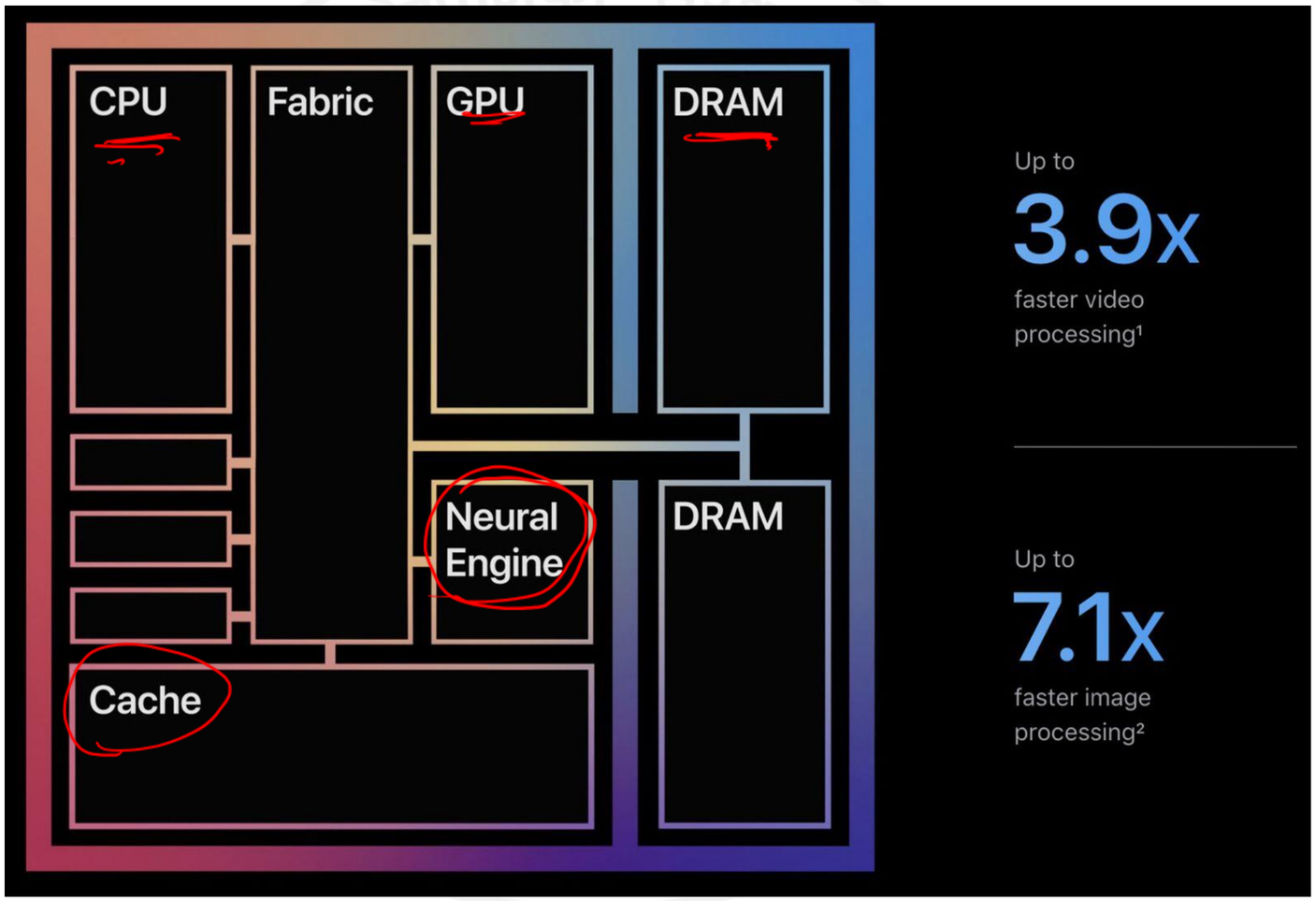
- Performs arithmetic and logical operations
- Performs I/O ops
- Interpret user commands

Inside CPU



- Arithmetic & Logic unit (ALU)
- Registers
- Program counter (PC)
- Control circuit
- L1 cache
- Modern procs: putting more memory components inside

