

# MPSoCs

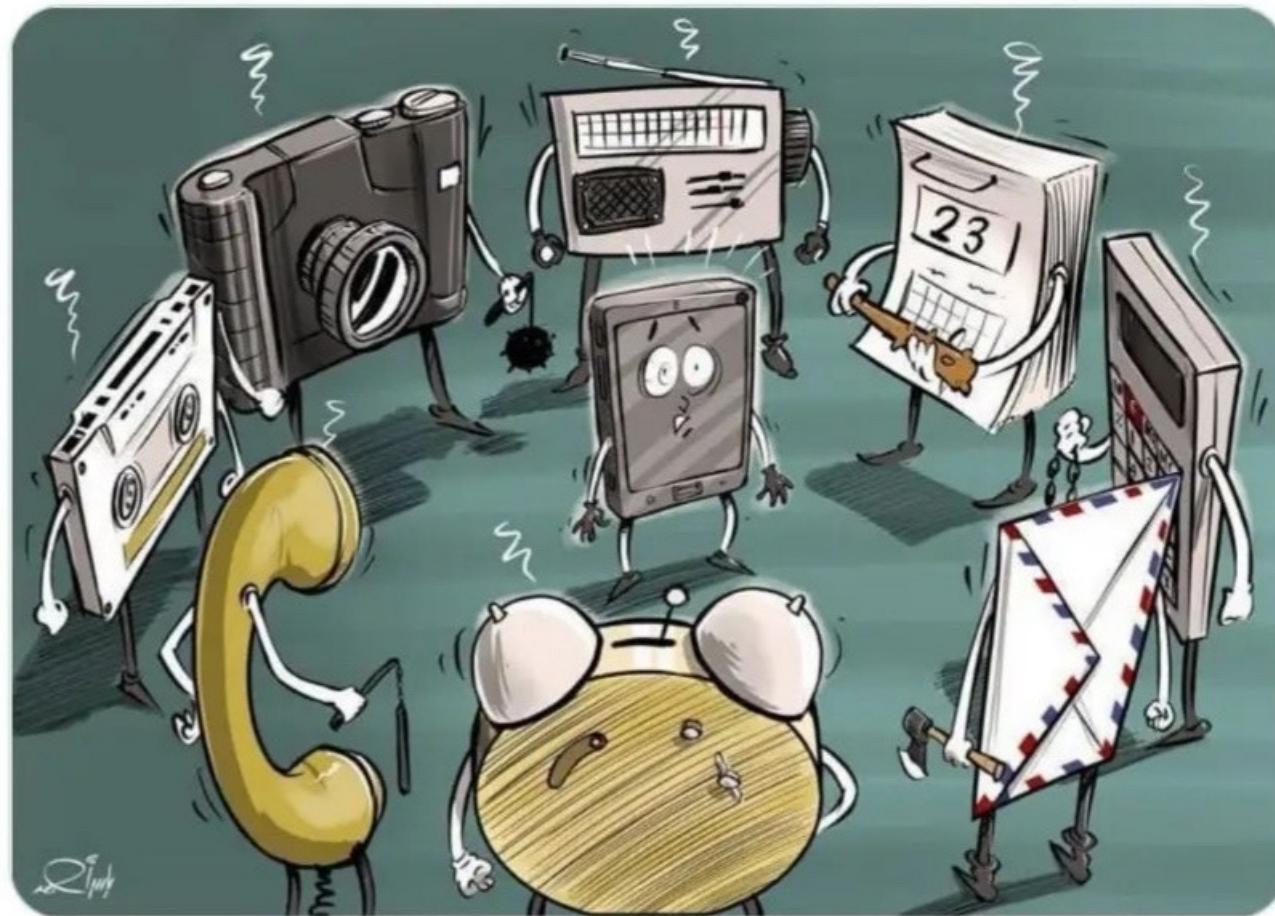
Fernando Gehm Moraes

PUCRS – Escola Politécnica

revisão: 30/abril/2022

# Múltiplas aplicações em um só dispositivo

Então você é o cara que tirou nossos empregos!

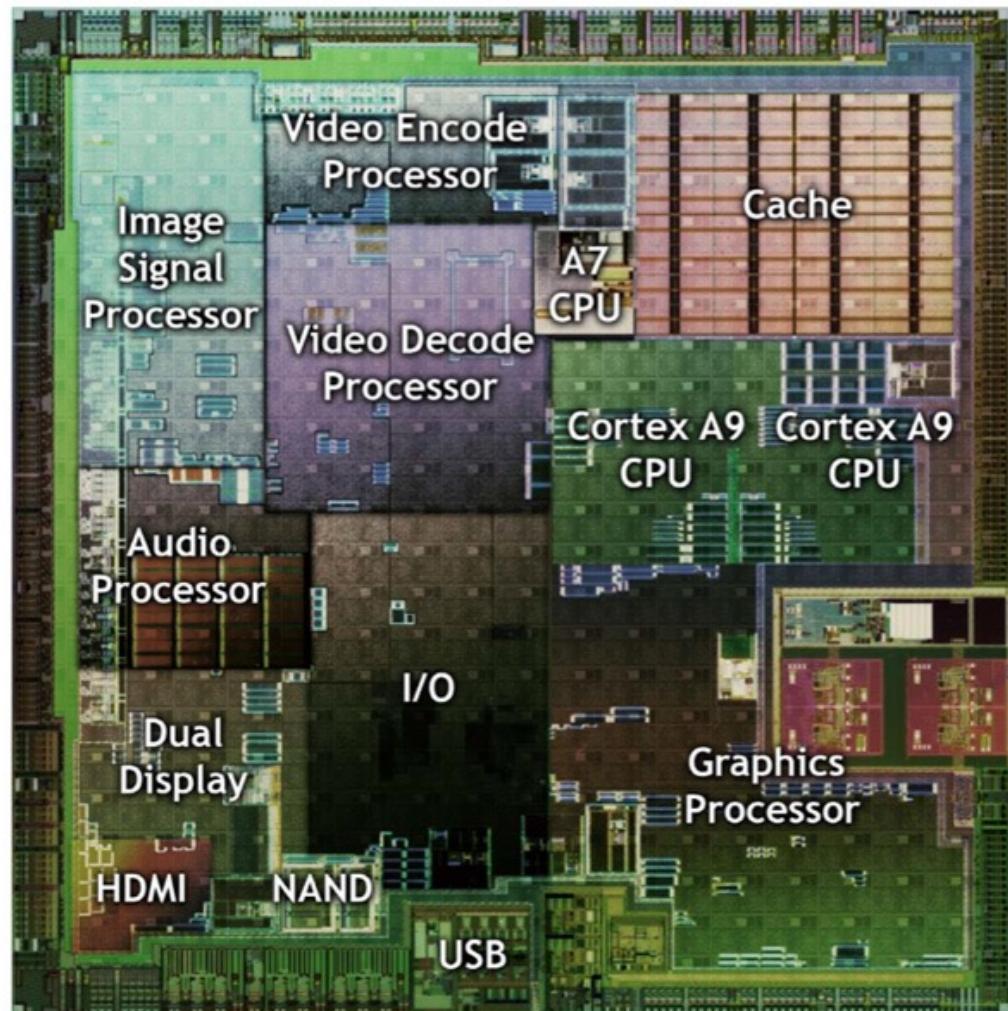


# O que é um SOC (System-on-a-chip)?

- Todo sistema integrado em um único chip
- Projeto modular
- Reuso de IPs de diferentes fabricantes
- Exemplo: Tegra 2 da Nvidia (2011)
- Foco em smartphones

## Um SOC pode conter:

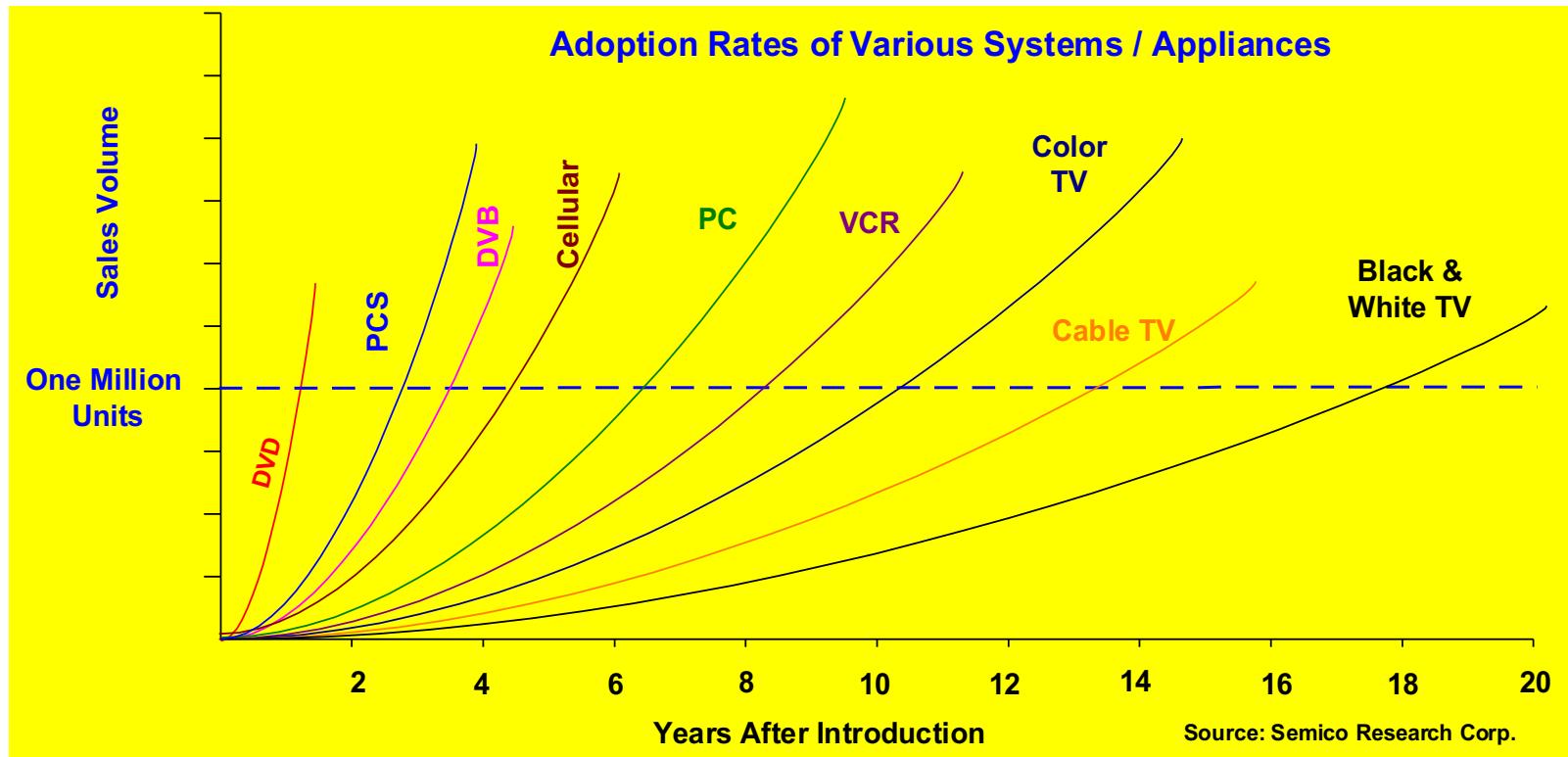
- Módulos de propriedade intelectual reutilizáveis
- Processador(es) embarcado(s)
- Memória embarcada
- Software
- Interfaces com o mundo externo (USB, PCI, Ethernet)
- Blocos analógicos
- Hardware programável (FPGAs)



Fonte: <https://www.bdti.com/InsideDSP/2011/10/20/NvidiaQualcomm>

# Porque SoCs?

- Time-to-market!



**E o IPAD? 28 dias para vender 1 milhão de unidades!**  
**PS4 sells over 300,000 in first two days on sale in Japan**

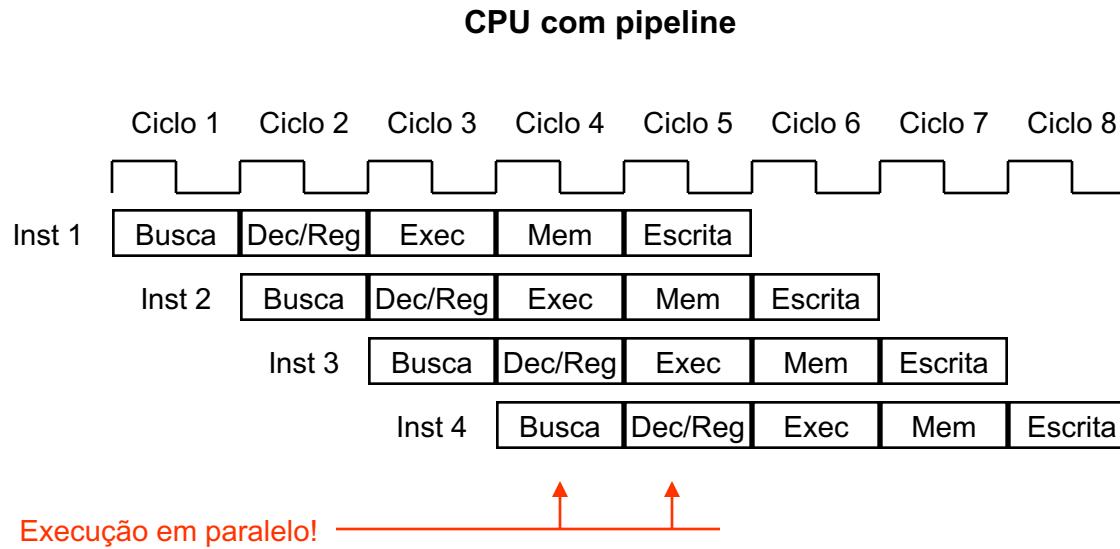
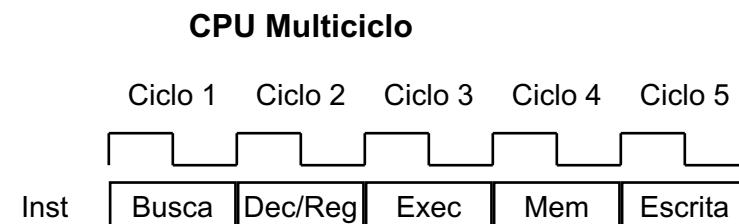
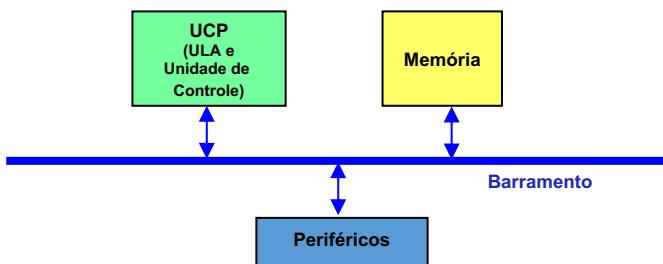
# Evolução dos Processadores

- Como ampliar o número de transístores em um CI?
  - Diminuir o tamanho dos transístores
  - Aumentar a área do chip
  - Aumentar a densidade (transistores/área)
  - Fazer chips tridimensionais
  - Usar vários dies (pedaços de silício)
- E quais problemas acompanham essas estratégias?
  - Dissipação de potência
  - Consumo de energia
  - Sincronismo (distribuição do clock)
  - Distribuição da alimentação
  - Defeitos de fabricação
  - Falhas transitórias

# Evolução dos Processadores

- E os processadores possuem problemas adicionais!
  - Explorar paralelismo de forma eficiente
  - Controlar todas as tarefas das múltiplas unidades
  - Evitar paradas nas unidades de processamento
  - Manter compatibilidade com arquiteturas anteriores
  - Definir política de gerência da hierarquia de memória
  - Viabilizar comunicação eficiente entre núcleos de processamento (...)

