

SISTEMAS PARALELOS

Clase 9 - Tendencias en HPC



FACULTAD DE INFORMATICA



UNIVERSIDAD
NACIONAL
DE LA PLATA

Agenda de esta clase

- Tendencias en HPC

TENDENCIAS EN HPC

Impacto del consumo energético

- El objetivo de HPC ha sido incrementar el rendimiento y ocasionalmente el cociente precio/rendimiento



TOP500

- Ranking que lista las 500 supercomputadoras más potentes del mundo.
 - Comenzó en 1993 manteniéndose hasta la actualidad.
 - Se actualiza 2 veces al año (junio y noviembre).
 - Para el cálculo de la potencia se emplea un benchmark específico llamado LINPACK.

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	DOE/NNSA/LLNL United States	El Capitan - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot- 11, TOSS HPE	11,039,616	1,742.00	2,746.38	29,581
2	DOE/SC/Oak Ridge National Laboratory United States	Frontier - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct	9,066,176	1,353.00	2,055.72	24,607

Impacto económico del consumo energético de HPC

- Notable incremento en el consumo de energía eléctrica



El Capitán
29.5MW
US\$ 50150000



Frontier
24.6MW
US\$ 41800000



Aurora
38.7MW
US\$ 65790000



Eagle
?
US\$?

Impacto económico del consumo energético de HPC



El Capitán

29.5MW

US\$ 50150000

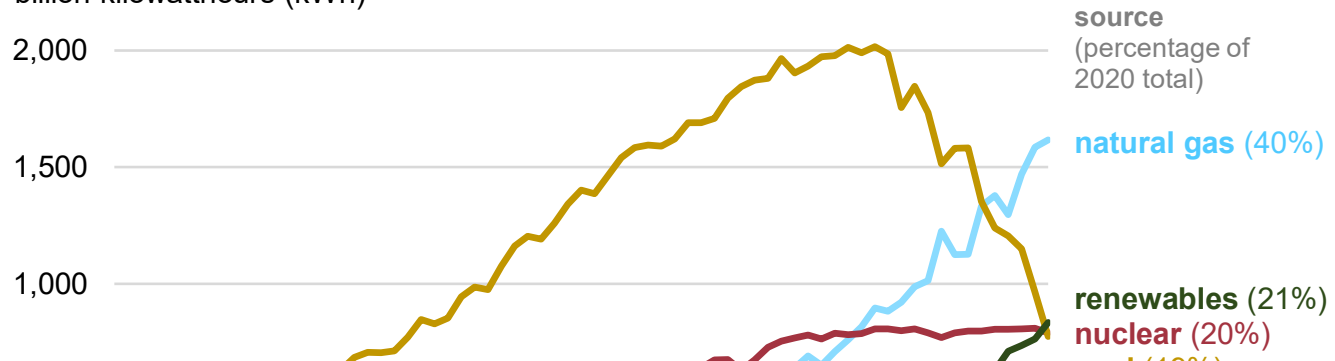
- Ciudad de Santa Fe, Argentina posee 391164 habitantes (INDEC, 2010)

- En Argentina, la potencia eléctrica demandada por un hogar es 0.25 kW (prom.)
 - La potencia demandada por El Capitán equivale a 118000 hogares
 - Asumiendo 4 residentes por hogar, Fugaku equivale al consumo de 472000 residentes.



Impacto social y medioambiental de HPC

Annual U.S. electricity generation from all sectors (1950–2020)
billion kilowatthours (kWh)



OMS estima que 7 millones de muertes ocurren cada año debido a la contaminación atmosférica

En las Américas, mueren más de 131 mil personas en países de bajos ingresos y 96 mil en países de altos ingresos por causas vinculadas a la polución del aire

GINEBRA | 25 de marzo de 2014 — En nuevas estimaciones publicadas este 25 de marzo, la Organización Mundial de la Salud (OMS) informa que en 2012 unos 7 millones de personas murieron —una de cada ocho del total de muertes en el mundo— como consecuencia de la exposición a la contaminación atmosférica. Esta conclusión duplica con creces las estimaciones anteriores y confirma que la contaminación atmosférica constituye en la actualidad, por sí sola, el riesgo ambiental para la salud más importante del mundo. Si se redujera la contaminación atmosférica podrían salvarse millones de vidas.

