

# 088949 - ADVANCED COMPUTER ARCHITECTURES COURSE OVERVIEW

Second Semester 2024/2025



POLITECNICO  
MILANO 1863

Prof. Cristina Silvano  
email: [cristina.silvano@polimi.it](mailto:cristina.silvano@polimi.it)  
Politecnico di Milano



# *What are the Advanced Computer Architectures?*

---



# ACA Course Objectives

---

- Main goal is to understand the major concepts used in **modern microprocessor architectures** by the end of the semester and main techniques on **how to improve performance**.
  
  - The course will cover **different forms of parallelism**:
    - **instruction-level**
    - **data-level**
    - **thread-level**
    - **core-level**
  
  - The course discusses how these forms of parallelism can be exploited by **parallel processor architectures**.
-



# Parallel Architectures

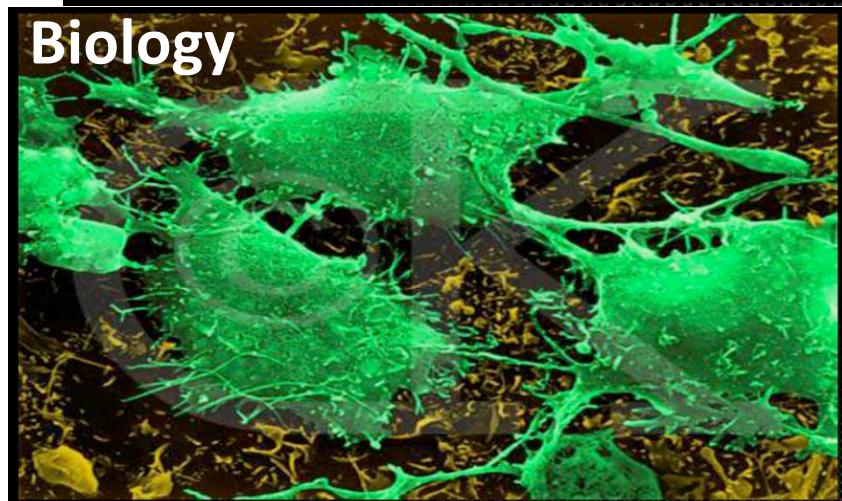
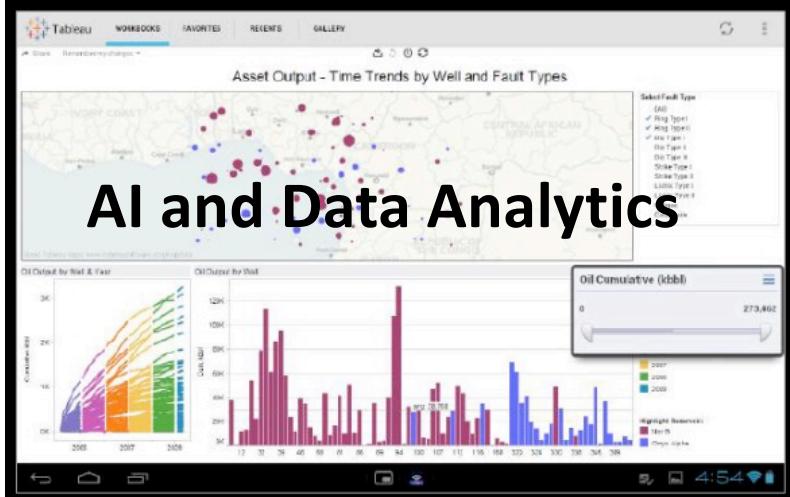
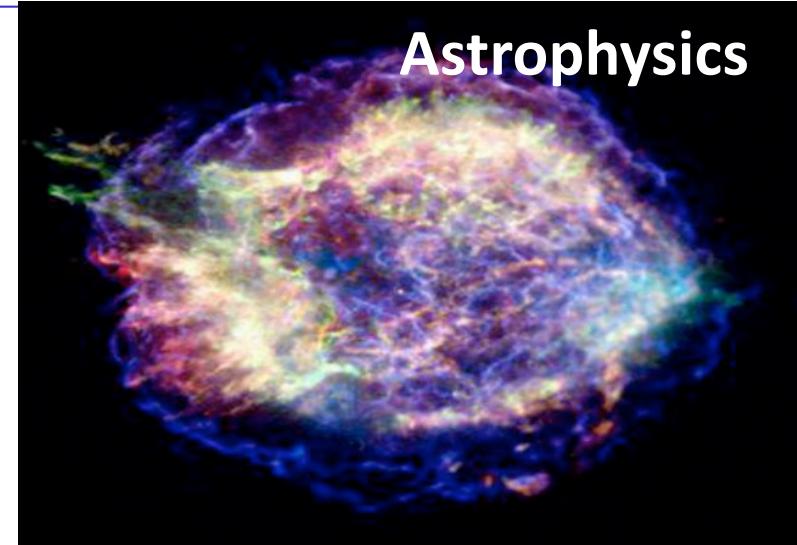
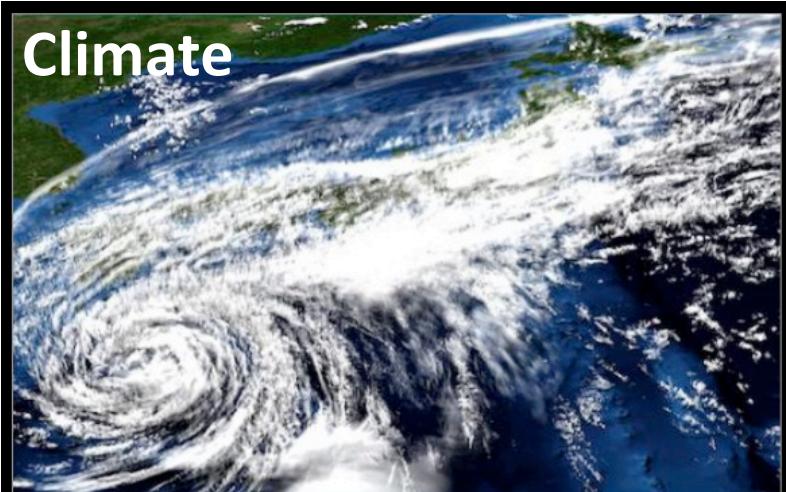
---

- Parallel Superscalar Processors
  - Multicore and Manycore Processors
  - Multithreading Processors
  - Data Parallel Vector Processors
  - Neural and Neuromorphic Processors
  - Reconfigurable Architectures
  - Heterogeneous Architectures
-

# *Systems based on Parallel Computing Architectures*



# Applications driving the demand for more computing performance and big data





# TOP500 List

- The **TOP500 list** is ranking the world's most powerful supercomputers.
- The TOP500 list is updated twice a year: in June and in November.
- The **LINPACK Benchmark** (introduced by Jack Dongarra) is used to measure the system's floating point computing power.
- LINPACK measures how fast a computer solves a dense  $n$  by  $n$  system of linear equations  $A x = b$ , which is a common task in engineering.



[www.top500.org](http://www.top500.org)





# How to measure performance?

Flop/s, Floating Point Operations per Second



Name	Flop/s
ExaFlop/s	$10^{18}$
PetaFlop/s	$10^{15}$
TeraFlop/s	$10^{12}$
GigaFlop/s	$10^9$
MegaFlop/s	$10^6$
KiloFlop/s	$10^3$
Flop/s	1



# High Performance Computing

- First supercomputer reaching the **Petascale** ( $10^{15}$  Flops) performance: IBM Roadrunner installed in **2008** at Los Alamos National Lab (New Mexico, USA).
- Research on supercomputing has pushed to the ***Exascale era of billions of billions ( $10^{18}$ ) FLOPS reached in 2022.***



*In 2022, TOP500 announced Frontier as the first Exascale machine up to 1.102 ExaFlop/s.*



# Top500 (November 2024)

**#1 El Capitan (USA)** reached 1.742 ExaFlop/s with 11M combined CPU and GPU cores based on AMD 4th gen. EPYC processors with 24 cores at 1.8GHz and AMD Instinct MI300A. Energy efficiency of 58.89 GigaFlops/Watt.

**Site:** Lawrence Livermore National Laboratory in California, USA.





## Top500 (November 2024)

**#2 Frontier (USA):** In June 2022, first Exascale supercomputer with 1,2 ExaFlop/s on Linpack.

In Nov. 2024, it reached 1.35 Exaflop/s (2024) with 9 M cores and 24.6 MW. Based on HPE Cray EX235a architecture, AMD EPYC 64C 2GHz processor and AMD Instinct MI 250X.

**Site:** Dept. of Energy's, Oak Ridge National Lab., Tennessee (USA).





## Where is Italy placed in the worldwide ranking?

### #5 in Top500 and #1 in Europe: HPC6

Same architecture as Frontier, but only 3 M cores.

**Pre-Exascale** system with 477.90 PFlop/s and 8.5 MW power dissipation.

**Site:** ENI center in Ferrera Erbognone (PV), Italy.





## *Where is Italy placed in the worldwide ranking?*

---



**#9 in Top500 and #4 in Europe: Leonardo Pre-Exascale** with 241 PetaFlop/s on Linpack benchmark with 1.8M cores and Atos BullSequana XH2000 system based on Xeon Platinum 8358 32C 2.6 GHz, NVIDIA A100 SXM 4 64GB.

**Site:** CINECA, Bologna (Italy).  
**Co-funding** 240 Meuro EuroHPC & Italian Ministry of Univ. and Research.

**Leonardo** entered as **#4** in Top500 in Nov. 2022.



# HPC vs Laptop: Performance comparison

---

2024



#1 Frontier 1.2 ExaFlop/s

$10^6 \times$

2024



Laptop at 4 GHz 3.6 TeraFlop/s



# HPC vs. Smartphone: Performance comparison

---

2004



IBM BlueGene 2004: 9.1 TeraFlop/s



2024



Apple iPhone 15 Ultra: 2.5 TeraFlop/s



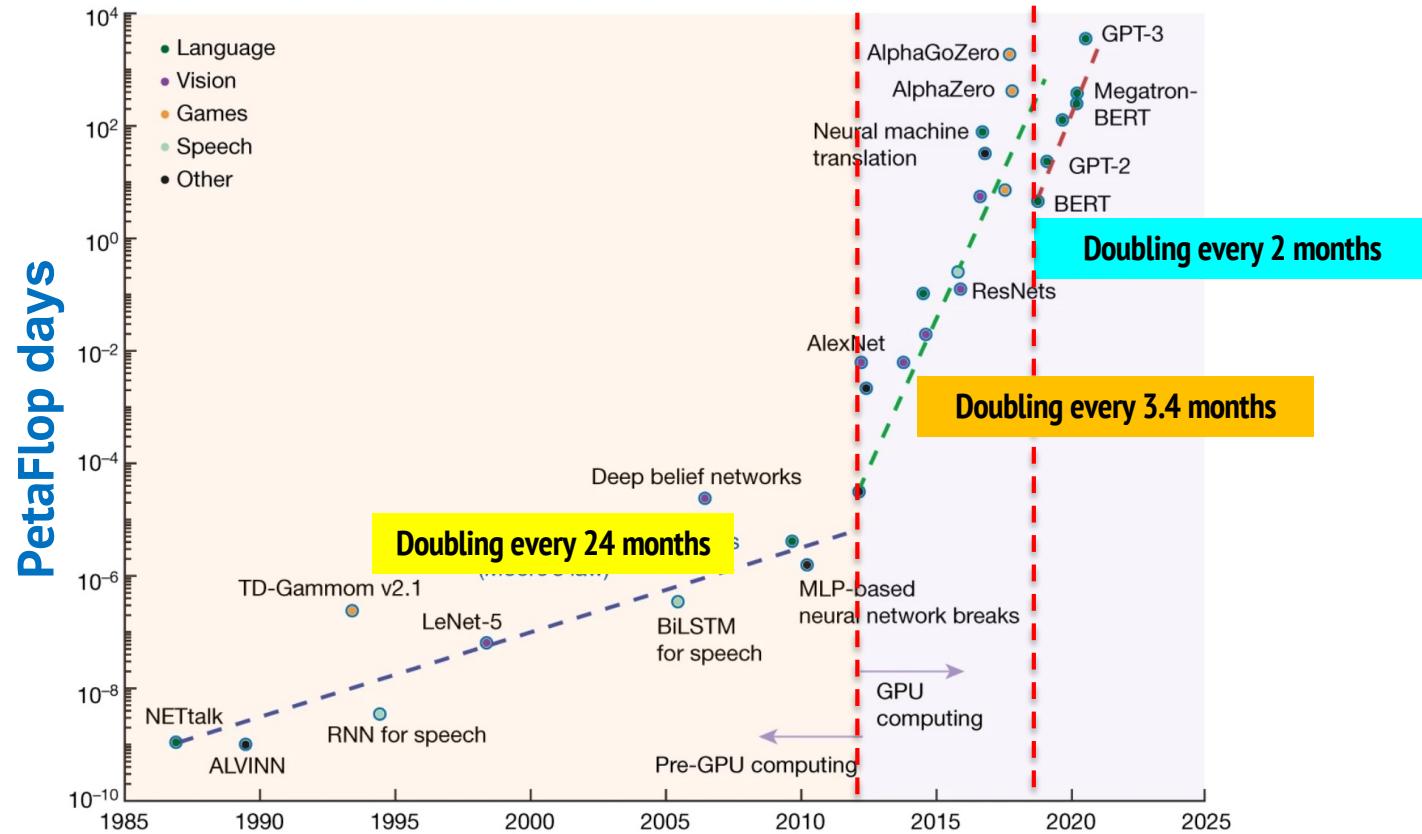
# *HPC and Artificial Intelligence: Some trends*