# Princípios básicos de paralelismo



# Pipeline

INSTRUCTION

CYCLE

1

2

3

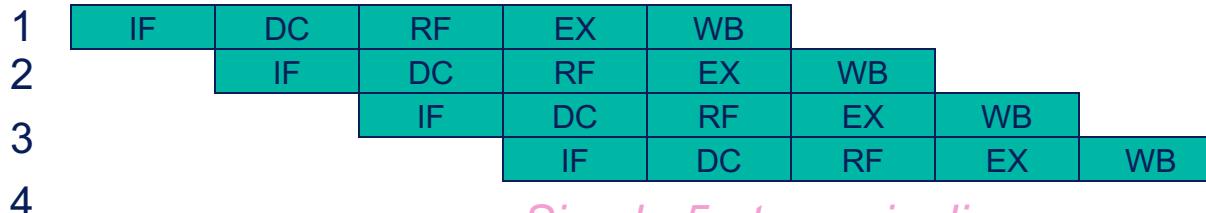
4

5

6

7

8



*Simple 5-stage pipeline*

IF: Instruction Fetch  
DC: Instruction Decode  
RF: Register Fetch  
EX: Execute instruction  
WB: Write Result Register

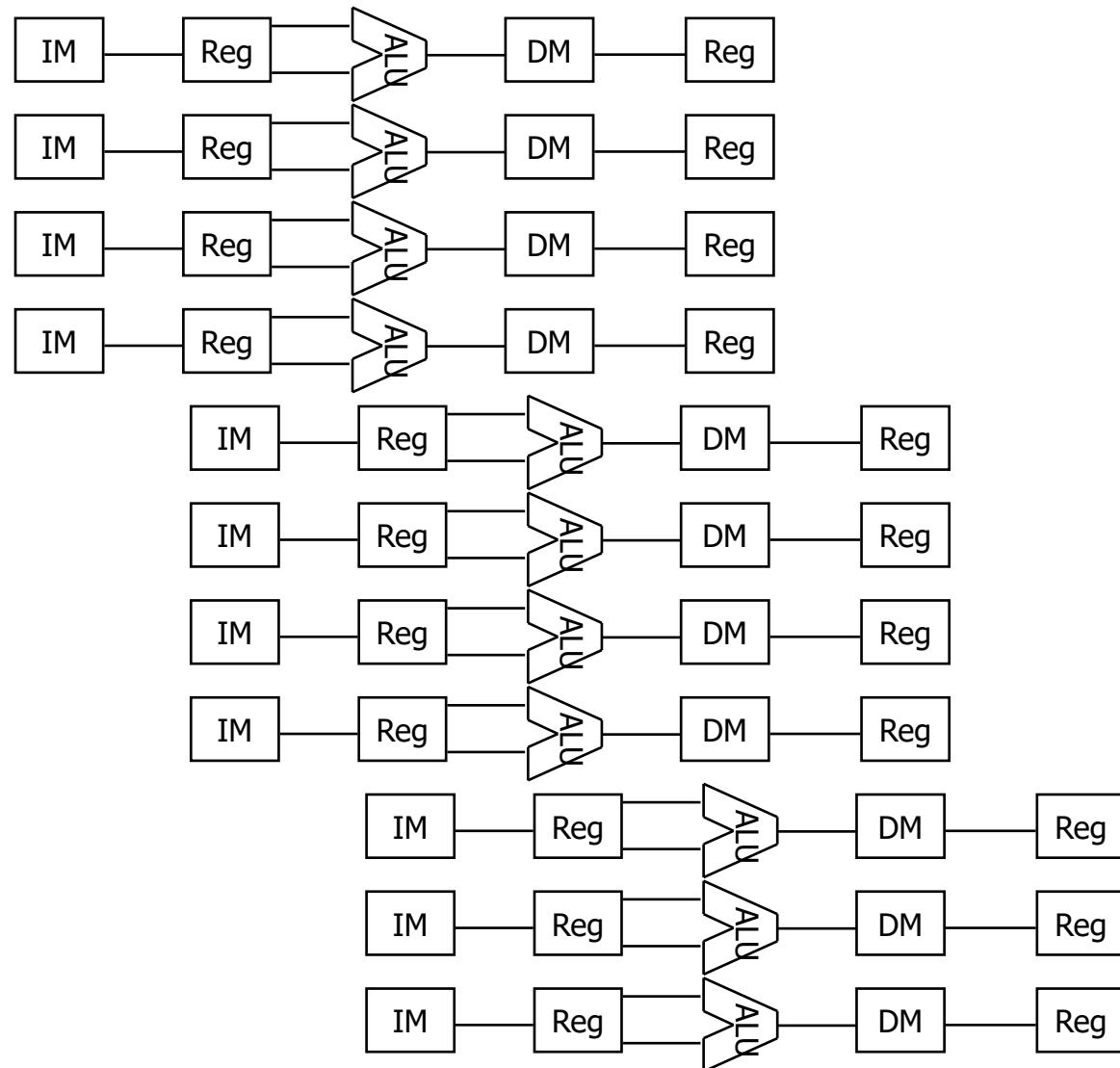
## ◆ Purpose of pipelining:

- Reduce #gate\_levels in critical path
- Reduce CPI close to one (instead of a large number for the multicycle machine)
- More efficient Hardware

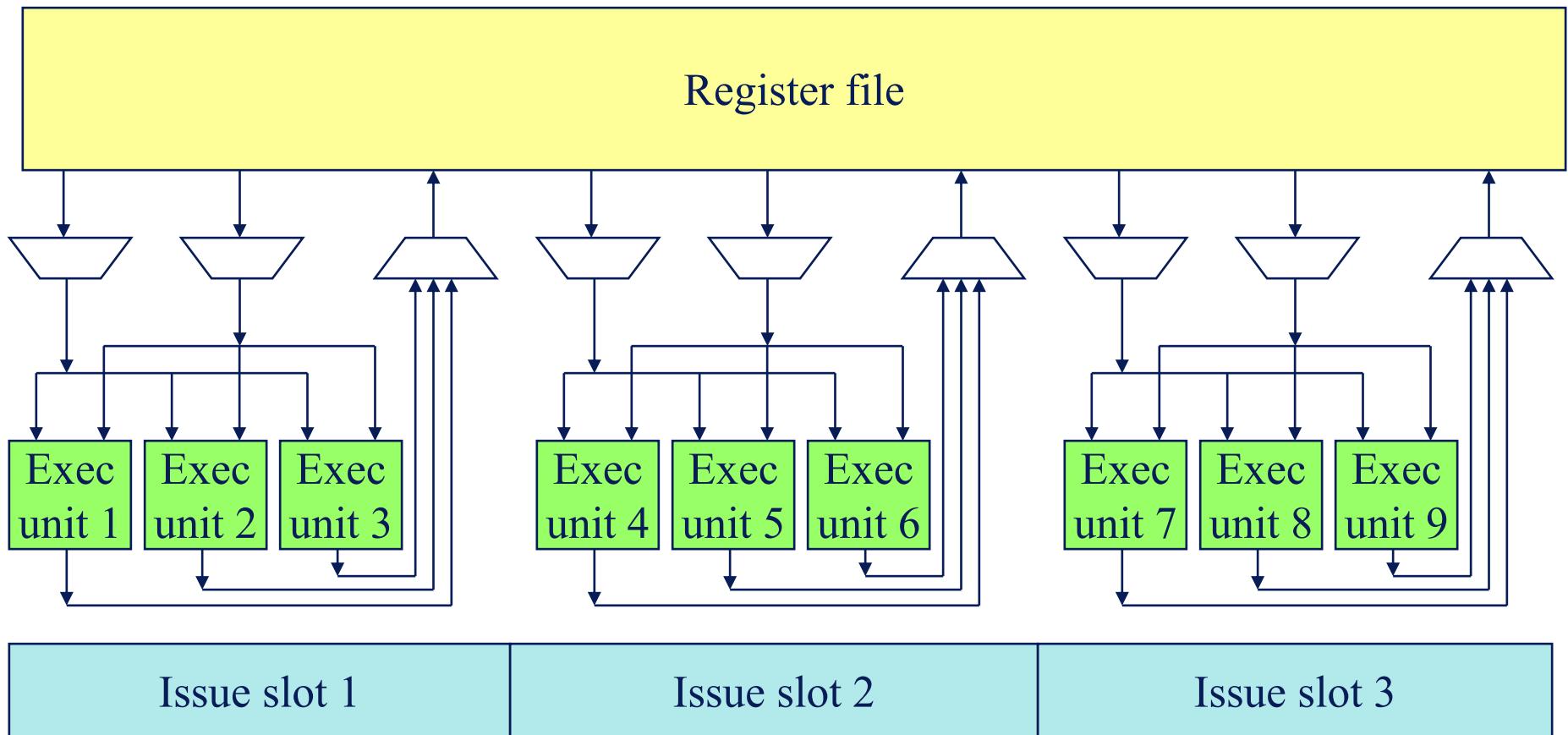
## ◆ Some bad news: Hazards or pipeline stalls

- Structural hazards: add more hardware
- Control hazards, branch penalties: use branch prediction
- Data hazards: by passing required

# Execução Superscalar



# Máquina VLIW



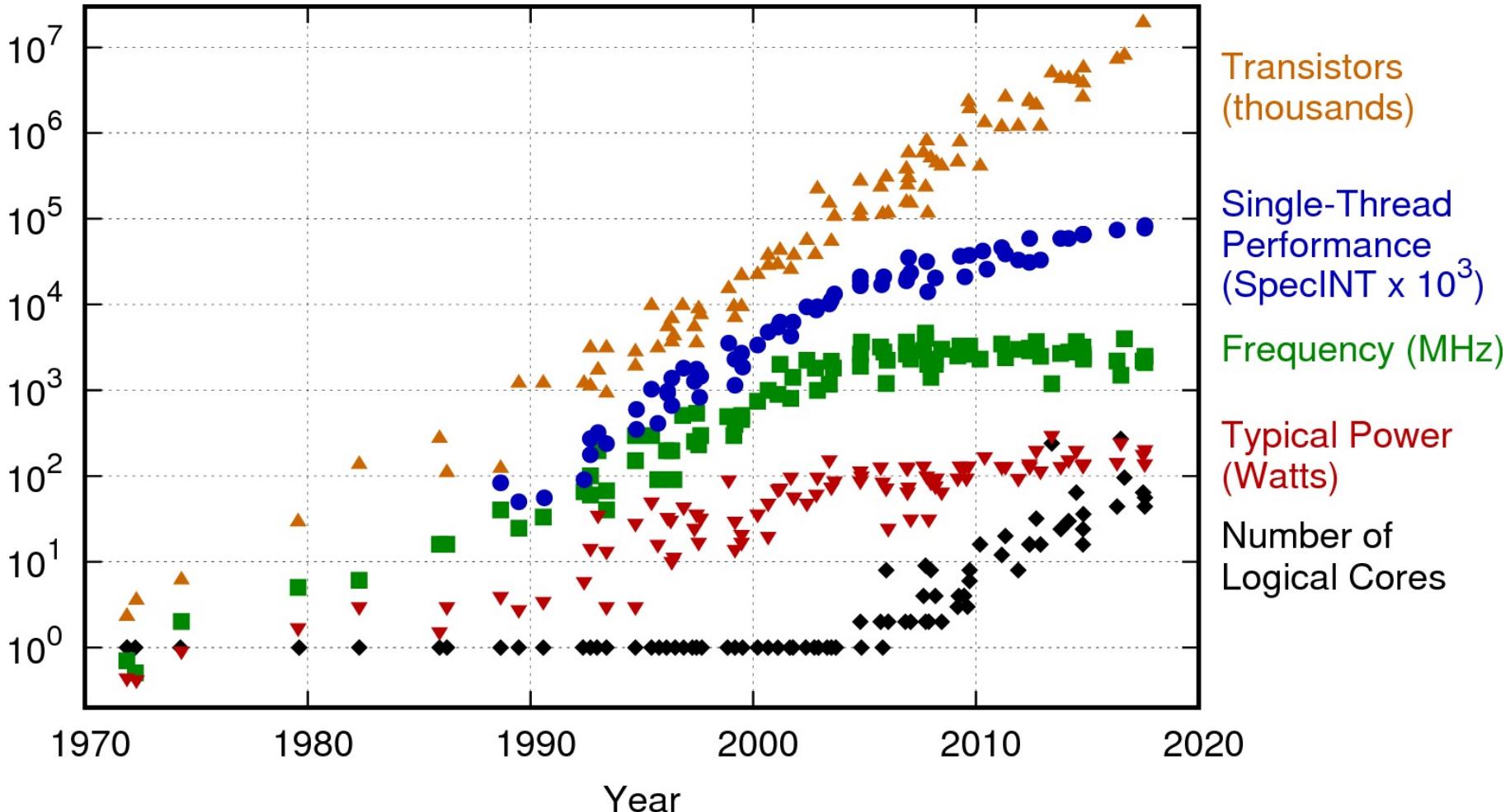
# Como processadores ficaram mais rápidos?

- superpipelining
- superscalar execution
- dynamic scheduling
- multilevel memory caching
- aggressive speculation
- DSP specialization
- fabrication technology



**LIMITAÇÕES TÉCNICAS E  
TECNOLÓGICAS!**

## 42 Years of Microprocessor Trend Data



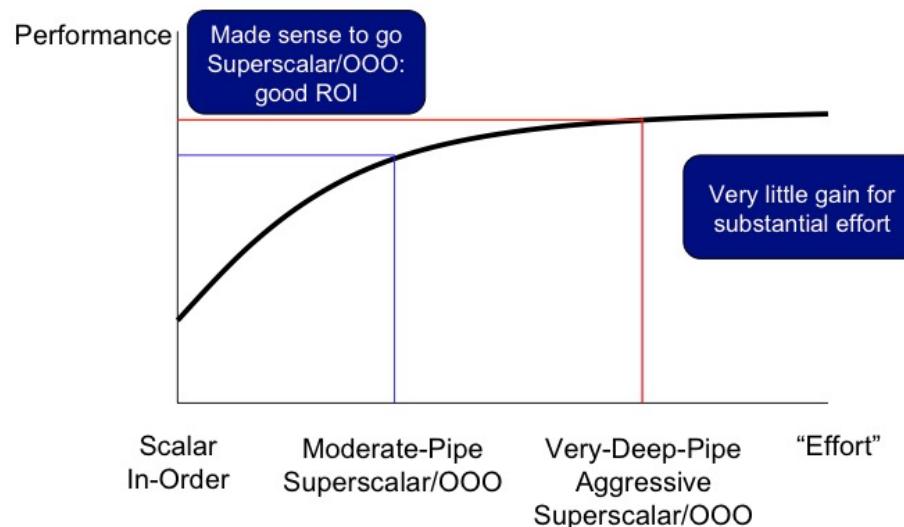
- #transistors follows Moore
- but **not** freq. and performance/core

# “1º MURO” - ILP WALL

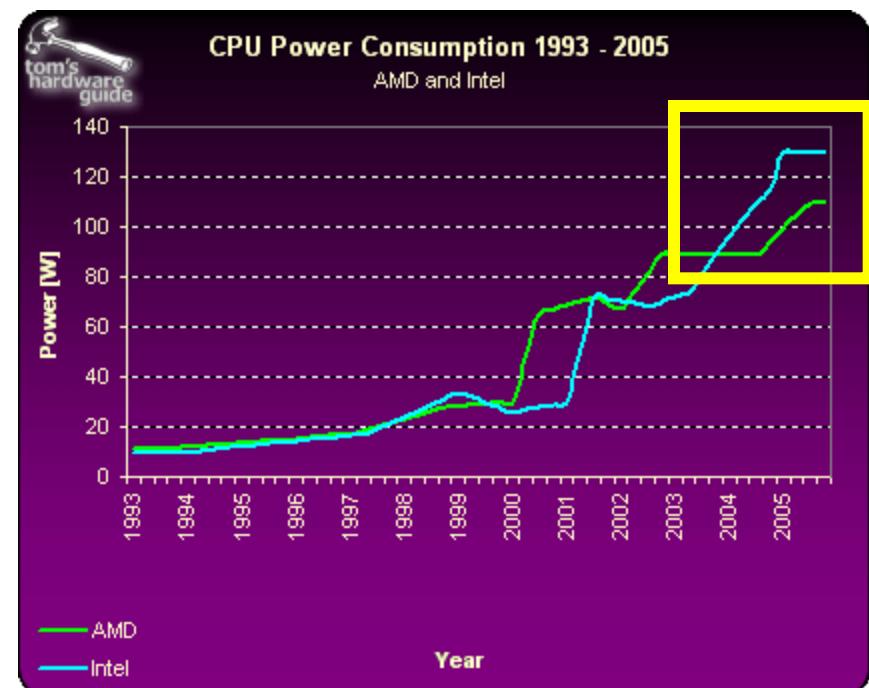
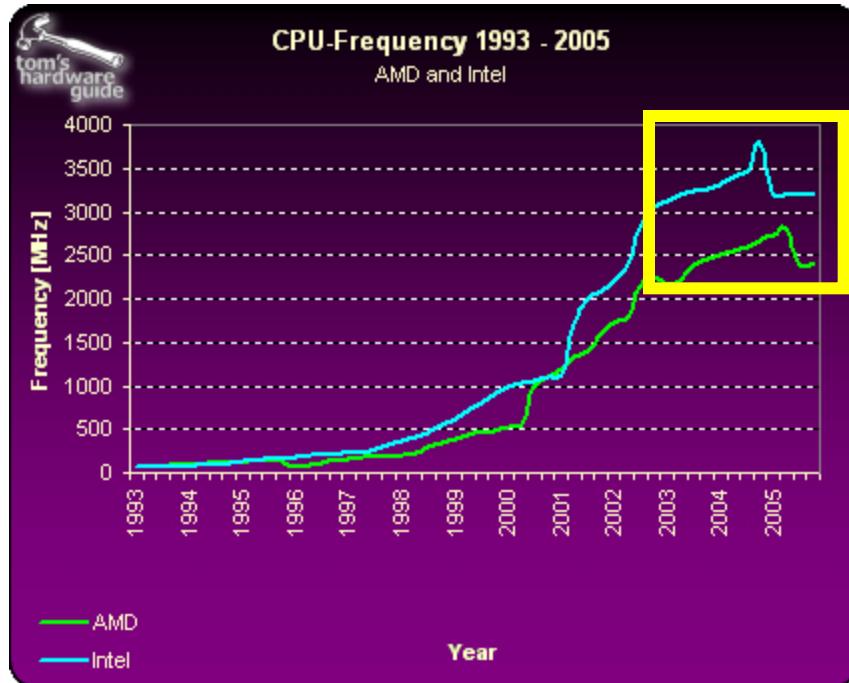
## ■ Influência do compilador!