- Representa uno de los mayores obstáculos para superar la escala de FLOPS

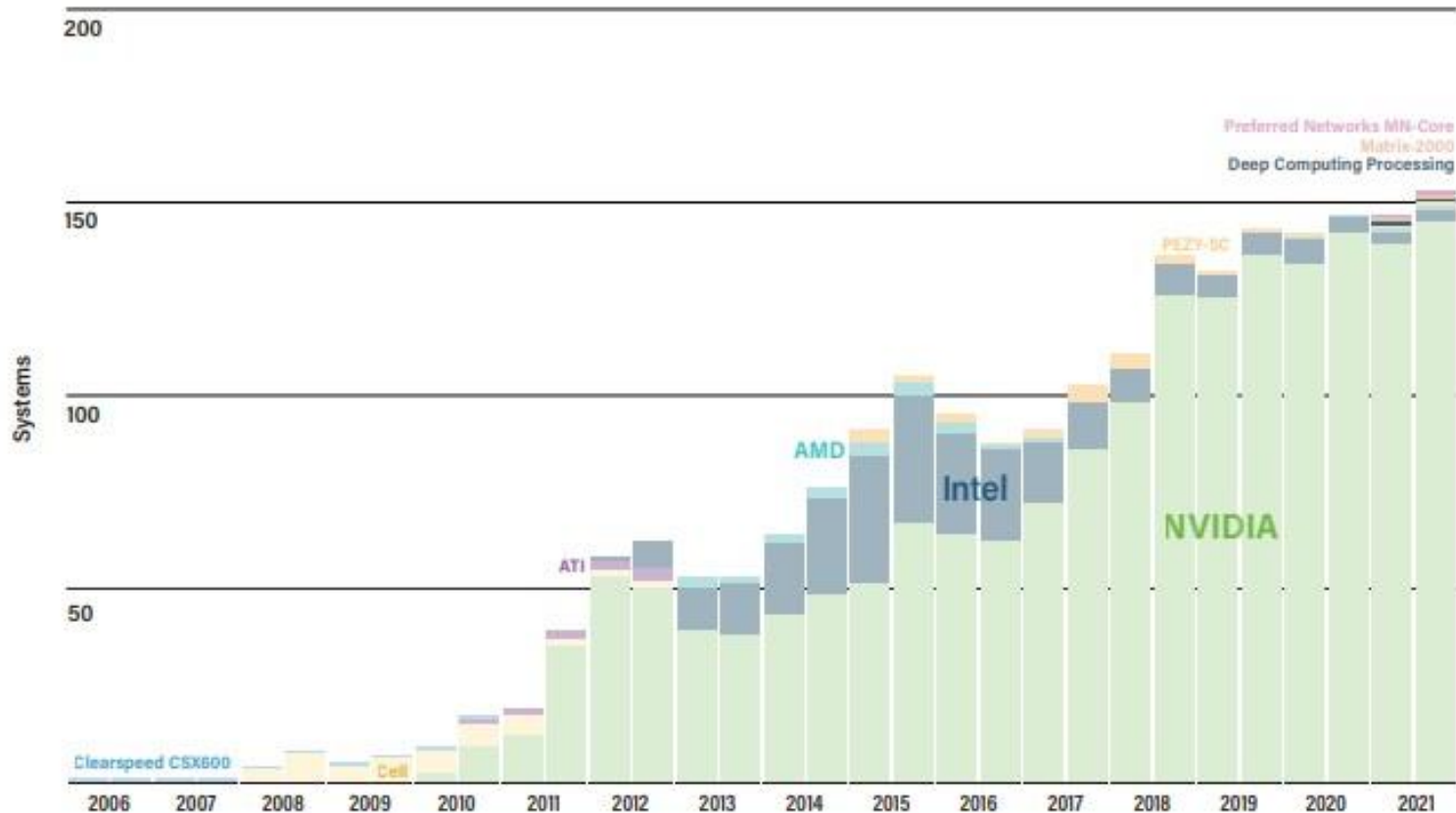


GREEN500

- Ranking que lista las 500 supercomputadoras más eficientes desde el punto de vista energético del mundo.
 - Comenzó en 2006 manteniéndose hasta la actualidad.
 - Se actualiza 2 veces al año (junio y noviembre).
 - Al igual que el TOP500, para el cálculo de la potencia se emplea un benchmark específico llamado LINPACK. Además, se debe medir el consumo energético durante su ejecución

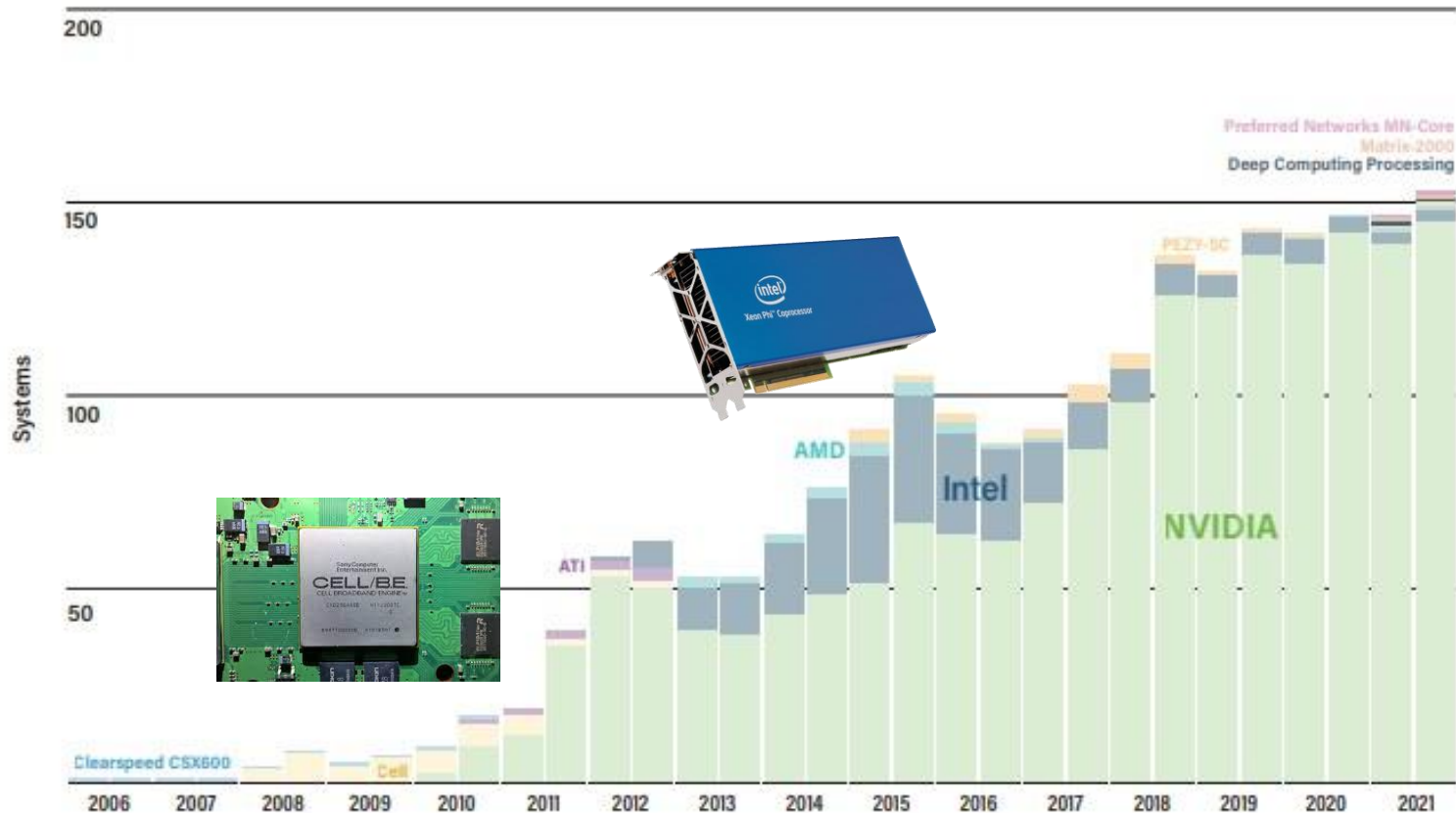
Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
1	222	JEDI - BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, ParTec/EVIDEN EuroHPC/FZJ Germany	19,584	4.50	67	72.733
2	122	ROMEO-2025 - BullSequana XH3000	47,328	9.86	160	70.912

