ICNPG 2023

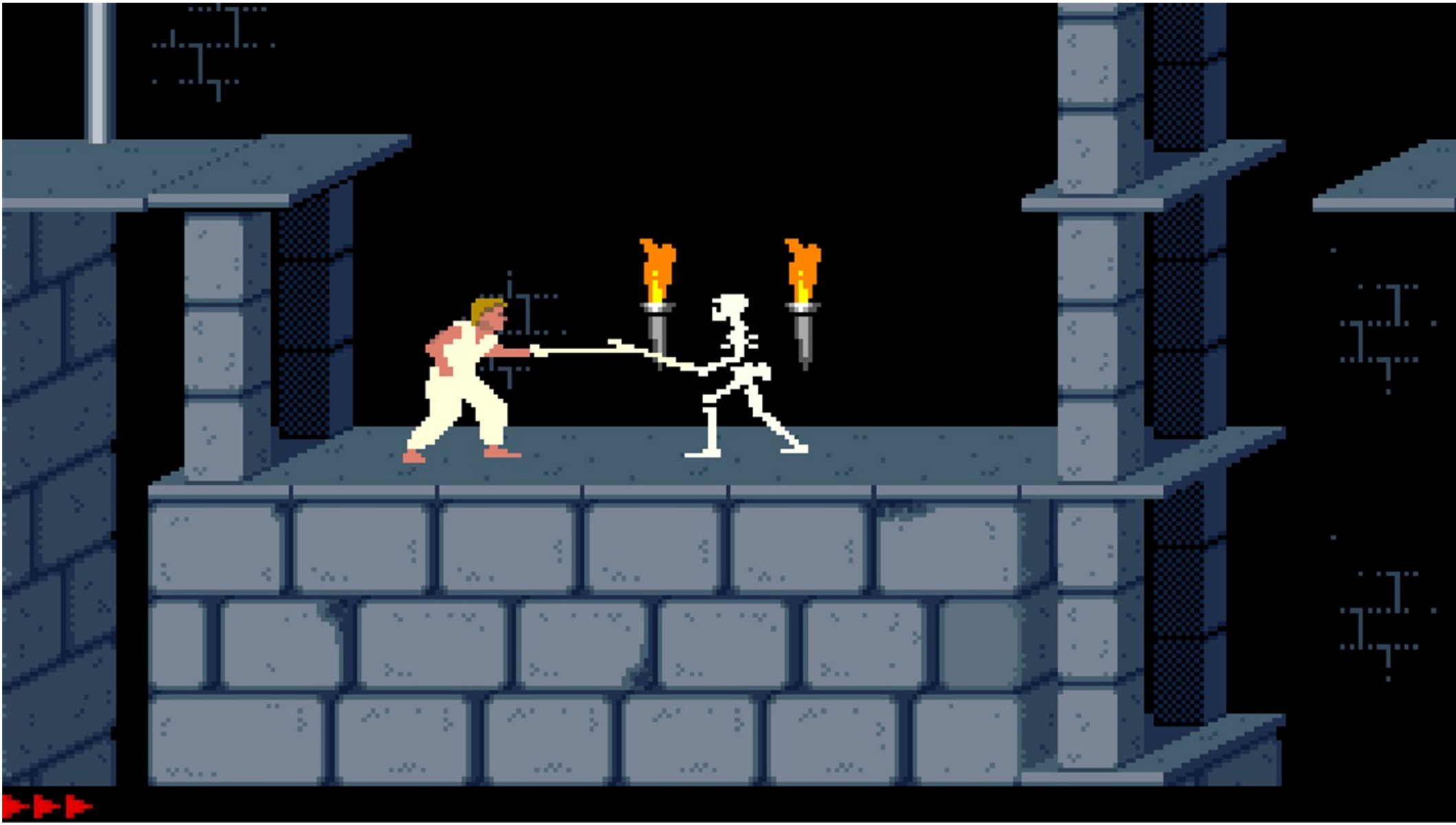
Clase 0

Introducción al Cálculo Numérico en Procesadores Gráficos

- 16 clases x 4hs = 64 hs. Martes y Miércoles de 9 a 13. Aula Egipcios.
- Laboratorio de computación paralela en GPUs, orientada a aplicaciones científicas y tecnológicas.
- Ejercicios ilustrativos en clase y para hacer en casa (para discutir la clase siguiente).
- Evaluación: Participación en clase, desarrollo y exposición/entrega de un miniproyecto usando GPGPU. El código tiene que correr en el cluster
- Obligatorio aprender a usar el cluster de GPUs.
- Consultas fuera de horario: google-chat.
- Todo el material en el google-classroom IB.

El principe de Persia (1990)

El desarrollo de placas gráficas comenzó con el desarrollo de los videojuegos. Se buscaba disminuir el trabajo en la CPU



El principe de Persia (2010)

Videos de Yt:

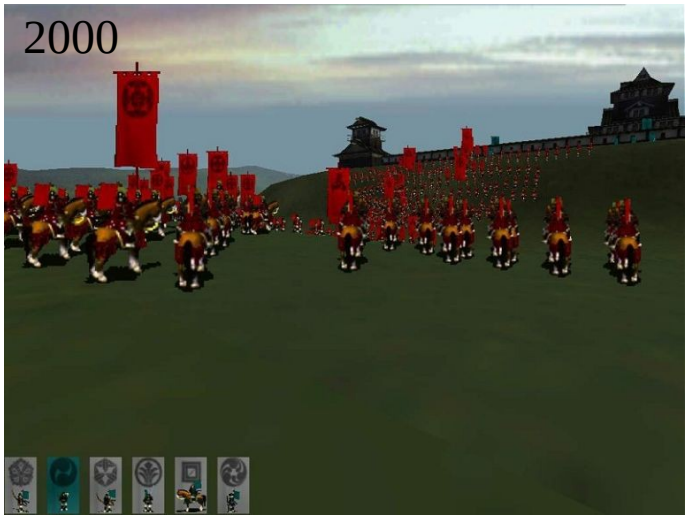
The Evolution Of Prince of Persia Games

https://youtu.be/yS6DEuuFJ6k

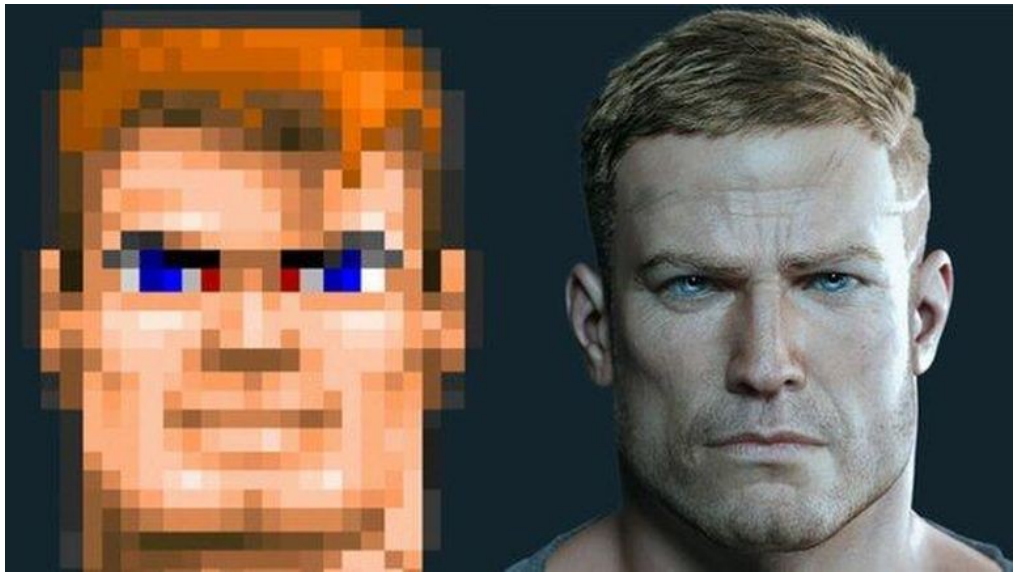
Evolution of Video Game Graphics 1958-2020

https://youtu.be/IsPPWWlV-T8


The Evolution of Mario
1981
1985
1987
1990
1991
1996
2002
2008
*Dates based on US release dates.


2000


2006

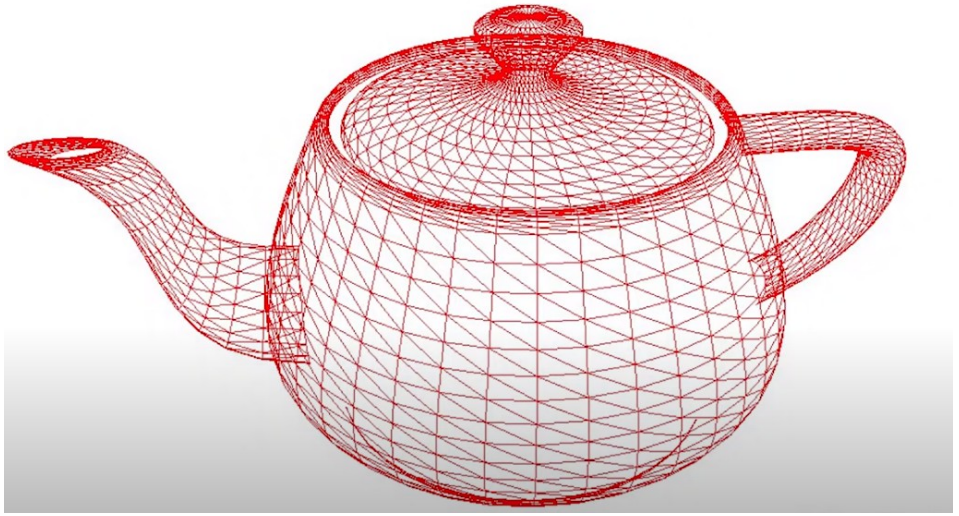



2013

Juegos (2018)