- descobrir o paralelismo em máquinas superescalar e VLIW
- a ordem das instruções (dependência) afeta o desempenho

## ■ “ILP wall”



# “2º/3º MUROs”:FREQ. WALL / POWER WALL

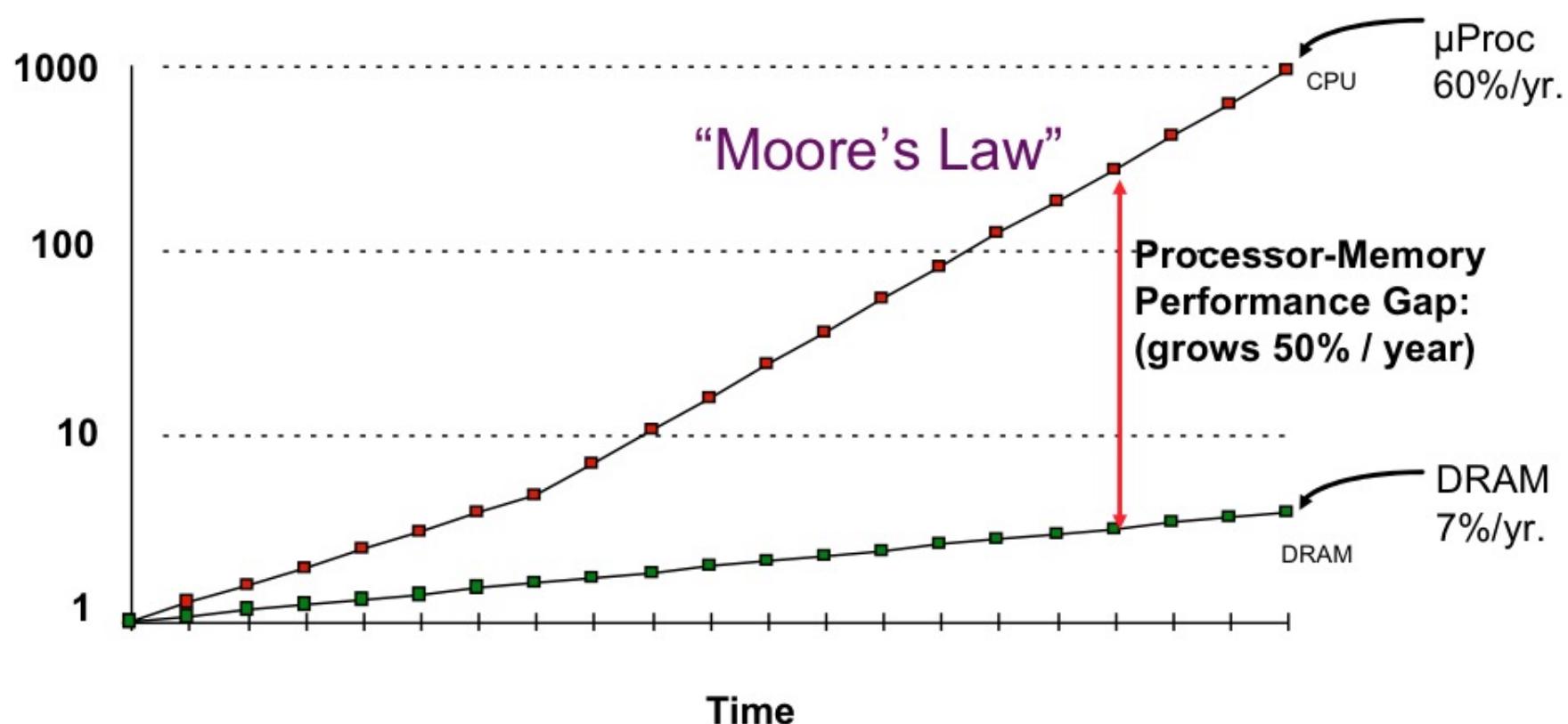


# “4º MURO” - MEMORY WALL

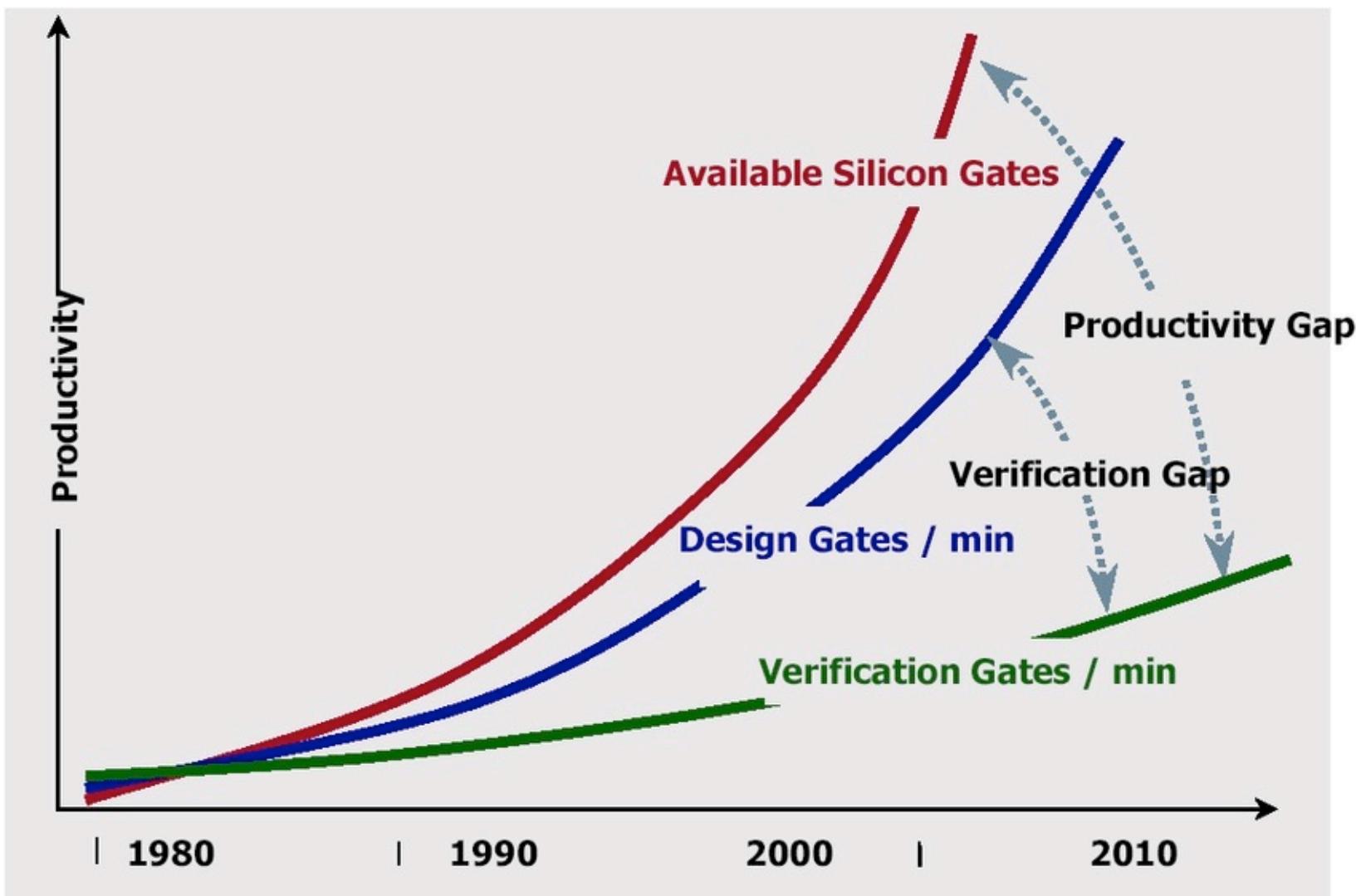
	Capacidade	Velocidade
Lógica:	2x em 3 anos	2x em 3 anos
DRAM:	4x em 3 anos	2x em 10 anos
Disco:	4x em 3 anos	2x em 10 anos

- Fato: memórias grandes são lentas, memórias rápidas são pequenas
- Como criar uma memória grande e rápida (pelo menos na maior parte do tempo)?
  - Hierarquia
  - Paralelismo

# “4º MURO” - MEMORY WALL

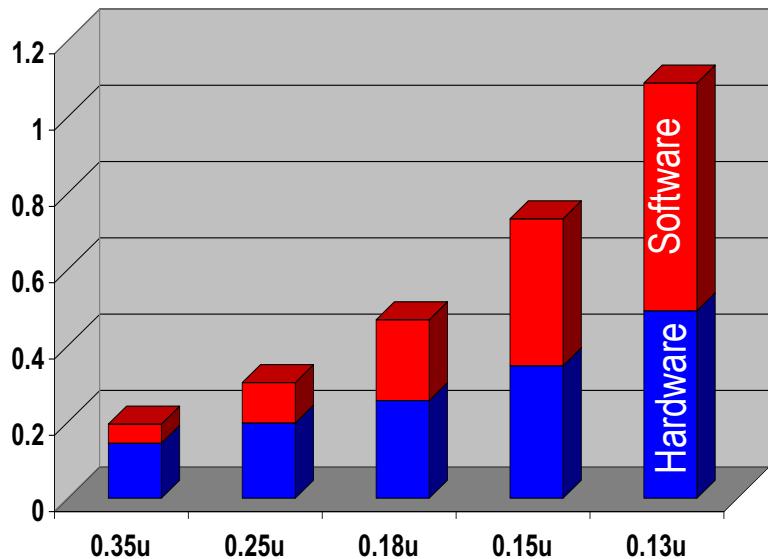


# DESIGN GAP

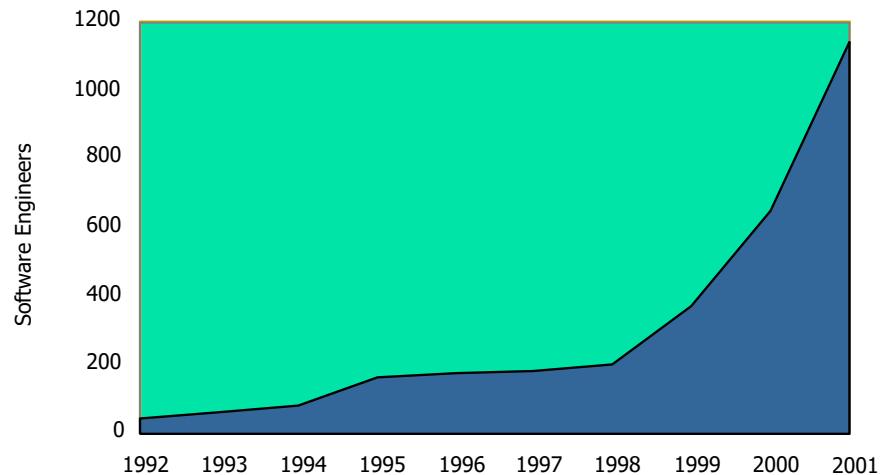


# MORE SOFTWARE THAN HARDWARE

## IC Hardware & Software Effort



## Growth in SW Engineers

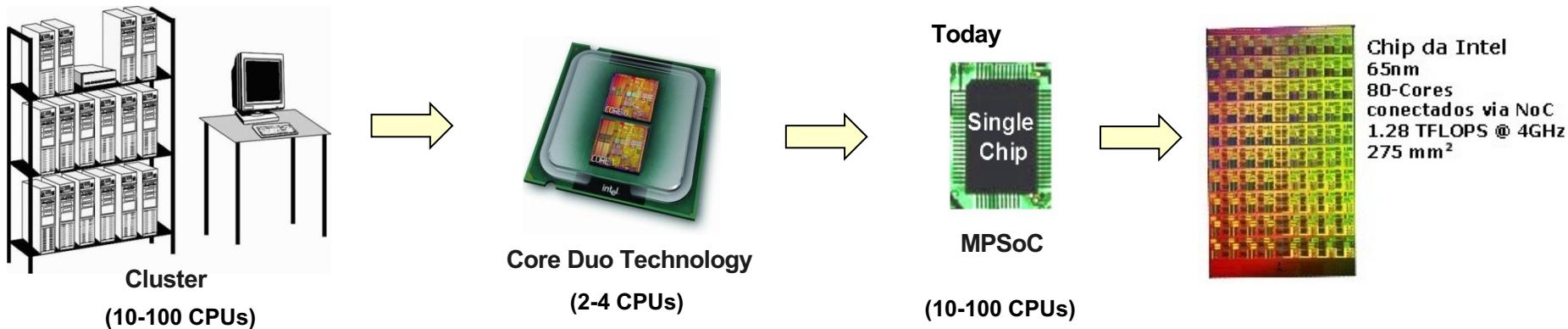


Source: Top Ten Semiconductor Supplier

**Solution? Probably one are many-cores/SMP/MPSoCs**

# MPSoC – Multiprocessor system-on-chip

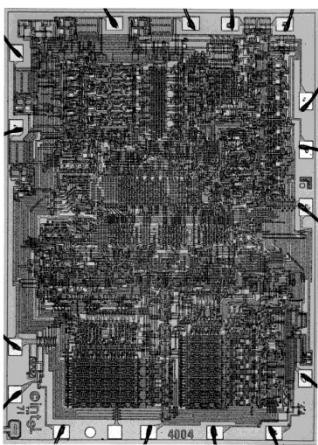
- Sistema multiprocessado integrado como um SoC
- Elementos de processamento + IPs + Infraestrutura de Comunicação
- Evolução de *Clusters* (+ recursos = + desempenho)



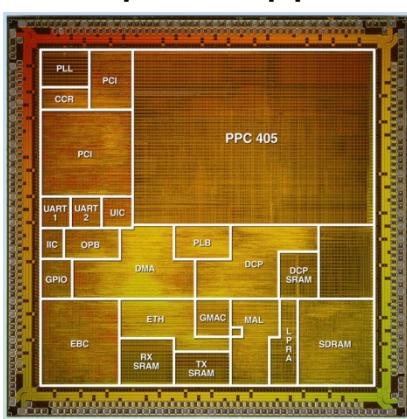
# Design Evolution

## ASICs:

Dedicated hw  
1 algorithm



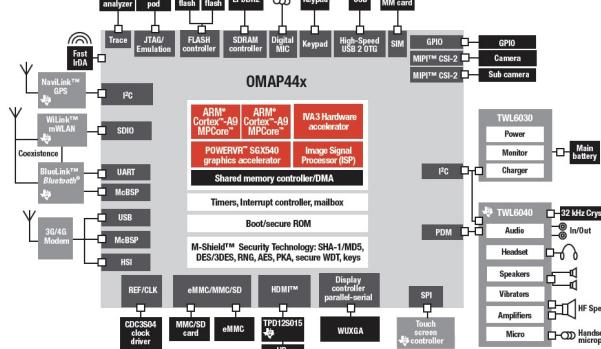
70's



80's

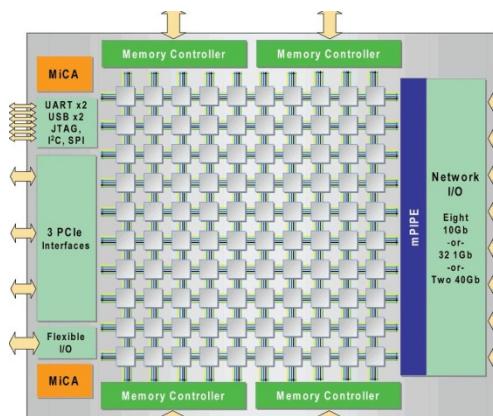
**Single microprocessor SoC:**  
Complete application

**Bus-based MPSoC**  
Platform design  
Target multiple applications



90's

**NoC-based MPSoC**  
Many applications  
Dynamic behavior  
100's PEs



00's

# Então, porque utilizar MPSoCs ?

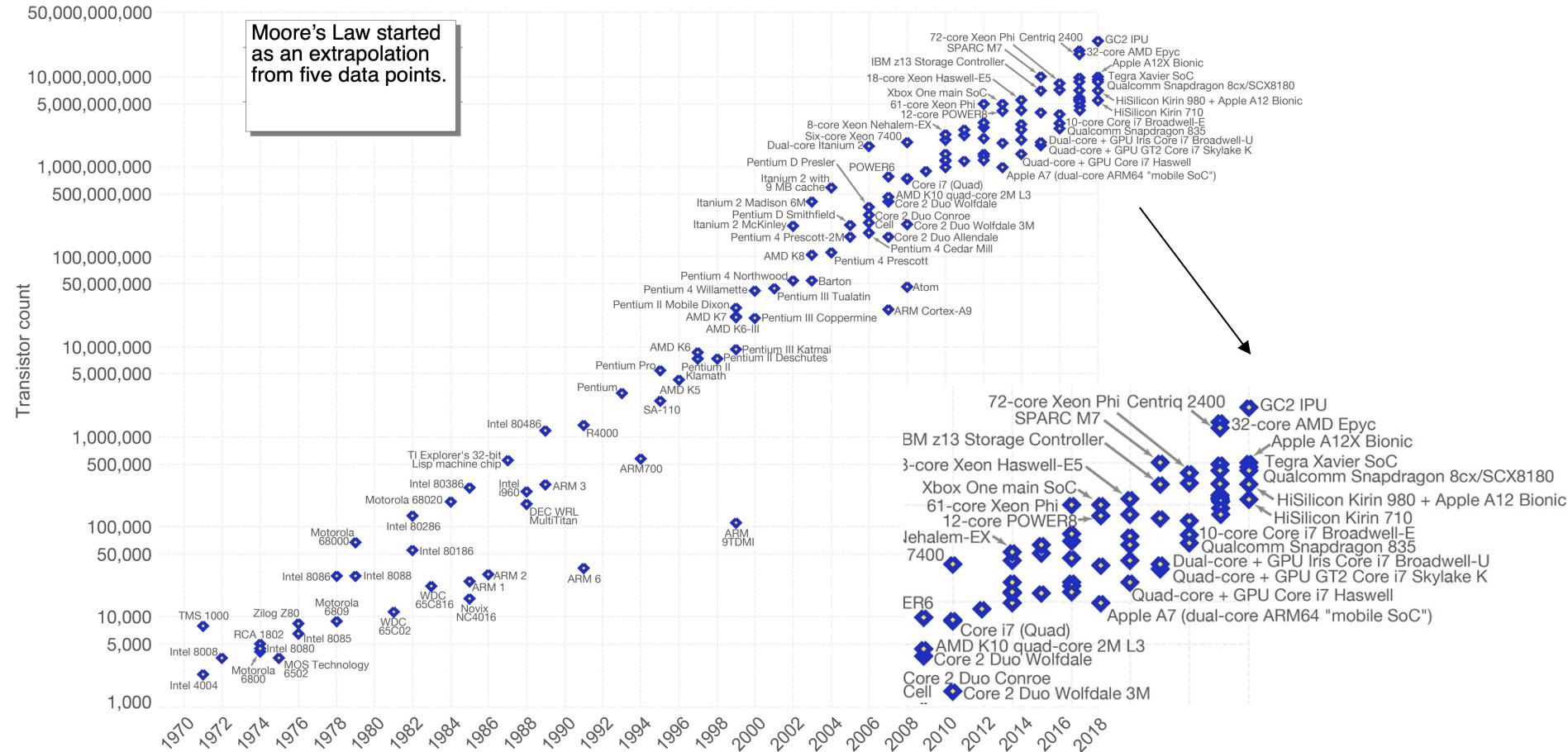
- Quebram o “ILP wall”
  - Múltiplas threads/tarefas executam simultaneamente
  - Paralelismo de *grão-grande* (não mais no nível de instrução)
- Quebram o “frequency wall” e o “power wall”
  - Múltiplos PEs mais lentos e mais simples
  - Escalabilidade obtida pelo aumento do número de PEs e não pelo aumento da freqüência
  - Uso de processamento heterogêneo pra aumento de desempenho/redução de potência
- Quebram o “memory wall”
  - Hierarquia de memórias distribuídas e adaptadas às aplicações
  - Replicação do sistema operacional nos PEs
- Diminuem o “design gap”
  - *Design gap*: capacidade de integração ultrapassa a capacidade de projeto
  - Replicação dos PEs diminui os custos de projeto de verificação
  - Aumenta a importância da rede de comunicação

# MPSoCs are here

## Moore's Law – The number of transistors on integrated circuit chips (1971-2018)

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

Our World  
in Data



Data source: Wikipedia ([https://en.wikipedia.org/wiki/Transistor\\_count](https://en.wikipedia.org/wiki/Transistor_count))

The data visualization is available at [OurWorldInData.org](http://OurWorldInData.org). There you find more visualizations and research on this topic.

Licensed under CC-BY-SA by the author Max Roser.