Consolidación de aceleradores en HPC

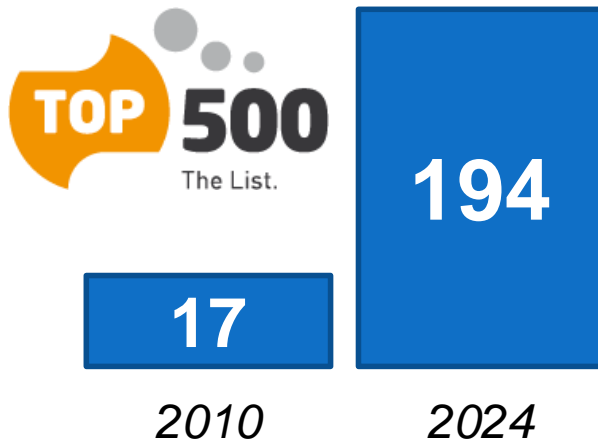


Acelerador: dispositivo de hardware diseñado para mejorar el rendimiento del sistema

Consolidación de aceleradores en HPC

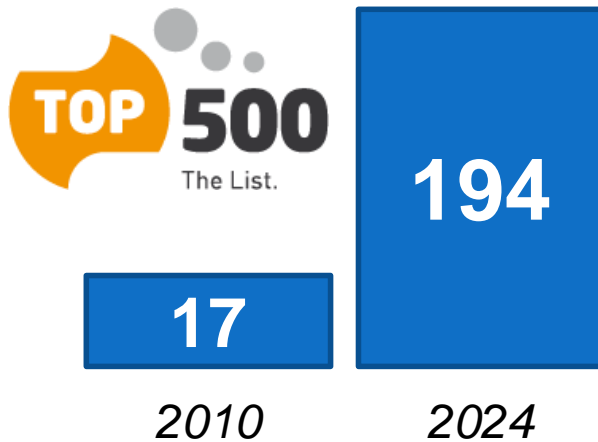


Reciente consolidación de aceleradores en HPC



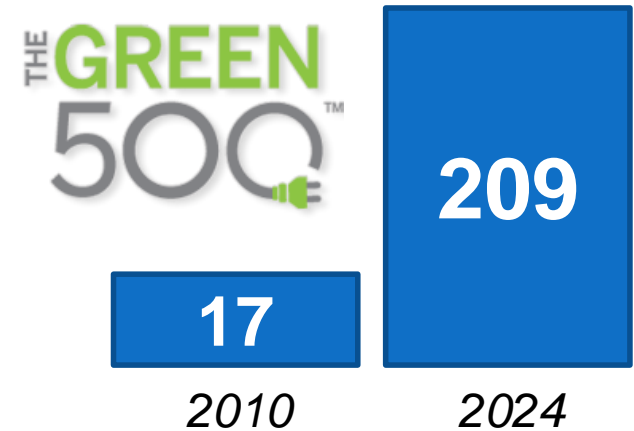
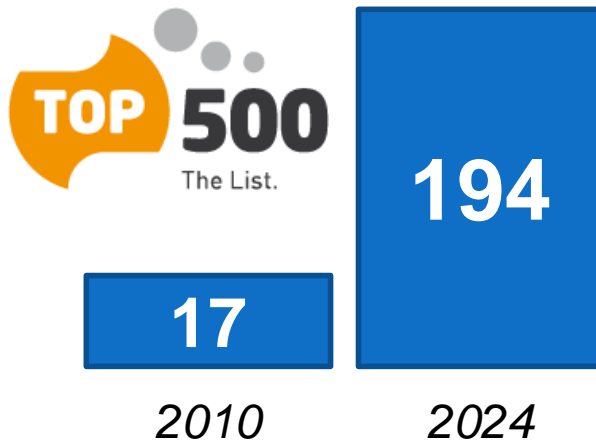
Dispositivo	GFLOPS (SP)
Intel® Xeon® Platinum 8360Y Processor (36 cores, 72 ht, 2.3-3.4 Ghz)	2650
NVIDIA A100 PCIe (6912 núcleos CUDA)	19960

Reciente consolidación de aceleradores en HPC



Dispositivo	GFLOPS (SP)	Watt (TDP)	GFLOPS/Watt
Intel® Xeon® Platinum 8360Y Processor (36 cores, 72 ht, 2.3-3.4 Ghz)	2650	250	10.6
NVIDIA A100 PCIe (6912 núcleos CUDA)	19960	250	79.84

Reciente consolidación de aceleradores en HPC



Dispositivo	GFLOPS (SP)	Watt (TDP)	GFLOPS/Watt
Intel® Xeon® Platinum 8360Y Processor (36 cores, 72 ht, 2.3-3.4 Ghz)			
NVIDIA A100 PCIe (6912 núcleos CUDA)	19960	250	79.84

Rendimiento mejora 8x
Consumo de potencia se mantiene
Eficiencia energética mejora 8x

Crecimiento del uso de hardware especializado en otras áreas

- FPGAs

HPC | wire

Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them

- Home
- Technologies
- Sectors
- Exascale

[Home](#) » [HPC Software](#) » [Cloud HPC](#) » Accelerating the Alibaba Cloud with Intel Arria 10 FPGAs

Accelerating the Alibaba Cloud with Intel Arria 10 FPGAs

March 11, 2017 by [Rich Brueckner](#) [Leave a Comment](#) [Print](#)

The Alibaba Cloud has announced a joint

ernet of
sion of

HPC | wire

Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them

- Home
- Technologies
- Sectors
- Exascale
- Specials
- Resource Library
- Events
- Job Bank



CERN openlab Explores New CPU/FPGA Processing Solutions

By Linda Barney

April 13, 2017

Editor's note: In this contributed feature, Linda Barney describes the ongoing technical collaboration between CERN and Intel to develop a co-packaged Xeon/FPGA processor.



Intel's FPGAs Target Datacenters, Networking

By George Leopold

[Home](#) » [Industry Segments](#) » [Datacenter](#) » FPGAs Accelerate Machine Learning at Baidu

FPGAs Accelerate Machine Learning at Baidu

October 30, 2016 by [Rich Brueckner](#) [Print](#)

- Home
- Technologies
- Sectors
- Exascale

[Home](#) » [HPC Software](#) » [High Performance Analytics](#) » Intel FPGAs Break Record for Deep Learning Facial Recognition

Intel FPGAs Break Record for Deep Learning Facial Recognition

[Home](#) / [News](#) / Microsoft Goes All in for FPGAs to Build Out AI Cloud

Microsoft Goes All in for FPGAs to Build Out AI Cloud

Michael Feldman | September 27, 2016 08:42 CEST

[@ E-mail](#) [Tweet](#) [f Like](#) [G +1](#) [in Share](#) 168

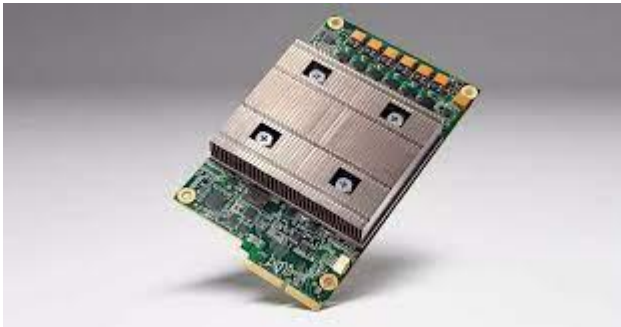
Software giant bets the (server) farm on reconfigurable computing

Microsoft has revealed that Altera FPGAs have been installed across every Azure cloud server, creating what the company is calling "the world's first AI supercomputer." The deployment spans 15 countries and represents an aggregate performance more than one exa-ops. The announcement was made by Microsoft CEO Satya Nadella and engineer Doug Burger during

Web Ser
promise
workloa
sors.