Las GPU aparecieron como un

Acelerador gráfico 3D

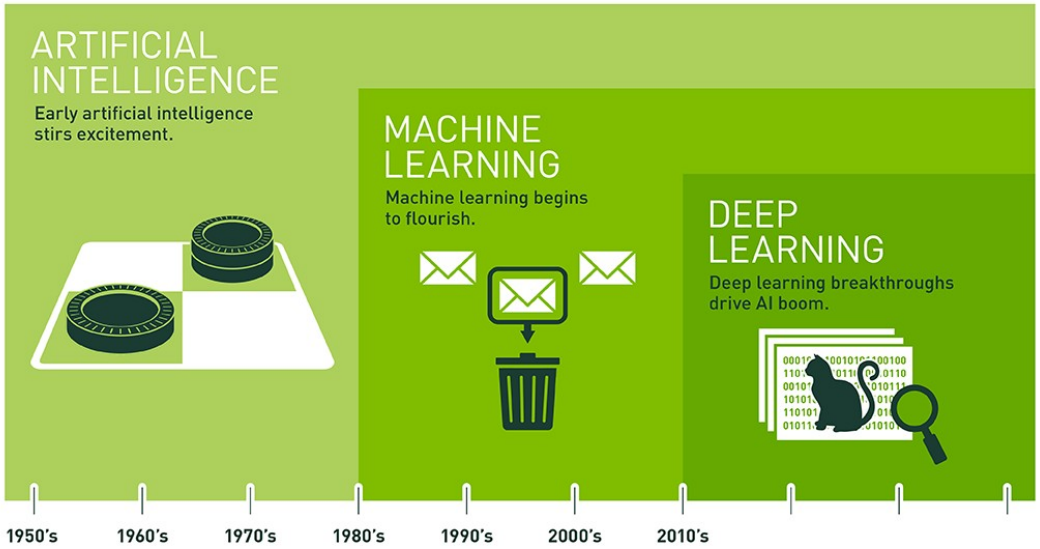


Modelo 3D basado en vértices → Pixels en la pantalla

Para mover, rotar, dar textura, iluminación, etc. es necesario realizar

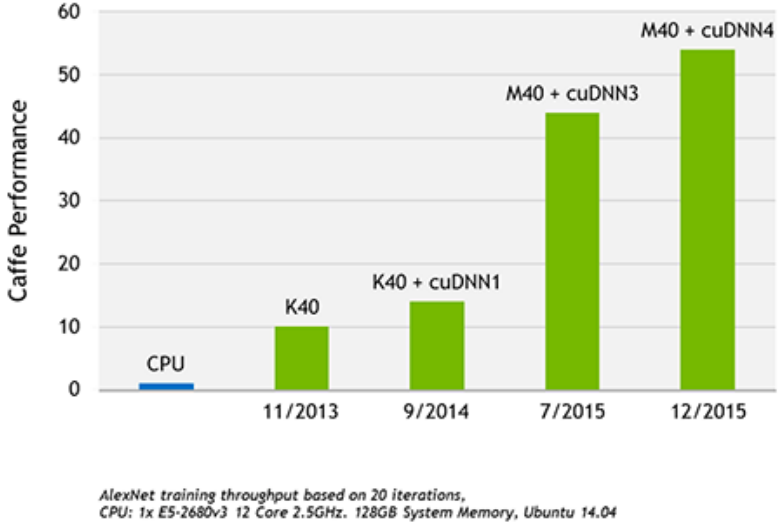
¡Un montón de operaciones!

que son altamente paralelizables



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Además, lo que vemos hoy en Deep Learning no sería posible si no tuviéramos las GPUs



PyTorch TensorFlow mxnet

PyG

PaddlePaddle

Delving Deep into Rectifiers:
Surpassing Human-Level Performance on ImageNet Classification

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun

Microsoft Research

(2015)

nature machine intelligence

Review Article | Published: 23 March 2022

The transformational role of GPU computing and deep learning in drug discovery

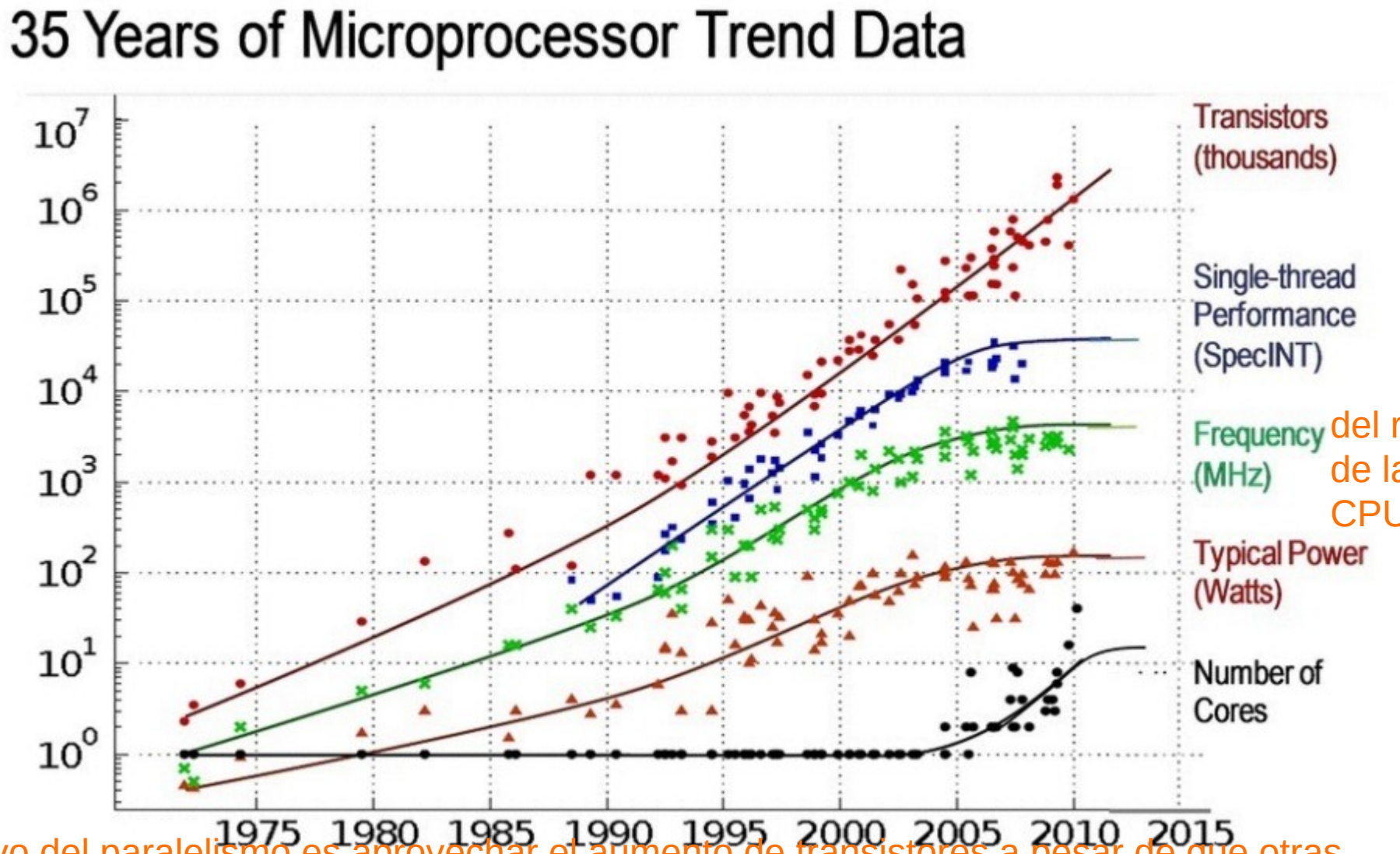
Mohit Pandey, Michael Fernandez, Francesco Gentile, Olexandr Isayev, Alexander Tropsha, Abraham C. Stern ✉ & Artem Cherkasov ✉

Nature Machine Intelligence **4**, 211–221 (2022) | Cite this article

https://www.nature.com/articles/s42256-022-00463-x

Se acabó: Mi viejo programa secuencial de CPU ya no correrá más rápido cuando cambie la compu...