---





# 1) Computing power demands for AI applications



Mehonic and Kenyon, Brain-inspired computing needs a master plan, Nature, 2022



## 2) Generative AI workloads are energy hungry

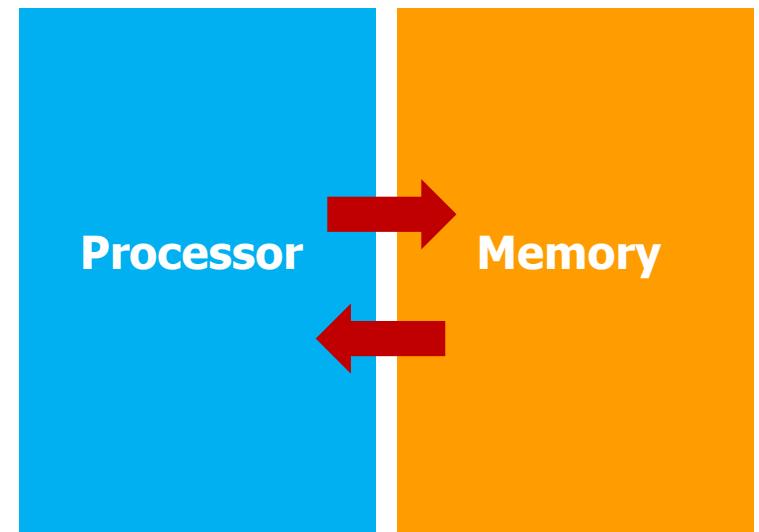
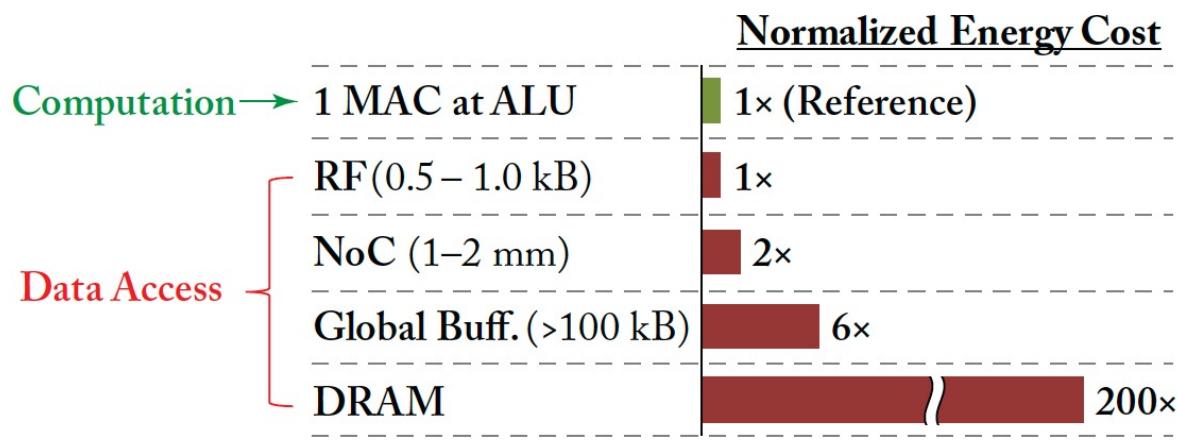
---

- ChatGPT-3 training required thousands of GPUs for about one month with **more than 200 MWh**.
- Estimating a family consumes on average **10 MWh per year** => GPT-3 training corresponds to the average energy consumed by a family in **20 years!**
- A single request to ChatGPT-3 requires about ten times more energy (**2.9 Wh**) than a search on Google (**0,3 Wh**).





### 3) Data transfer dominates energy consumption



Most of the energy consumption is spent to move data.



## 4) Brain-inspired or Neuromorphic Computing

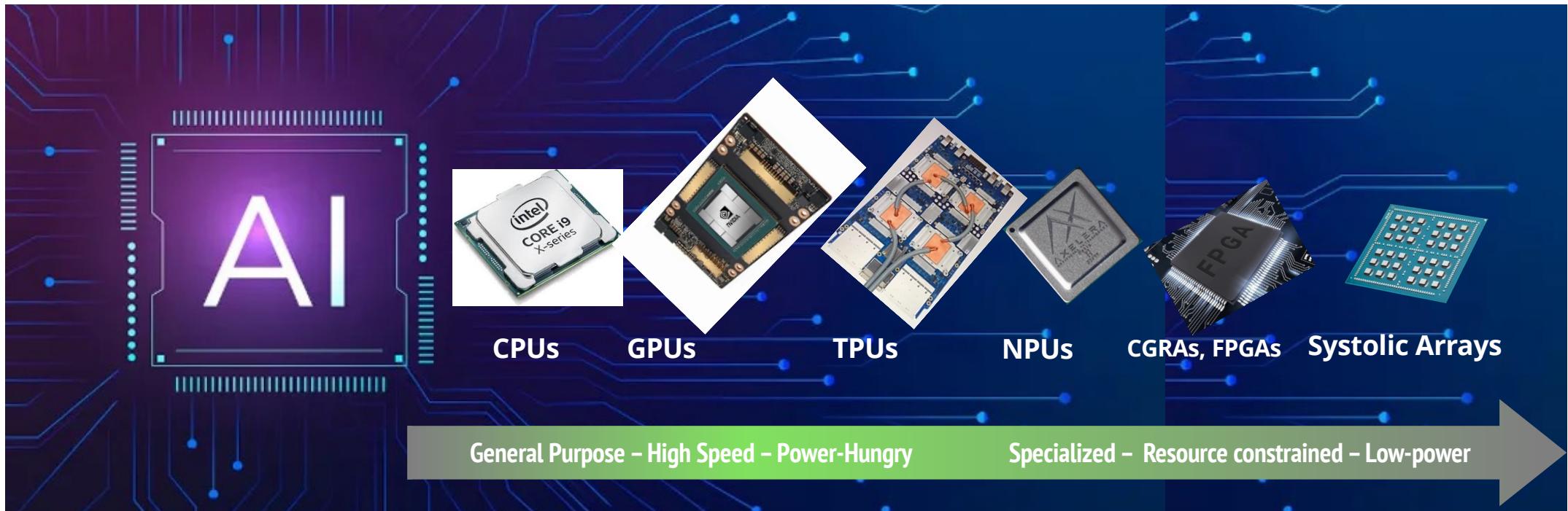
---

- To take inspiration from **biology**, where data storage is not separate from processing.
- In the **human brain**, neurons and synapses perform both functions in massively parallel and adaptable structures with extreme energy efficiency.
- Neuromorphic Computing systems transform the way of processing signals and data by **co-locating memory and processing**.





# AI Accelerators from HPC to the Edge



D. Garisto, "Accelerating AI: The cutting-edge chips powering the computing revolution", Nature, News Feature, Vol.630, June 2024.



# *How did we get here?*

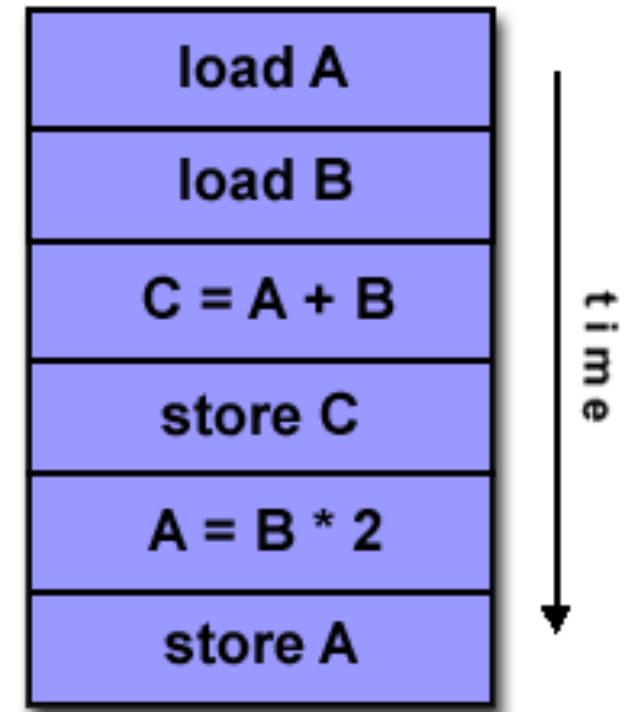
---

- *Technology?*
- *Parallelism?*
- *Which kind of parallelism?*



# SISD: Single Instruction Single Data

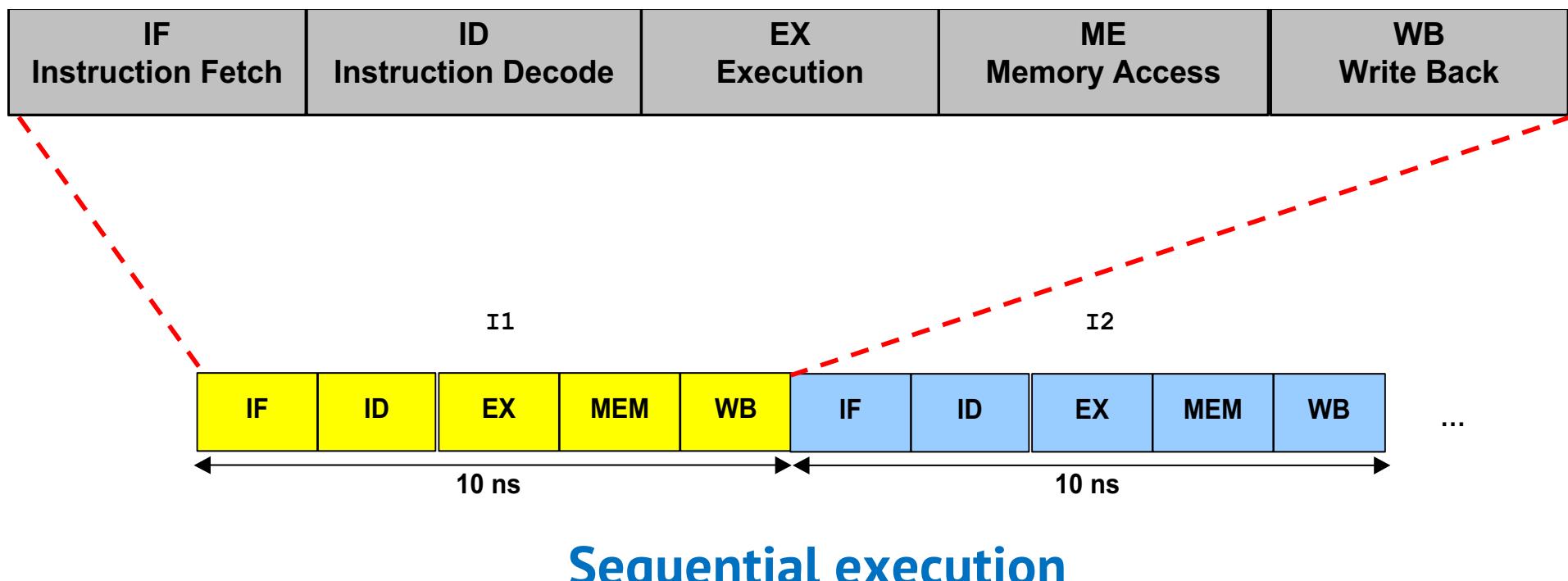
- A serial (non-parallel) computer
- **Single instruction:** only one instruction stream is executed by the CPU
- **Single data:** only one data stream is being used as input
- Each instruction is executed by the CPU during each clock cycle