# INTEL – evolução histórica

Year	Processor	Transistors	Feature size	Data Width	Frequency	Features
1971	4004	2300	10000nM	4	740 KHz	First Microprocessor
1978	8086	29000	3000nm	16	10MHz	IBM PC/AT
1985	80386	275000	1000nm	32	33MHZ	Pipelining
1989	80486	1200000	800nm	32	100MHz	Integral FPU
1993	Pentium	3100000	800nm	32	150 MHz	On-Chip L1 Cache; Superscalar
1995	Pentium Pro	5500000	600nm	32	200MHz	Out-of-order execution
1997	Pentium MMX P55C	4500000	350nm	32	450MHz	Dynamic branch prediction; MMX (SIMD) instructions
1999	Pentium III	28000000	180nm	32	1.1GHz	On-chip L2 Cache
2004	Pentium 4E	125000000	90nm	32	3.8 GHz	Hyper-threading
2006	Xeon Tulsa	167000000	65nm	64	3.4 GHz	Dual-Core
2010	Xeon 7500 Nehalem	2300000000	45nm	64	2.26GHz	Eight-Cores

# TOP 500

<https://www.top500.org/statistics/treemaps/>

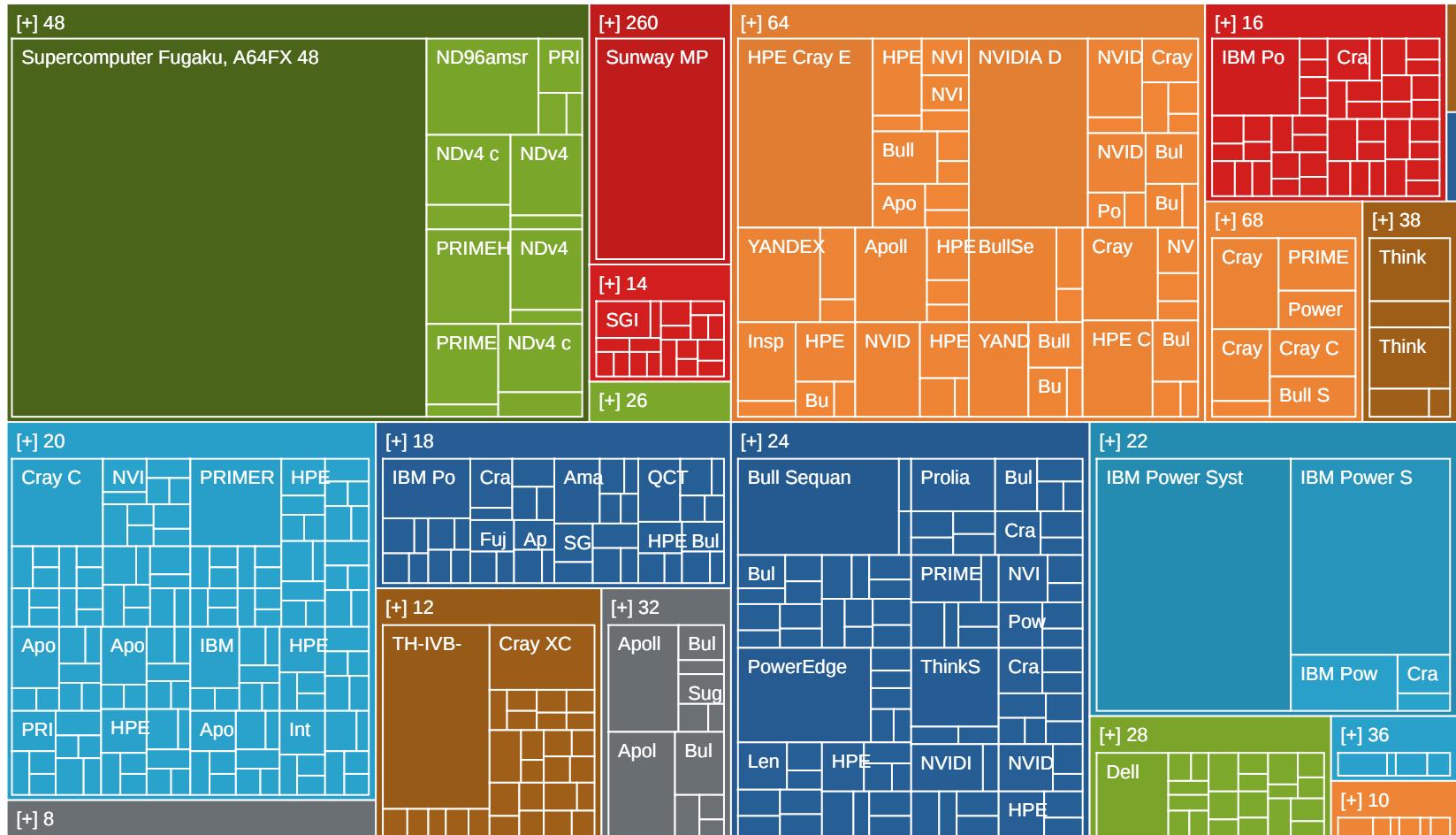
The TOP500 list the the 500 most powerful commercially available computer systems known



# TOP 500

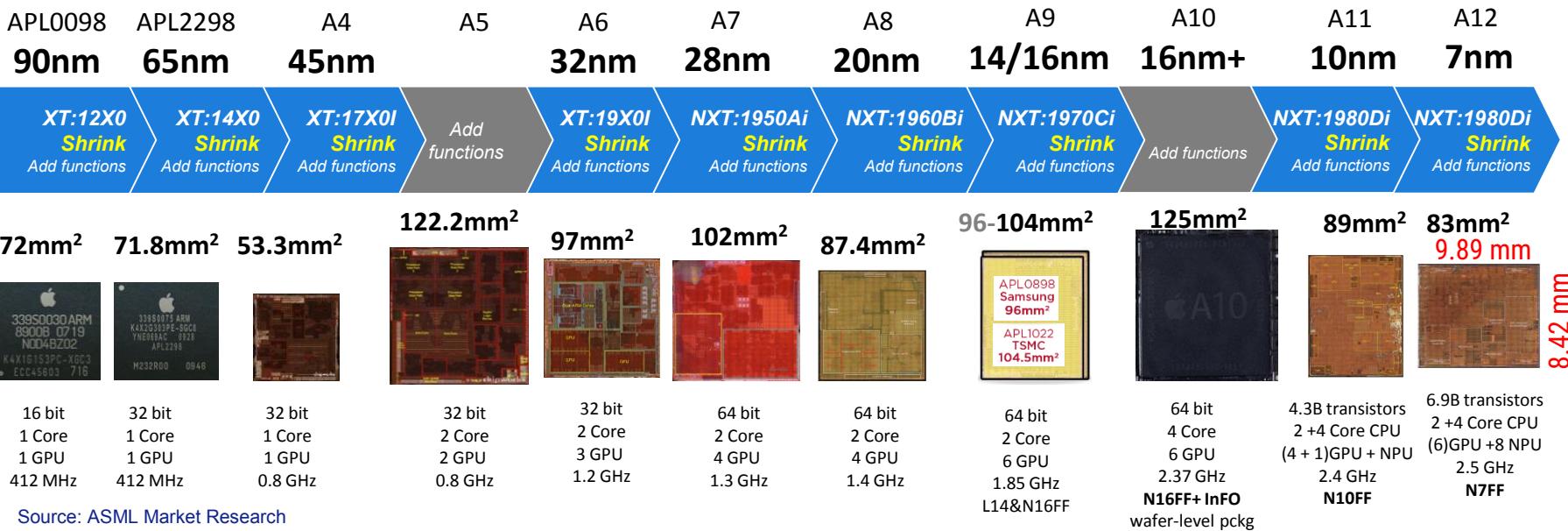
<https://www.top500.org/statistics/treemaps/>

## CORES POR CHIP



# Transistor size continues to shrink

Public  
Slide 21  
20 March 2019



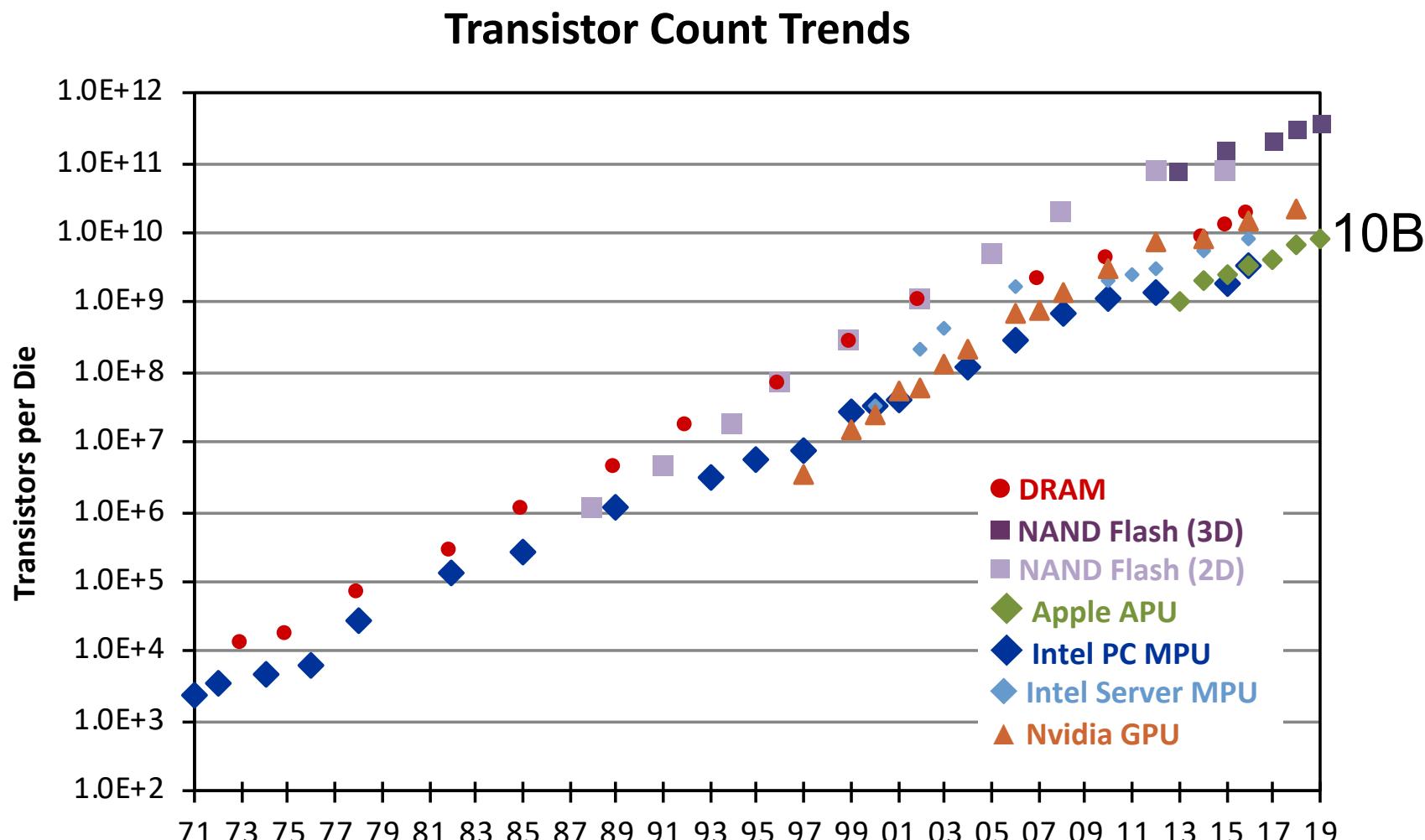
Source: ASML Market Research

TWINSCAN NXT:1980Di  
193-nm Step and Scan (Resolution: ≤ 38 nm)

<https://en.wikichip.org/wiki/apple/ax/a12>

[https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=2ahUKEwjA6o\\_AxP\\_oAhWNGLkGHaeID5kQFjABegQIAhAB&url=https%3A%2F%2Fwww.asml.com%2F-%2Fmedia%2Fasml%2Ffiles%2Finvestors%2Fpast-events-and-presentations%2Fasml\\_20190319\\_2019-03-20\\_baml\\_taiwan\\_mar\\_2019\\_v1\\_final.pdf&usg=AOvVaw3rEbc4uaCfGIHnU6z4CY5H](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=2ahUKEwjA6o_AxP_oAhWNGLkGHaeID5kQFjABegQIAhAB&url=https%3A%2F%2Fwww.asml.com%2F-%2Fmedia%2Fasml%2Ffiles%2Finvestors%2Fpast-events-and-presentations%2Fasml_20190319_2019-03-20_baml_taiwan_mar_2019_v1_final.pdf&usg=AOvVaw3rEbc4uaCfGIHnU6z4CY5H)

# Designs with one trillion transistors?

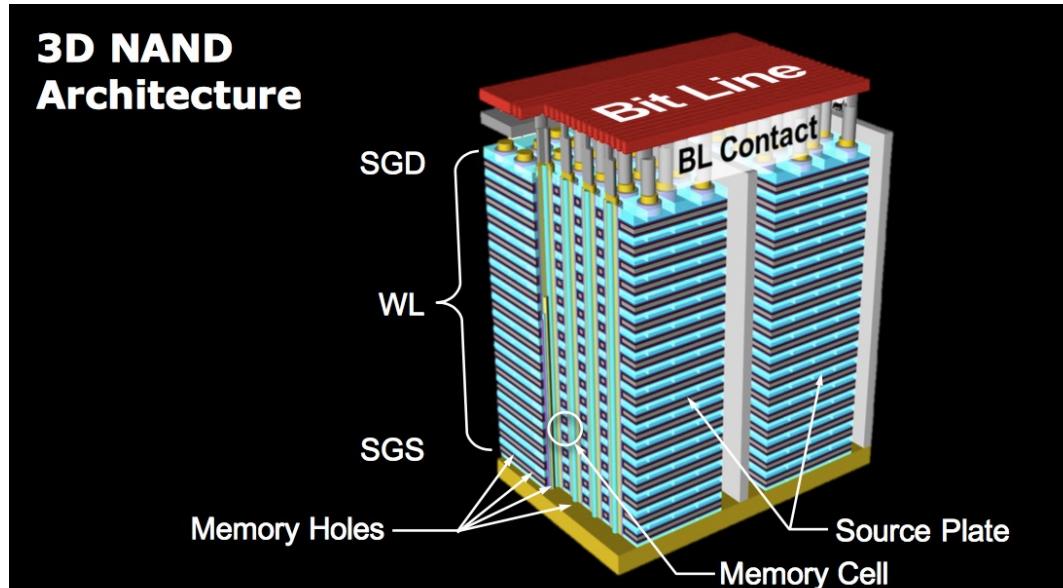


Sources: Intel, SIA, Wikichip, IC Insights

# 3D NAND architecture

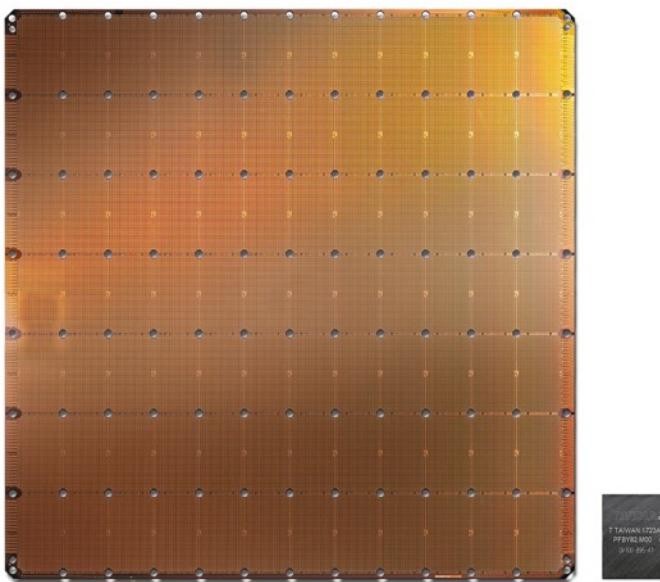
3D NAND is quantified by the number of layers stacked in a device. As more layers are added, the bit density increases. Today, 3D NAND suppliers are shipping 64-layer devices, although they are now ramping up the next technology generation, which has 96 layers. And behind the scenes vendors are racing to develop and ship the next iteration, 128-layer products, by mid-2019, analysts said.

<https://semiengineering.com/3d-nand-flash-wars-begin/>

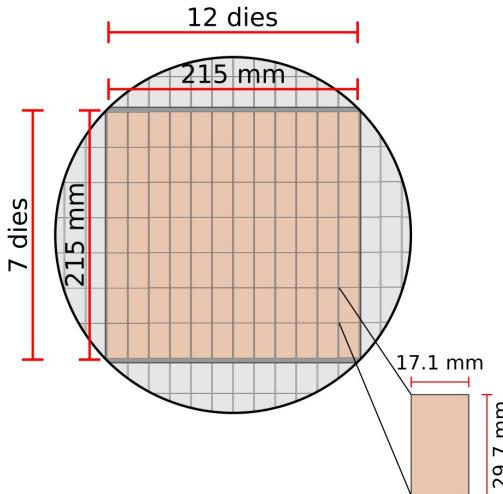


Year	2016-2017	2018-2019		2020-2021	2022-2023
<b>Generation 3D</b>	L48	L64	L96	L128	512
<b>Die size (3b/cell)</b>	256-512 Gb	512Gb – 1Tb	512Gb-2Tb	1-3Tb	2-6 Tb
<b>Hole CD</b>	65-100	65-100	65-100	65-100	65-100
<b>Slit pitch (# holes)</b>	4	4	4-8	8	8
<b>Vertical pitch</b>	50-70nm	40-60	40-60	40-50	40-50
<b>BL CD</b>	20	20	20 - 40	~40	~40
<b>Multiple stacks</b>	No	No	No	No	Yes (2-4)
					Yes (4-8)

# Cerebras – <https://www.cerebras.net>



- TSMC 16nm, 84 dies
- The WSE (Wafer Scale Engine ) is 215 mm by 215 mm



**CS-1 is powered by the Cerebras Wafer Scale Engine - the largest chip ever built**

**56x the size of the largest Graphics Processing Unit**

The Cerebras Wafer Scale Engine is 46,225 mm<sup>2</sup> with 1.2 Trillion transistors and 400,000 AI-optimized cores.

By comparison, the largest Graphics Processing Unit is 815 mm<sup>2</sup> and has 21.1 Billion transistors.