M1 chip



https://www.macrumors.com/guide/m1/

Power Consumptions in CPU

$$P = cV^2 f$$

V increases with linearly f

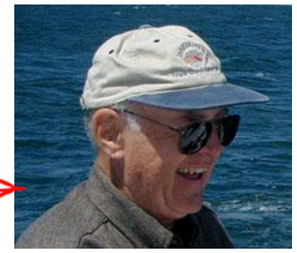
Hence, $P = cf^3$

V ≈ 1 Volt
f ≈ 1-4 GHz

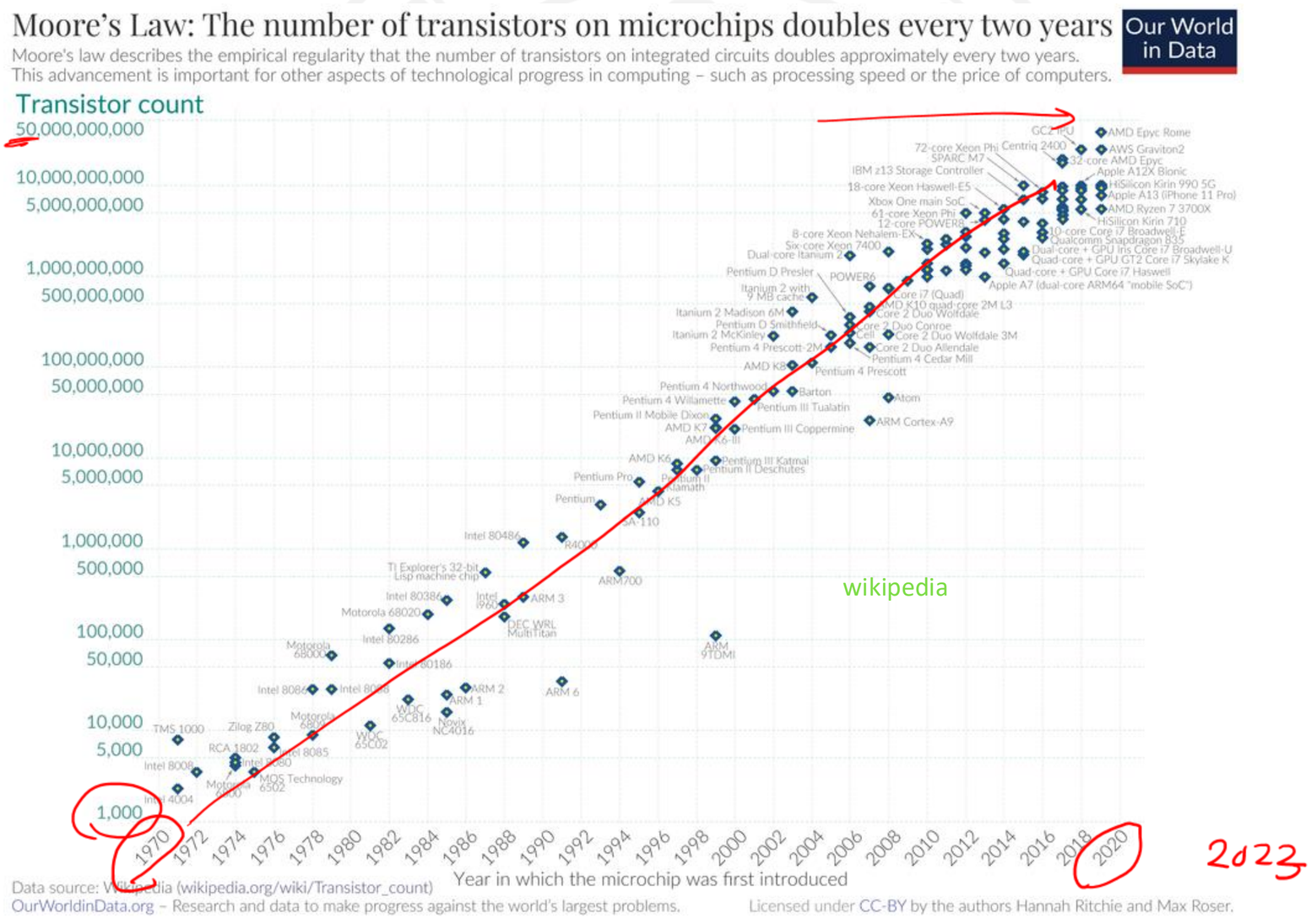
Keep lower V and f

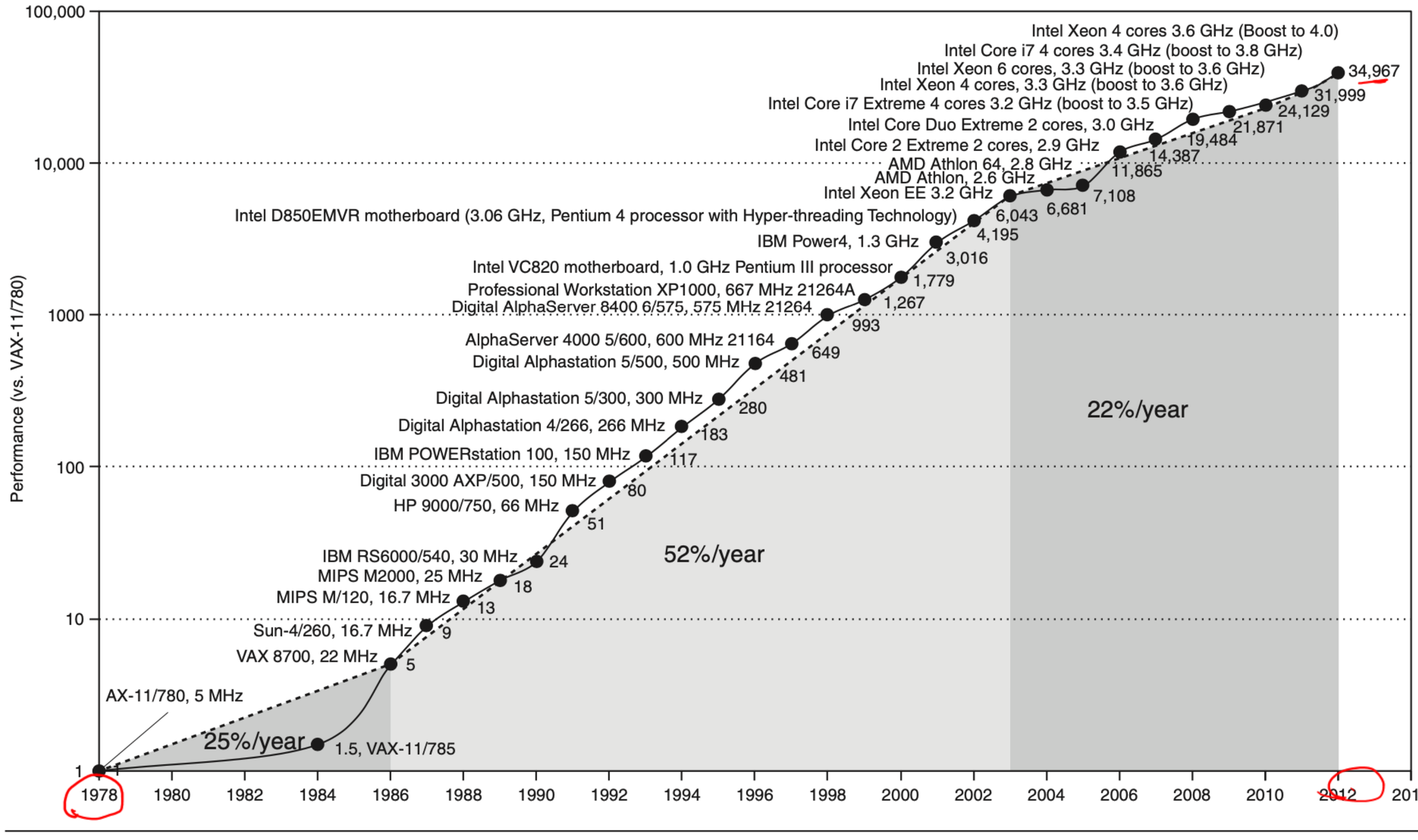
Moore’s law (1965)

The number of transistors in a IC doubles about every two years.