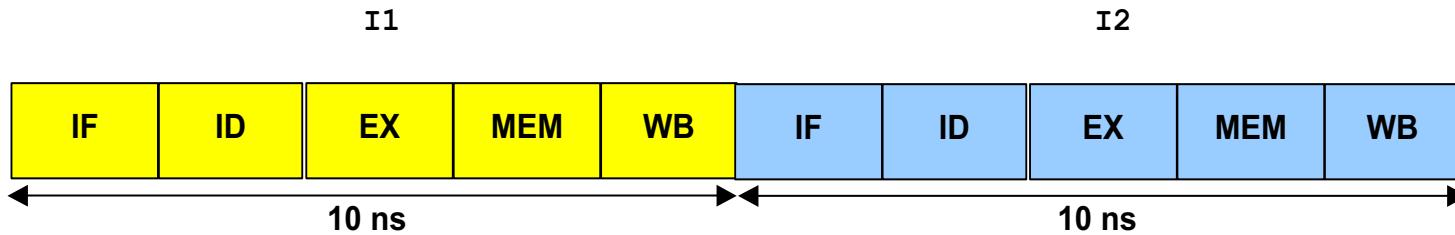
# Instruction execution into phases

Each instruction is executed in **5 clock cycles (phases)**

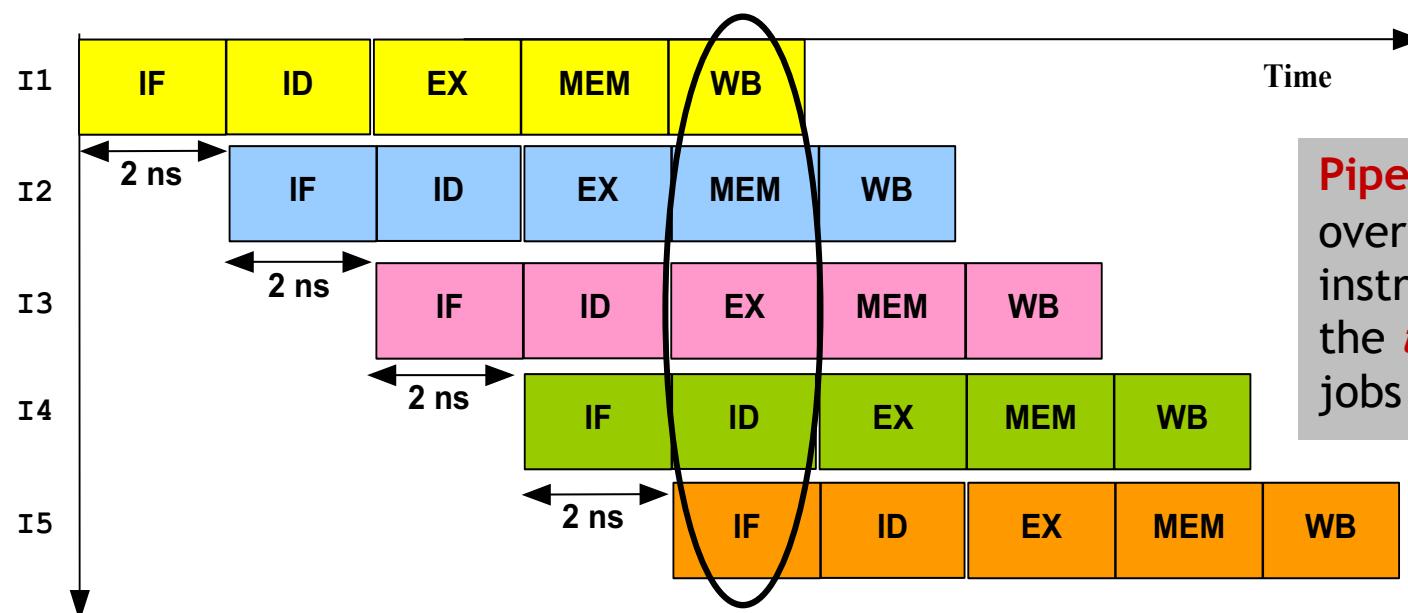




# Sequential vs. Pipelining Execution: Instruction Level Parallelism



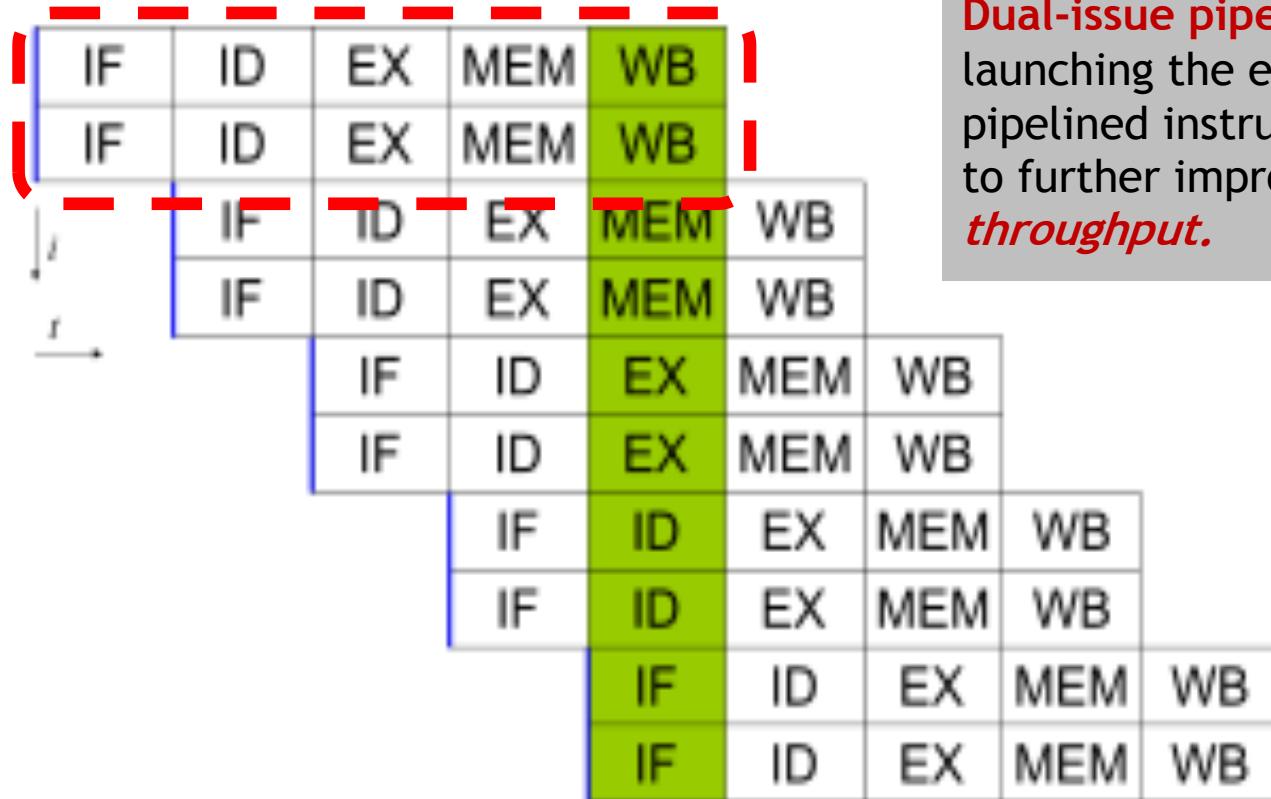
**Sequential execution**  
of instruction stages.



**Pipelined execution:**  
overlapping execution of  
instruction stages to improve  
the **throughput**: number of  
jobs executed per time unit



# Multiple-issue pipeline to double the throughput



## Dual-issue pipeline:

launching the execution of two pipelined instructions per cycle to further improve the *throughput*.



# *Parallelism? Which kind?*

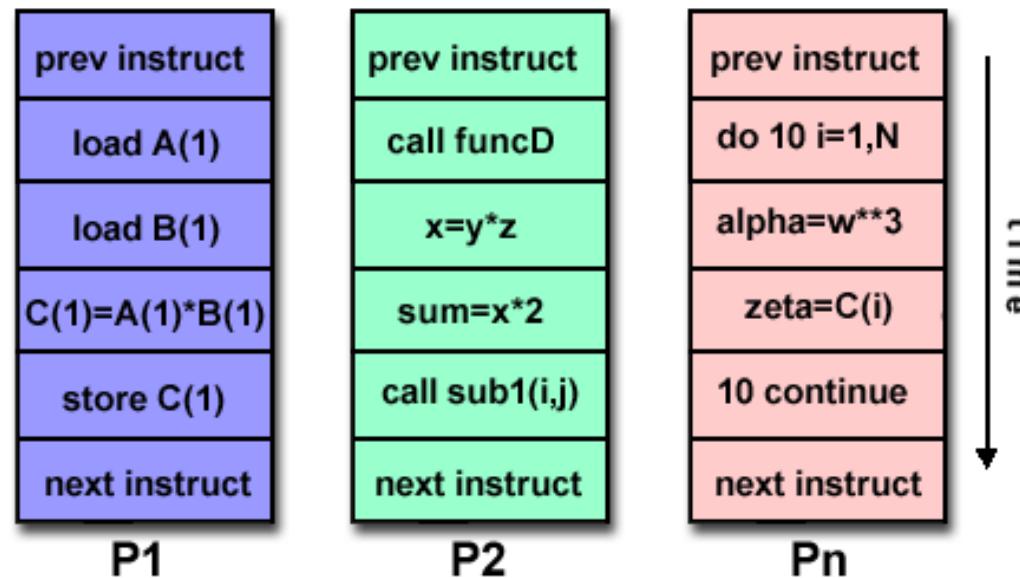
---

- *Instruction Level Parallelism: Pipelining, Multiple Issue*
  
- *What's next? Multiprocessing*



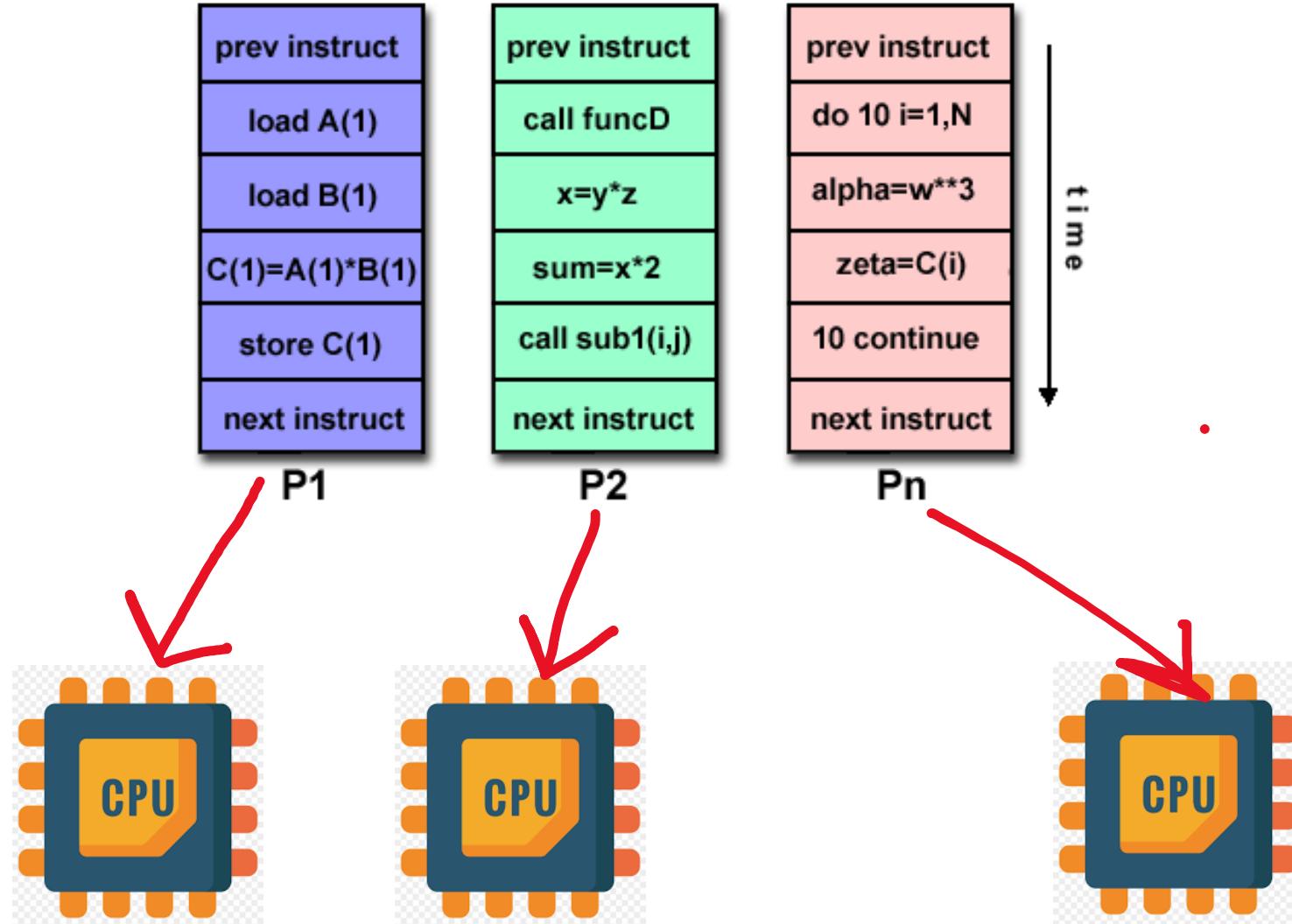
# MIMD: Multiple Instruction Multiple Data

- **Multiple Instruction:** every processor may be executing a different instruction stream
- **Multiple Data:** every processor may be working with a different data stream