**Consumo de potência máxima: 20 kW**

**Purpose-built for Deep Learning: enormous compute, fast memory and communication bandwidth**

**46,225 mm<sup>2</sup> chip**

56x larger than the biggest GPU ever made

**400,000 core**

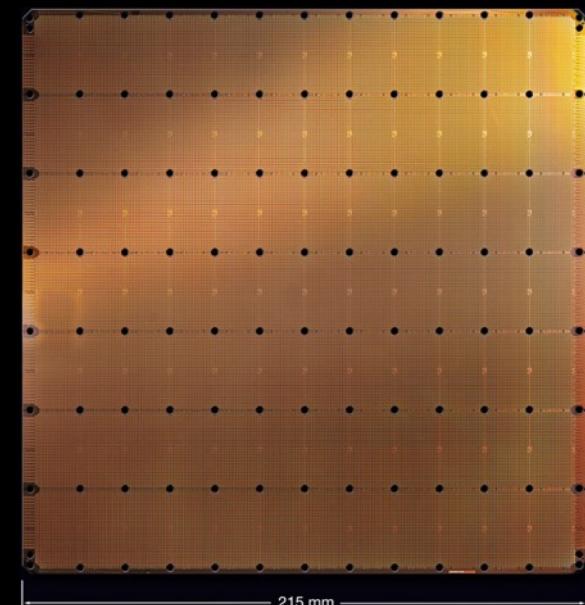
78x more cores

**18 GB on-chip SRAM**

3000x more on-chip memory

**100 Pb/s interconnect**

33,000x more bandwidth



# Previsão da evolução dos MPSoCs

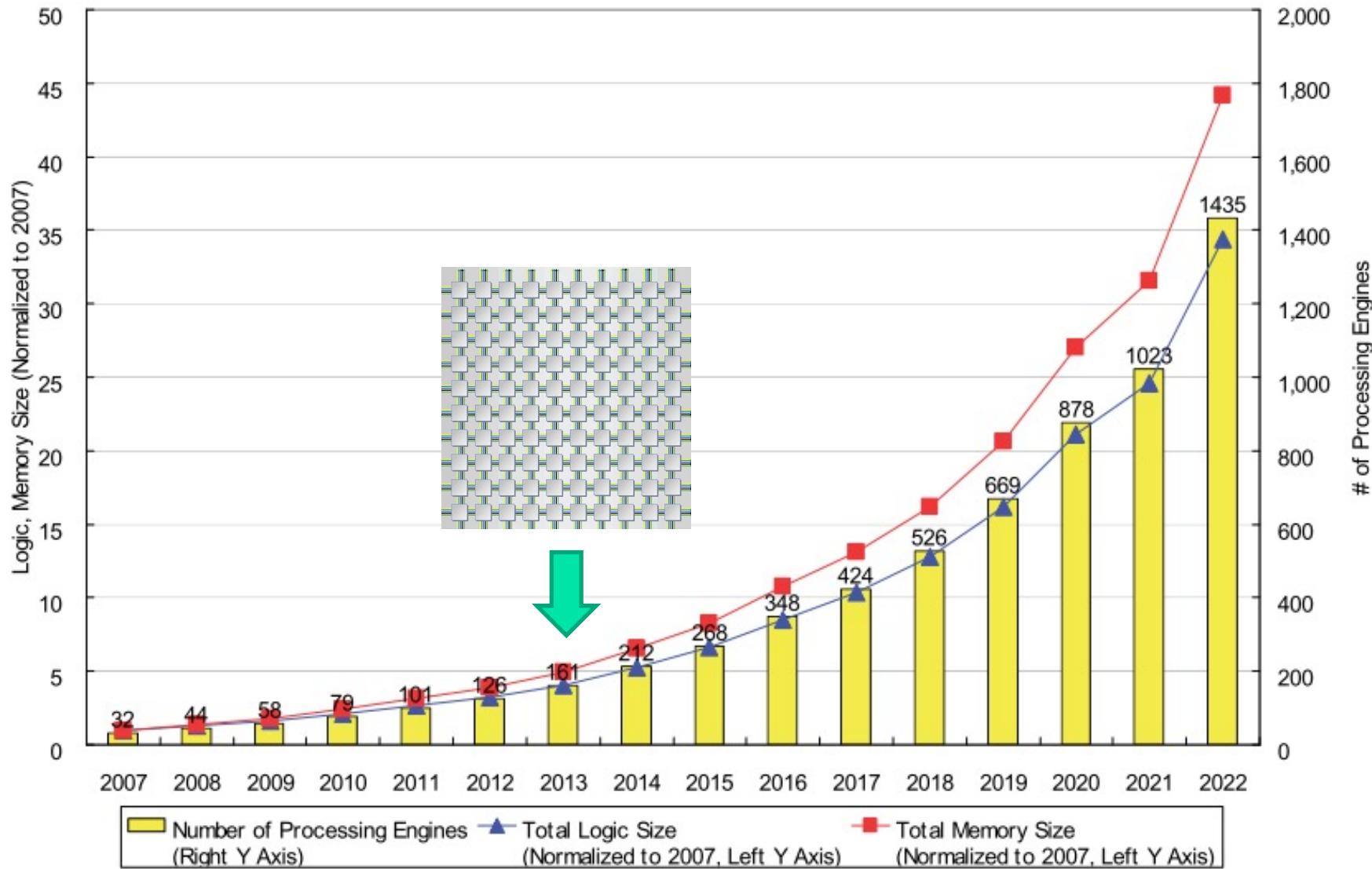


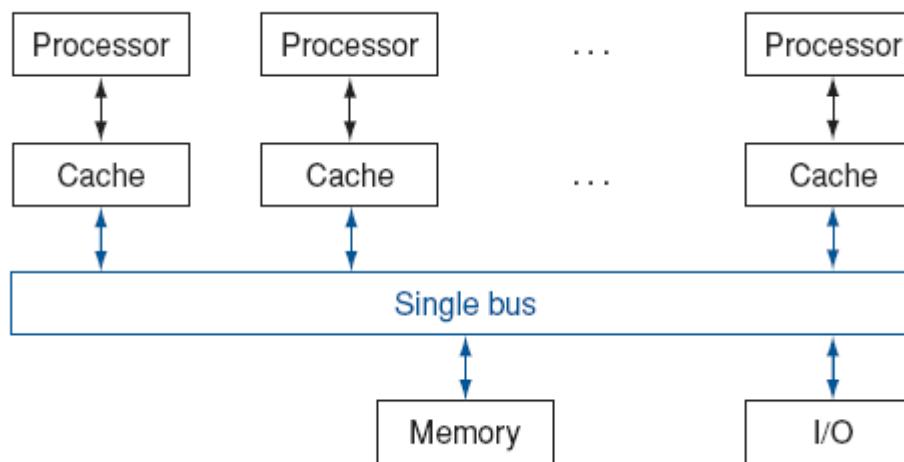
Figure SYSD5 SOC Consumer Portable Design Complexity Trends

# Organizações básicas de MPSoCs

- A forma como a comunicação entre os processadores é realizada **define a organização** do MPSoC
- Modelo de comunicação com a memória
  - Memória Unificada – UMA / NUMA
  - Troca de mensagens
- Modelo físico de comunicação
  - Barramento
  - Rede intra-chip, ou NoC (Network-on-chip)

# MPSoC conectado por barramento

- Número típico de processadores: 2-32
- Espaço único de endereçamento
- Coerência de *cache* é simples (protocolos *snoop*)

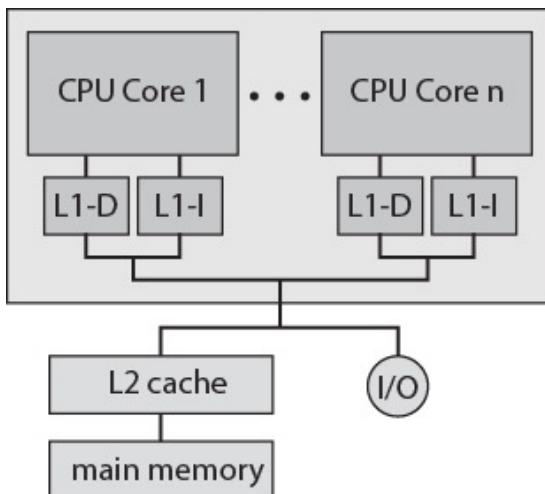


**cache coherency** Consistency in the value of data between the versions in the caches of several processors.

Programação mais simples, mas não escalável

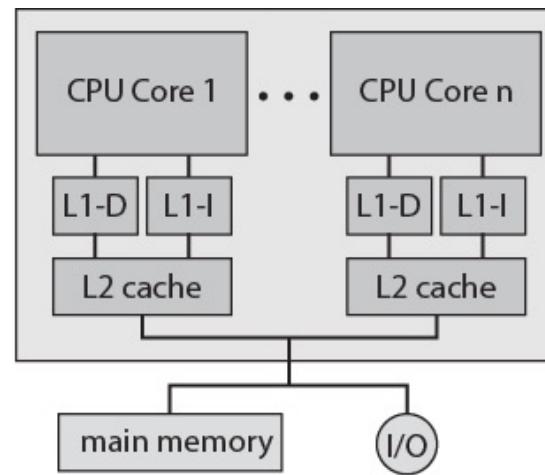
# Barramentos e hierarquia de memória

ARM11 MPCore



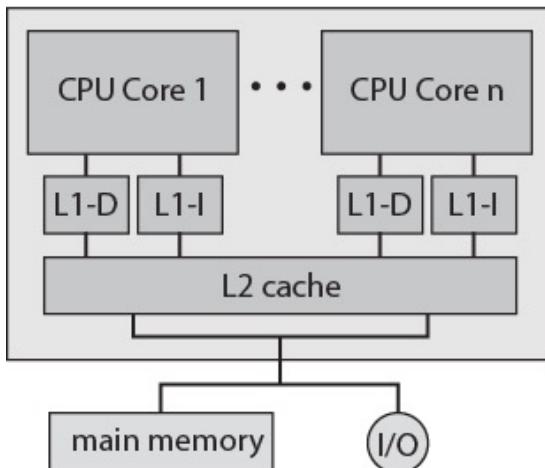
(a) Dedicated L1 cache

AMD Opteron



(b) Dedicated L2 cache

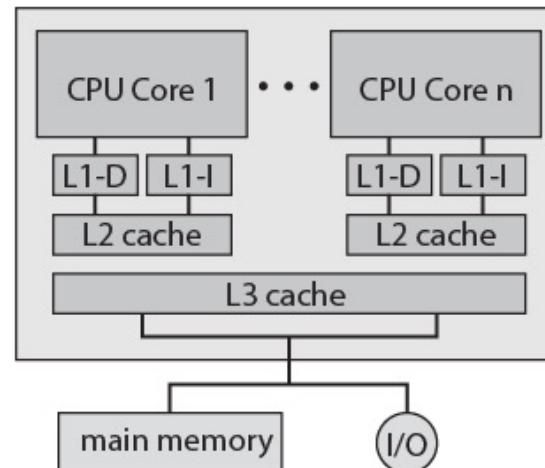
Intel Core Duo



(c) Shared L2 cache

No shared

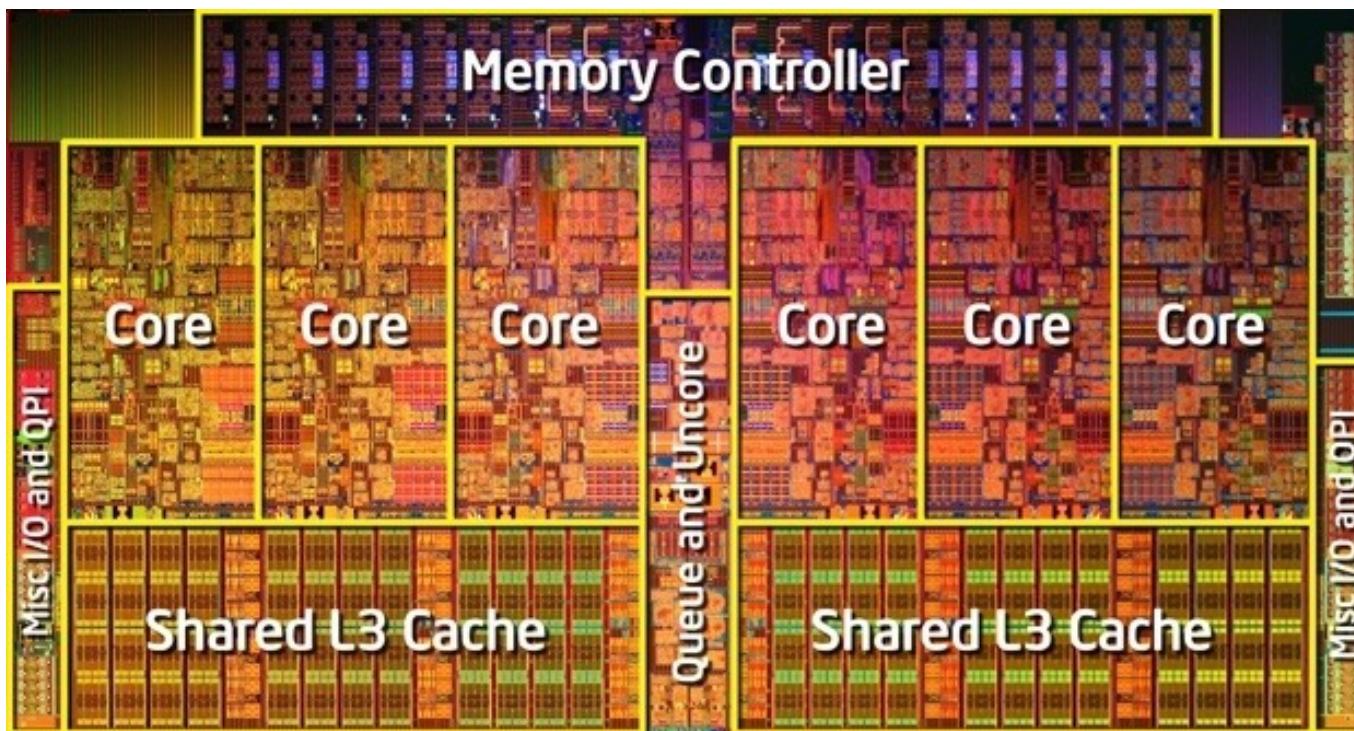
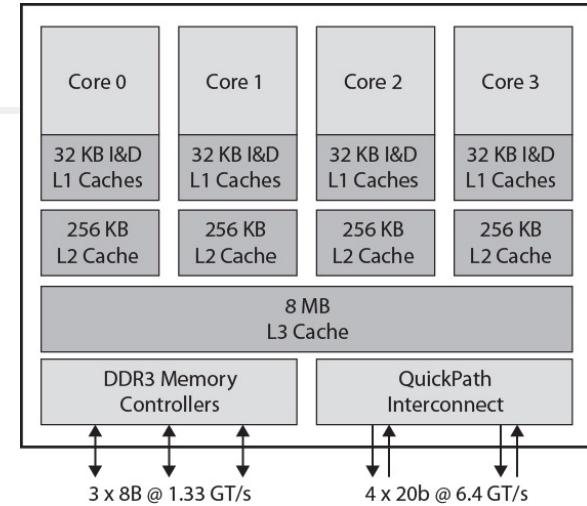
Intel Core i7



(d) Shared L3 cache

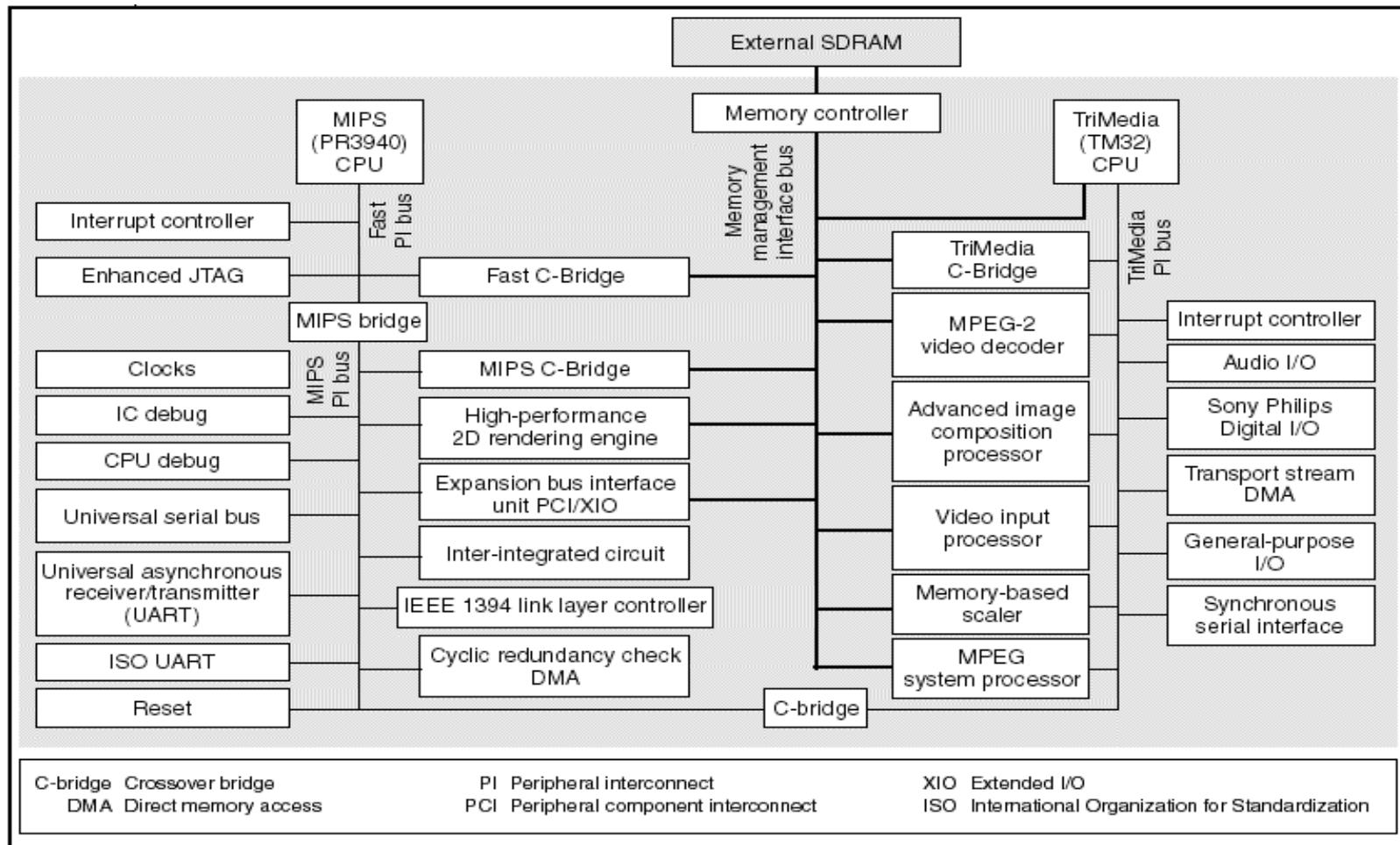
# Intel Core i7 Block Diagram

- Intel Core i7
- Shared L3
- Layout for 6 cores:



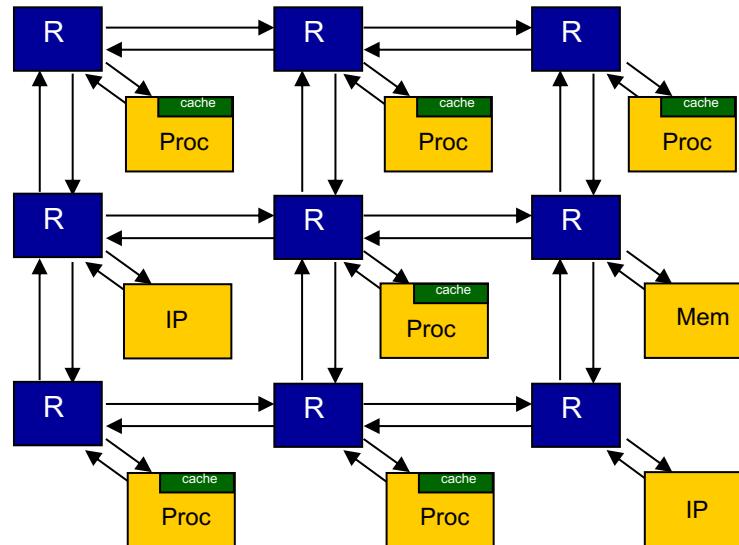
# Barramento

- Complexidade crescente da interconexão por barramento



# MPSoC conectado por rede (NoC)

- Arranjo de roteadores conforme uma *topologia*
  - Abaixo uma rede malha (*mesh*)
- Elementos de processamento, memórias, IPs conectados aos roteadores (redes diretas)
- Comunicação por troca de mensagens



Programação mais complexa, porém escalável

# Organizações básicas de MPSoCs

Category	Choice		Number of processors
Communication model	Message passing		8–2048
	Shared address	NUMA	8–256
		UMA	2–64
Physical connection	Network		8–256
	Bus		2–36

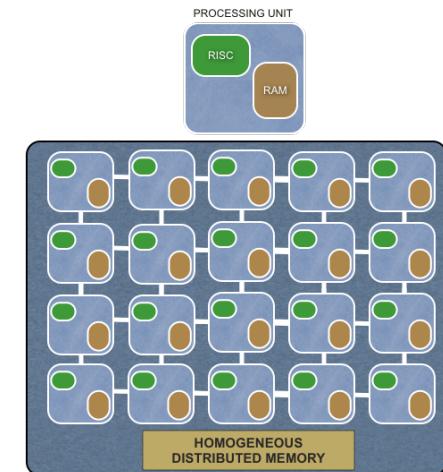
Logo:

- 1) Utilizar comunicação por troca de mensagens
- 2) Utilizar redes intra-chip (NoCs) ao invés de barramento

# MPSoC – tipo de processamento

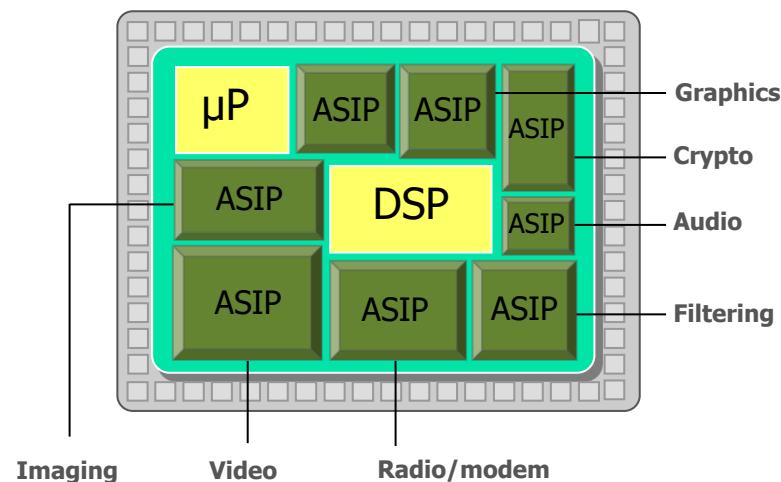
## ■ Homogêneo

- elementos de processamento (PEs) idênticos
- Pros
  - tarefa de programação – mesma aplicação pode executar em qualquer PE
  - mapeamento das aplicações no MPSoC
- Cons
  - tarefas mais intensivas podem não ter seus requisitos de desempenho atendidos



## ■ Heterogêneo

- PEs distintos (GPPs, DSPs, NPUs)
- Pros
  - PEs dedicados para tarefas específicas
- Cons
  - gerência do sistema
  - programação



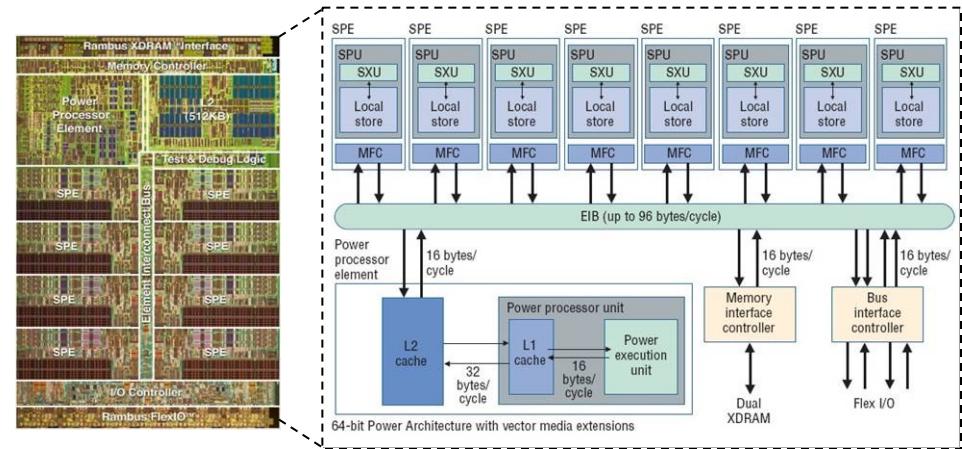
- **Acquisitions**
  - **Facebook** bought Sonics (2019)
    - [https://www.eetimes.com/document.asp?doc\\_id=1334429#](https://www.eetimes.com/document.asp?doc_id=1334429#)
  - **Nvidia** bought Mellanox
    - \$6.9 bi (2019)
  - **Intel** bought NetSpeed (2018)
  - **Qualcomm** bought Arteris (2013)
- Google, Microsoft, Amazon, ARM, Samsung, etc. are still missing. All of them involved in doing SoCs.

# IBM – Cell Processor

MPSoC proposto pela IBM para o PlayStation3 (2006)

## ■ Processador central

- Power Processing Element
  - IBM PowerPC
  - Multithreaded (2 Threads)
  - Cache L1 (32kb I e 32 Kb D)
  - Cache L2- 512 Kb
  - Freqüência: 3.2 GHz
  - Atua como controlador dos processadores periféricos



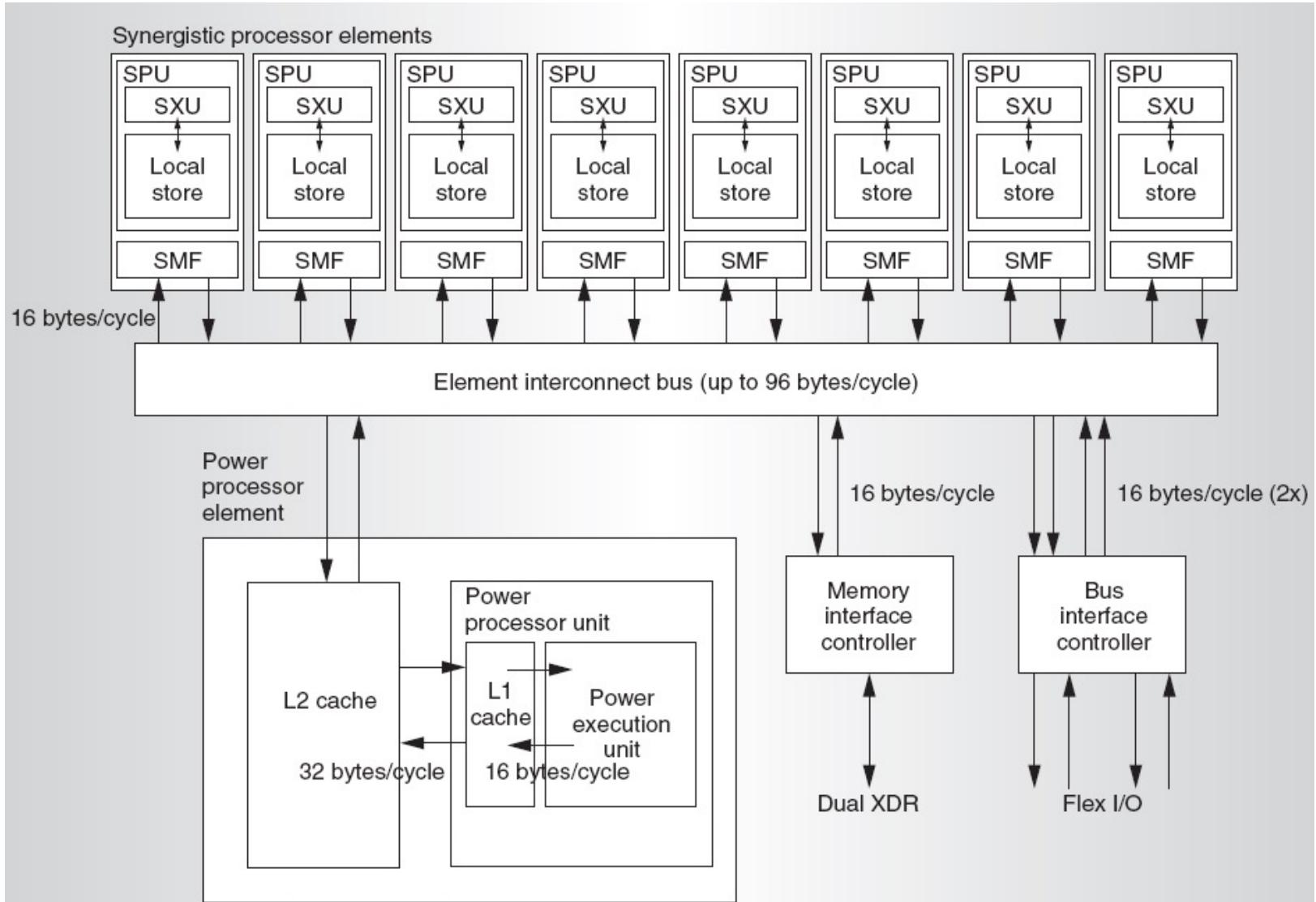
## ■ Oito processadores periféricos

- Synergistic Processing Elements (SPE)
  - Baseados em Processamento Vetorial (SIMD)
  - 256 Kb Local Storage
  - Controlados por software

## ■ Dificuldades para codificar software

- Desenvolvedores de SW são responsáveis pelo gerenciamento dos dados das Local Storages

# IBM - Cell Processor



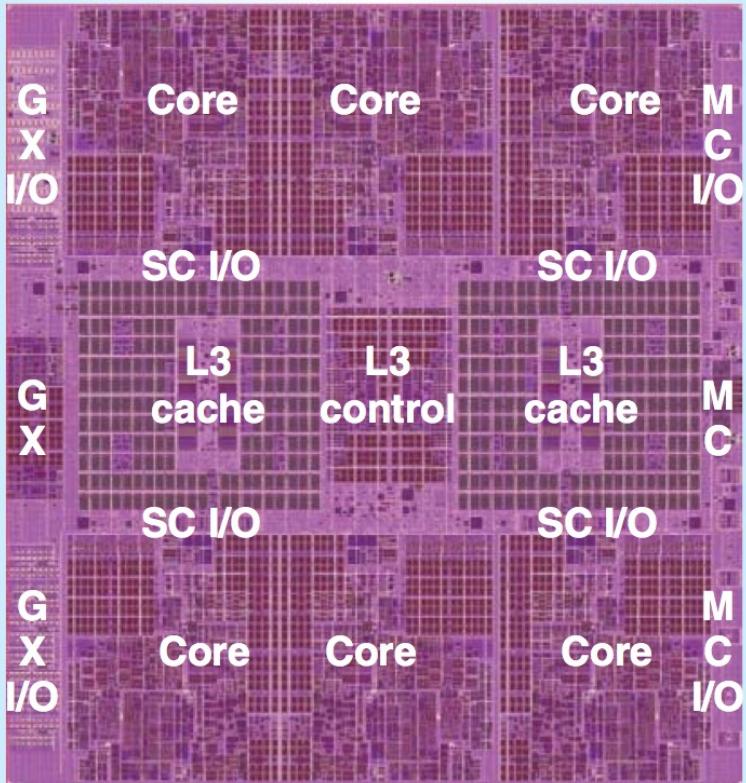


Figure 1. The zEnterprise EC12 (zEC12) microprocessor chip with six processor cores, 48 Mbytes of level-3 (L3) cache, and logic and interfaces connecting to the rest of the system. The chip was fabricated using 32-nm silicon-on-insulator (SOI) technology with roughly 2.75 billion transistors.

Fonte: IEEE Micro Março/Abril 2013

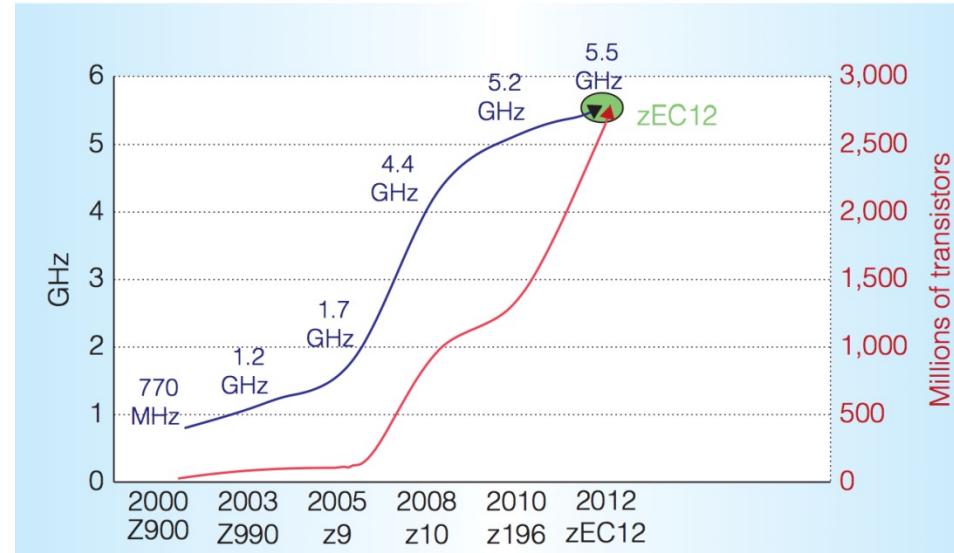
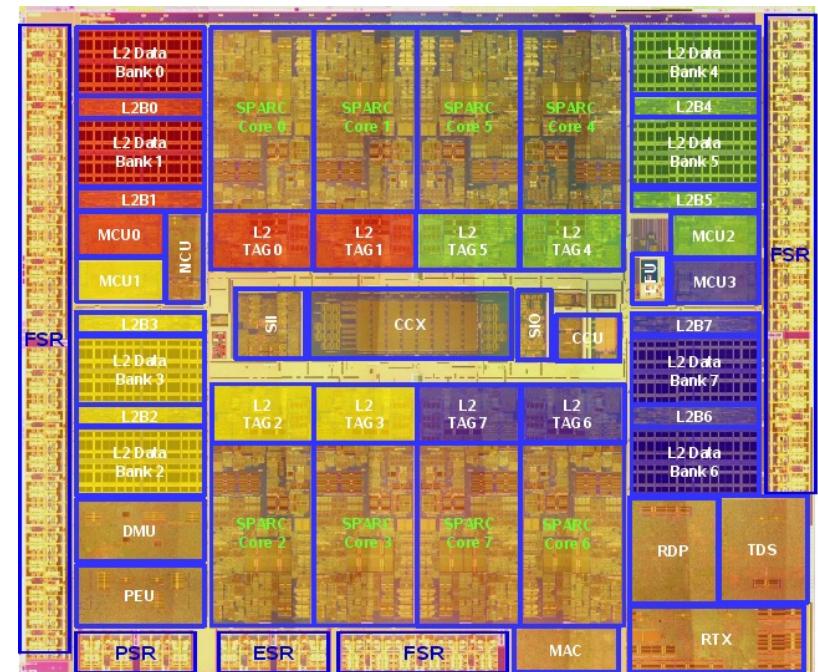


Figure 2. Frequency and transistor counts for the last six generations of microprocessor chips used in IBM's mainframes. System z10 was designed with a deep pipeline microarchitecture to operate at an ultra-high frequency, allowing the next two generations, z196 and zEC12, to continue their frequency improvements.