Crecimiento del uso de hardware especializado en otras áreas

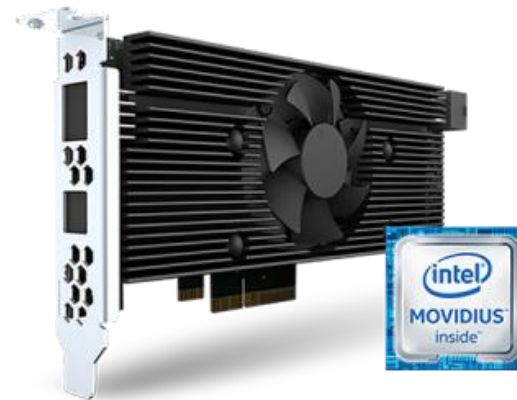
- ASICs (especialmente para IA)



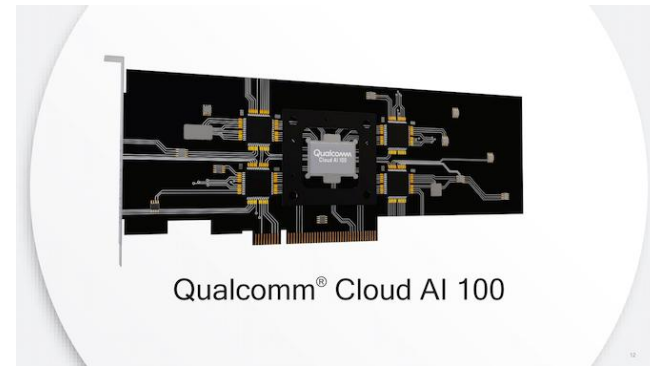
Google TPUs



Apple neural engine



Intel VPUs



Qualcomm® Cloud AI 100

Popularización de hardware especializado

CPUs



GPUs



Xeon Phi's



FPGAs



ASICs



Desafío de programación

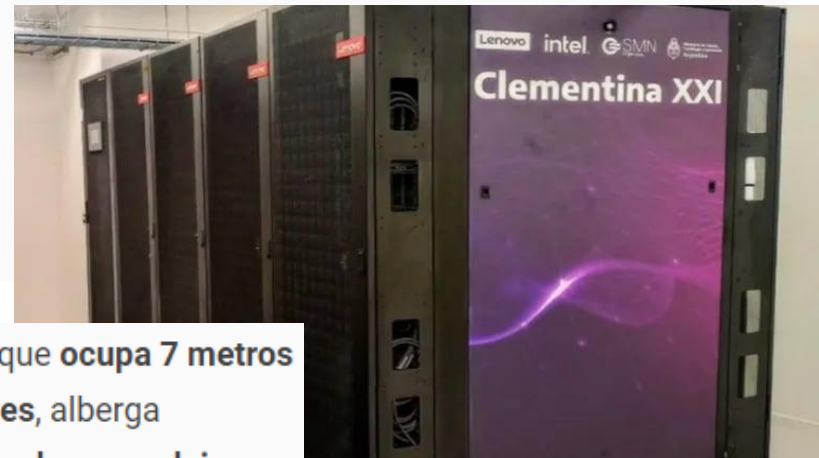
- × Aumenta costos de programación y complejidad
- × Complica mantenimiento y extensión de código a futuro

¿Hay supercomputadoras en Argentina?

La última supercomputadora argentina

Así es Clementina XXI, la computadora argentina que está entre las 100 más poderosas del mundo

Con un poder de procesamiento equivalente a 1672 PlayStation 5 o a 2903 MacBook Pro, sirve para realizar tareas que demandan una enorme cantidad de cálculos, como investigaciones de genómica, biomedicina, estudios de inteligencia artificial y ciencia de datos.



Clementina XXI es un monstruo de la tecnología. Con un tamaño que **ocupa 7 metros cuadrados** y un **consumo energético** equivalente al de 300 hogares, alberga múltiples procesadores que trabajan en conjunto para realizar **cálculos complejos a velocidades inimaginables**.

74 nodos con 4 GPUs c/u

Clementina XXI tiene un **rendimiento máximo de 15,7 petaFLOPS**, 5120 núcleos de la serie Intel® Xeon® CPU Max y 37,888 núcleos Intel® Data Center GPU Max Series. Eso la convierte en **la segunda supercomputadora más poderosa de América Latina**.

<https://tn.com.ar/tecno/novedades/2024/02/20/asi-es-clementina-xxi-la-computadora-argentina-que-esta-entre-las-100-mas-poderosas-del-mundo/>

<https://ccad.unc.edu.ar/files/Clementina-XXI.pdf>

Otras supercomputadoras de la UNC

Equipos en producción

Se consigna el rango de años desde que se instalaron hasta que se terminaron de ampliar.

- [Mendieta Fase 2](#), 2022- .
- [Searfín](#), 2021- .
- [Eulogia](#), 2018-2021.
- [Mulatona](#), 2018.