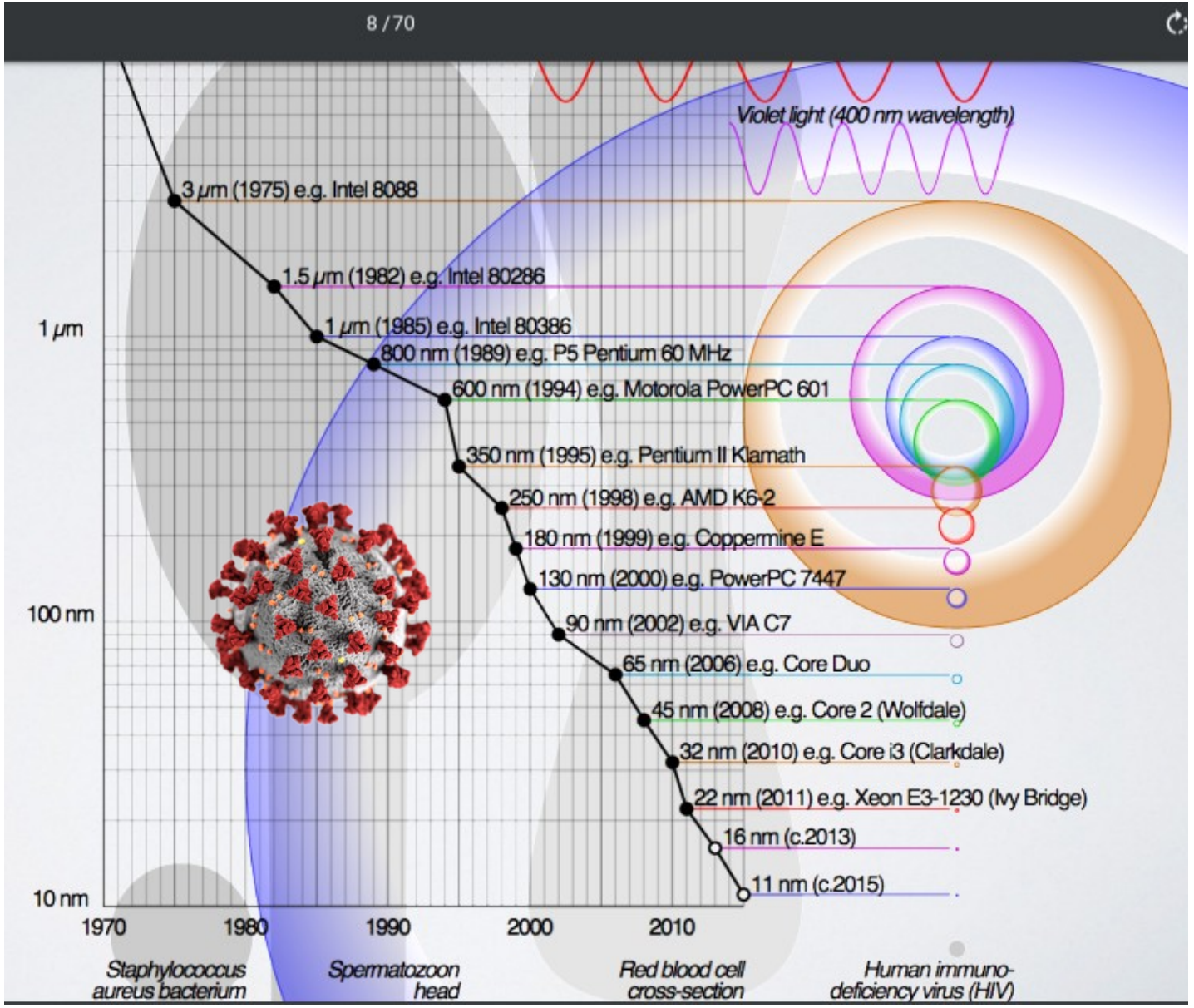
porque la frecuencia en el tiempo satura y por razones físicas no es posible aumentarla

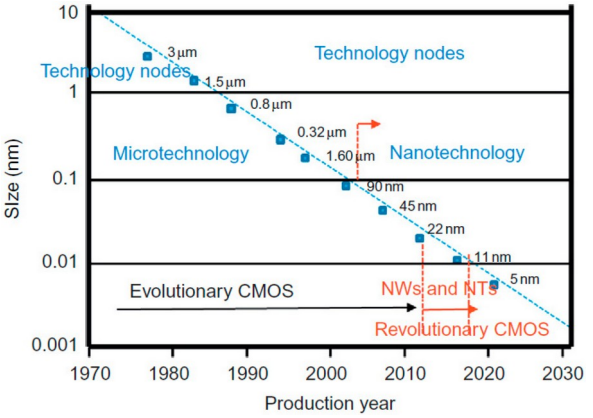


del reloj de la CPU

El objetivo del paralelismo es aprovechar el aumento de transistores a pesar de que otras características están saturando

Gracias a la miniaturización (nanociencia/nanotecnología)