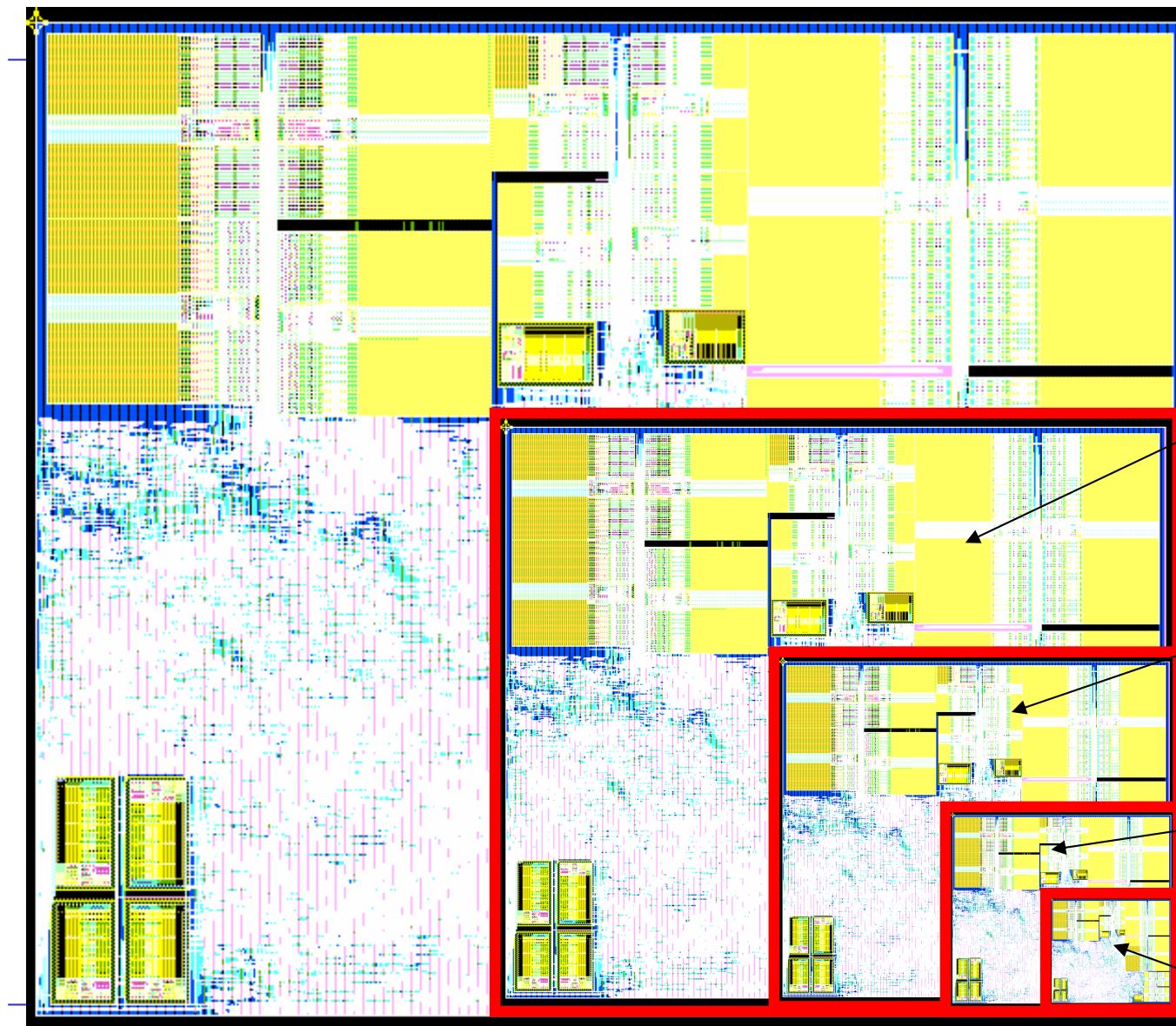


# MIMD: Multiprocessors





# Squeezing of computing cores



2005  
65 nm  
 $1.4 \text{ mm}^2$



Source:  
ARM9 STMicroelectronics

2007  
45 nm

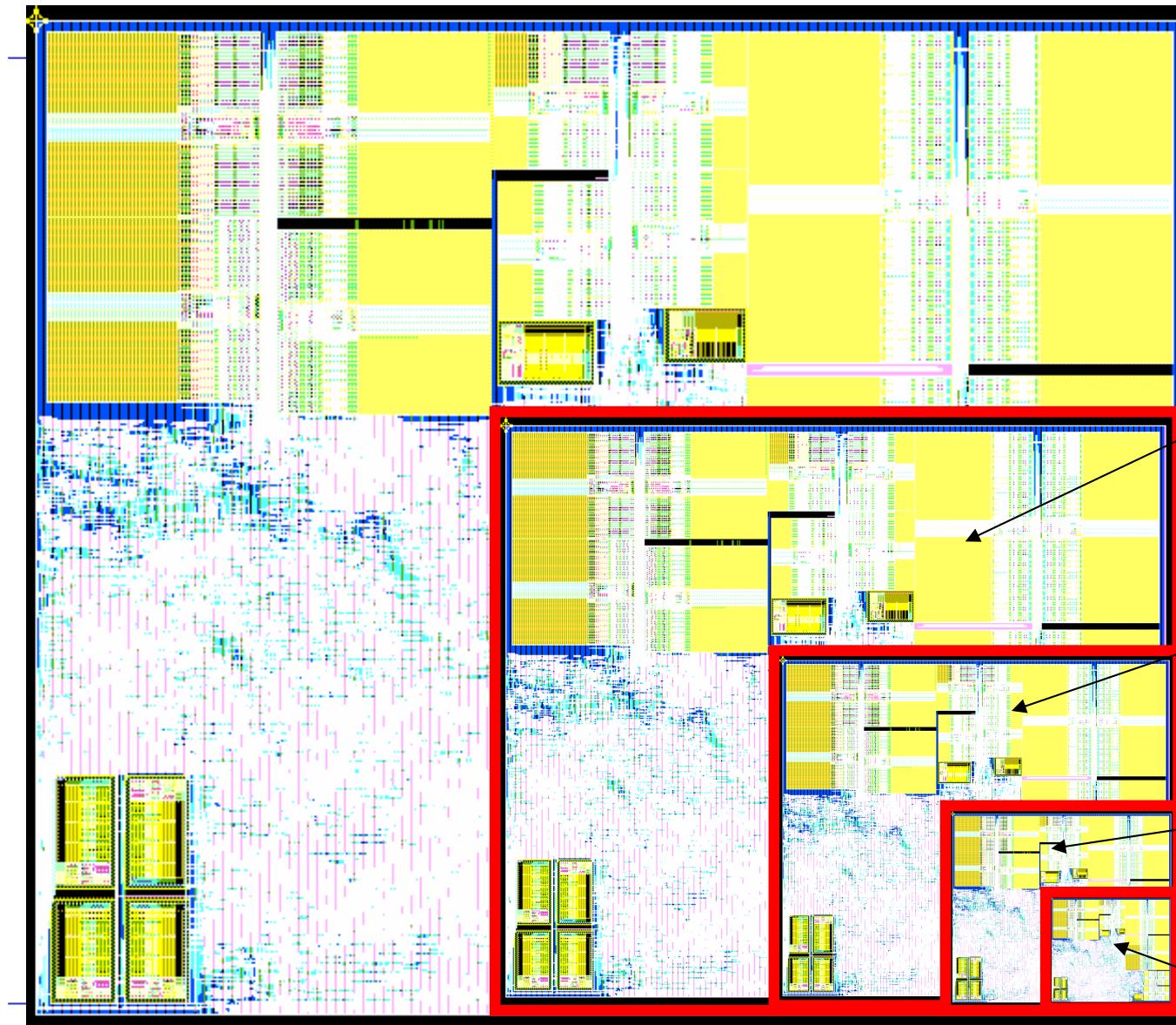
2009  
32 nm

2011  
22 nm

2013  
14 nm



# ... entering the multi/many-core era



2005  
65 nm  
 $1.4 \text{ mm}^2$



Source:  
ARM9 STmicroelectronics

2007  
45 nm

2009  
32 nm

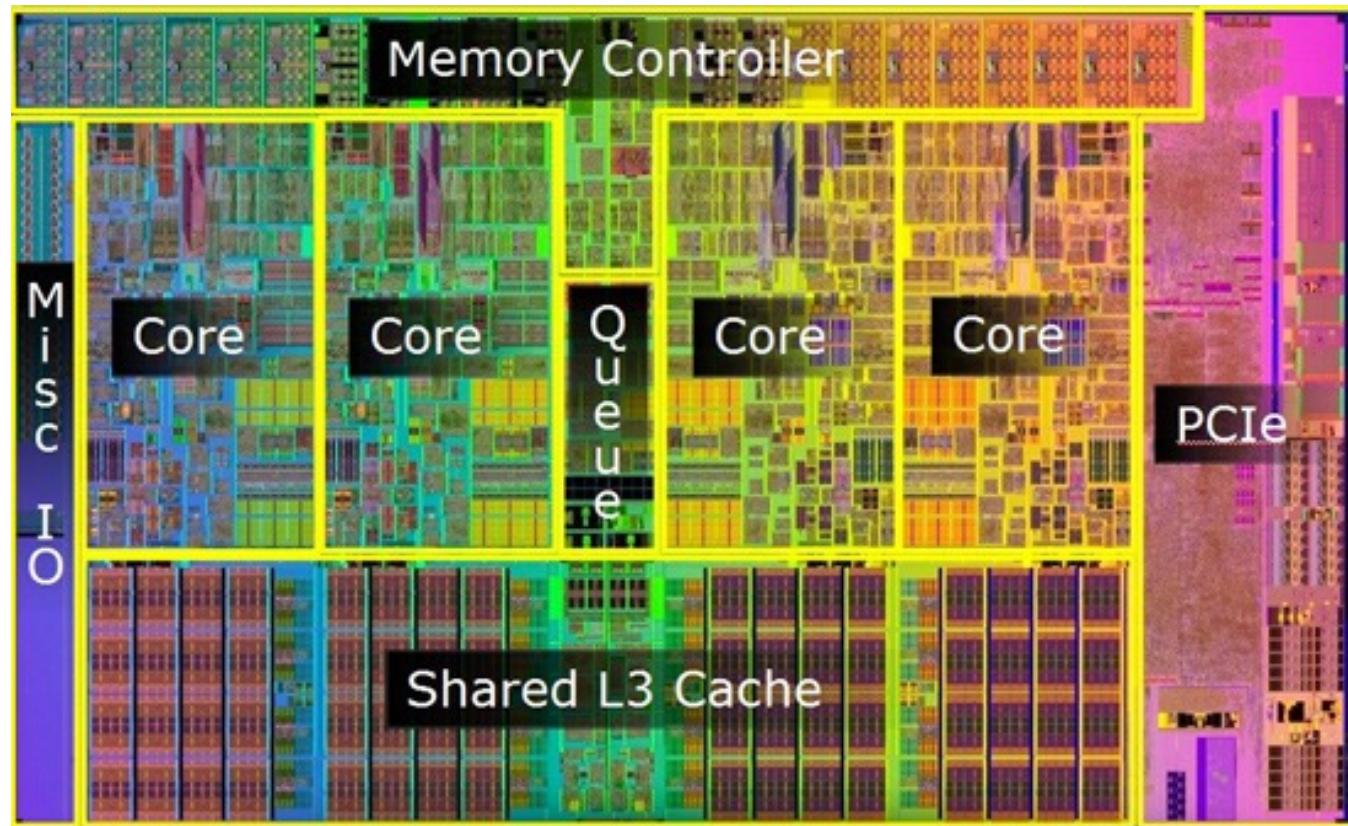
2011  
22 nm

2013  
14 nm



# Multicore: Multiprocessor on a single-chip

---



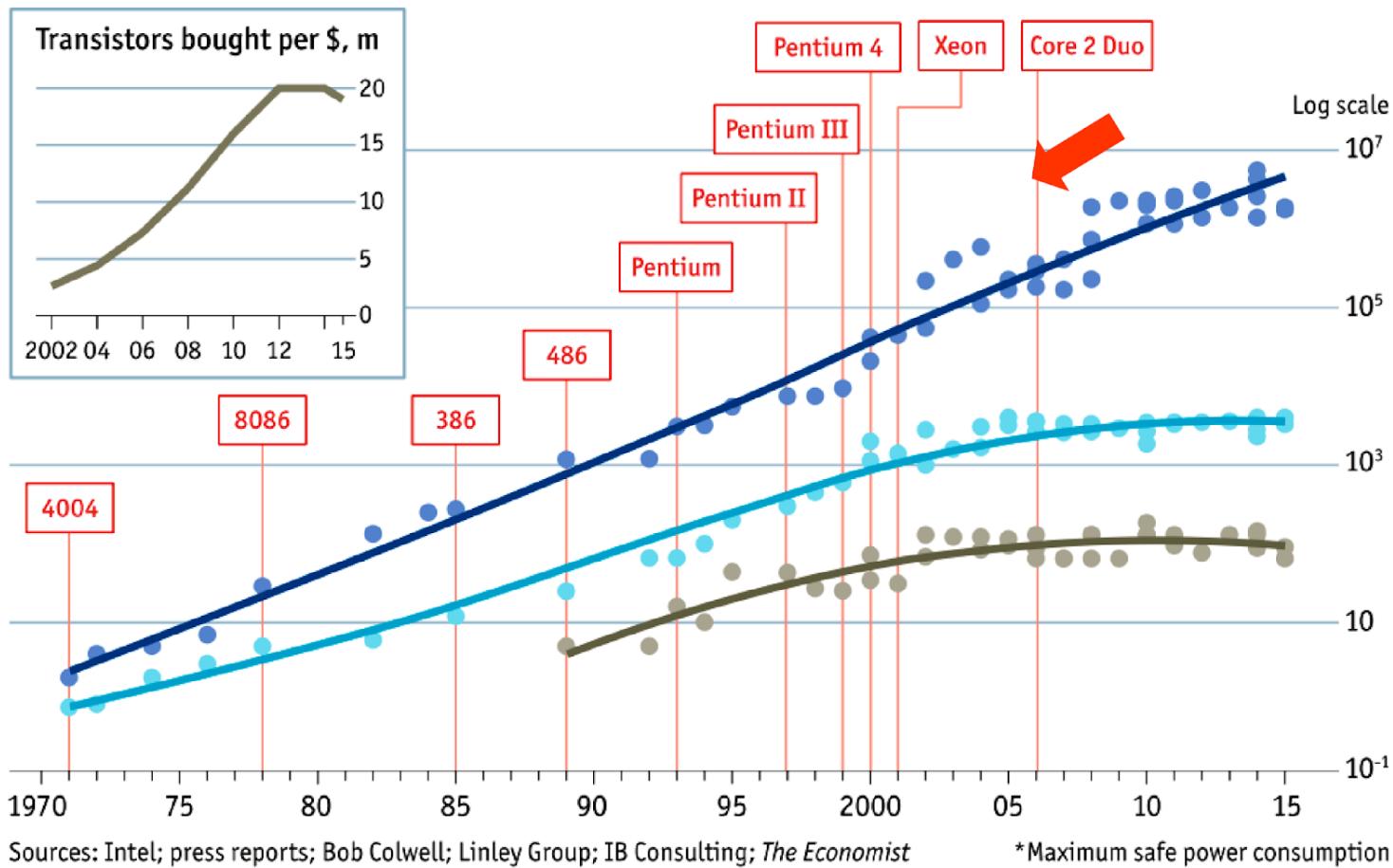


# Technology Trends

## Stuttering

● Transistors per chip, '000 ● Clock speed (max), MHz ● Thermal design power\*, W