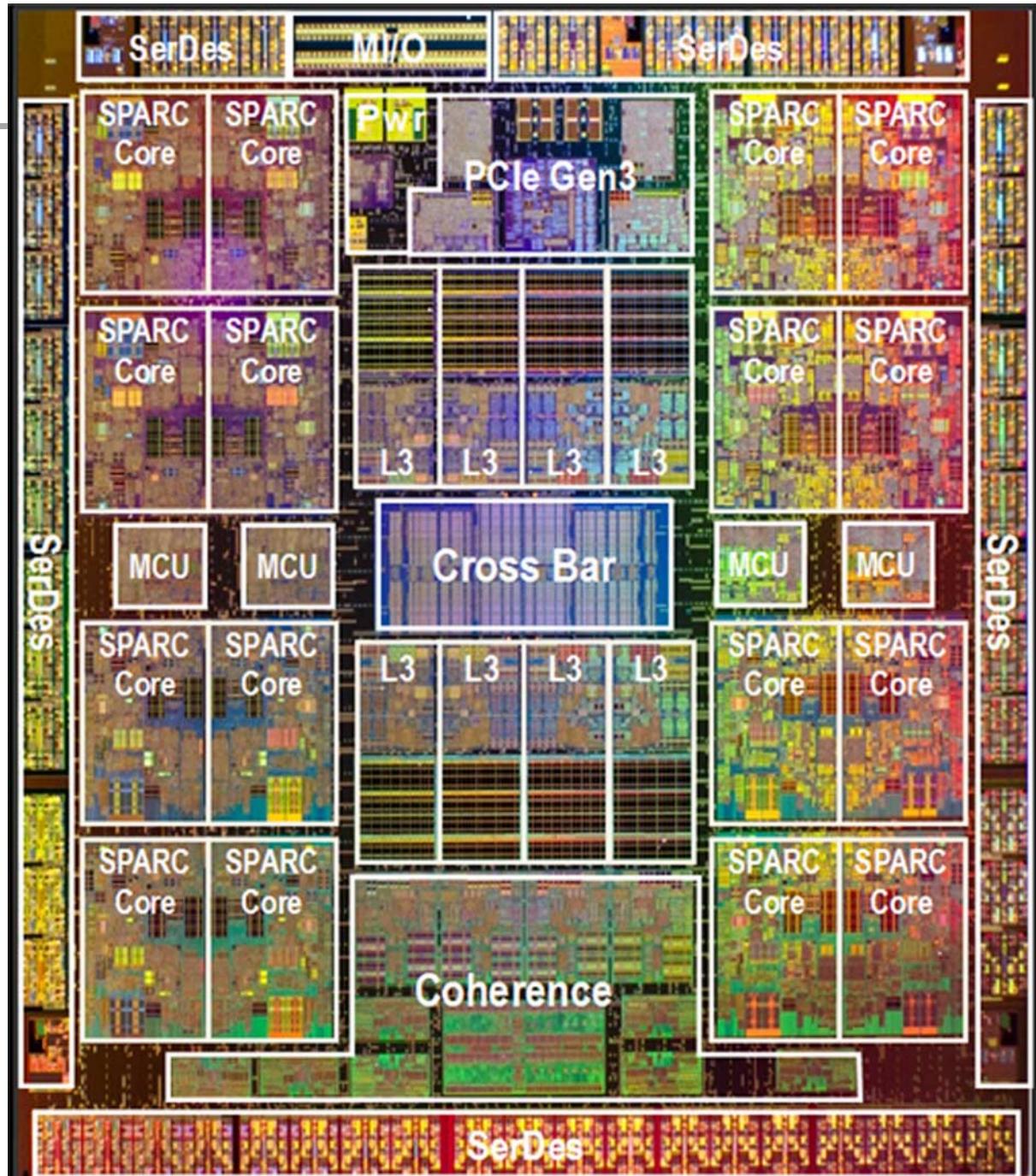
## ▶ UltraSparc T2 (Niagara 2)

- 8 processadores baseados no instruction set Sparc-V9
  - Cada processador é capaz de executar 8 threads concorrentemente (total de 64 threads)
  - “Server on a Chip”
    - PCI express
    - Duas portas 10 Gigabit Ethernet
    - 4 controladores dual-channel FBDIMM
- Pipeline de 8 estágios (1.6Ghz)
- Cache L2 4MB
  - 8 Bancos
  - Associativa por conjunto 16
- Aplicações
  - WebServers
  - Database



## ▶ UltraSparc T5 (Oracle)

IEEE JOURNAL OF SOLID-  
STATE CIRCUITS, 2014  
**A 3.6 GHz 16-Core SPARC  
SoC Processor in 28 nm**



# ORACLE

[https://www.best.de/wp-content/uploads/T8M8\\_Architecture\\_WP\\_20170914-1.pdf](https://www.best.de/wp-content/uploads/T8M8_Architecture_WP_20170914-1.pdf)

TABLE 2. SPARC M8, SPARC M7, SPARC M6, AND SPARC T5 PROCESSOR FEATURE COMPARISON.

Feature	SPARC M8 Processor	SPARC M7 Processor	SPARC M6 Processor	SPARC T5 Processor
CPU frequency	5.0 GHz	4.13 GHz	3.6 GHz	3.6 GHz
Out-of-order execution	Yes	Yes	Yes	Yes
Instruction issue width	4	2	2	2
Data/instruction prefetch	Yes	Yes	Yes	Yes
SPARC core	Fifth generation	Fourth generation	Third generation	Third generation
Cores per processor	32	32	12	16
Threads per core	8	8	8	8
Threads per processor	256	256	96	128
Sockets in systems	Up to 8	Up to 16	Up to 32	Up to 8
Memory per processor	Up to 16 DDR4 DIMMs	Up to 16 DDR4 DIMMs	Up to 32 DDR3 DIMMs	Up to 16 DDR3 DIMMs
Caches	32 KB L1 four-way instruction cache 16 KB L1 four-way data cache Shared 256 KB L2 four-way instruction cache (per quad cores) 128 KB L2 eight-way data cache (per core) Shared 64 MB (L3) cache	16 KB L1 four-way instruction cache 16 KB L1 four-way data cache Shared 256 KB L2 four-way instruction cache (per quad cores) Shared 256 KB L2 eight-way data cache (per core pair) Shared 64 MB (L3) cache	16 KB L1 four-way instruction cache 16 KB L1 four-way data cache 128 KB L2 eight-way cache Shared 48 MB L3 twelve-way cache	16 KB L1 four-way instruction cache 16 KB L1 four-way data cache 128 KB L2 eight-way cache Shared 8 MB L3 sixteen-way cache
Large page support <sup>1</sup>	16 GB	16 GB	2 GB	2 GB
Power management granularity	Half of the chip	A quarter of the chip	Entire chip	Entire chip
Technology	20 nm technology	20 nm technology	28 nm technology	28 nm technology

# ORACLE SPARC M8

[https://www.best.de/wp-content/uploads/T8M8\\_Architecture\\_WP\\_20170914-1.pdf](https://www.best.de/wp-content/uploads/T8M8_Architecture_WP_20170914-1.pdf)

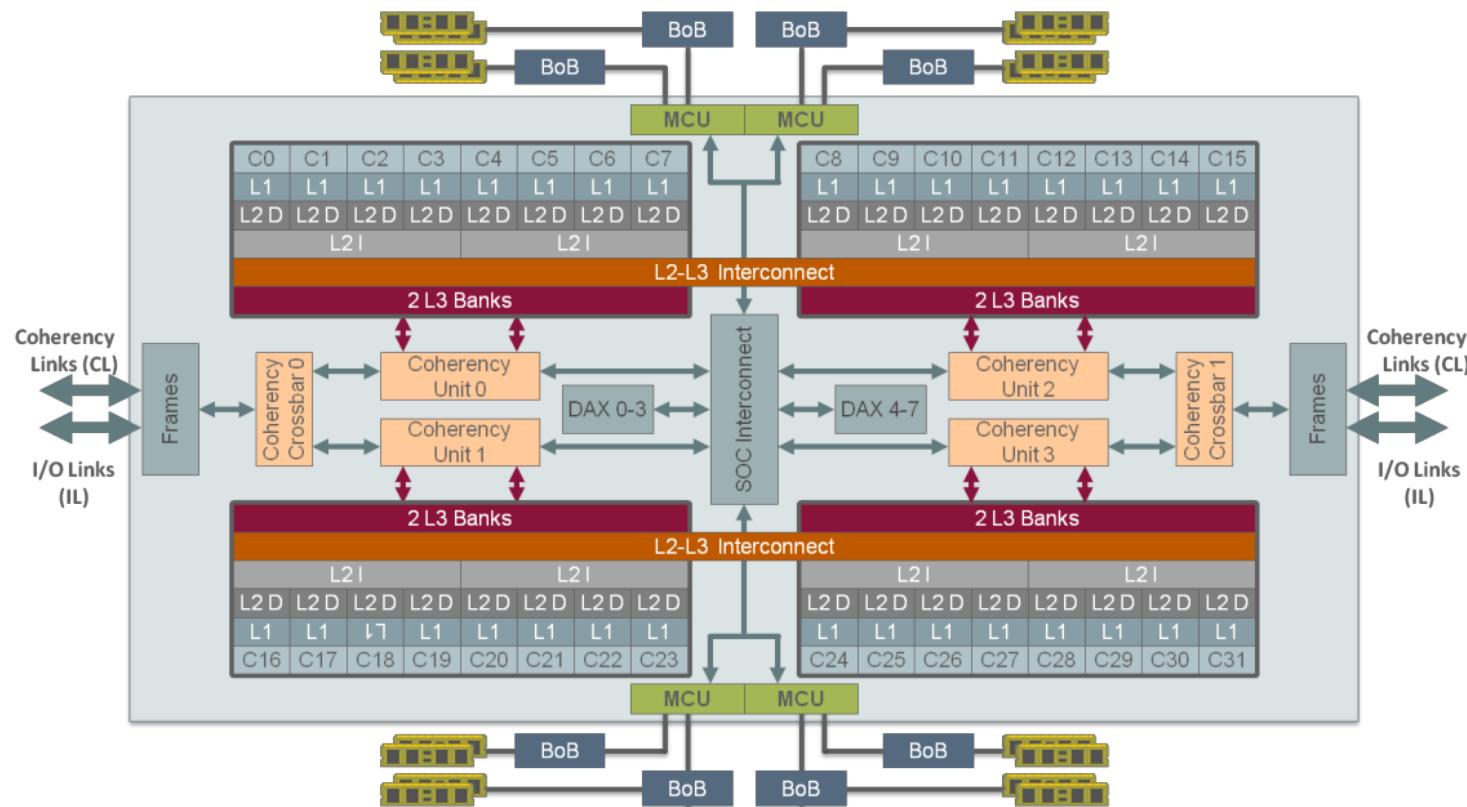
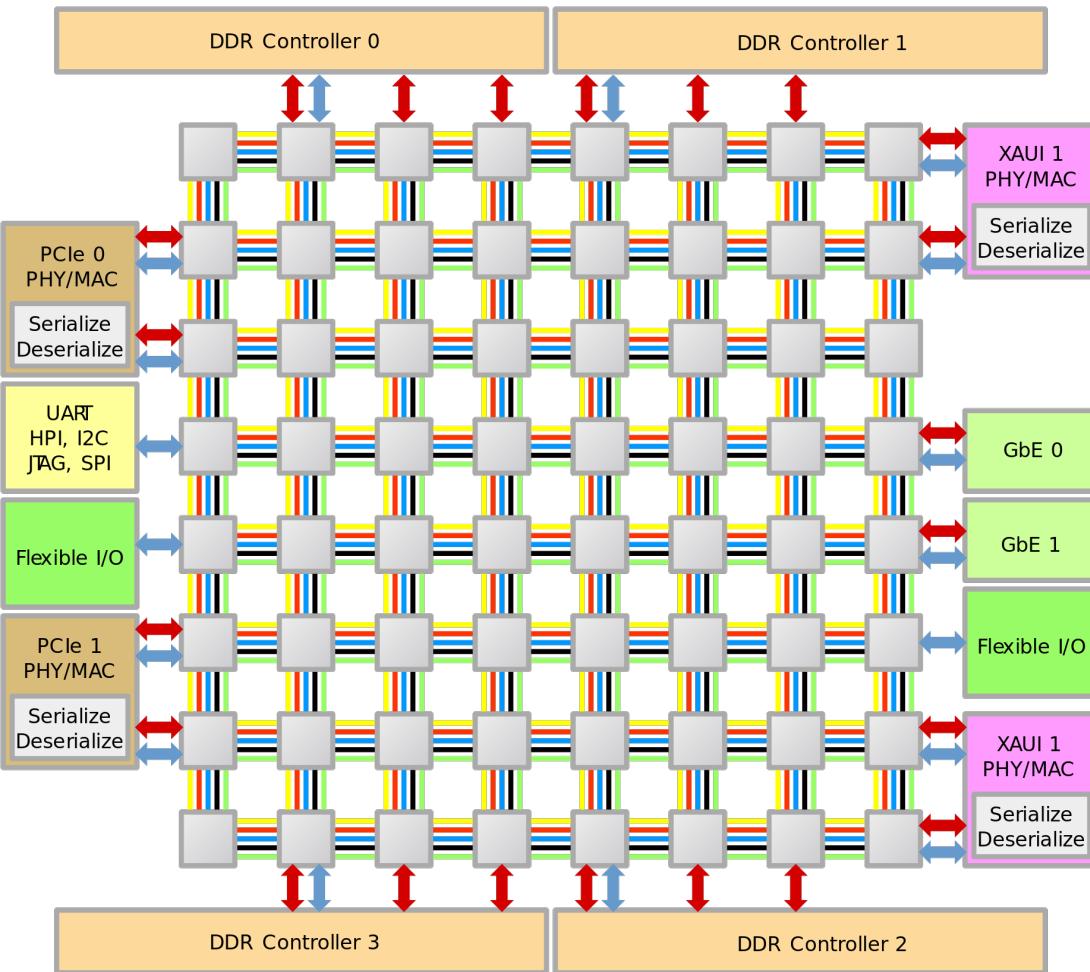


Figure 3. The SPARC M8 processor features 32 cores, which are grouped in two partitions, four memory controller units (MCUs), and eight Data Analytics Accelerators (DAX) units.

## Tile GX– Tilera (100 Núcleos)

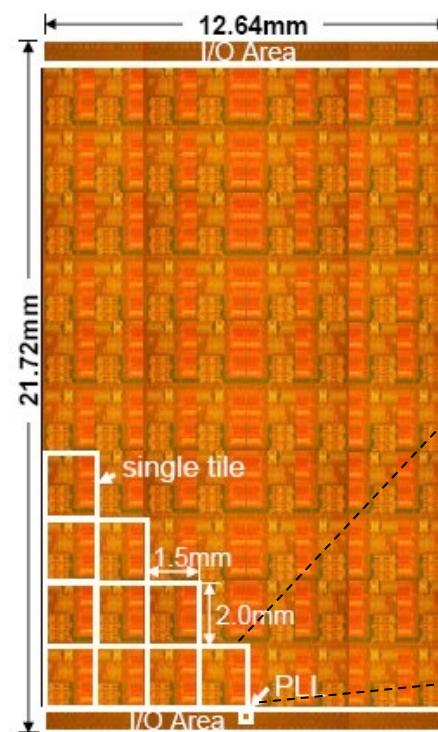


- **100 cores on a single chip**
  - 40 nm technology
- **64-bit VLIW processors**
  - Up to 3 instructions/cycle
  - 3-stage pipeline
  - Up to total of 750 BOPS
  - 32K L1i cache, 32K L1d cache, 256K L2 cache per tile
- **5 on-chip Mesh networks**
  - Over 200 Tbps
- **1 to 1.5 GHz clock frequency**
- **Supports SMP Linux and virtualization**
- **Other members: 16, 32, 64 tiles**

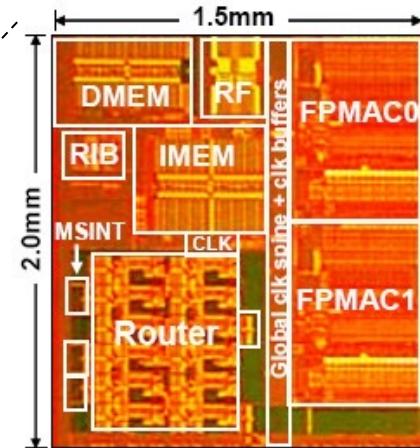
# INTEL - POLARIS

## Intel – (80 Núcleos)

- MPSoC Homogêneo
- Baseado em NoC
- 1 trilhão de operações de ponto flutuante por segundo
- Consumo de potência: 62 Watts
- Baseado em uma arquitetura VLIW
- Exploração do ILP via compilador



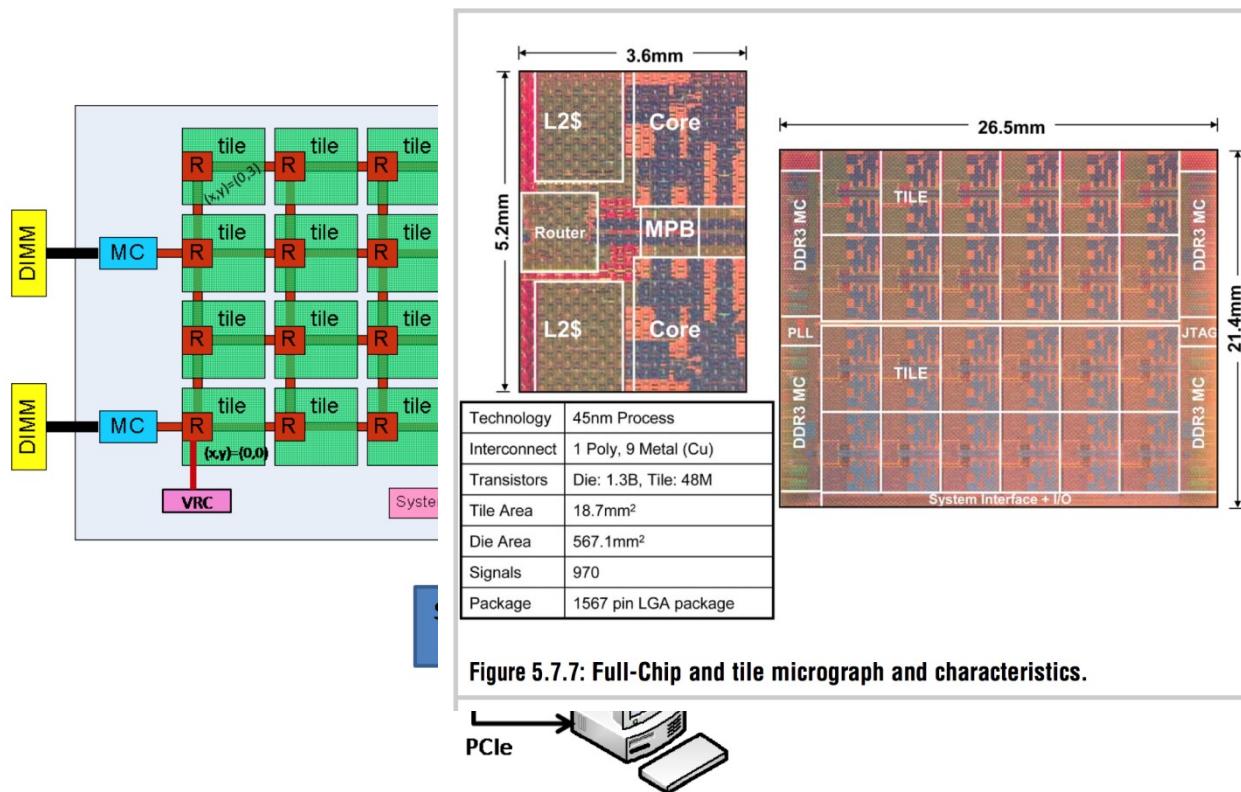
**MPSoC proposto pela Intel (2007)**



Simplicidade dos Cores:  
Remoção do HW superescalar

# INTEL - Single-chip Cloud Computer (SCC)

- 24 “tiles” with two IA cores per tile
- A 24-router mesh network with 256 GB/s bisection bandwidth
- 4 integrated DDR3 memory controllers
- Hardware support for message-passing