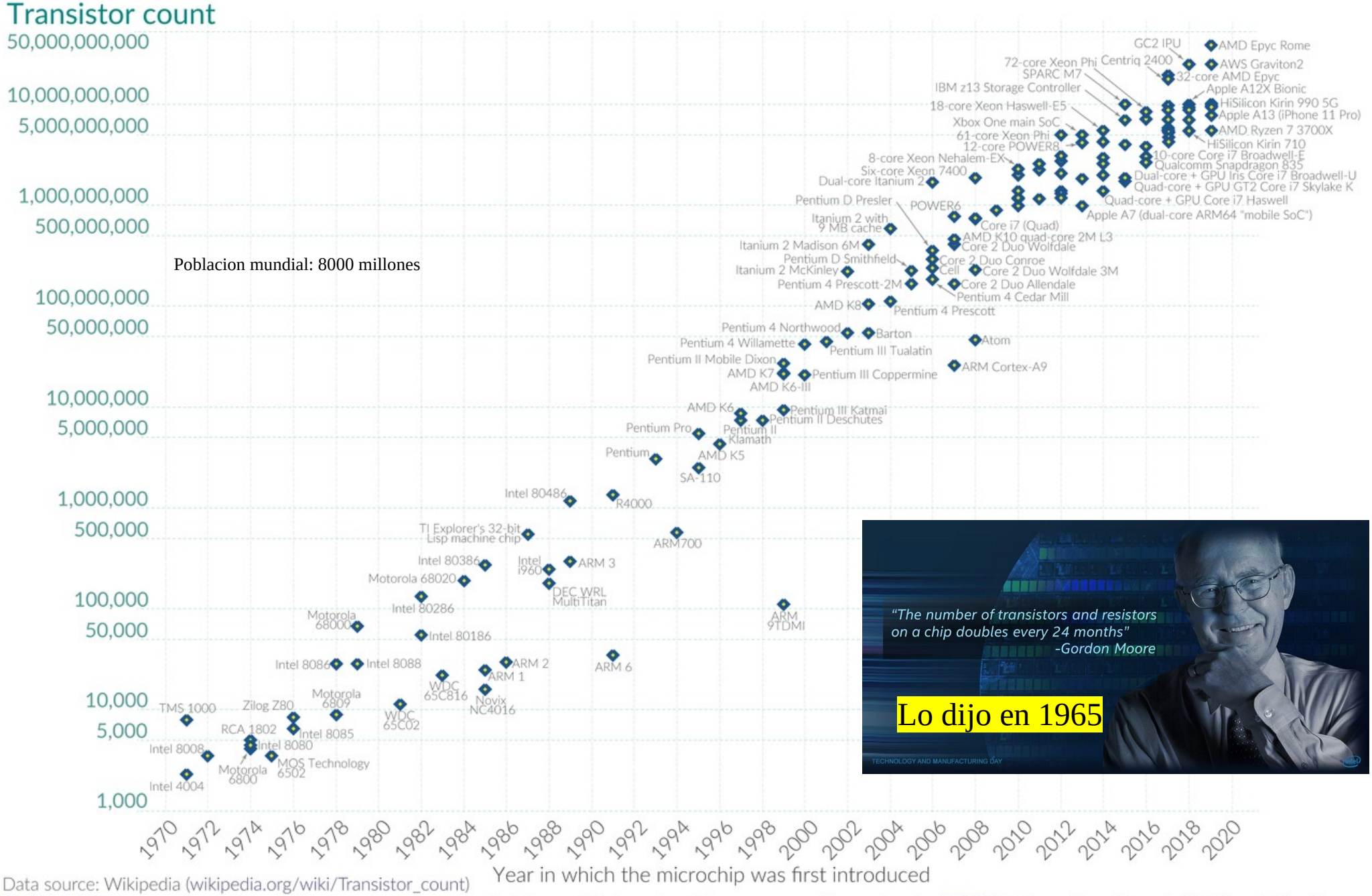
2023: ~5 nm

No se puede ver con luz visible

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

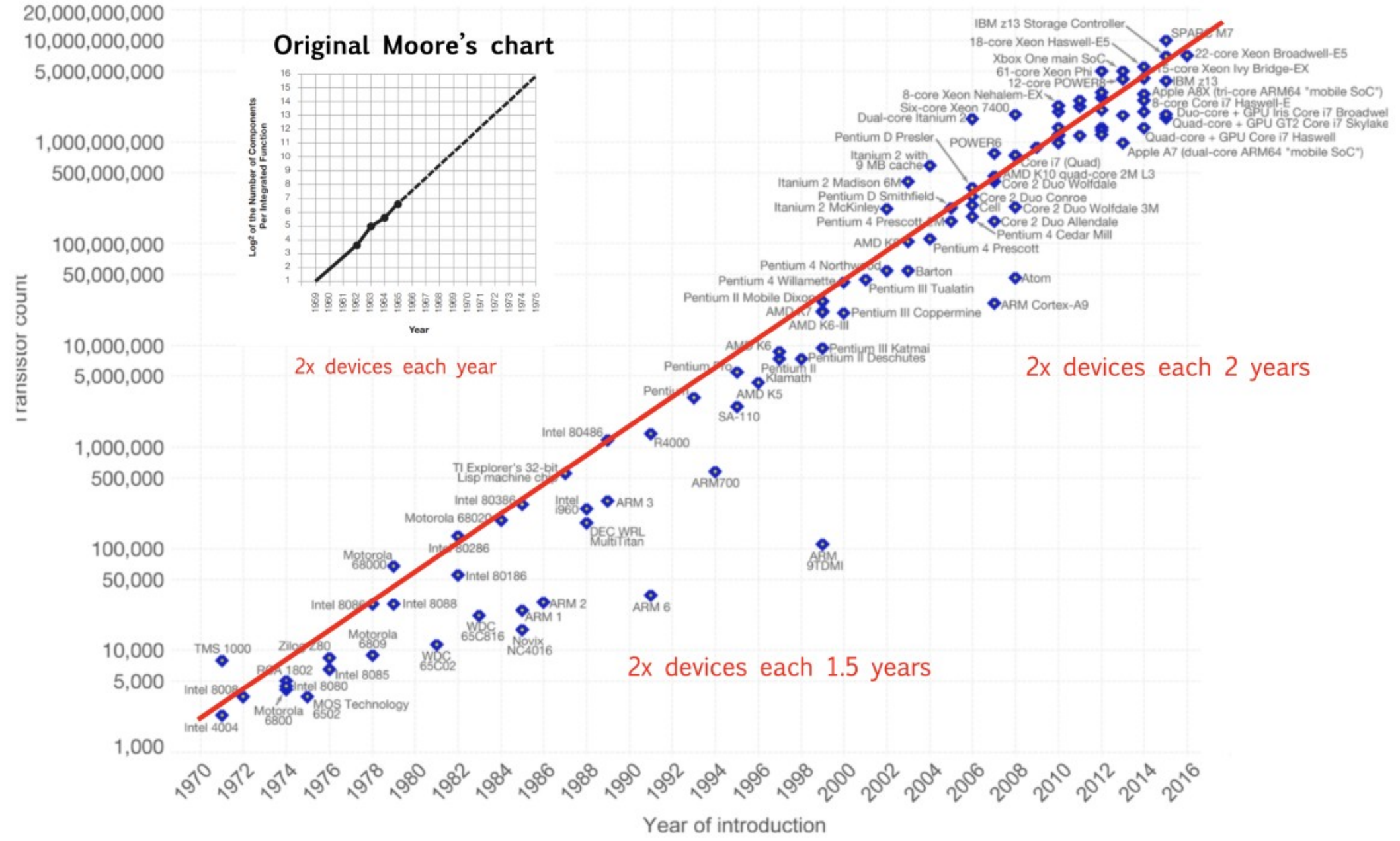
Our World in Data

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.


Original Moore's chart
Log² of the Number of Components Per Integrated Function
Year
2x devices each year
2x devices each 2 years
2x devices each 1.5 years
Transistor count
Year of introduction
20,000,000,000
10,000,000,000
5,000,000,000
1,000,000,000
500,000,000
100,000,000
50,000,000
10,000,000
5,000,000
1,000,000
500,000
100,000
50,000
10,000
5,000
1,000
1970
1972
1974
1976
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
IBM z13 Storage Controller
SPARC M7
18-core Xeon Haswell-E5
22-core Xeon Broadwell-E5
Xbox One main SoC
61-core Xeon Phi
15-core Xeon Ivy Bridge-EX
12-core POWER8
IBM z13
Apple A8X (tri-core ARM64 "mobile SoC")
8-core Xeon Nehalem-EX
8-core Core i7 Haswell-E
Six-core Xeon 7400
Duo-core + GPU Iris Core i7 Broadwel
Dual-core Itanium 2
Quad-core + GPU GT2 Core i7 Skylake
Pentium D Presler
POWER6
Quad-core + GPU Core i7 Haswell
Itanium 2 with 9 MB cache
Apple A7 (dual-core ARM64 "mobile SoC")
Core i7 (Quad)
AMD K10 quad-core 2M L3
Itanium 2 Madison 6M
Core 2 Duo Wolfdale
Pentium D Smithfield
Core 2 Duo Conroe
Itanium 2 McKinley
Cell
Core 2 Duo Wolfdale 3M
Pentium 4 Prescott
Core 2 Duo Allendale
Pentium 4 Cedar Mill
Pentium 4 Prescott
Barton
Pentium 4 Northwood
Pentium 4 Willamette
Atom
Pentium III Tualatin
Pentium II Mobile Dixon
ARM Cortex-A9
Pentium III Coppermine
AMD K6-III
Pentium III Katmai
Pentium II Deschutes
Pentium II Klamath
AMD K5
SA-110
Intel 80486
R4000
ARM700
TI Explorer's 32-bit Lisp machine chip
Intel 80386
Intel i960
ARM 3
Motorola 68020
DEC WRL MultiTitan
Motorola 68000
Intel 80286
ARM 9TDMI
Intel 80186
Intel 8086
Intel 8088
ARM 2
ARM 1
ARM 6
WDC 65C816
Novix NC4016
Motorola 6809
TMS 1000
Zilog Z80
WDC 65C02
RCA 1802
Intel 8085
Intel 8008
Intel 8080
MOS Technology 6502
Motorola 6800
Intel 4004

‘Moore’s Law’s dead,’ Nvidia CEO Jensen Huang says in justifying gaming-card price hike

Last Updated: Sept. 22, 2022 at 7:43 a.m. ET
First Published: Sept. 21, 2022 at 6:16 p.m. ET

MORE ABOUT MOORE —

Intel: “Moore’s law is not dead” as Arc A770 GPU is priced at $329

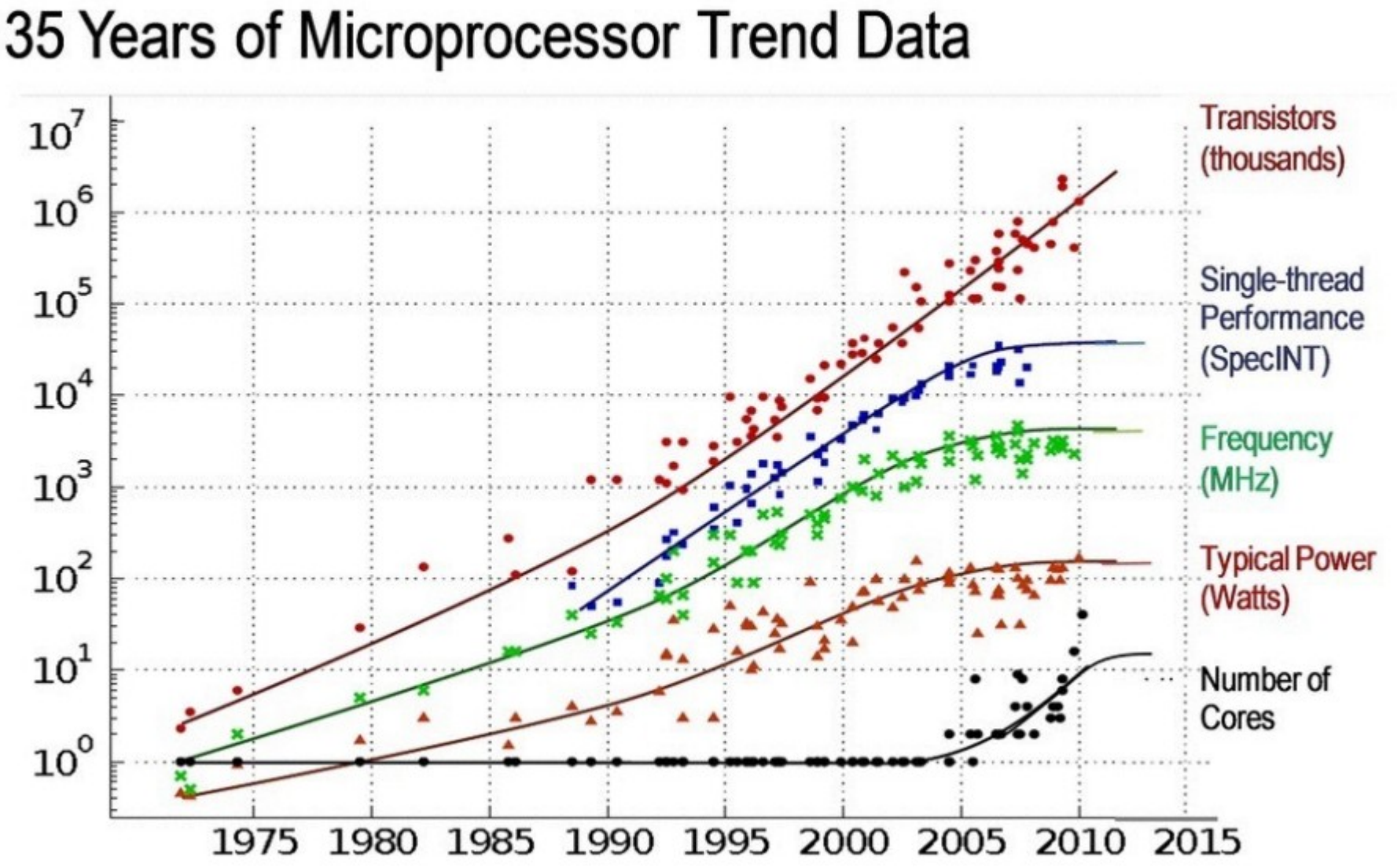
Expected performance somewhere near Nvidia's RTX 3060 Ti—at least, for DirectX 12.