Chip introduction dates, selected





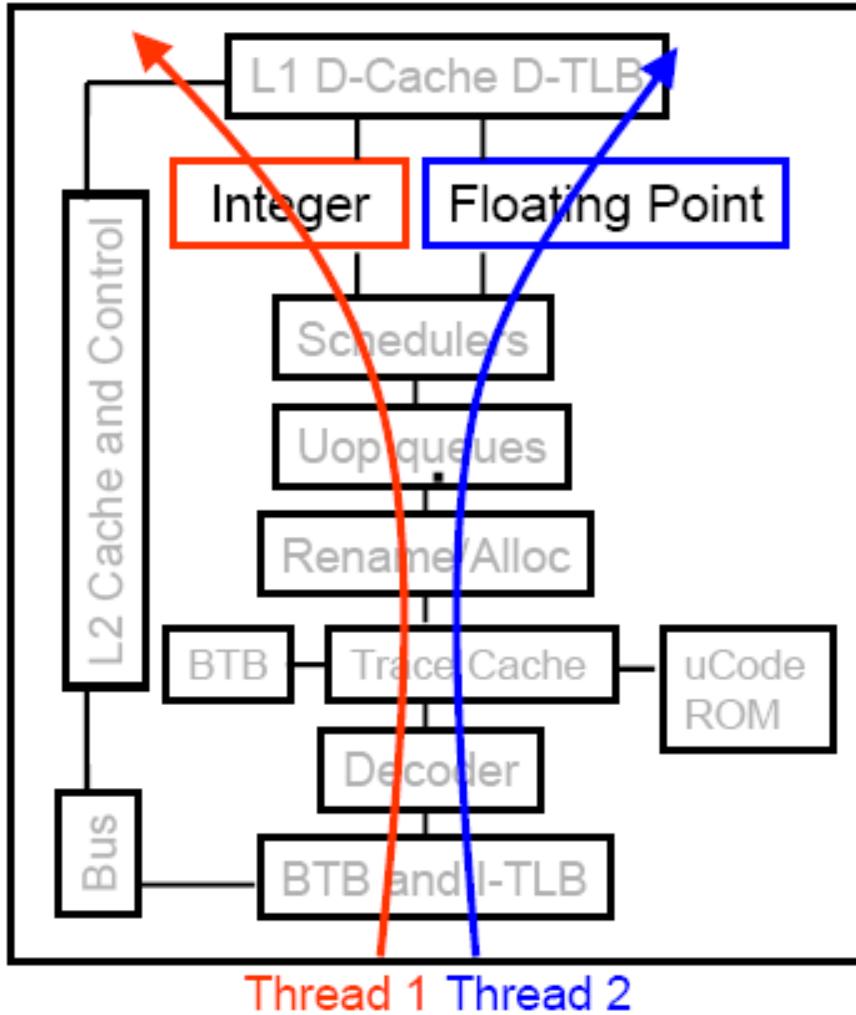
# *Parallelism? Which kind?*

---

- ILP: Pipelining, Multiple Issue (SISD)
- Multiprocessing and Multicores (MIMD)
- *What's next?*
  - *Multithreading*

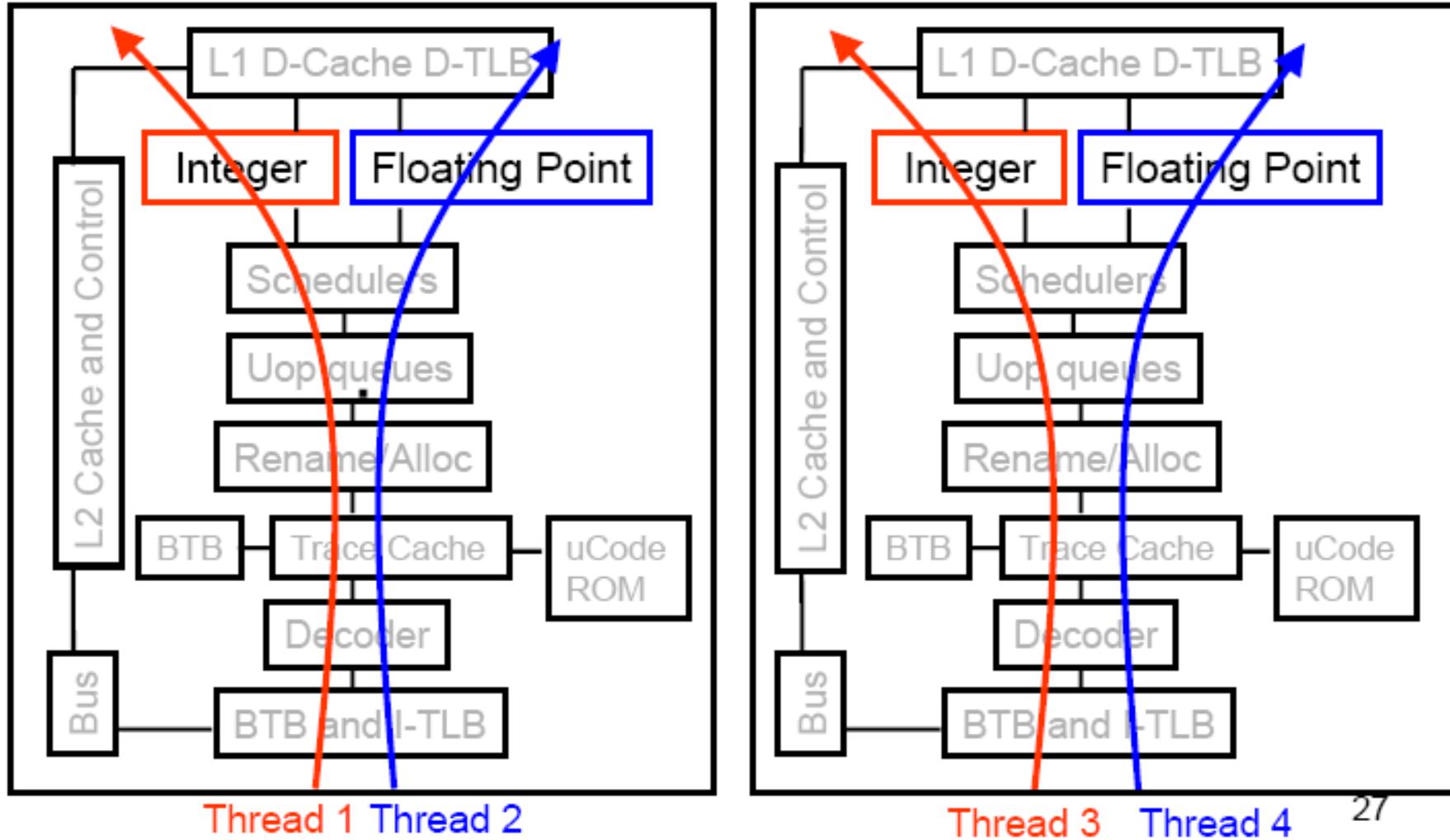


# Two threads per processor core



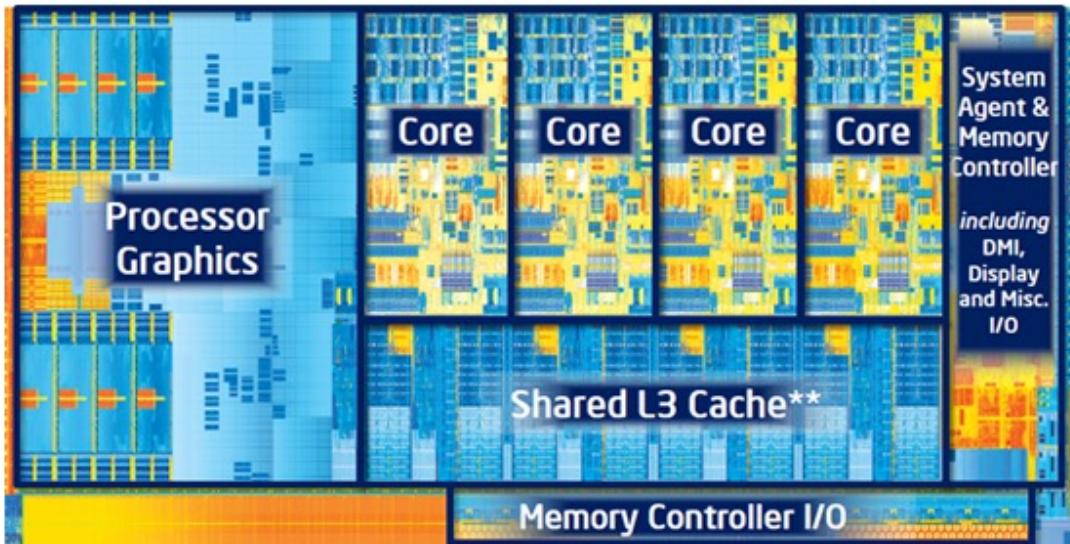


# Dual-core processor with 2 threads per core





# Intel® Core™ i7-3770T Hyperthreading Processor



# of Cores	4
# of Threads	8
Clock Speed	2.5 GHz
Max Turbo Frequency	3.7 GHz
Intel® Smart Cache	8 MB
Instruction Set	64-bit
Instruction Set Extensions	SSE4.1/4.2, AVX
Number of transistors	1.4 B transistors
Lithography	22 nm
Max TDP	45 W
Recomm. Customer Price	TRAY: \$294.00
Max Memory Size	32 GB
Memory Types	DDR3-1333/1600
# of Memory Channels	2
Max Memory Bandwidth	25.6 GB/s

- First generation launched in 2008.
- Next generations: Broadwell, Skylake, Kaby Lake at 14nm (2014); Cannonlake at 10nm (2H 2017); Ice Lake 10nm (2018).



# *Parallelism? Which kind?*

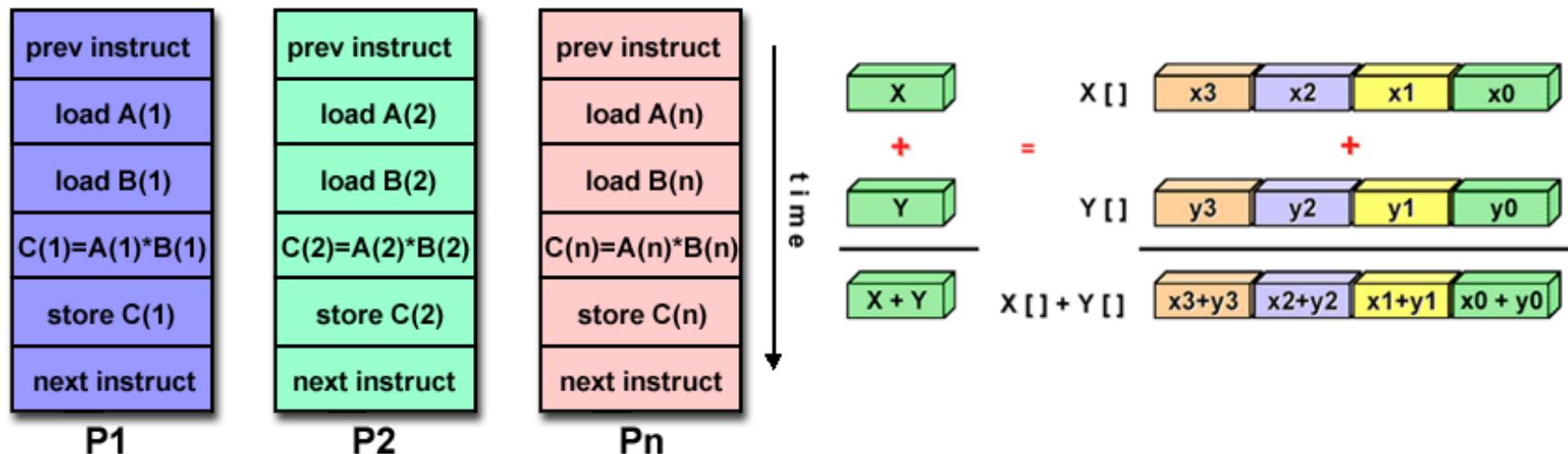
---

- ILP: Pipelining, Multiple Issue (SISD)
- Multiprocessing and Multicores (MIMD)
- Multithreading
- *What's next? Data-Level Parallelism (SIMD)*