Equipos retirados

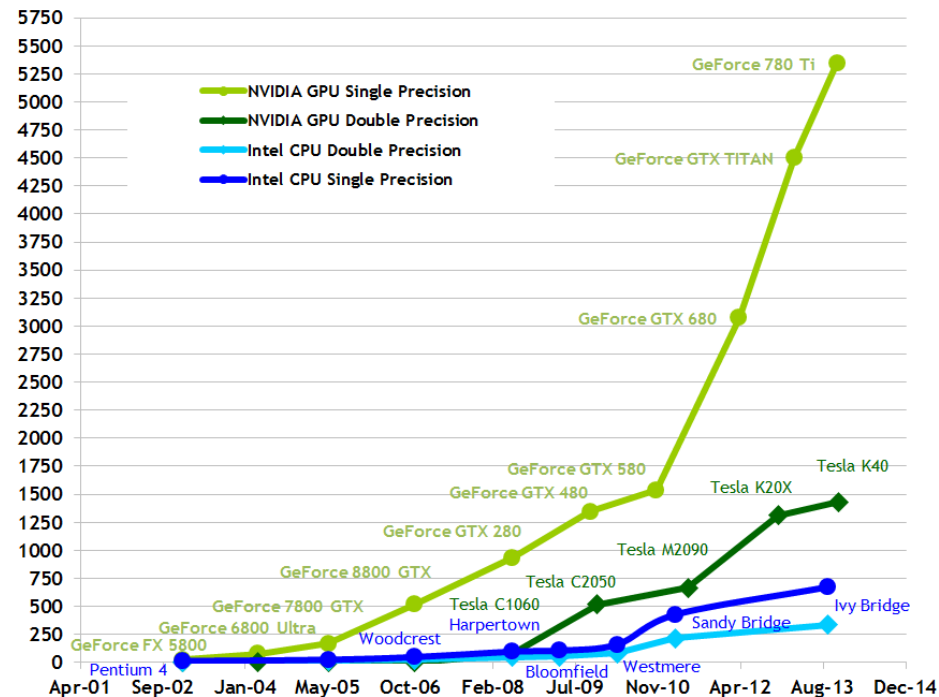
- [Mendieta](#), 2012-2013.
- [Cluster SeCyT](#) (Facultad de Ciencias Exactas Físicas y Naturales).
- [Cristina](#) (Facultad de Ciencias Químicas – Instituto de Investigaciones Fisicoquímicas de Córdoba).

GPUs

GPUs

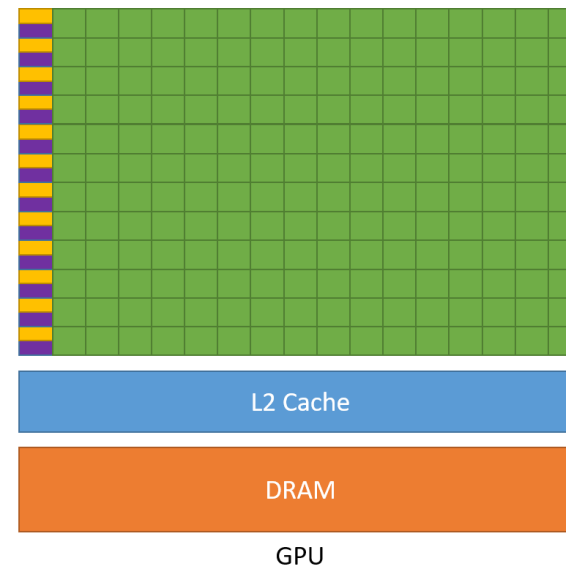
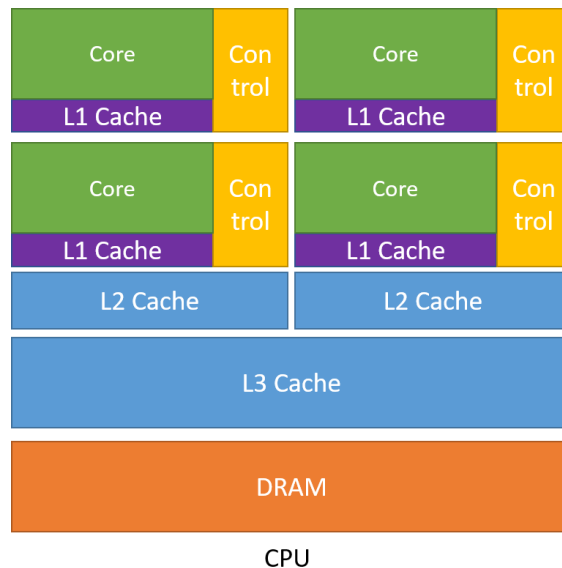
- Originalmente diseñadas para procesamiento de gráficos. Debido a su gran potencia de cálculo, las empresas fabricantes comenzaron a aumentar su grado de programación
- Motivó el surgimiento de nuevas técnicas, lenguajes y herramientas para la programación de GPUs, lo cual permite utilizar a las mismas como arquitecturas paralelas para resolver problemas de propósito general

Theoretical GFLOP/s



GPUs

- La significativa diferencia de rendimiento que existe entre las CPUs y las GPUs se debe a que sus filosofías de diseño son muy distintas
 - CPUs destinan los recursos de silicio principalmente a memorias caché y a núcleos de compleja organización que permitan explotar ILP.
 - GPUs emplean la mayor parte del silicio disponible en unidades funcionales. Cada una de ellas tiene un conjunto de núcleos simples que comparten lógica de control, ejecutan instrucciones en orden y operan en grupos como si fueran un procesador vectorial.

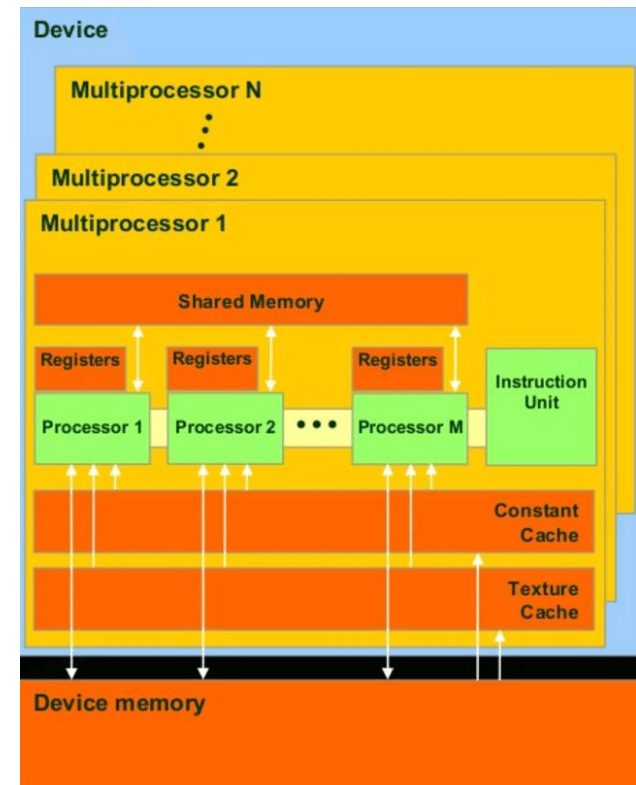


GPUs

<https://youtu.be/-P28LKWTzrI>

GPUs

- A diferencia de las CPUs, las GPUs tienen una jerarquía de memoria compleja
 - **Memoria global:** es una memoria off-chip que sirve de memoria principal. Ancho de banda limitado y latencia alta comparado a memorias on-chip o caché.
 - **Memoria compartida:** Memoria on-chip caracterizada por alto ancho de banda y baja latencia. Se administra por software y es accesible por todos los hilos activos de un multiprocesador.
 - **Memoria de constantes:** Memoria rápida pero *pequeña* y de sólo lectura, ubicada dentro de la memoria global. Es visible por todos los hilos.
 - **Memoria de texturas:** Similar a la de constantes. Memoria off-chip optimizada para localidad espacial 2D.