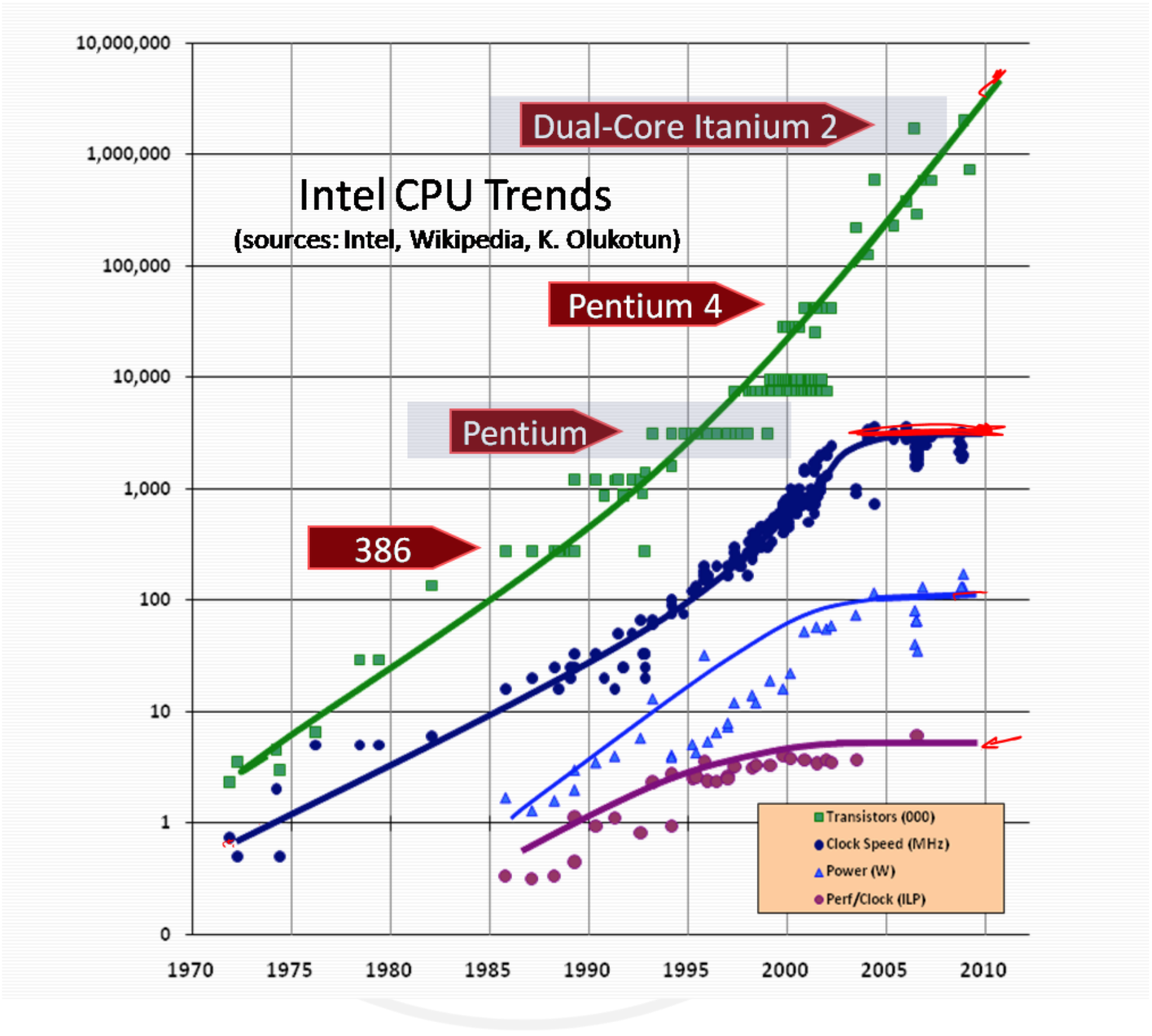
wikipedia





Patterson & Hennessy, Computer Organization and Design

ILP
Instruction-level
Parallelism



Source:!http://www.gotw.ca/publications/concurrency-ddj.htm!

- We can’t pack more when we reach atomic or quantum scale.
- We are close to saturation (?)
- 2-nanometer tech: 50 billion transistors each of size of ~5 atoms. Size ≈ fingernail.
- 1-nanometer next..
- Quantum computer, Graphene, Spintronics…

CISC vs RISC

- Complex instruction set computer (CISC)
- Reduced instruction set computer (RISC)

CISC

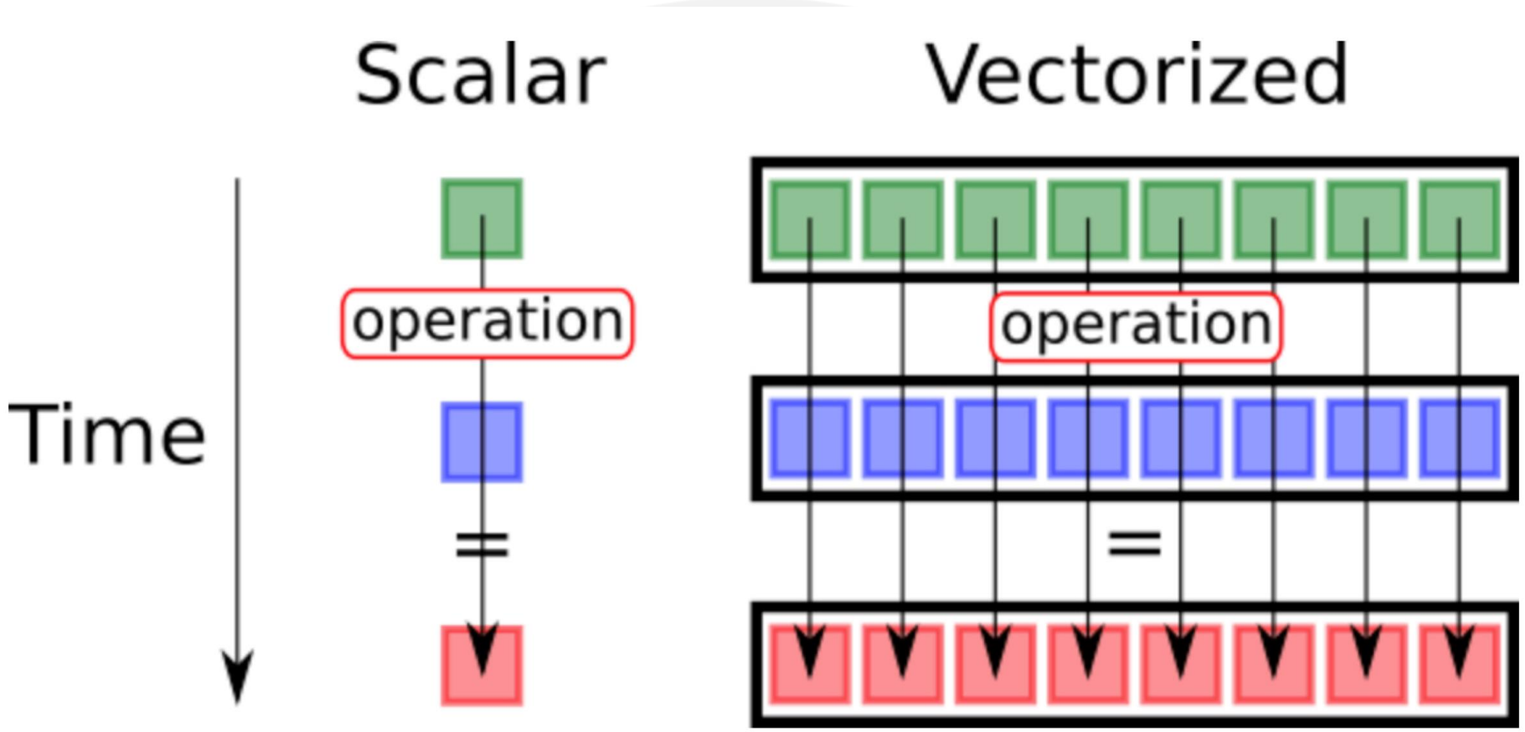
- A single instruction can execute several low-level operations,
- Loading the data from the memory to the register, arithmetic operations on the data, and the send back the result to the memory
- Smaller size of assembly code
- Less memory
- Intel processors employ CISC