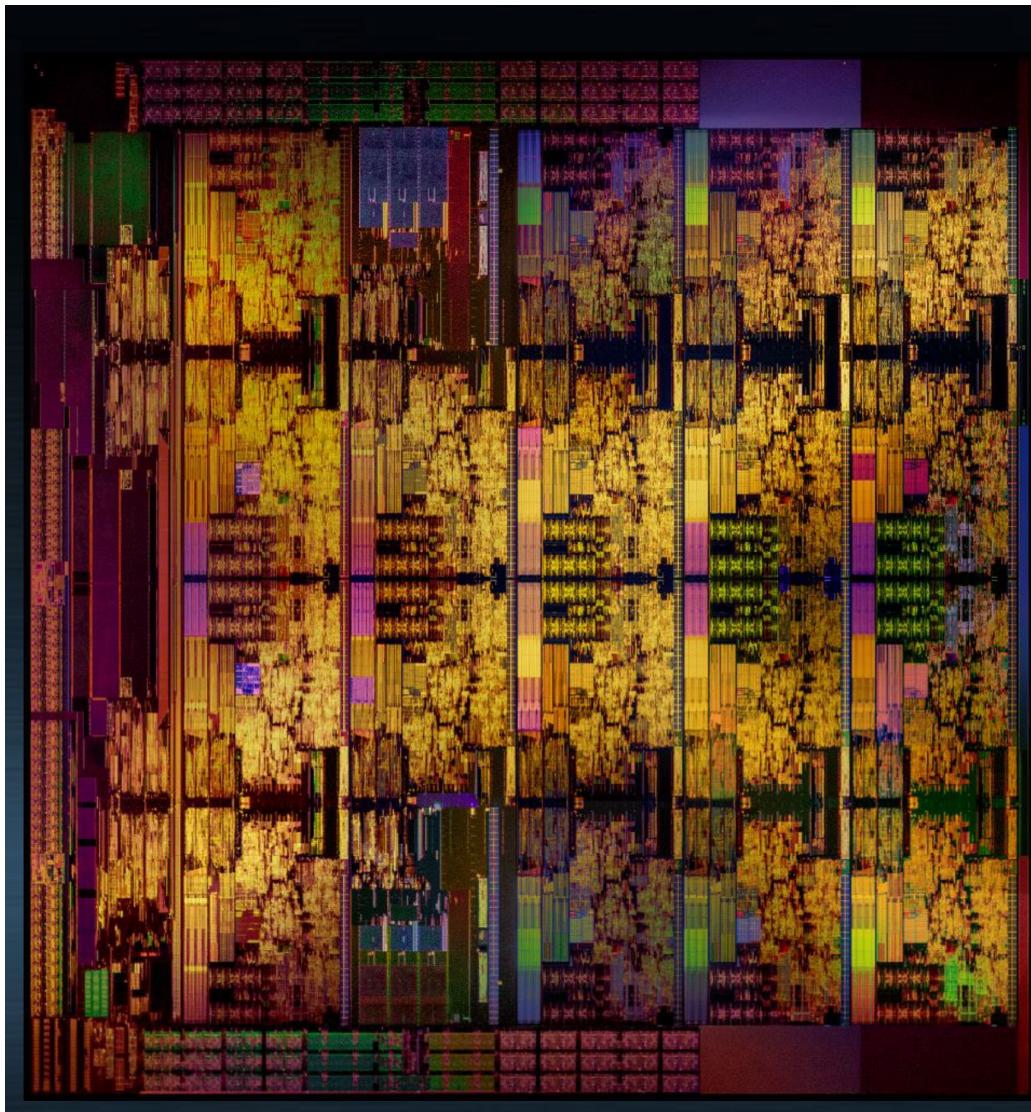
<http://www.intel.com/content/www/us/en/research/intel-labs-single-chip-cloud-program-guide.html>  
<http://communities.intel.com/community/marc>



## Intel Core X Series Processor Family Specifications:

CPU Name	i9-7980XE	i9-7960X	i9-7940X	i9-7920X	i9-7900X	i7-7820X	i7-7800X	i7-7740X	i5-7640X
CPU Process	14nm+								
Architecture	SKL-X	KBL-X	KBL-X						
Cores/Threads	18/36	16/32	14/28	12/24	10/20	8/16	6/12	4/8	4/4
Base Clock	2.6 GHz	2.8 GHz	3.1 GHz	2.9 GHz	3.3 GHz	3.6 GHz	3.5 GHz	4.3 GHz	4.0 GHz
(Turbo Boost 2.0)	4.2 GHz	4.2 GHz	4.3 GHz	4.3 GHz	4.3 GHz	4.3 GHz	4.0 GHz	4.5 GHz	4.2 GHz
(Turbo Boost Max 3.0)	4.4 GHz	4.4 GHz	4.4 GHz	4.4 GHz	4.5 GHz	4.5 GHz	N/A	N/A	N/A
L3 Cache	24.75 MB	22 MB	19.25 MB	16.5 MB	13.75 MB	11 MB	8.25 MB	6 MB	6 MB
L2 Cache	18 MB	16 MB	14 MB	12 MB	10 MB	8 MB	6 MB	4 MB	4 MB
Memory	Quad DDR4	Dual DDR4	Dual DDR4						
PCIe Lanes	44	44	44	44	44	28	28	16	16
Socket Type	LGA 2066								
TDP	165W	165W	165W	140W	140W	140W	140W	112W	112W
Price	\$1999 US	\$1699 US	\$1399 US	\$1189 US	\$999 US	\$599 US	\$389 US	\$349	\$242

# INTEL Core i9-7980XE - 18-Core



<https://wccftech.com/review/intels-massive-core-i9-7980xe-18-core-reviewed/2/>

## Core H-series de décima geração (i9-10980HK)

---

- Lançado em abril de 2020
- Tecnologia de 10 nm
- Densidade de 100,78 milhões de transistores por mm<sup>2</sup>
- Até oito núcleos e 16 threads
- Frequência de até 5.3GHz
- Cerca de 7 bilhões de transistores (estimativa)
- Foco em laptops

# INTRODUCING ICE LAKE: 10NM CPU

## NEW SUNNYCOVE CORES

Up to 4 Cores / 8 Threads  
Up to 4.1GHz

## NEW CONVERGED CHASSIS FABRIC

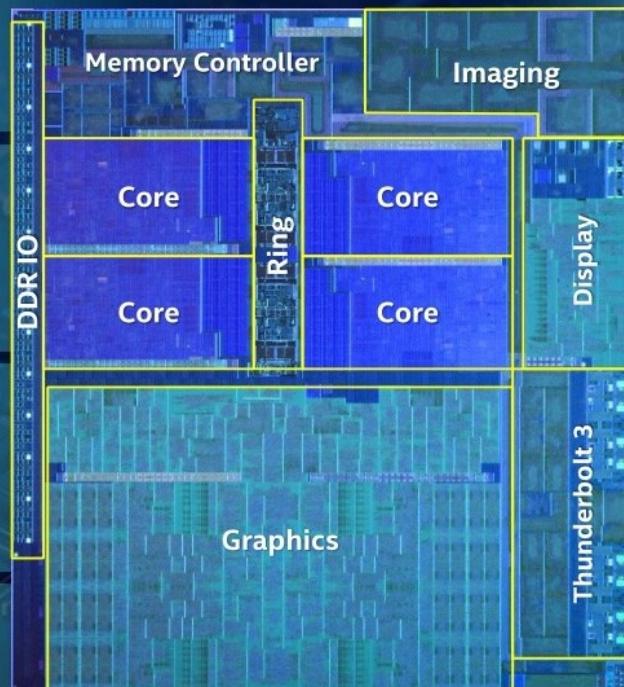
High Bandwidth / Low Latency  
IP and Core Scalable

## NEW MEMORY CONTROLLER

LP4/x-3733 4x32b up to 32GB  
DDR4-3200 2x64b up to 64GB

## FIRST INTEGRATED THUNDERBOLT™ 3

Full 4x DP/USB/PCIe mux on-die  
Up to 40Gbps bi-directional per port



## NEW GEN11 GRAPHICS

Up to 64EU and 1.1GHz  
>1TFLOP

## NEW 2X MEDIA ENCODERS

Up to 4K60 10b 4:4:4  
Up to 8K30 10b 4:2:0

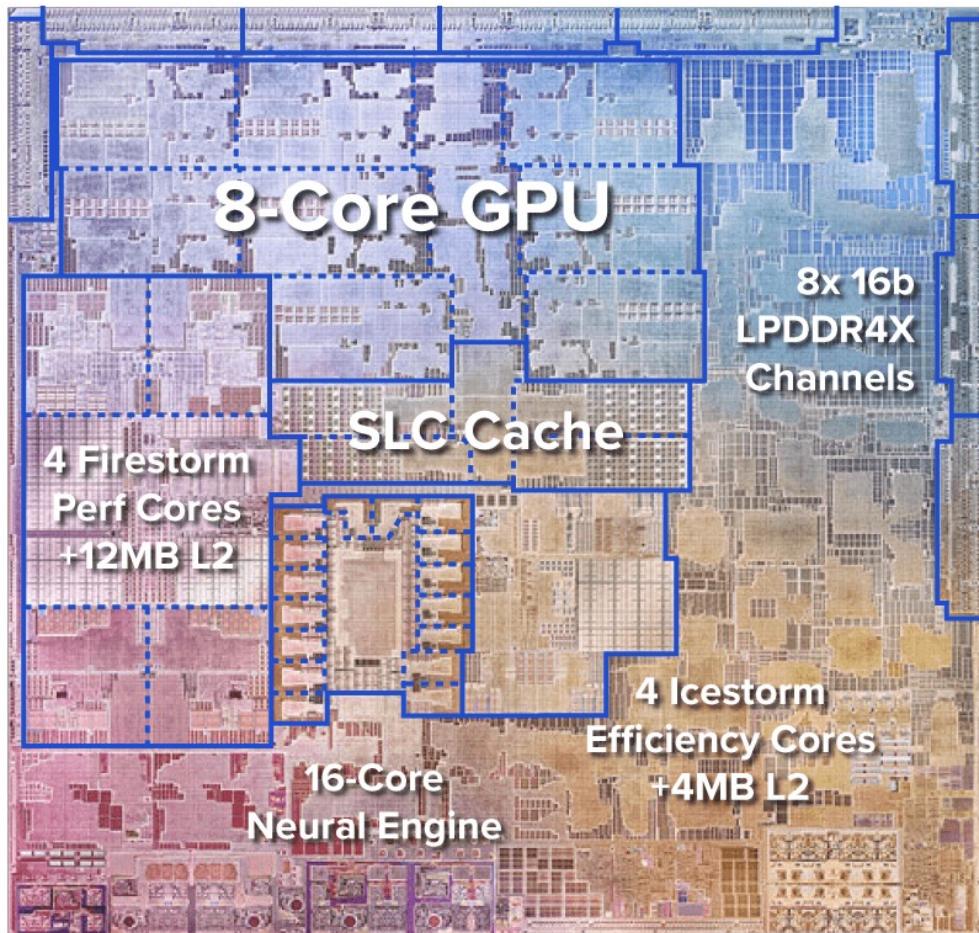
## NEW 3X DISPLAY PIPES

Up to 5K60 or 4K120  
DP1.4, BT.2020

## NEW IMAGE PROCESSING UNIT 4

Up to 16MP  
Up to 1080p120, 4K30

# Apple M1



Architecture:	Arm-based
CPU Cores:	8-core CPU
Nm Process:	5nm
Graphics:	Integrated 8-core GPU with 2.6 teraflops of throughput
Memory:	8GB or 16GB of LPDDR4X-4266 MHz SDRAM

Asymmetric multiprocessing (AMP):

- 4 CPUs 'Firestorm': performance
- 4 CPUs 'Icestorm': low power

# Esperanto SOC - 2021

Accelerating ML Recommendation with over a Thousand RISC-V/Tensor Processors on Esperanto's ET-SoC-1 Chip  
<https://doi.org/10.1109/HCS52781.2021.9566904>

## Summary Statistics of ET-SoC-1

The ET-SoC-1 is fabricated in TSMC 7nm

- 24 billion transistors
- Die-area: 570 mm<sup>2</sup>
- 89 Mask Layers

1088 ET-Minion energy-efficient 64-bit RISC-V processors

- Each with an attached vector/tensor unit
- Typical operation 500 MHz to 1.5 GHz expected

4 ET-Maxion 64-bit high-performance RISC-V out-of-order processors

- Typical operation 500 MHz to 2 GHz expected

1 RISC-V service processor

Over 160 million bytes of on-die SRAM used for caches and scratchpad memory

Root of trust for secure boot

Power typically < 20 watts, can be adjusted for 10 to 60+ watts under SW control

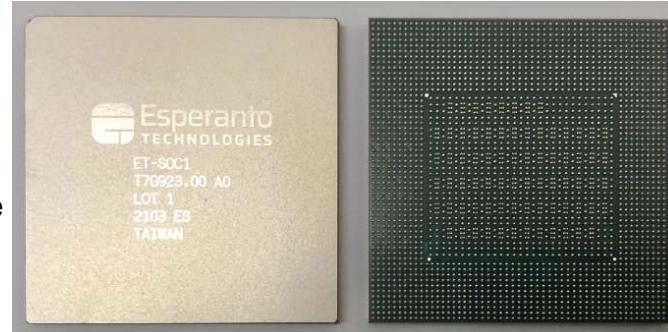
Package: 45x45mm with 2494 balls to PCB, over 30,000 bumps to die

- Each Minion Shire has independent low voltage power supply inputs that can be finely adjusted to mitigate V<sub>t</sub> variation effects and enable DVFS

Status: Silicon currently undergoing bring-up and characterization



ET-SoC-1 Die Plot



ET-SoC-1 Package

# SoC FPGA

## Processor

- Dual-core ARM® Cortex™-A9 MPCore™ processor
- 4,000 MIPS (up to 800 MHz per core)
- NEON coprocessor with double-precision FPU
- 32-KB/32-KB L1 caches per core
- 512-KB shared L2 cache

## Multiport SDRAM controller

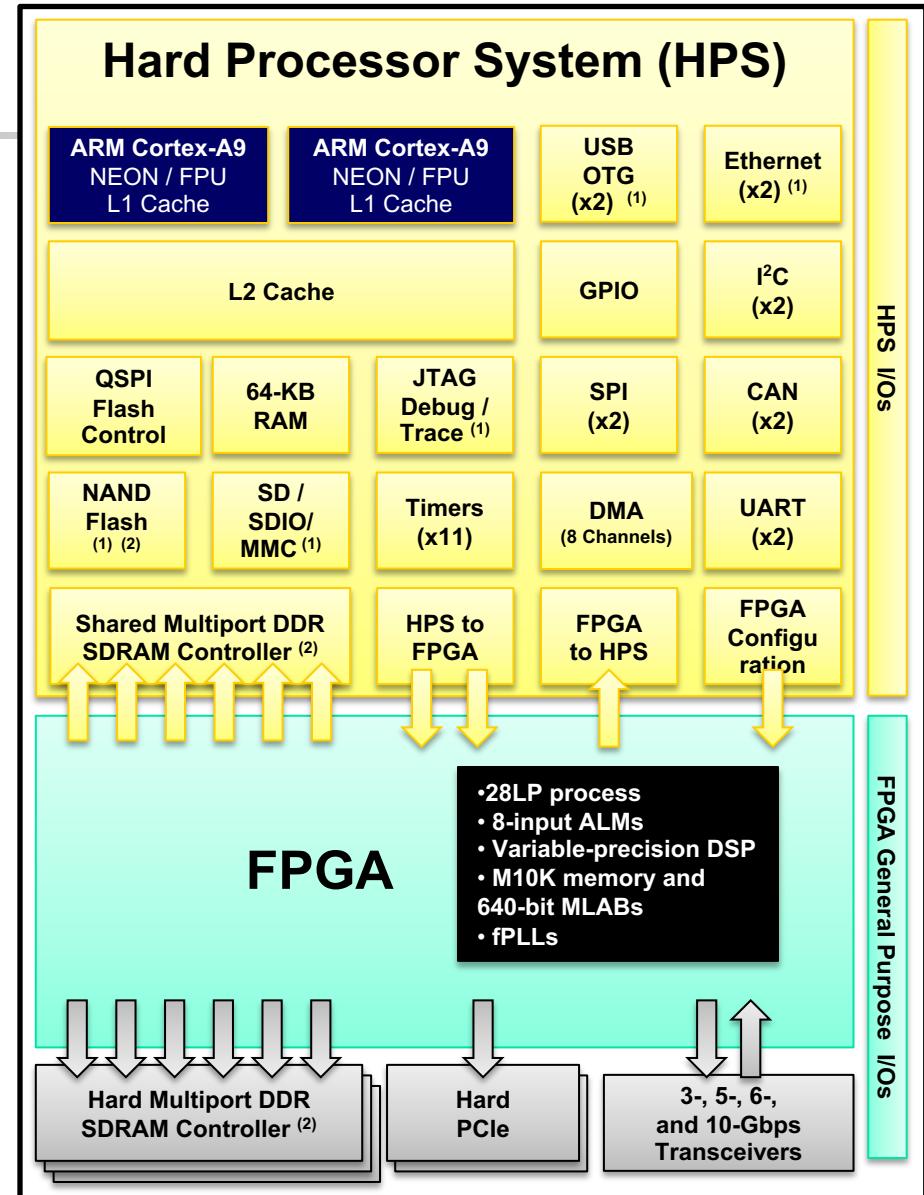
- Up to 533-MHz DDR3 and LPDDR2
- Up to 400-MHz DDR2
- Up to 200-MHz Mobile DDR
- Integrated ECC support

## High-bandwidth on-chip interfaces

- > 125-Gbps HPS-to-FPGA interface
- > 125-Gbps FPGA-to-SDRAM interface

## Cost- and power-optimized FPGA fabric

- Lowest power transceivers
- Up to 1,600 GMACS, 300 GFLOPS
- Up to 25Mb on-chip RAM
- More hard intellectual property (IP): PCIe® and memory controllers



### Notes:

(1) Integrated direct memory access (DMA)

(2) Integrated ECC