# SIMD: Single Instruction Multiple Data

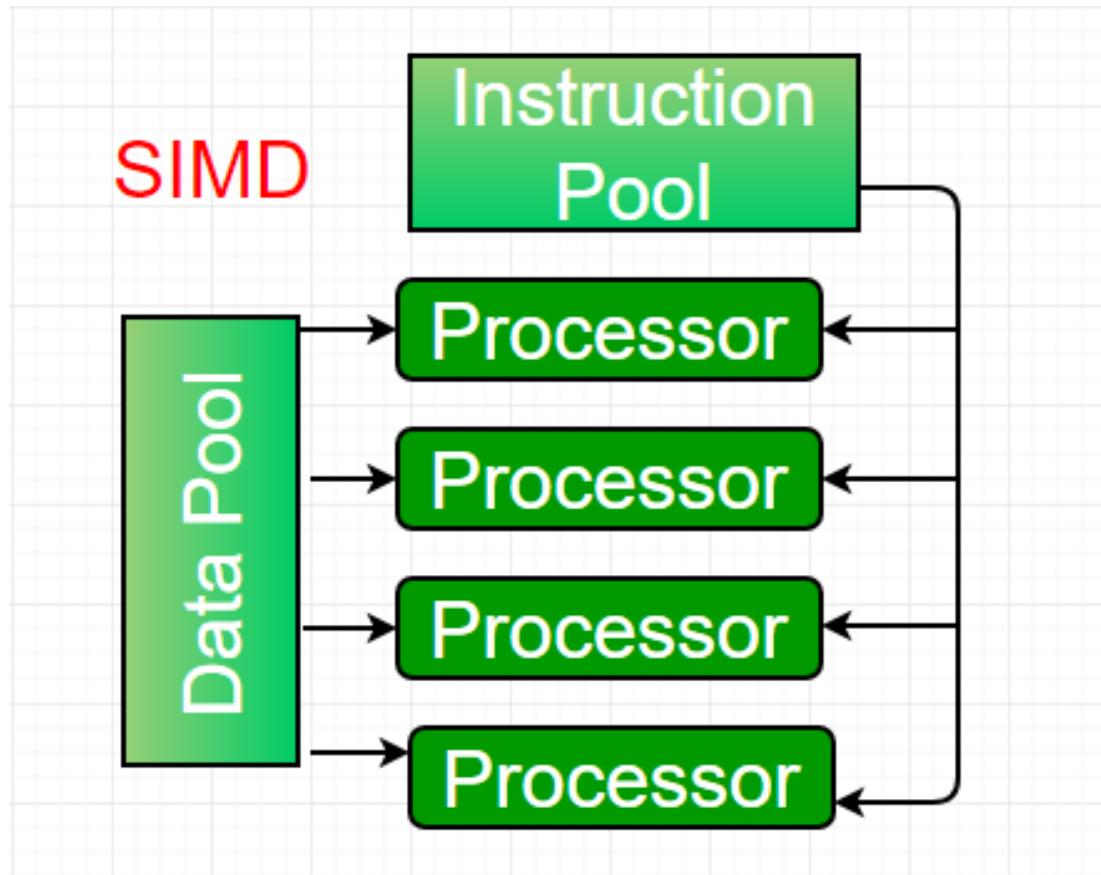
- **Single instruction:** all processing units execute the same instruction at any given clock cycle
- **Multiple data:** each processing unit can operate on a different data element



- Suitable for specialized problems managing vectors

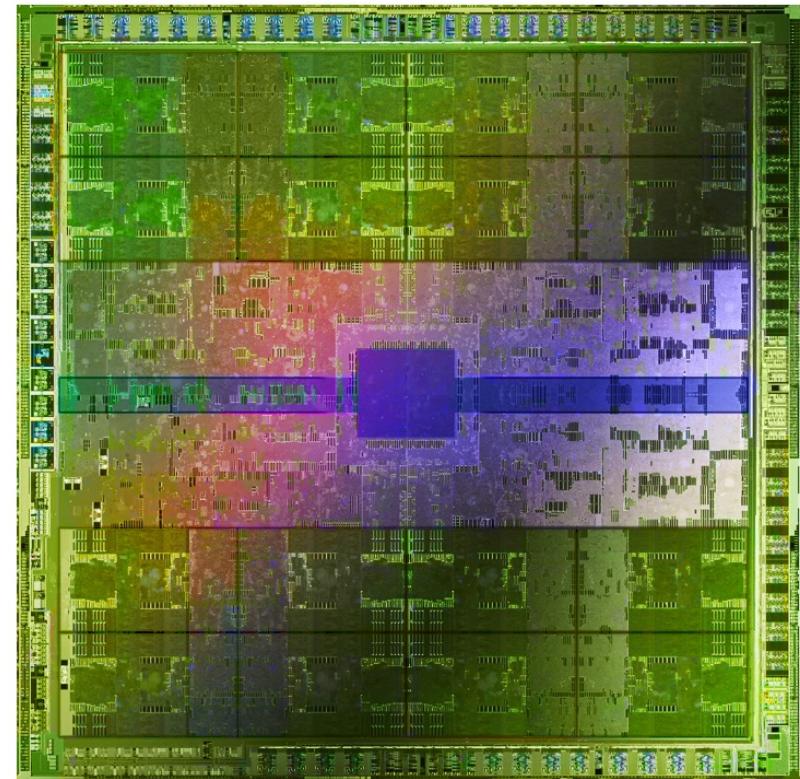
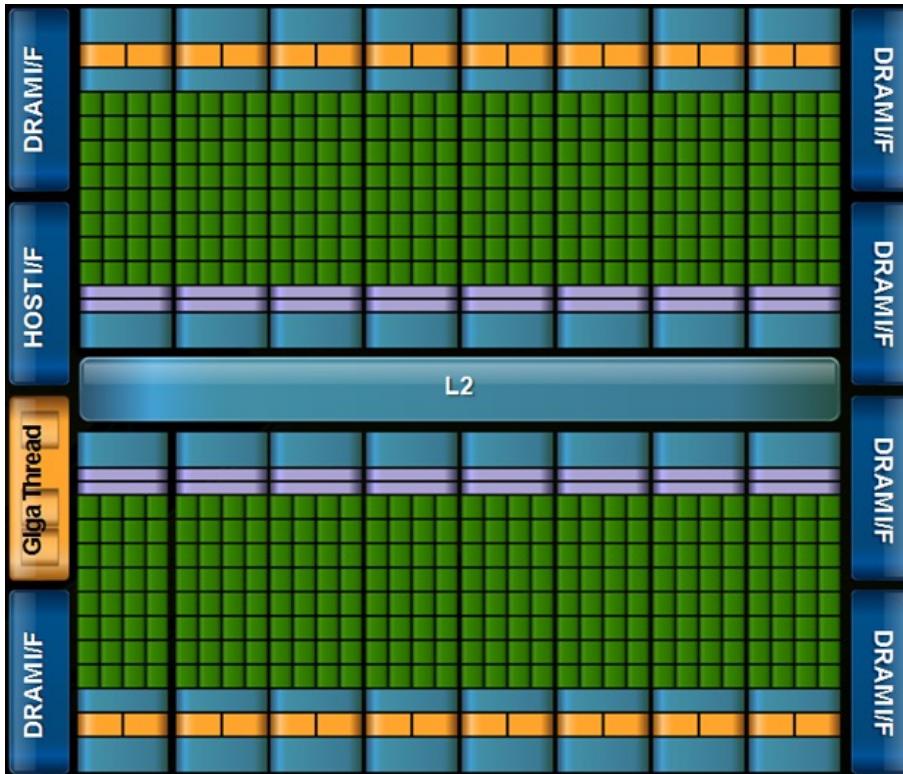


# SIMD: Single Instruction Multiple Data





# NVIDIA Fermi GPU (2010)



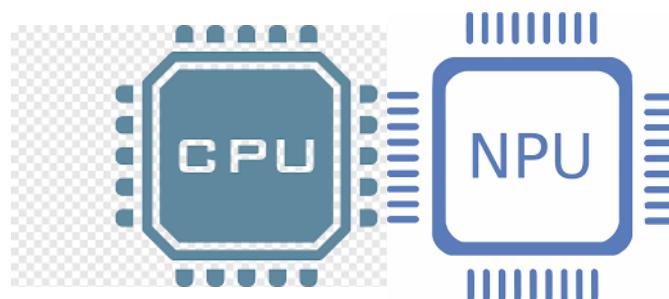
**16 Streaming Multiprocessors with 32 cores each => 512 cores and 3B transistors**



# Parallelism? Which kind?

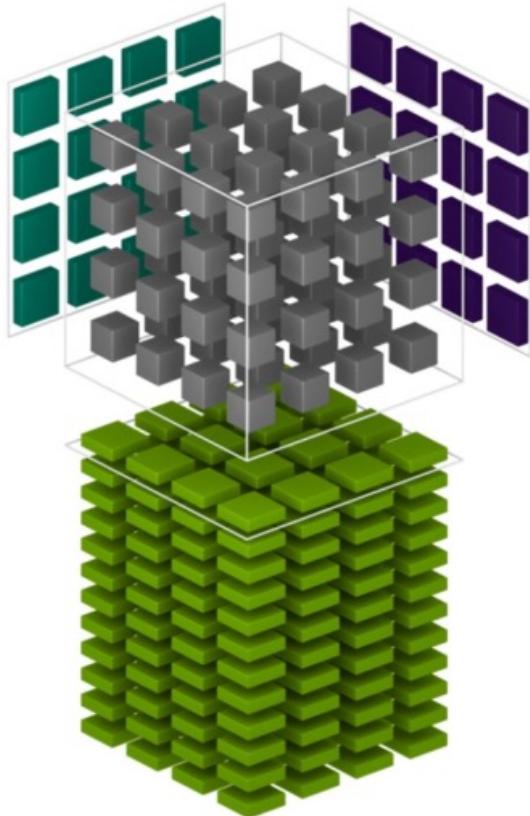
---

- ILP: Pipelining, Multiple Issue (SISD)
- Multiprocessing and Multicores (MIMD)
- Multithreading (MT)
- Data-Level Parallelism: SIMD and Vector Processors
- ***What's next?***
  - ***Heterogeneity: Processors + Accelerators such as Neural Processing Unit and Tensor Cores for machine learning and deep learning***





# NVIDIA Tensor Cores



$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix}_{\text{FP16 or FP32}} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix}_{\text{FP16}} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}_{\text{FP16 or FP32}}$$



# NVIDIA Volta GV100 with 84 SMs (2017)



**Full GV100 GPU with 84 Streaming Multiprocessors, for a total of 5376 FP32 cores, 5376 INT32 cores, 2688 FP64 cores, 672 Tensor Cores, and 336 texture units.**



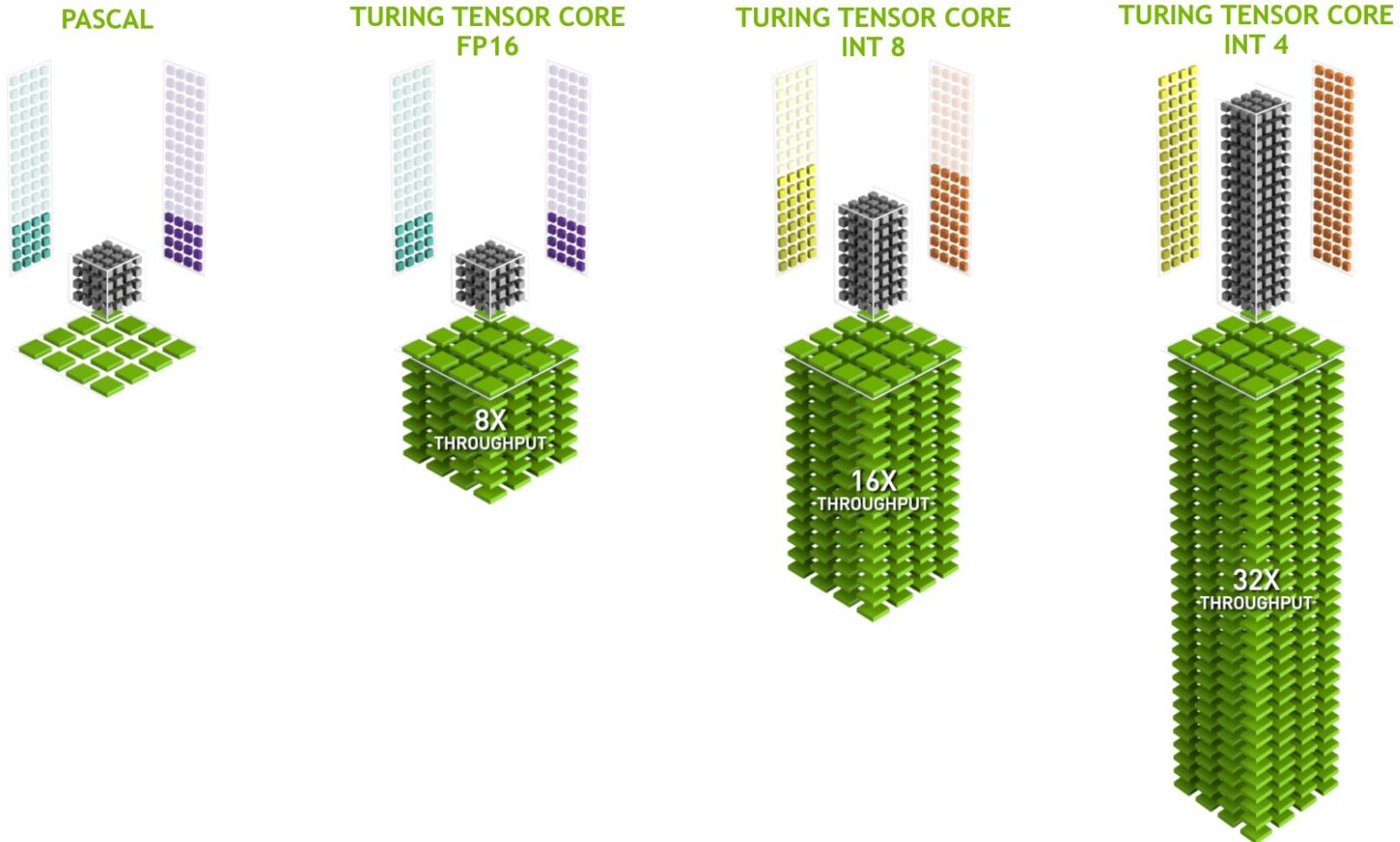
# Volta GV100 SM

- NVIDIA Volta GV100 Streaming Multiprocessor
- NVIDIA's first chip to feature Tensor Cores to accelerate Deep Learning