SAM MACHKOVECH - 9/27/2022, 2:40 PM

Actualmente las placas más conocidas son Nvidia, Intel y AMD. Con CUDA se puede aprender a usar las Nvidia y para las demás hay que usar otros lenguajes de características similares

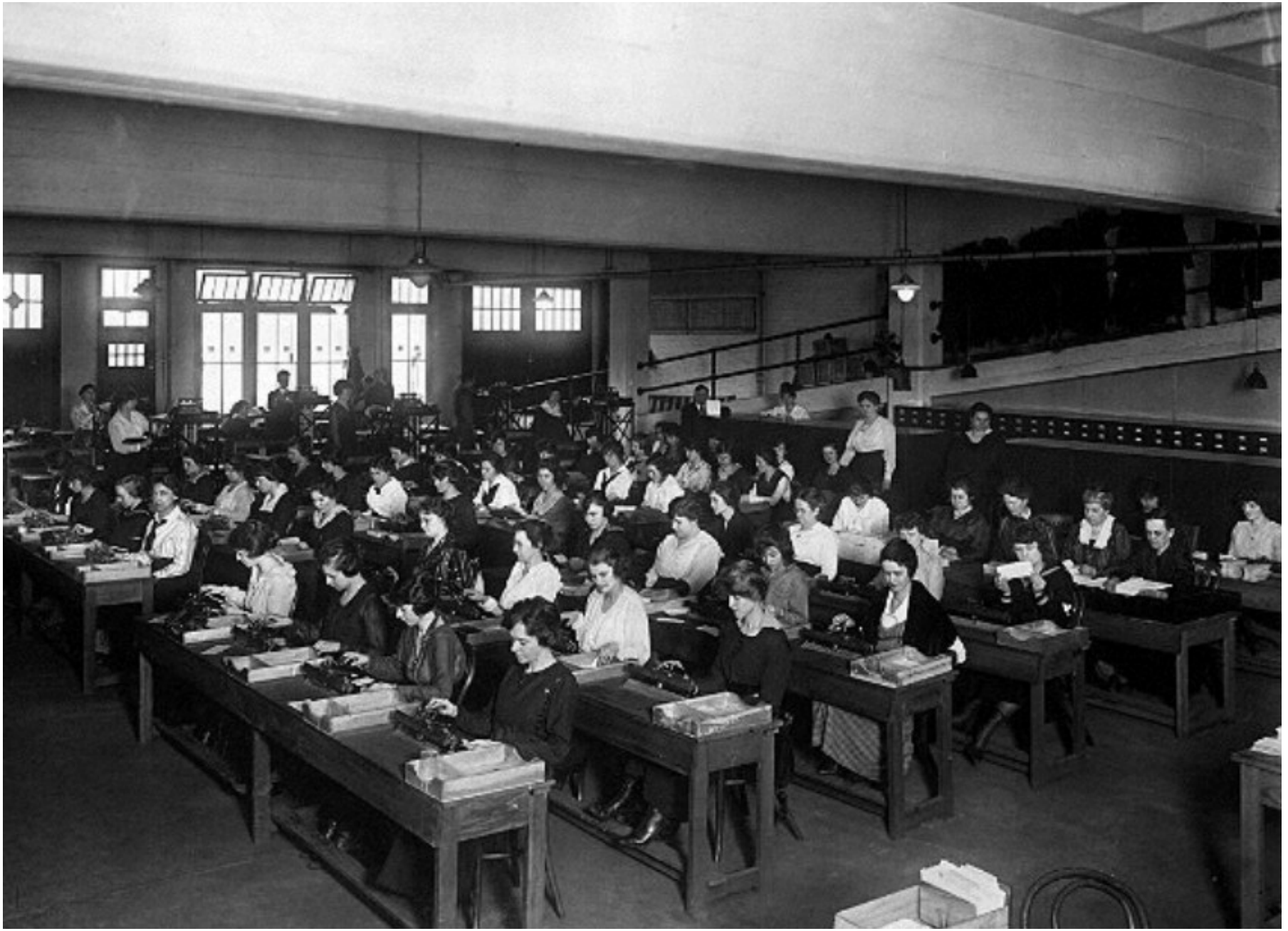
(al final del curso vamos a ver brevemente como se programan las gpus de intel!)

Mi nuevo programa paralelo si correrá más rápido cuando cambie mi GPU



Performance → paralelismo

Women at work tabulating during World War II



https://www.atomicheritage.org/history/human-computers-los-alamos

Realismo → paralelismo

- For example, imagine modeling these serially:


Galaxy Formation


Planetary Movments


Climate Change

Rush Hour Traffic


Plate Tectonics

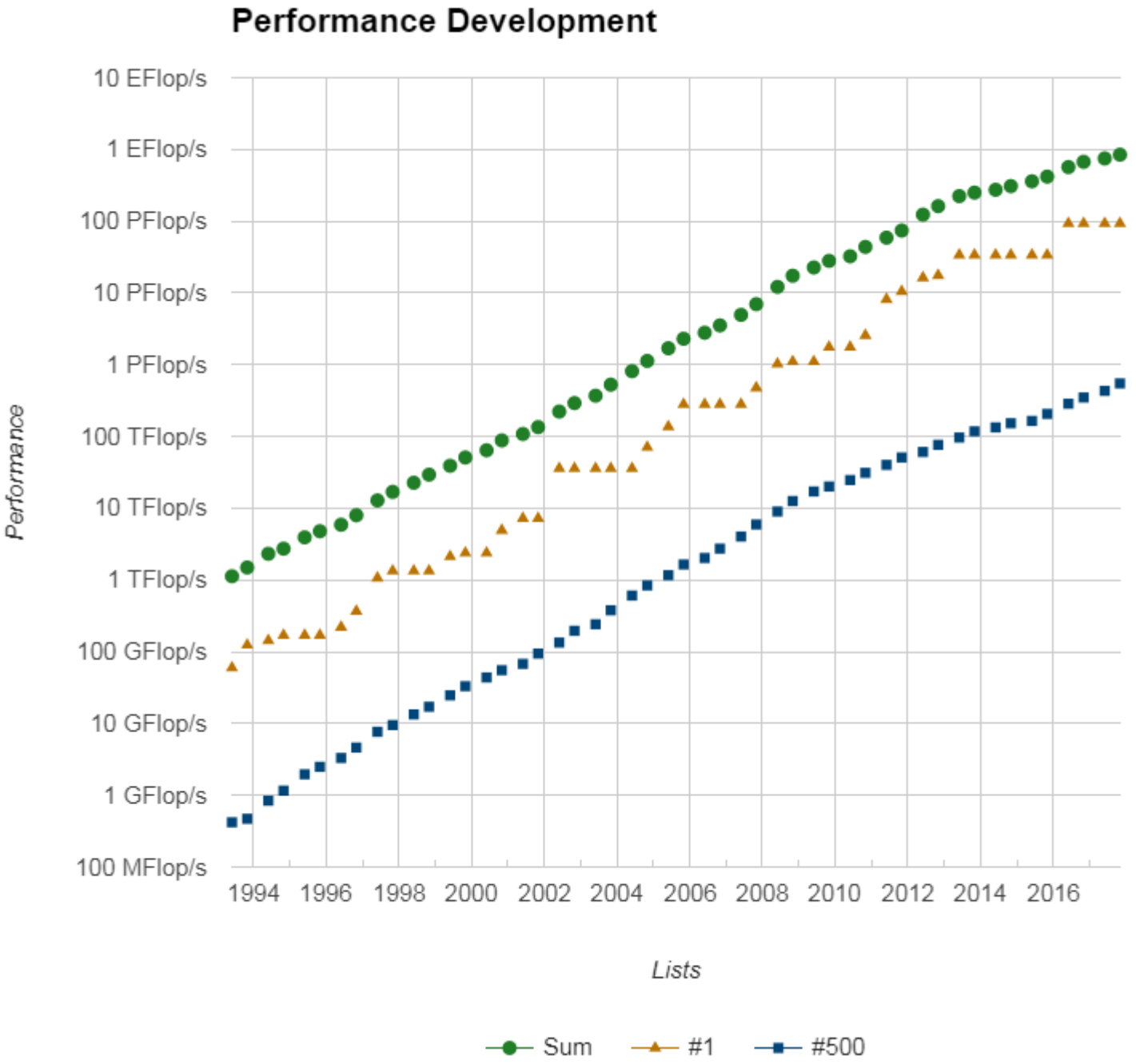

Weather


Auto Assembly


Jet Construction


Drive-thru Lunch

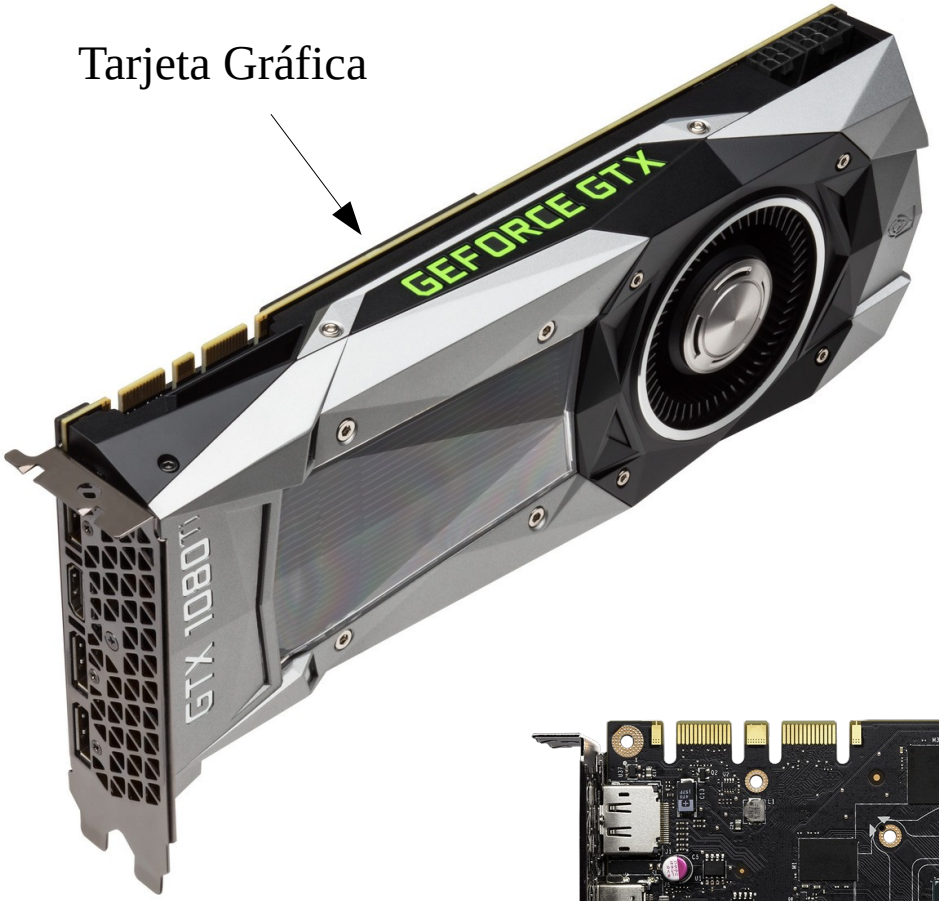
https://computing.llnl.gov/tutorials/parallel_comp/