RISC

- Instructions are simple, and they can be executed in a single clock cycle.
- Hardware is simple, and a code block can be executed in a single cycle.
- Requires larger memory.
- ARM processor; M1/M2 Chips of Apple
- “Simple is beautiful”

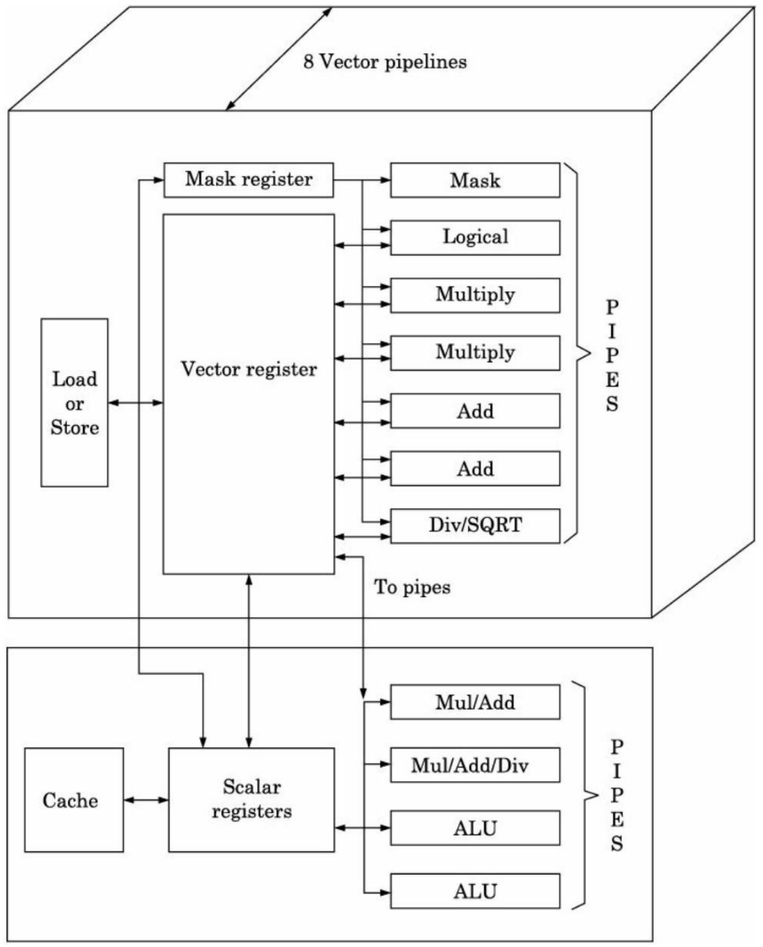
Vectorization

Cray, …



https://lappweb.in2p3.fr/~paubert/ASTERICS_HPC/6-6-1-985.html

Vector operations



Combination of add & multiplication

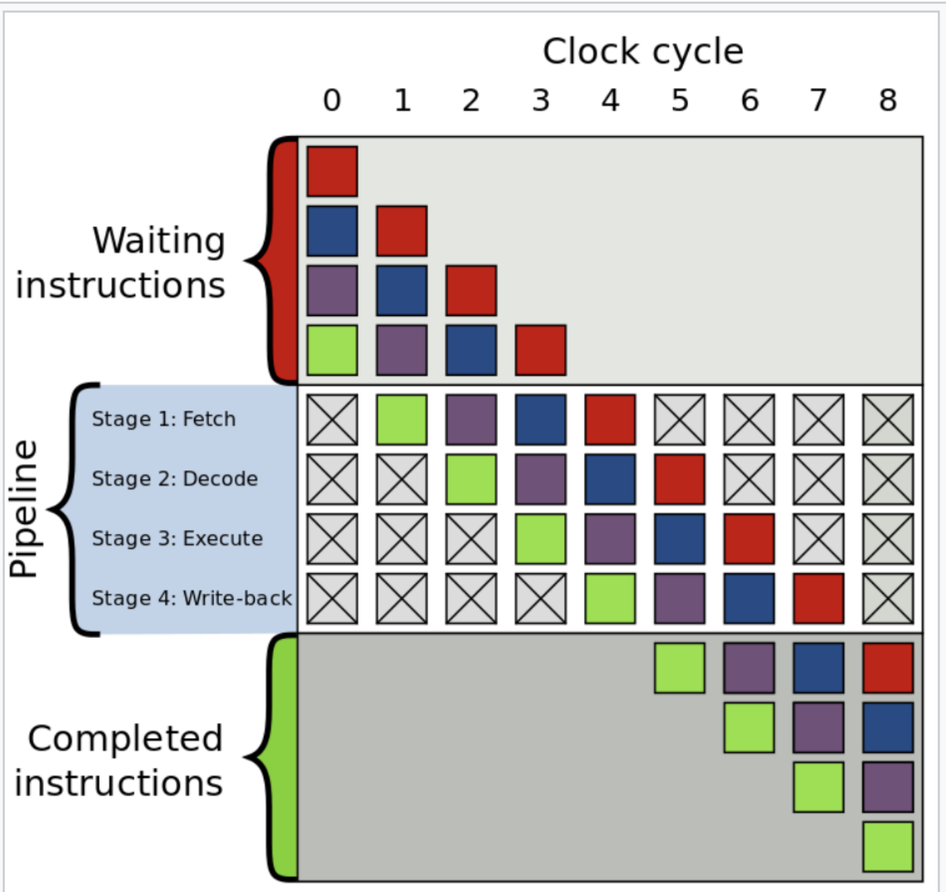
Rajaraman & Murthy, Parallel Computers

Pipelining

IBM, Cray, …

The Ford assembly line in 1913. Wikimedia Commons/public domain

Clock	Execution
0	• Four instructions are waiting to be executed
1	• The green instruction is fetched from memory
2	• The green instruction is decoded • The purple instruction is fetched from memory
3	• The green instruction is executed (actual operation is performed) • The purple instruction is decoded • The blue instruction is fetched
4	• The green instruction's results are written back to the register file or memory • The purple instruction is executed • The blue instruction is decoded • The red instruction is fetched