# NVIDIA Tensor Cores Evolution





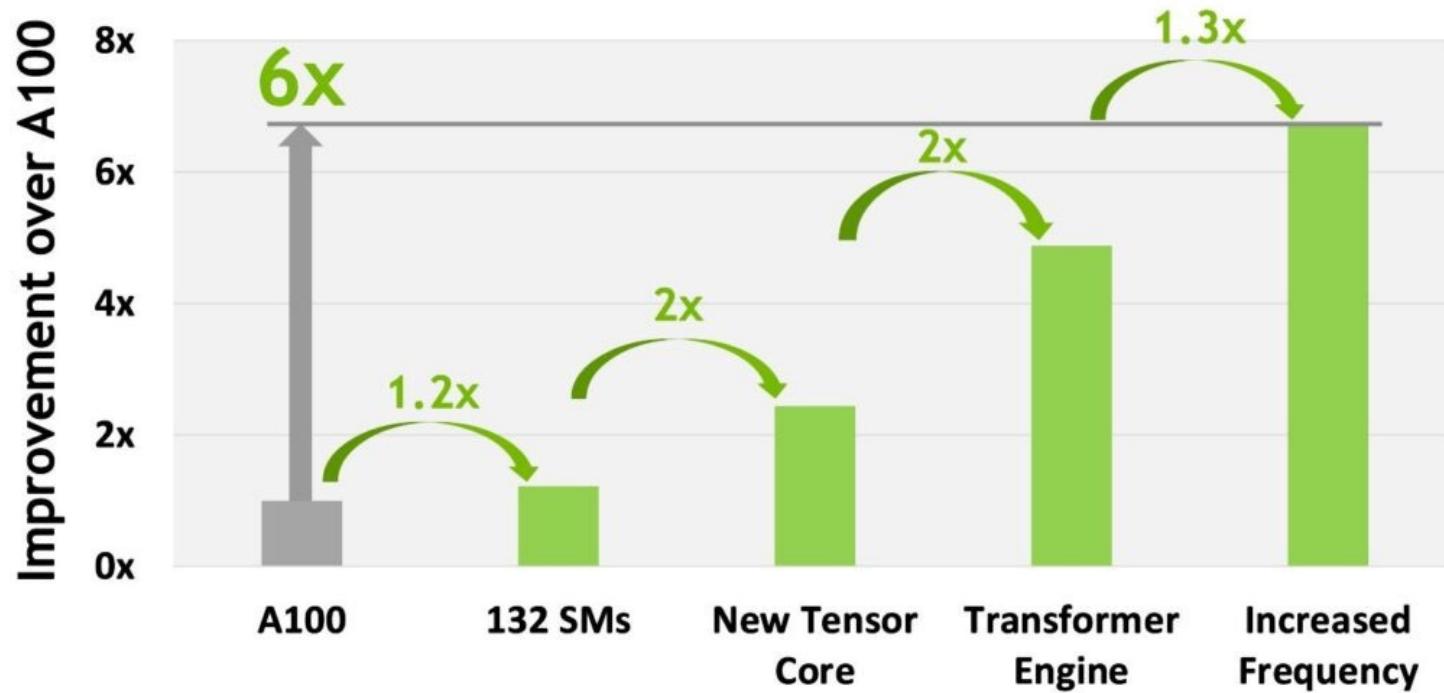
# NVIDIA AMPERE GA100 Tensor Core GPU with 128 SMs and 54.2 billion trans. (20x faster than Volta)





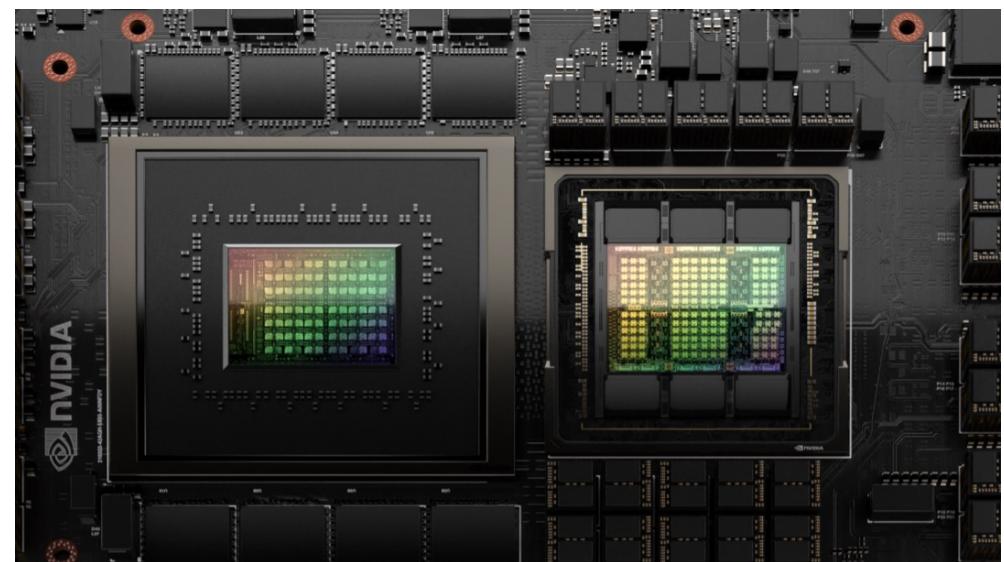
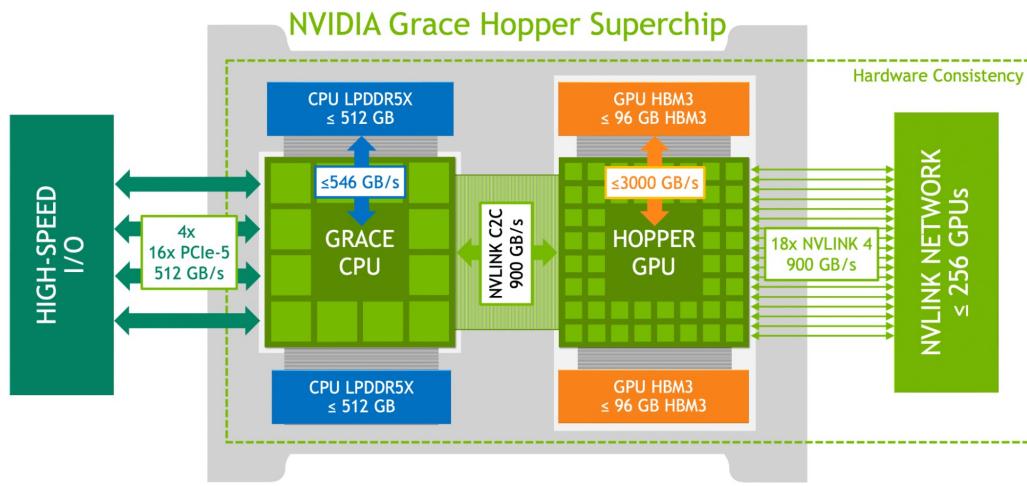
# NVIDIA Hopper H100

- **Next generation:** NVIDIA Hopper H100 with 144 Streaming Multiprocessors and 80 billion transistors with 4nm TSMC technology process





# NVIDIA Grace Hopper Superchip (2024)



<https://developer.nvidia.com/blog/nvidia-grace-hopper-superchip-architecture-in-depth/>

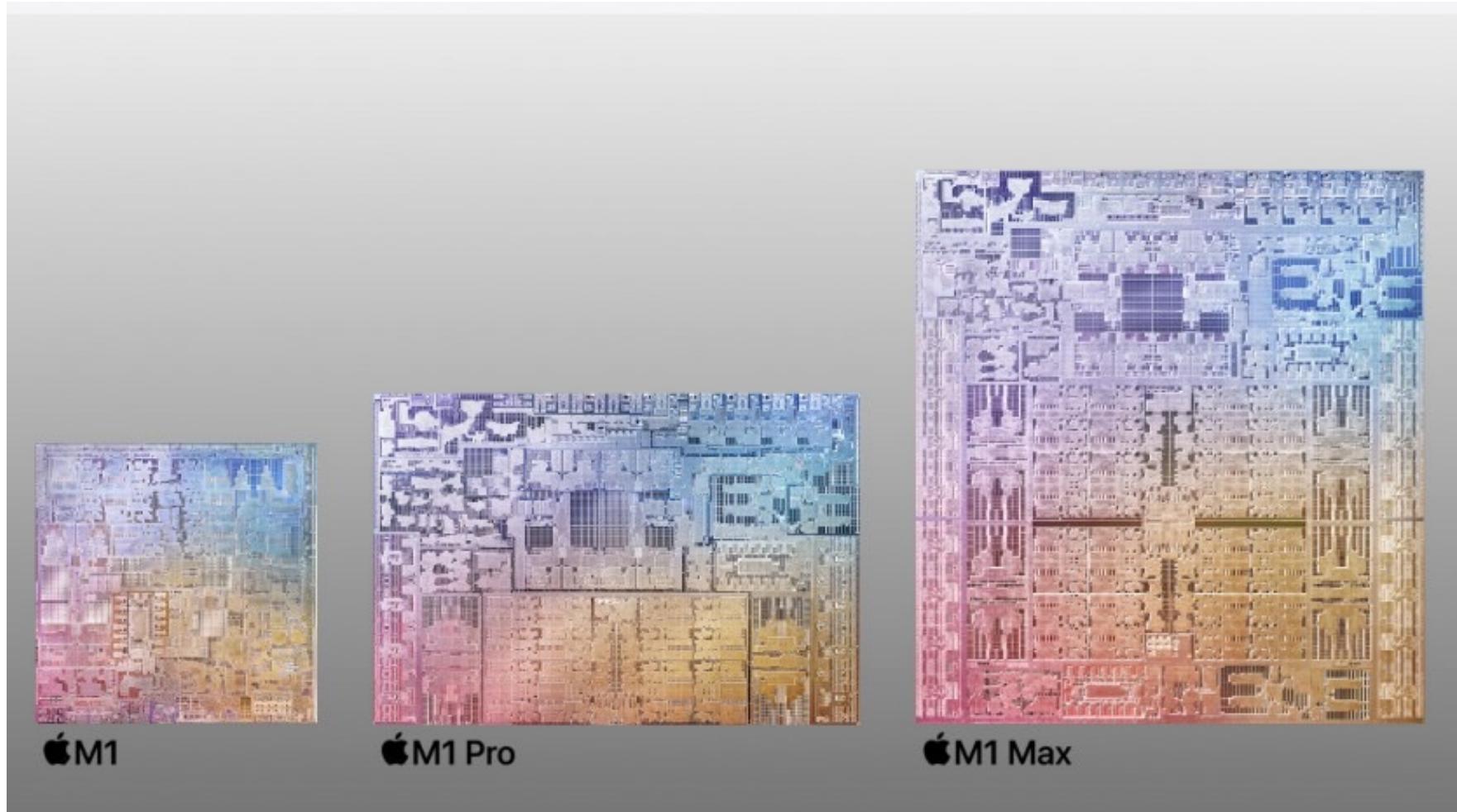


## *Putting all together ....*

---

- *ILP: Pipelining, Multiple Issue (SISD)*
- *Multiprocessing and Multicores (MIMD)*
- *Multithreading (MT)*
- *Data-Level Parallelism: SIMD and Vector Processors*
- *Heterogeneity*

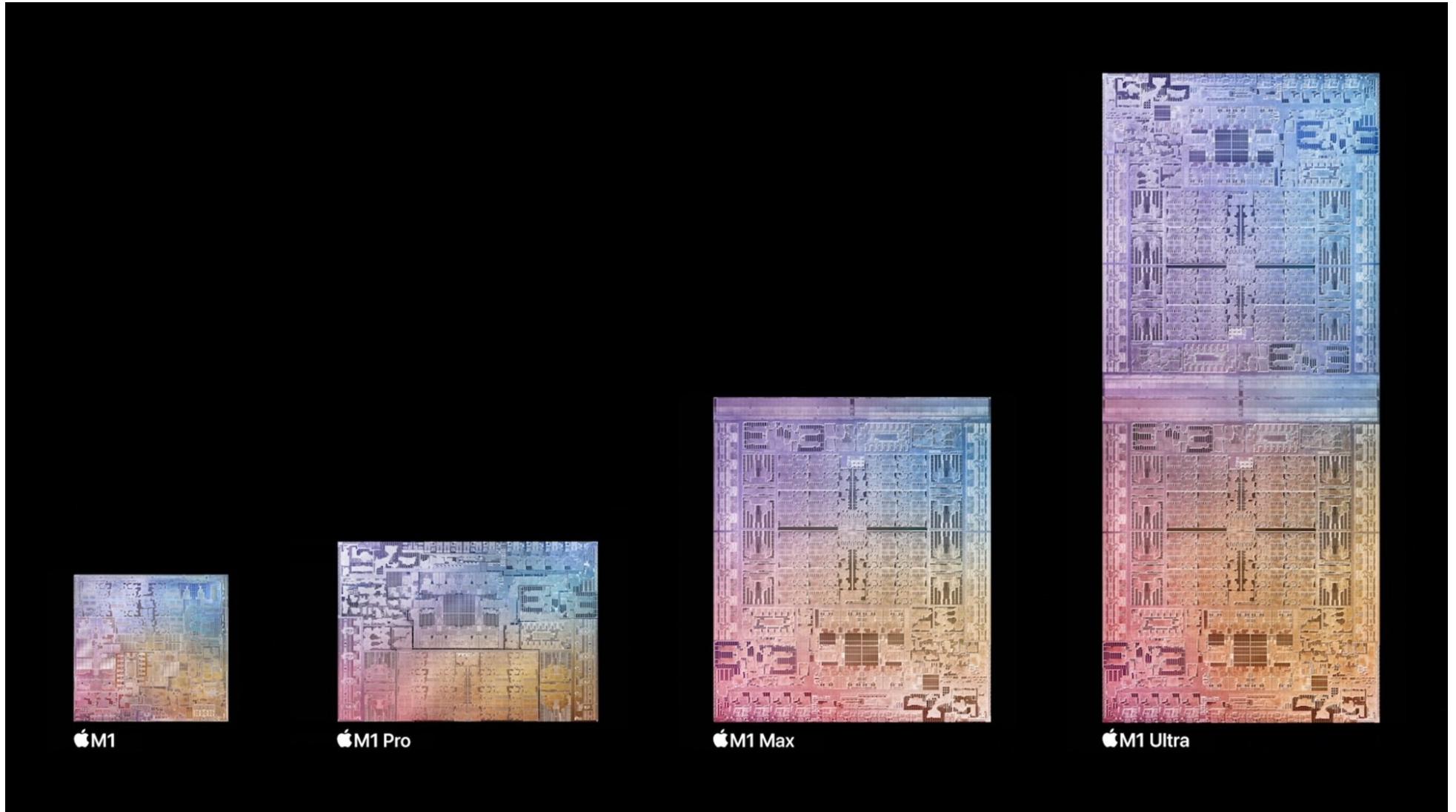
# Apple M1, M1 Pro and M1 Max (2020-2021)