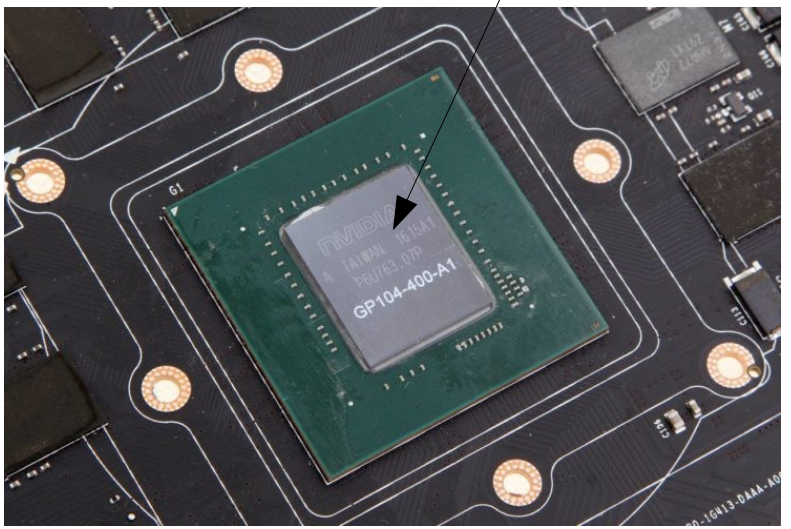
Source: Top500.org

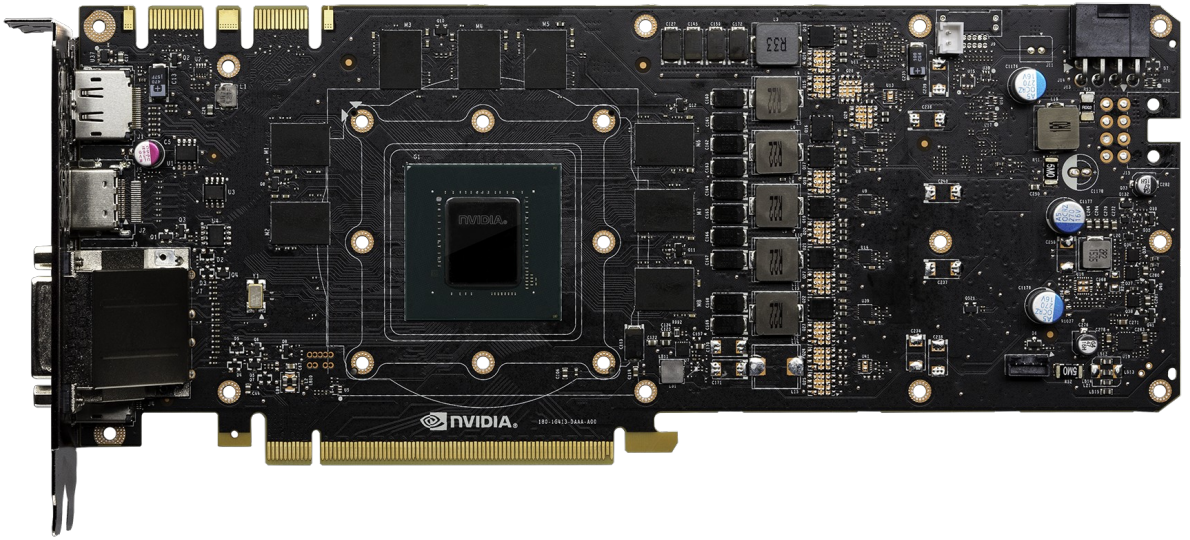
GPGPU: “computadoras paralelas”

Tarjeta Gráfica



GPU





Tipos de GPUs

- **Dedicadas**: aplicaciones graficas, juegos. Alta performance. Precios accesibles...
- **Integradas**: están integradas con la CPU. Menos poderosas que las dedicadas, pero más baratas.
- **Móviles**: precesadores gráficos para laptops, tablets, celulares. Menos poderosas que una dedicada pero optimizadas para consumir poco.
- **En la nube**: son gpus virtualizadas que pueden ser accedidas en forma remota. Se usa mucho en machine learning, etc.
- **Workstation**: para aplicaciones profesionales, modelado 3D, edición de video, computación científica. Alta performance, confiables, y permiten cosas avanzadas como multi-GPU y disponer de mucha memoria, etc. Caras.

Marcas

- **NVIDIA**: es la líder, para consumo y profesional. Línea GeForce para juegos. Quadro y Tesla para aplicaciones profesionales (modelado 3D, ciencia, inteligencia artificial, etc).
- **AMD**: Consumo y profesional. Línea Radeon para juegos, Radeon Pro, Instinct para aplicaciones profesionales.
- **Intel**: aunque es conocida principalmente por sus CPUs, hacen GPUs integradas con sus CPUs. Línea Iris and Iris Pro, y Arc, se usan en laptops and desktops. Ponte Vecchio para HPC (¡vamos a tener un cluster con estas últimas en Argentina pronto!).
- **Qualcomm**: GPUs para “mobile devices”, smartphones y tablets. Línea Adreno es popular para Android.
- **Imagination Technologies**: mobile devices, automotive y IoT (Internet of Things). Smartphones y tablets.
- **ARM**: mobile devices, automotive y IoT. Línea Mali usada en smartphones y tablets.

Mas operaciones por segundo...

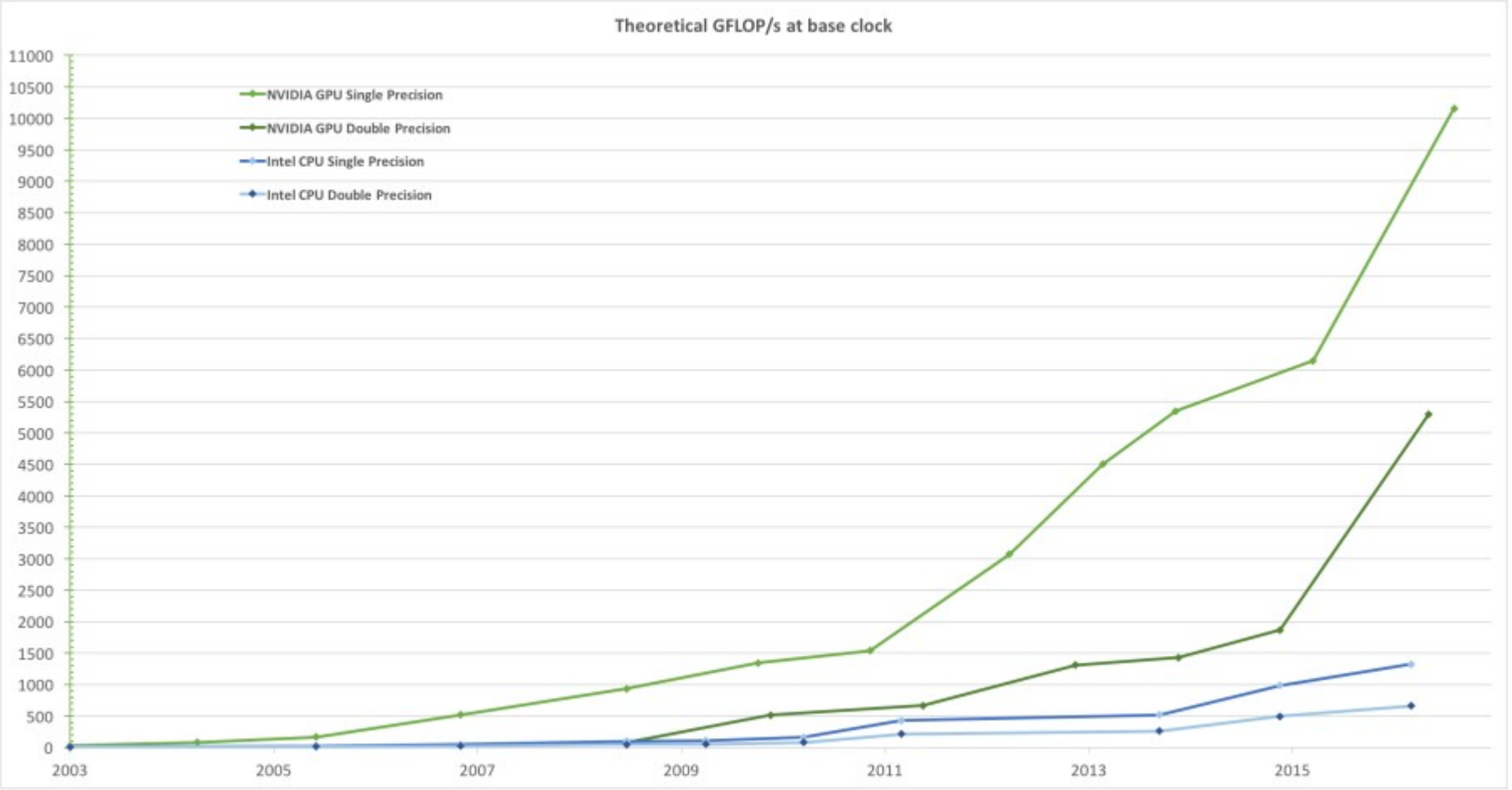


Figure 1 Floating-Point Operations per Second for the CPU and GPU

Memorias más rápidas...