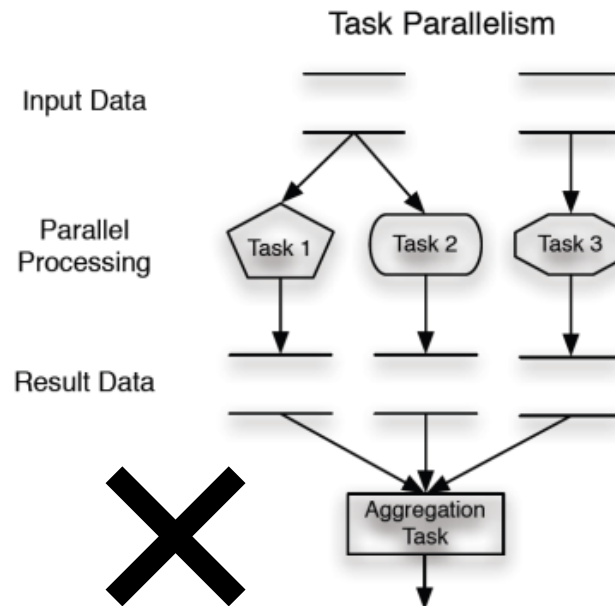
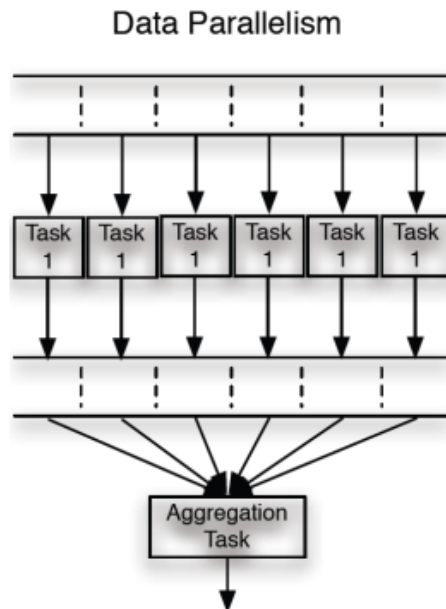


GPUs

- Las primeras aplicaciones no gráficas eran programadas en término de operaciones gráficas usando lenguajes como OpenGL o DirectX → Resultaba engorroso y propenso a errores
- Tanto la industria como la academia propusieron varios lenguajes que permiten abstraerse de los gráficos.
 - CUDA, OpenCL, OpenACC, SYCL
- Al día de hoy, son 3 las empresas que comparten el mercado de las GPUs.
 - Aunque Intel es la más grande, sólo domina el segmento correspondiente a placas integradas y de bajo rendimiento.
 - En el segmento de alto rendimiento, hasta hace 2 años AMD y NVIDIA eran los únicos proveedores (NVIDIA supera ampliamente a AMD)

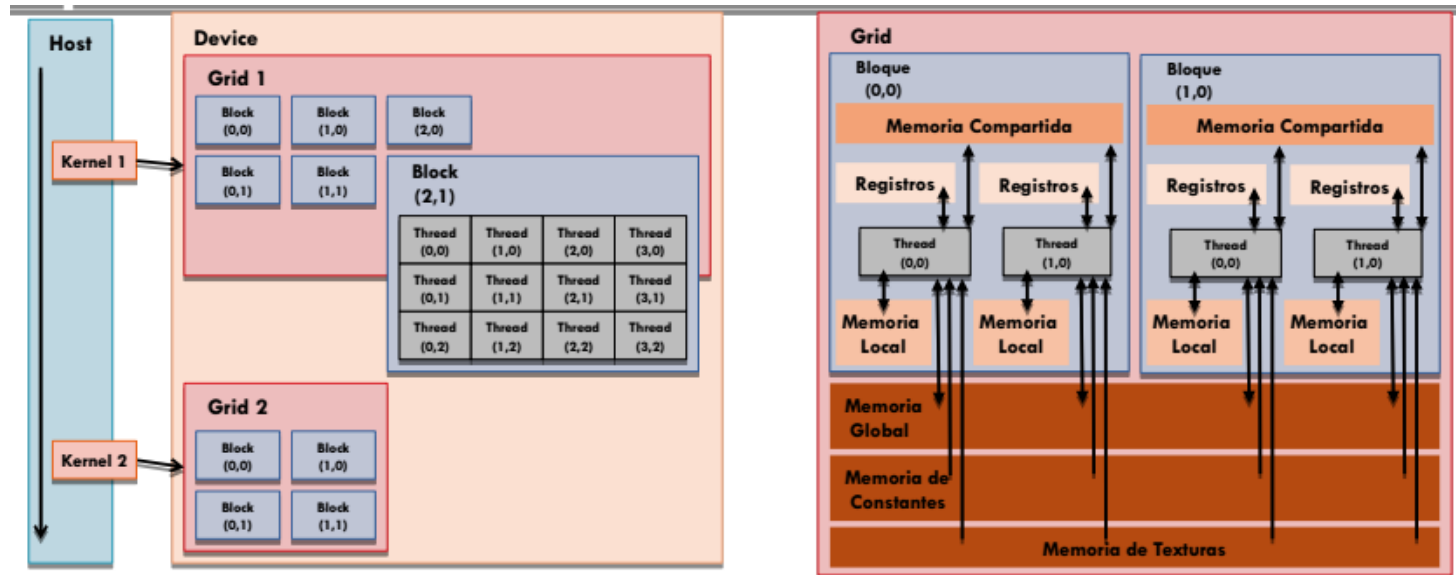
GPUs

- Las GPUs son arquitecturas de memoria compartida, inspiradas en el modelo SIMD de Flynn
- Por sus características, se adaptan mejor para aplicaciones que admiten paralelismo de datos, especialmente aquellas que son intensivas en cómputo (*CPU-bound*)