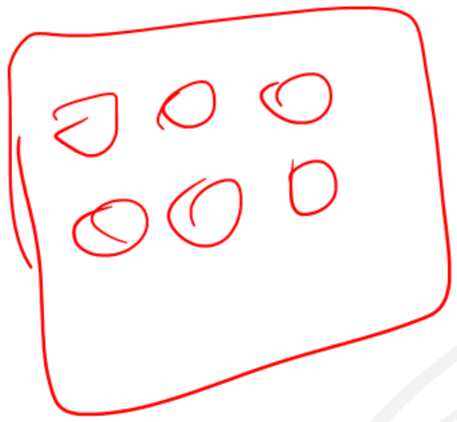
Generic 4-stage pipeline; the colored boxes represent instructions independent of each other

Wikipedia

A*B+C using pipelining

Clock Cycle	Segment 1 R1, R2	Segment 2 R3, R4	Segment 3 R5
1	A1, B1		
2	A2, B2	A1*B1, C1	
3	A3, B3	A2*B2, C2	A1*B1+C1
4	A4, B4	A3*B3, C3	A2*B2+C2
5	A5, B5	A4*B4, C4	A3*B3+C3

3. Multicore processors



Server processor

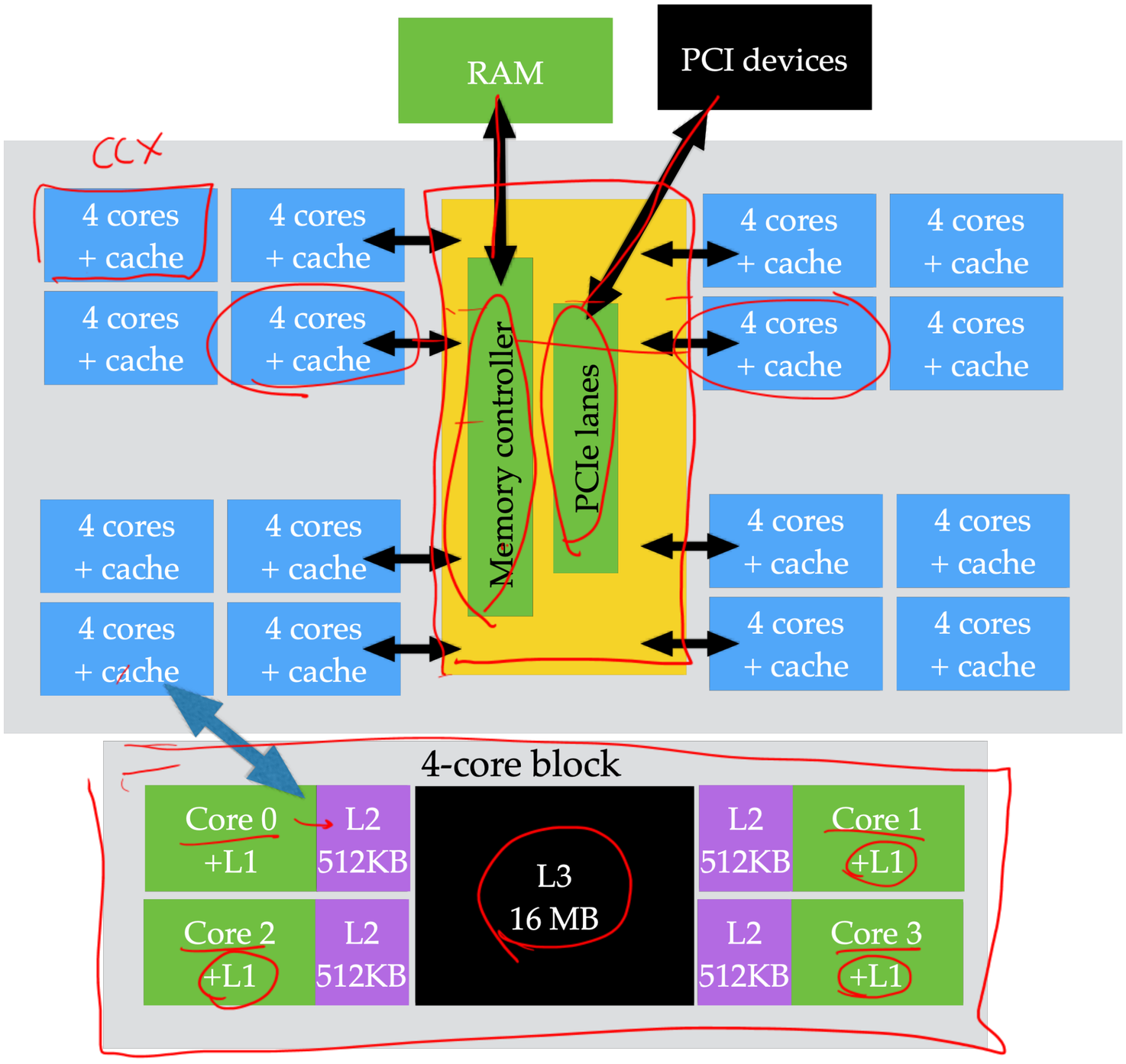
AMD EPYC 9754 (Bergamo)

128 Cores

AMD EPYC 7742 (Rome)

64 Cores

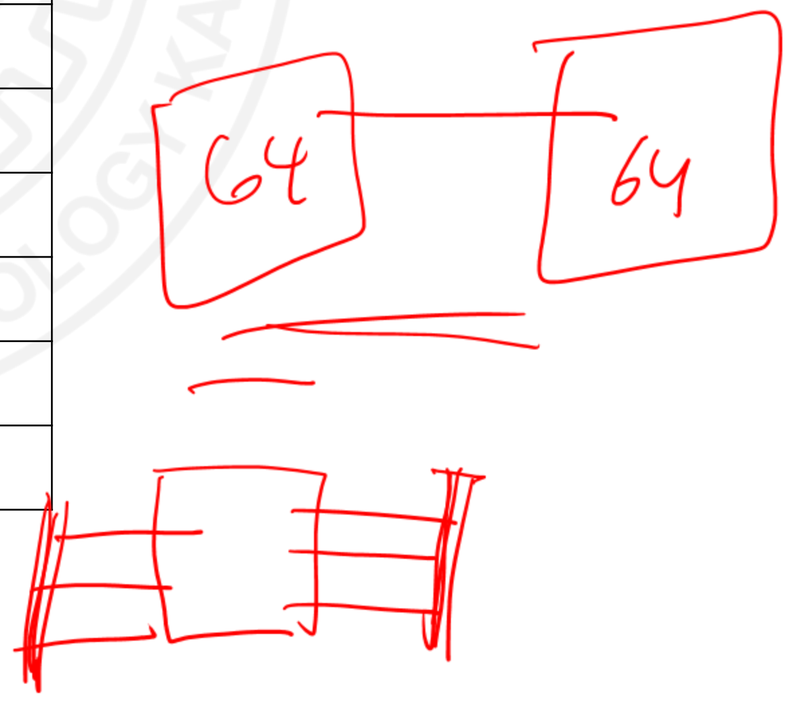
https://www.youtube.com/watch?v=BxLBLEeq6yg