Aumentó la velocidad de transferencia de datos

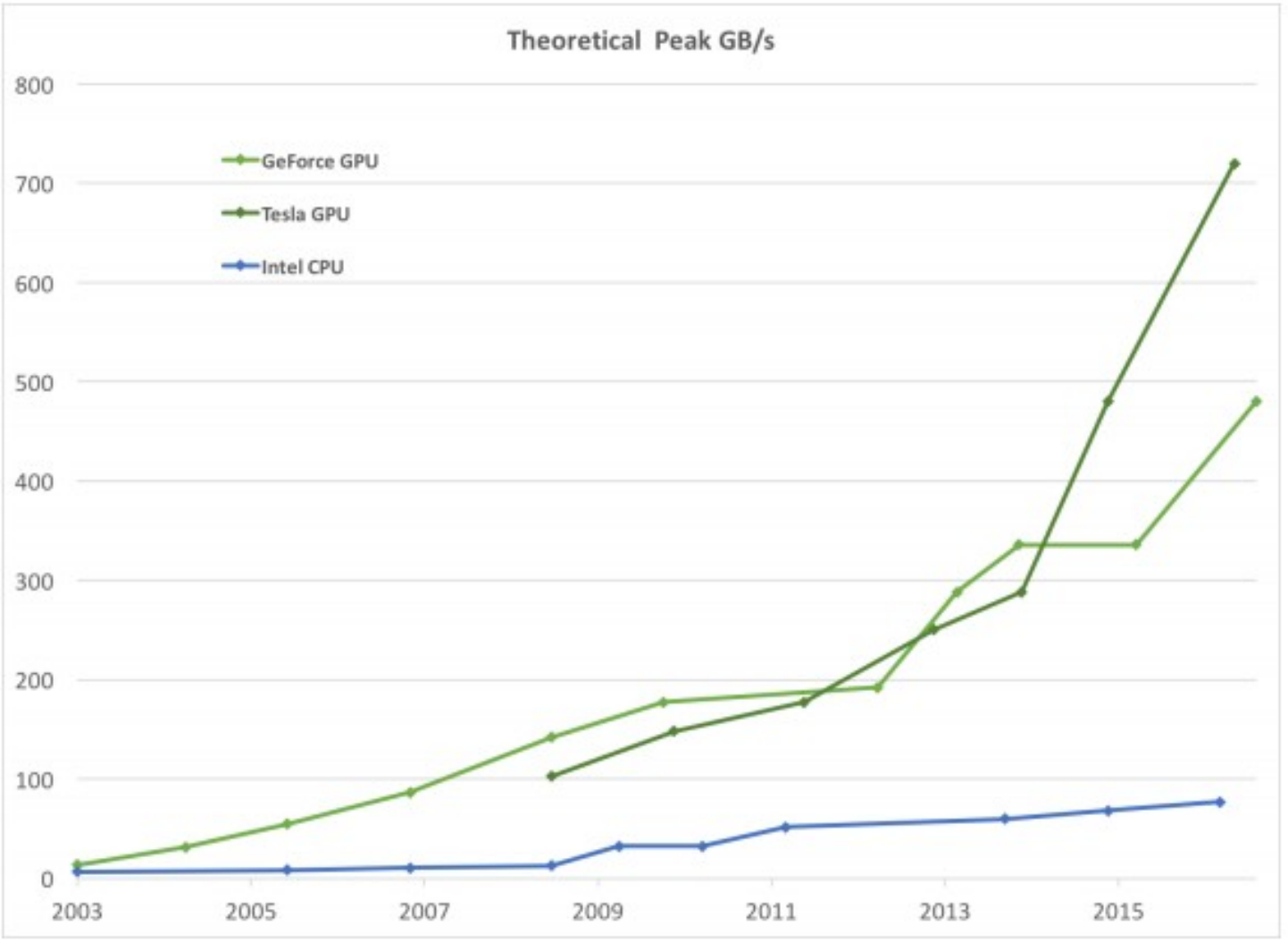


Figure 2 Memory Bandwidth for the CPU and GPU

¿Porque sirven como aceleradoras de cálculo?

The reason behind the discrepancy in floating-point capability between the CPU and the GPU is that the GPU is specialized for compute-intensive, highly parallel computation - exactly what graphics rendering is about - and therefore designed such that more transistors are devoted to data processing rather than data caching and flow control, as schematically illustrated by Figure 3.

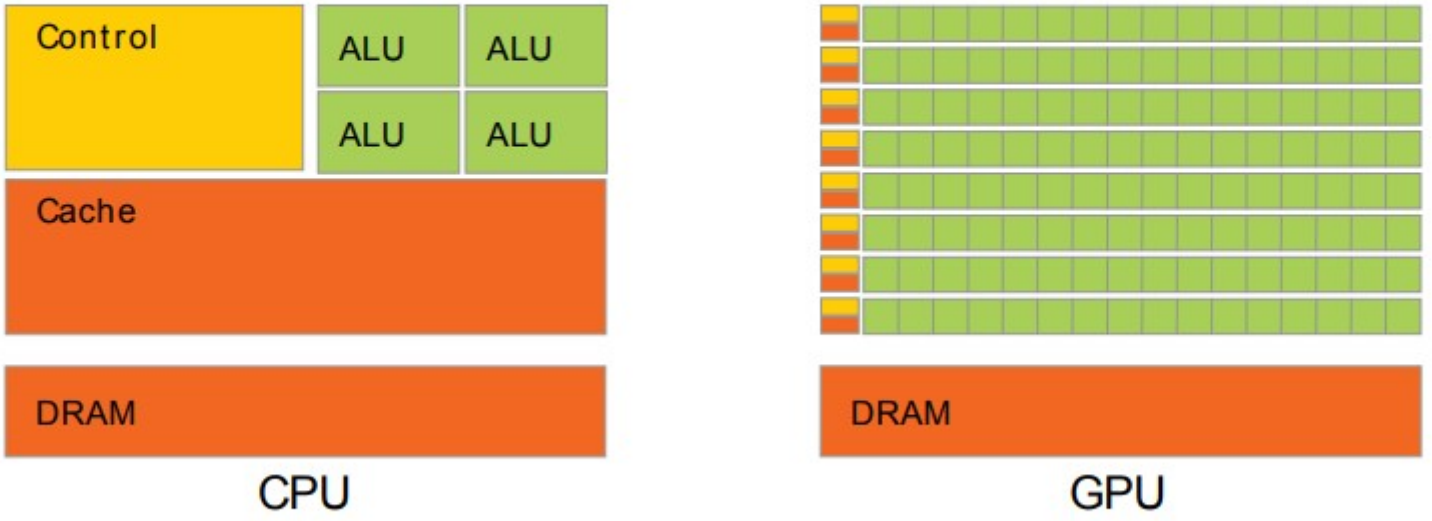
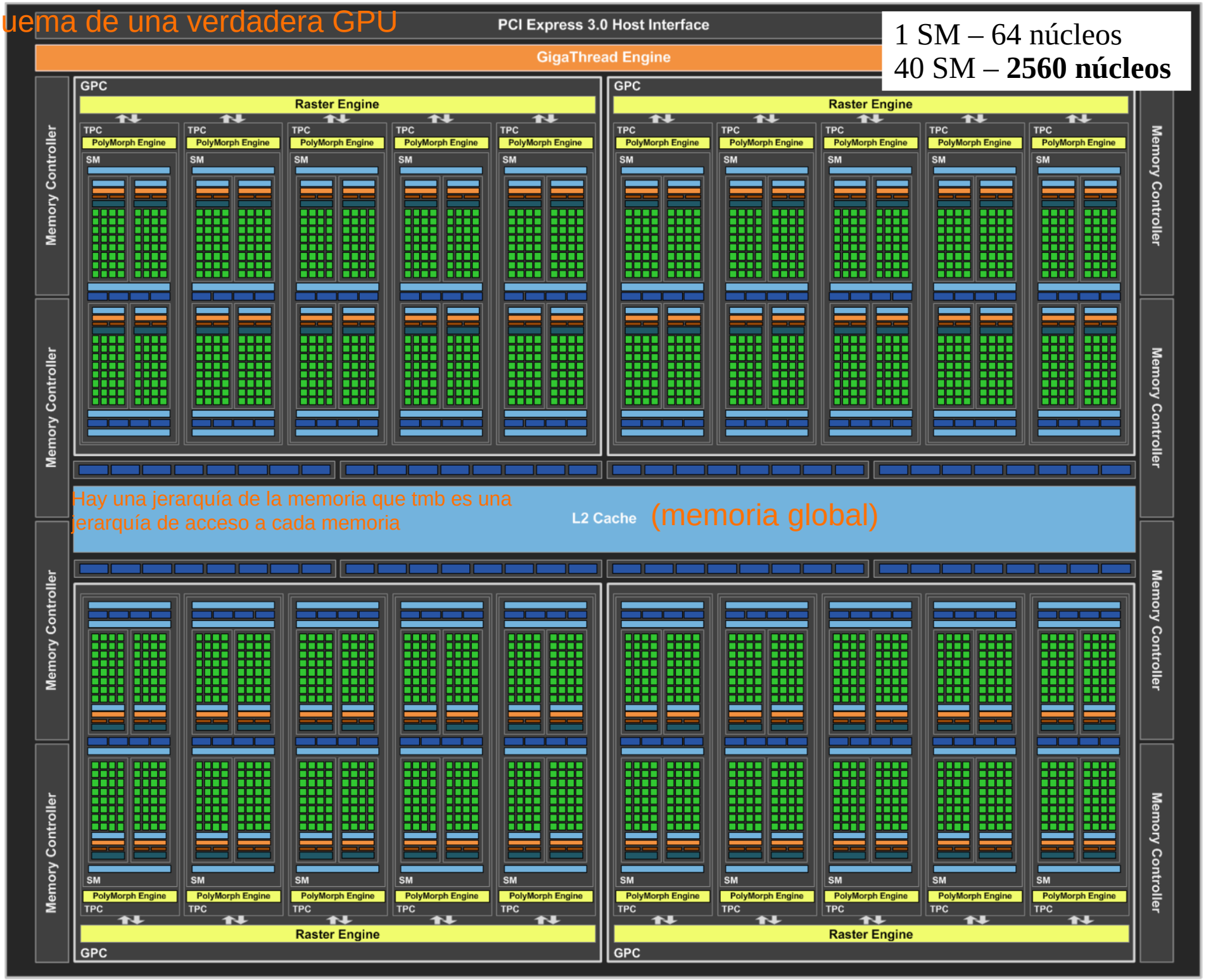