CUDA

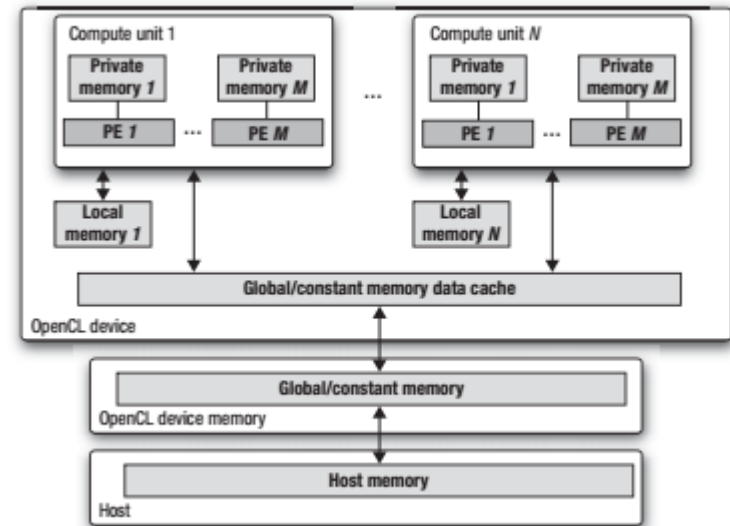
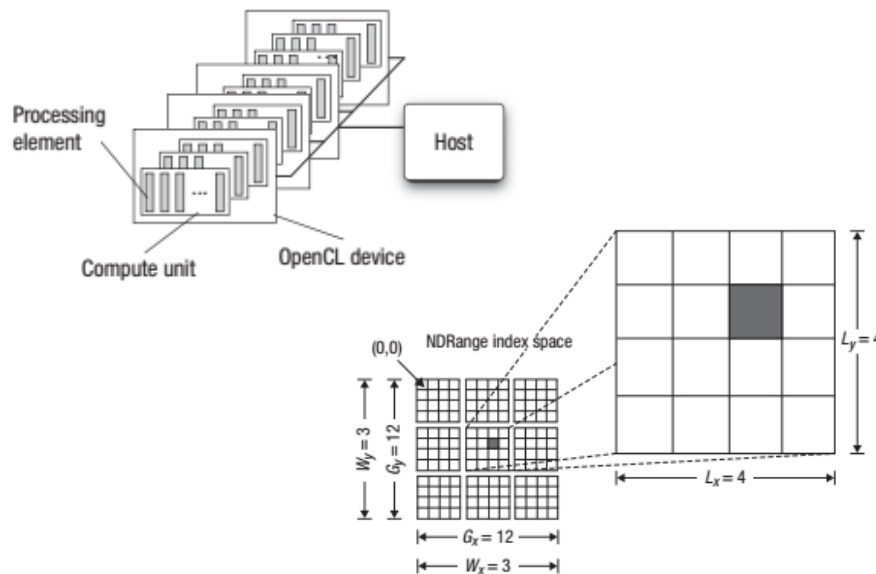
- Estándar de facto para programación de GPUs en HPC
- Modelo de ejecución (*host-device*) y arquitectura de memoria de CUDA



- El host es el responsable de administrar la memoria del dispositivos y sus transferencias, además de invocar la ejecución de los kernels.
- Un kernel es un trozo de código que ejecutan miles de hilos primitivos en paralelo en la GPU.

OpenCL

- Estándar para programación paralela multi-plataforma
- Modelo de ejecución (*host-device*) y arquitectura de memoria de OpenCL



- Similar a CUDA pero con un enfoque más general ya que no sólo aplica a GPUs (CPUs, Xeon Phi's, FPGAs, DSPs, entre otros)

SYCL

- Estándar para programación paralela multi-plataforma
- Basado en OpenCL pero buscando reducir esfuerzo de programación
 - Memoria Compartida Unificada
 - Reducciones paralelas (integradas)
 - Funciones a nivel de work-groups y sub-groups
 - Accessors
 - Interoperabilidad con otras APIs

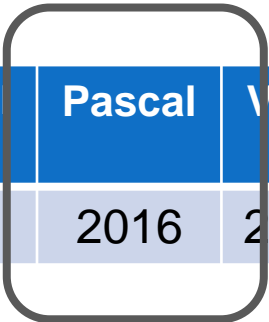
GPUs de NVIDIA

Tesla (G80)	Tesla 2.0 (GT200)	Fermi	Kepler	Maxwell	Pascal	Volta	Turing	Ampere	Hopper/Lovelace
2006	2008	2010	2012	2014	2016	2017	2018	2020	2022

- Algunas características destacadas
 - Rediseño de los SMXs pasando a llamarse SMMs → Hasta 1.35x de mejora en rendimiento por core y 2x de mejora en eficiencia energética
 - Integración con CPUs de ARM
 - Esquema de Memoria Unificada entre CPU y GPU para evitar reservas de memoria individuales (por software)

GPUs de NVIDIA

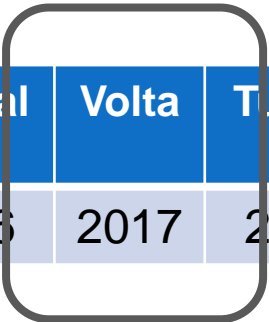
Tesla (G80)	Tesla 2.0 (GT200)	Fermi	Kepler	Maxwell	Pascal	Volta	Turing	Ampere	Hopper/Lovelace
2006	2008	2010	2012	2014	2016	2017	2018	2020	2022



- Algunas características destacadas
 - Orientada a mejorar la organización de la memoria y los buses de interconexión (adoptó HBM).
 - NVLINK, bus de alta velocidad (80 Gb/s) que reemplaza al PCIe (16 Gb/s).
 - Esquema de Memoria Unificada entre CPU y GPU para evitar reservas de memoria individuales (por hardware)
 - Soporte para precisión mixta → método que utiliza diferentes niveles de precisión dentro de una sola operación para lograr eficiencia computacional sin afectar el resultado final

GPUs de NVIDIA


Tesla (G80)	Tesla 2.0 (GT200)	Fermi	Kepler	Maxwell	Pascal	Volta	Turing	Ampere	Hopper/Lovelace
2006	2008	2010	2012	2014	2016	2017	2018	2020	2022



- Algunas características destacadas
 - Orientada al uso de Machine Learning → Incorpora soporte para precisión media (float de 16 bits) mediante núcleos específicos (Tensor cores)
 - Adopta HBM2
 - NVLINK 2.0

GPUs de NVIDIA

Tesla (G80)	Tesla 2.0 (GT200)	Fermi	Kepler	Maxwell	Pascal	Volta	Turing	Ampere	Hopper/Lovelace
2006	2008	2010	2012	2014	2016	2017	2018	2020	2022