RAM
PCI devices
CCX
4 cores
+ cache
Memory controller
PCIe lanes
4-core block
Core 0
+L1
L2
512KB
L3
16 MB
Core 1
Core 2
Core 3

	Epic 7742 Rome	Epic 9754 Bergamo
Lithography	7 nm	5 nm
Transistors	32 billions	71 billions
No of CPU cores	64	128
No of threads	128	256
Clock speed	2.3 GHz (boost 3.4 GHz)	2.25 GHz (boost 3.1 GHz)
L1 Cache	4 MiB (64x32KiB)	16 MiB (128x64KiB)
L2 Cache	32 MiB (64x512KiB)	128 MiB (128x1MiB)
L3 Cache	256 MiB (16x16 MiB)	256 MiB (16x16 MiB)
PCI Express version	PCIe 4.0x128	PCIe 5.0x128
System memory type	DDR4-3200	DDR5
Memory channel	8	12
Per socket mem BW	204.8 GB/s	460.8 GB/s
Power	225 W	320 W

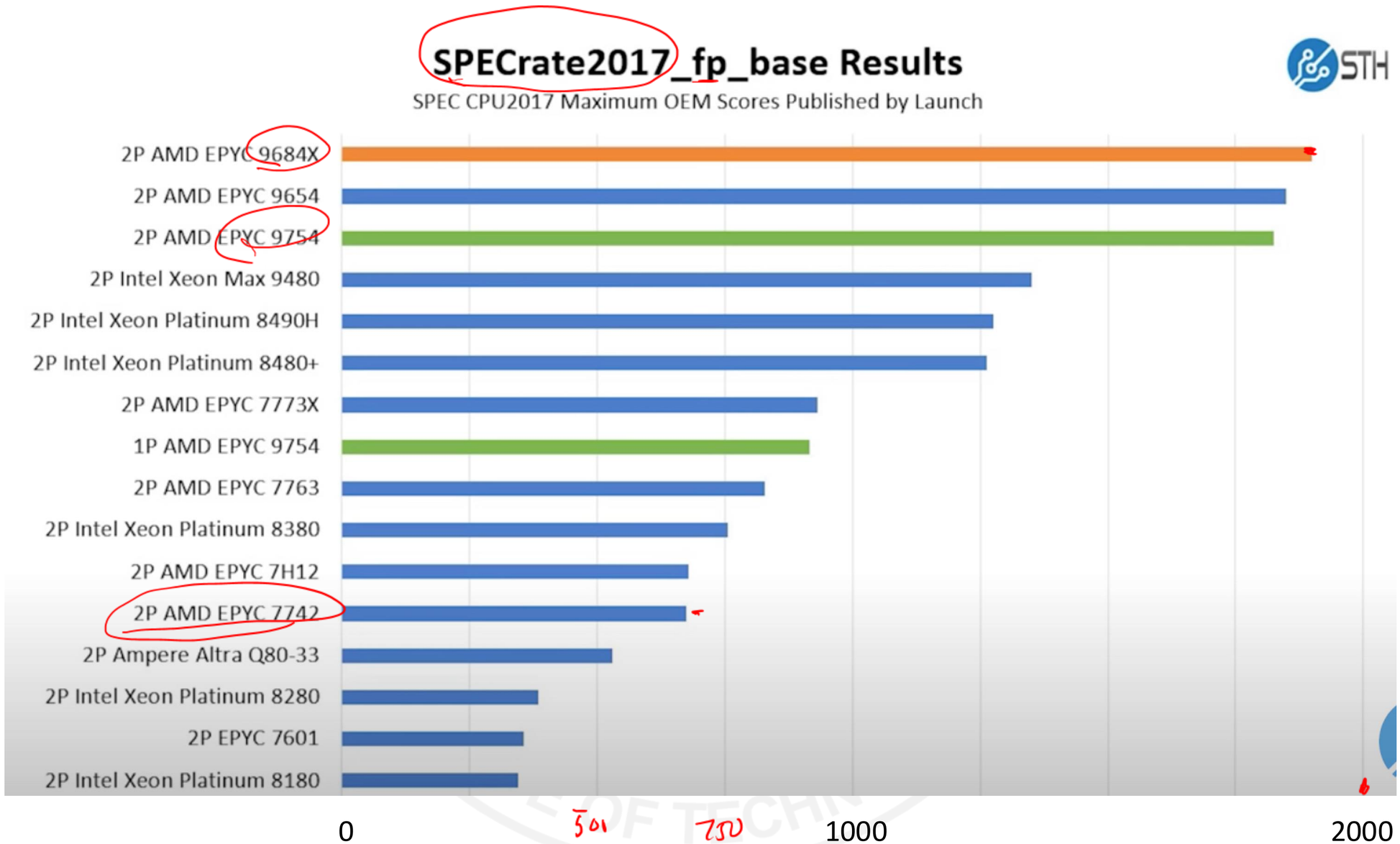
AMD EPYC 7742 (Rome)

AMD website



Computing power of Rome in FLOPS

- 16 double-precision FLOPS/core/cycle
- Maximum FLOPS = 64x16x2.5 ≈ 2.5 Teraflops
- Some tests yield approximately 2 Teraflops.
- Bergamo's peak speed is 5.376 TFs.