Source: <https://www.golem.de/news/m1-pro-max-dieses-apple-silicon-ist-gigantisch-2110-160415.html>



# Up to Apple M1 Ultra (2022)





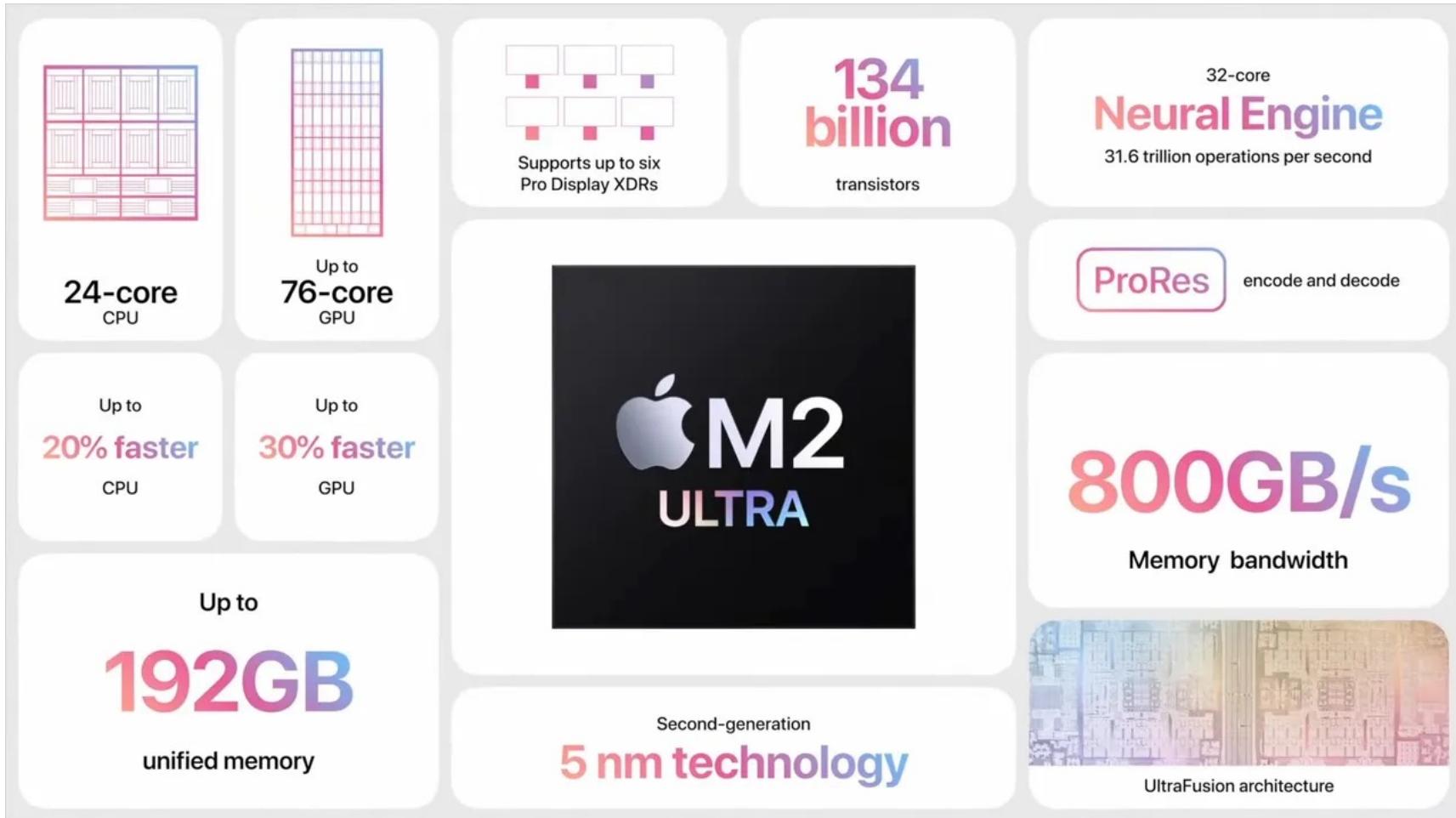
# Apple M1 Ultra (2022)

The infographic highlights the following features of the Apple M1 Ultra chip:

- ProRes**: Encode and decode.
- Thunderbolt 4**
- 5 nm process**
- 114 billion Transistors**
- 2.5TB/s interprocessor bandwidth** via a **Silicon interposer**.
- 800GB/s Memory bandwidth**.
- Up to 20-core CPU** and **64-core GPU**.
- 32-core Neural Engine** at **22 trillion operations per second**.
- Secure Enclave** (represented by a lock icon).
- Industry-leading performance per watt**.
- Up to 128GB unified memory**.
- UltraFusion architecture** (shown with a photograph of the die).



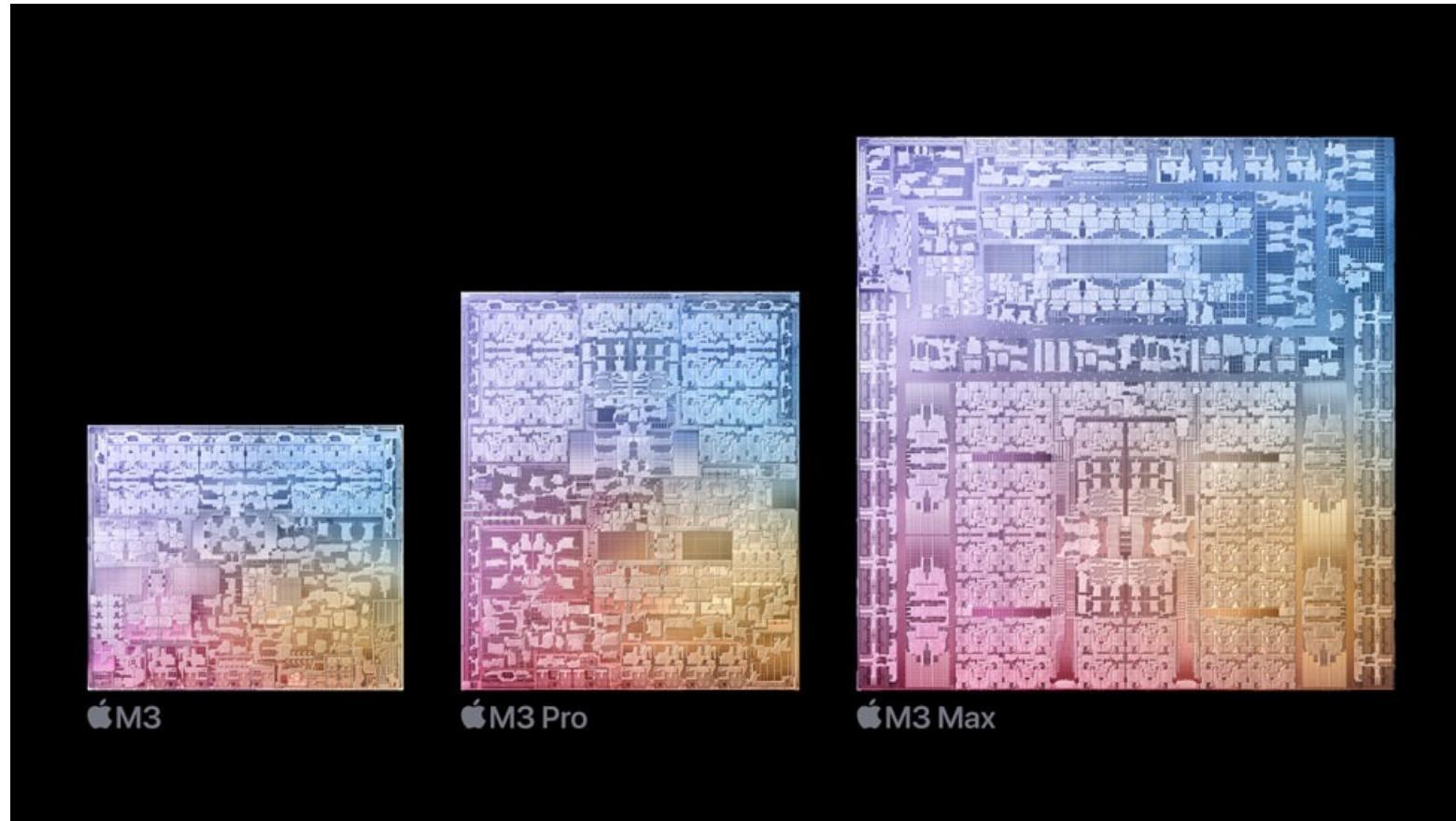
# Apple M2 Ultra (2023)





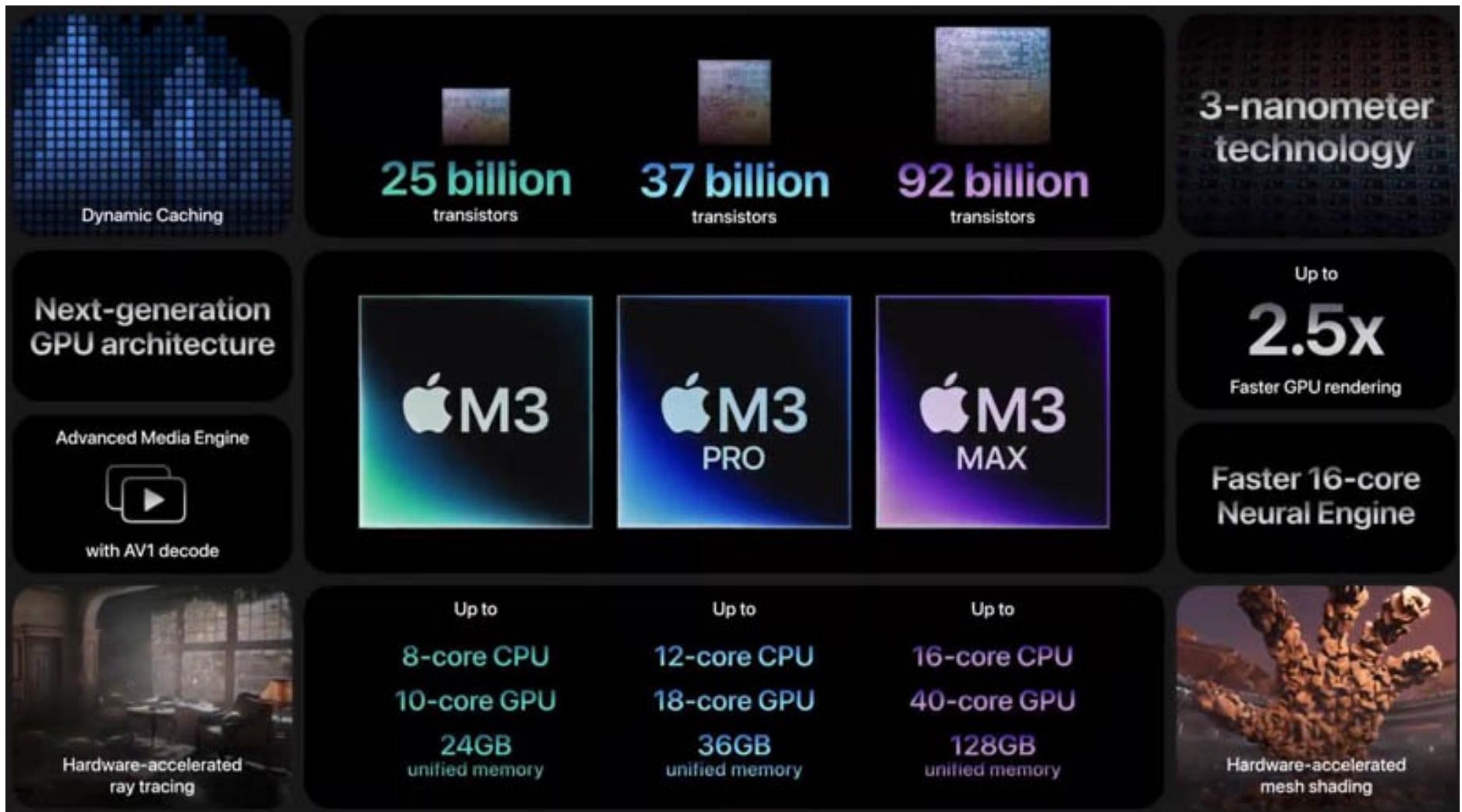
# Apple M3, M3 Pro and M3 Max (2023)

---





# Apple M3, M3 Pro and M3 Max (2023)





# Apple M4 announced in May 2024





# Apple M4 Pro and M4 Max announced in Oct. 2024

---