Figure 3 The GPU Devotes More Transistors to Data Processing

Los núcleos de una GPU son menos poderosos que los de una CPU pero la ventaja es que son muchos

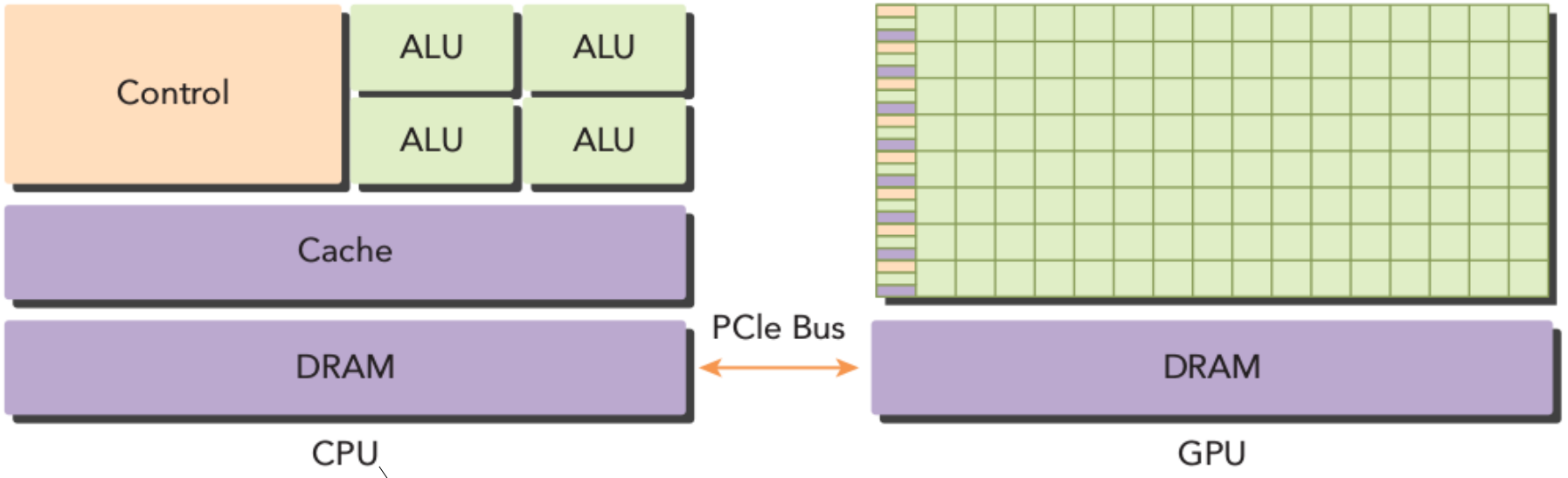
Esquema de una verdadera GPU

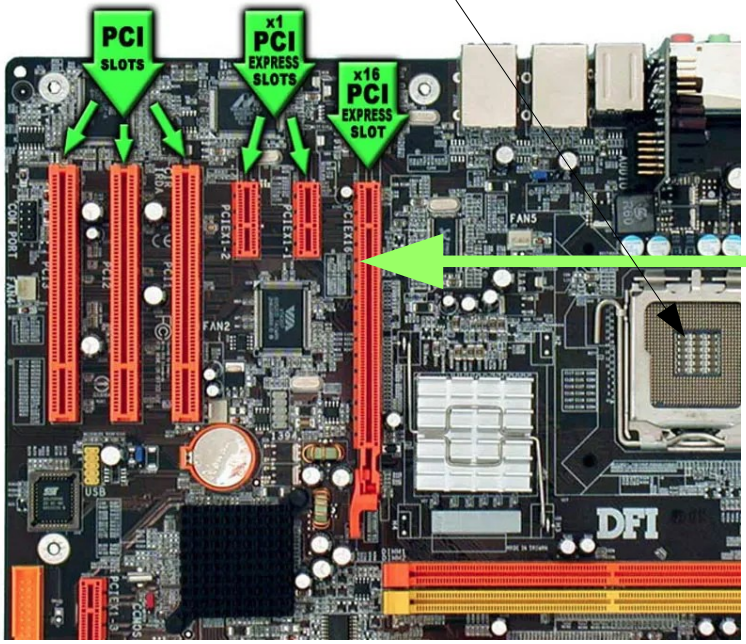


1 SM – 64 núcleos
40 SM – **2560 núcleos**

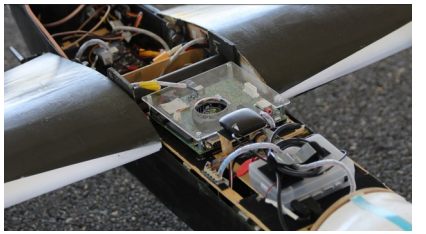
La conexión entre GPU y CPU se da a través de un puerto que es lento. Además, las GPU consumen mucho y es necesario tener una fuente acorde

Arquitectura paralela híbrida





Todas se programan igual



Las GPU tmb se pueden colocar en sistemas envevidos

Procesamiento paralelo con GPU sobre un robot o drone

Jetson TK1

127 x 127 mm, 5 Watts, 70 UDS

Jetson nano

70 x 45 mm

90USD

GPU: NVIDIA Kepler "GK20a" GPU with 192 SM3.2 CUDA cores (upto 326 GFLOPS)

CPU: NVIDIA "4-Plus-1" 2.32GHz ARM quad-core Cortex-A15 CPU with Cortex-A15 battery-saving shadow-core

Inception Spotlight: New Skydio 2 Drone Powered by NVIDIA Jetson

(2019)

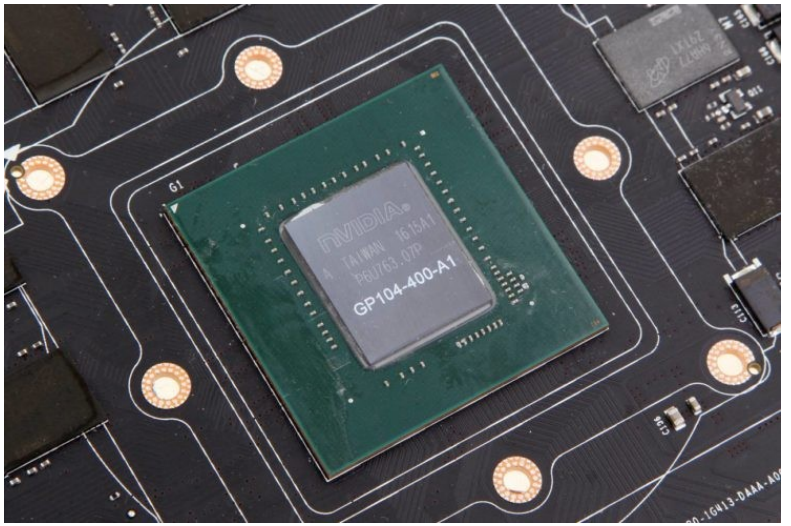
https://youtu.be/imt2qZ7uw1s

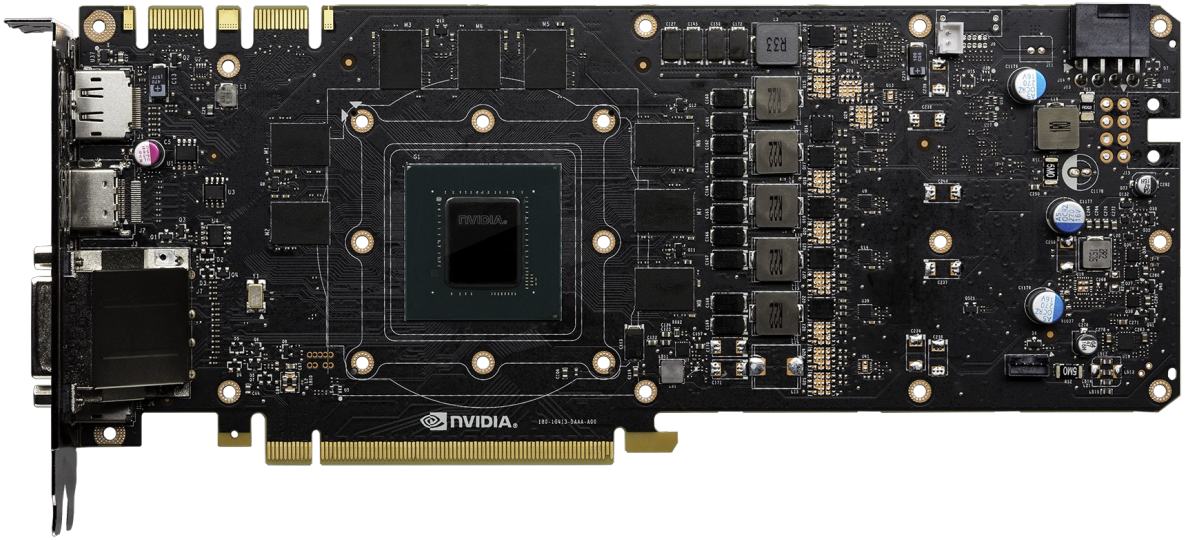
https://youtu.be/USYlt9t0lZY

Autonomous Drone Flight Over 1 Kilometer Forest Trail (2017)

ETC

¿Como se programan estas cosas?





Complicado hasta que salió CUDA en 2007

(estaba demasiado orientado a gráficos)