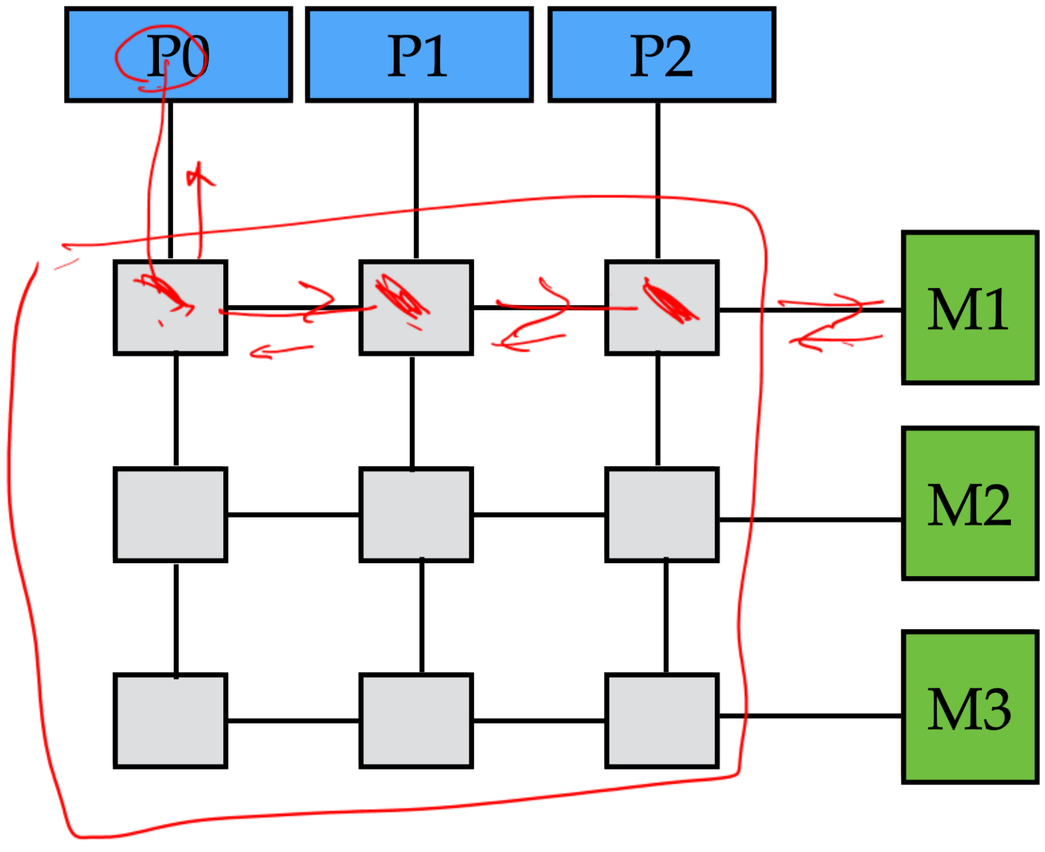


https://www.youtube.com/watch?v=BxLBLEeq6yg

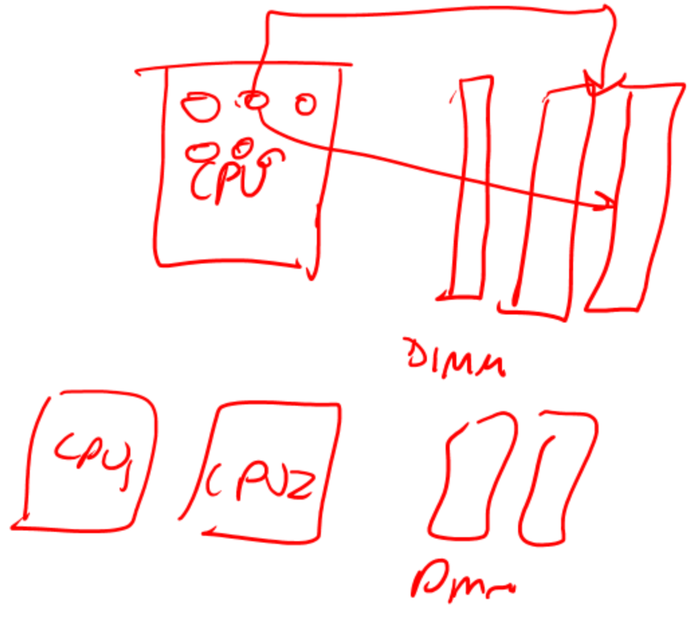
Intel Xeon Platinum (8352Y)

Lithography	10 nm
Transistors	16 billions
Cores	32
Clock speed	2.20-3.40 GHz
Cache	48 MB
Max memory	6 TB
System memory type	128-bit LPDDR4X
Mem BW	68.25 GB/s
Power	205 W

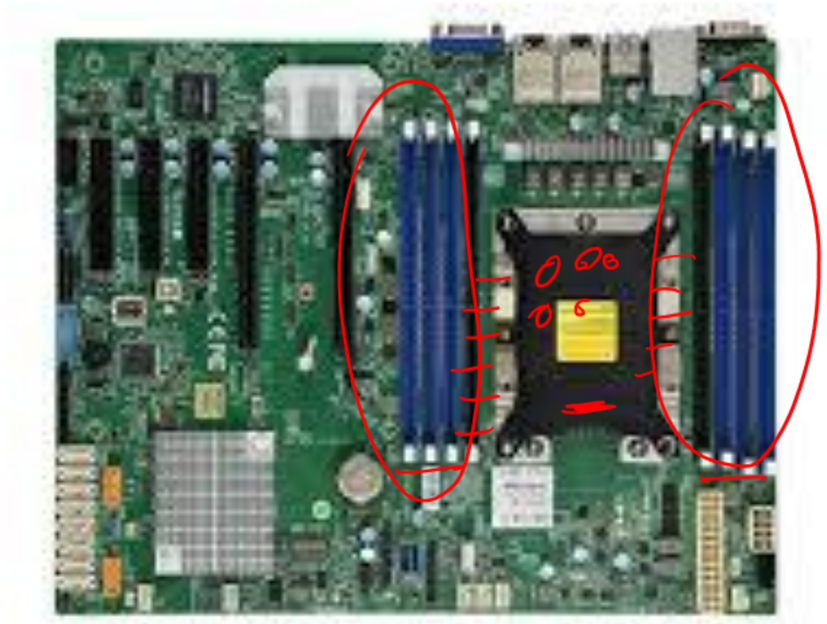
Connecting procs & mem

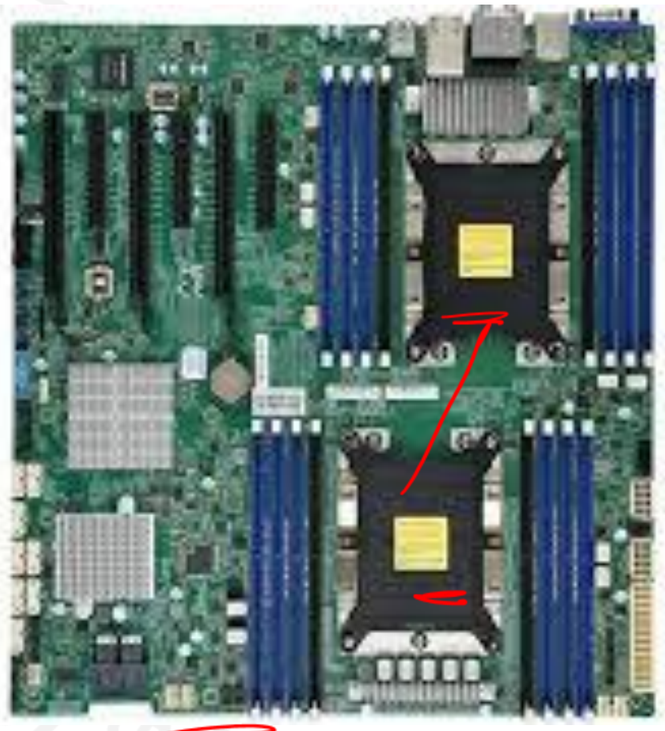


Crossbar switch



Supermicro motherboards



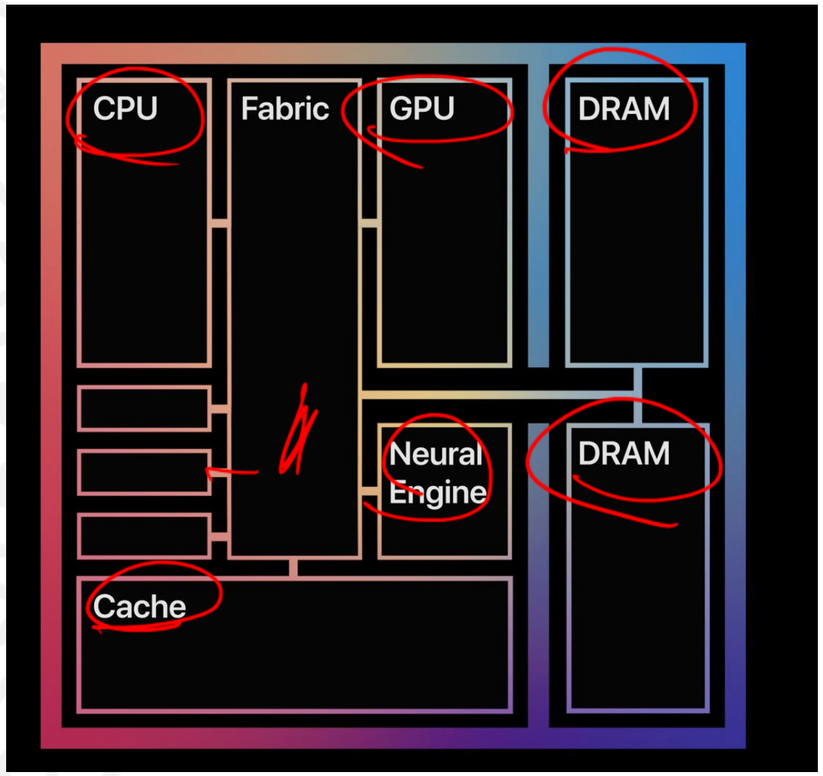


https://www.supermicro.com/en/products/motherboards/

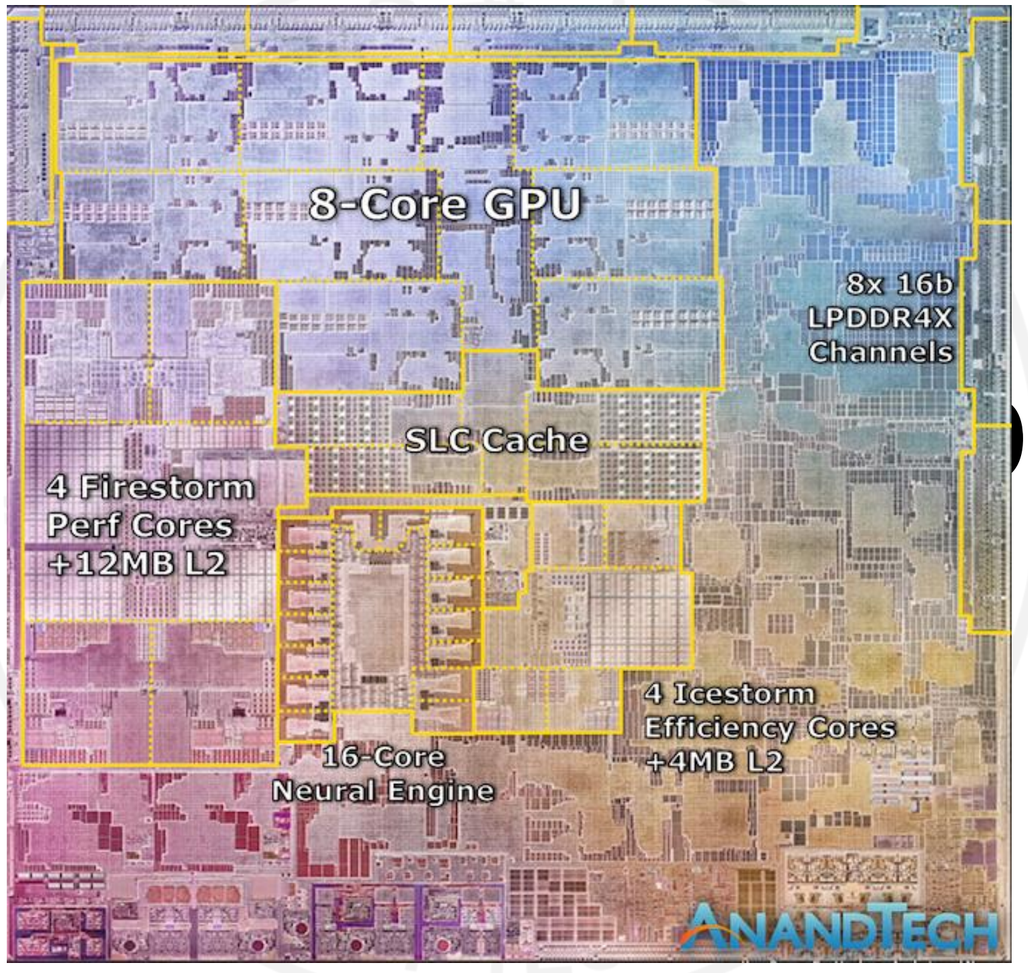
Apple’s processors

UMA

- GPU & Neural Engine in the same chip (SoC)
- DRAM (8-16 GB) is placed on the same package, but not on the same silicon
- Uniform memory access (UMA)
- 16-134 Billion transistors
- Efficient CPU
- Likely to impact HPC hardware



https://www.macrumors.com/guide/m1/



https://www.anandtech.com/show/16226/apple-silicon-m1-a14-deep-dive

	M1 chip	M2 Ultra Chip
Lithography	5 nm	5 nm
Transistors	16 billions	134 billions
High-performance core	4	16
Energy-efficient cores	4	8
Clock speed	3.2 GHz	3.49 GHz
GPU cores	8	60-76
L1 Cache	192+128 KB/core (High-perf)	192+128 KB/core (High-perf)
	128+128 KB/core (Energy-eff)	128+128 KB/core (Energy-eff)
L2 Cache	12 MB (High-perf)	64 MB (High-perf)
	4 MB (Energy-eff)	8 MB (Energy-eff)
Last level cache	8 MB	96 MB
Unified memory	8-16 GB	Up to 32 GB
System memory type	128-bit LPDDR4X	DDR5
Mem BW	68.25 GB/s	100 GB/s
Power	28 W	24-36 W

SP 7 32 bits

DP 15 64 bits

Thank you!