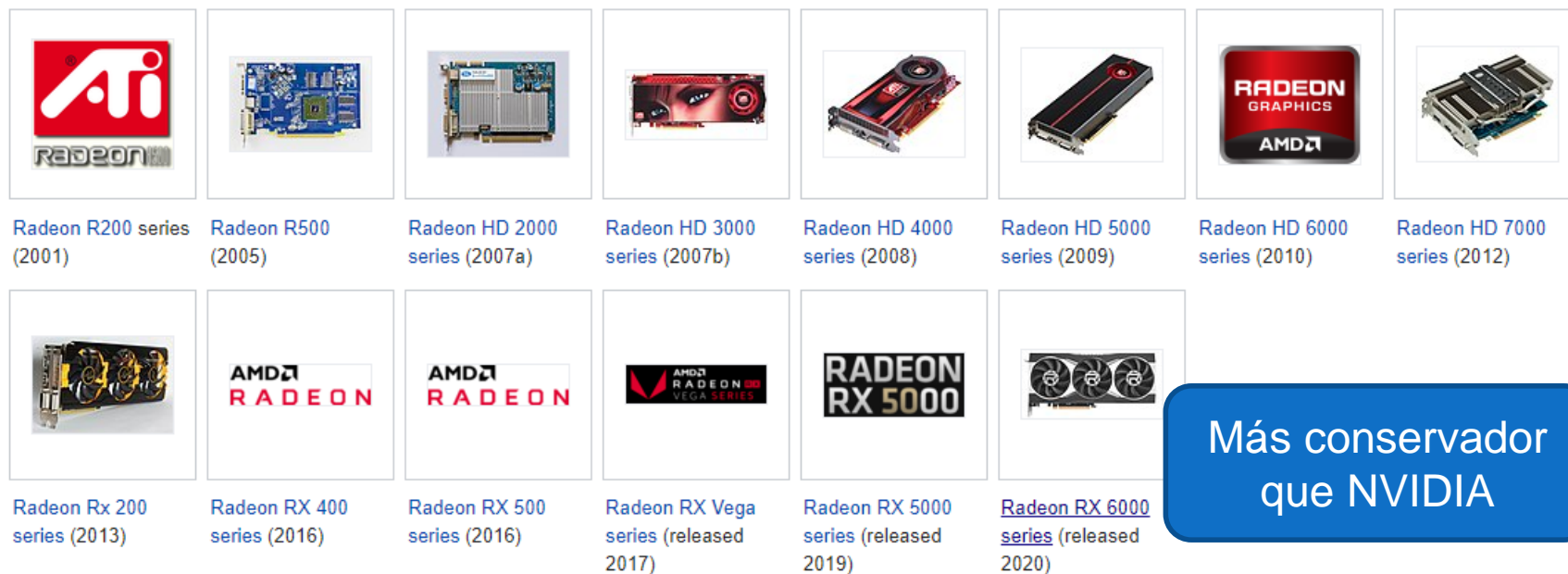


- Algunas características destacadas
 - Muy similar a Volta: Volta orientada al sector de alto rendimiento, Turing orientada al sector consumidor
 - Incorpora soporte específico para Ray-Tracing → Núcleos dedicados

GPUs de AMD

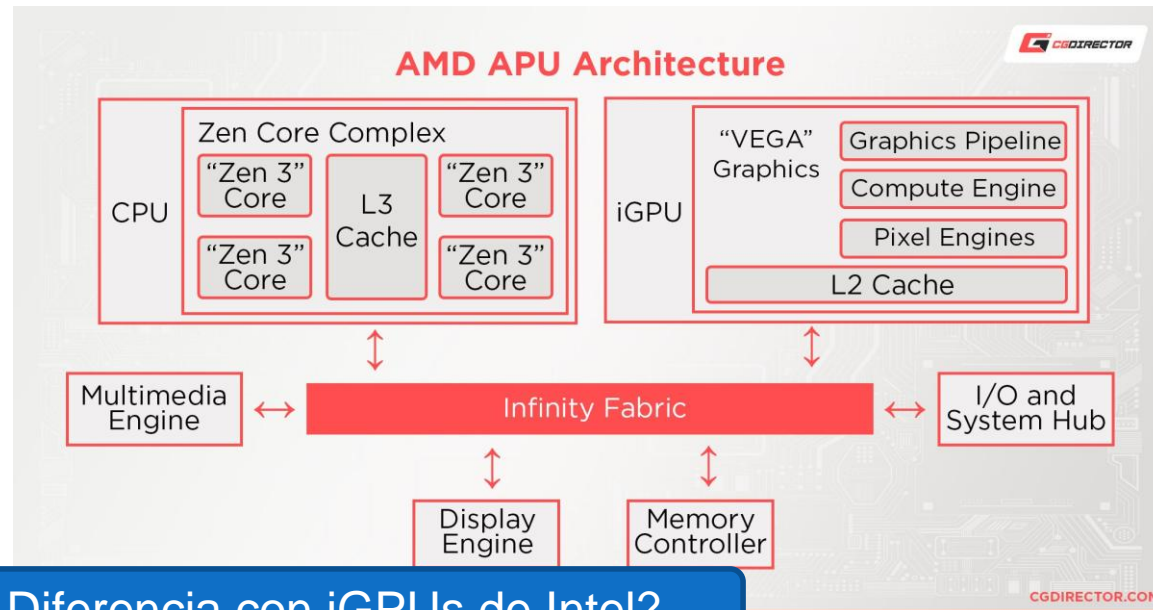
- En el año 2006, AMD compró la empresa de placas gráficas ATI, lo que le permitió incrementar notablemente su capacidad de producir e innovar hardware gráfico.
- Al igual que NVIDIA, para cada arquitectura, ofrece placas para diferentes segmentos (escritorio, integradas, *mobile*, *workstation*, servidores)

AMD Advanced Micro Devices **graphics processing unit**



GPUs de AMD

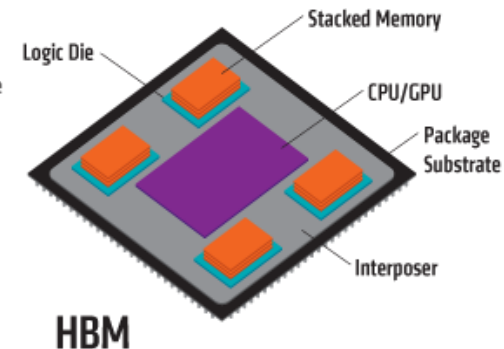
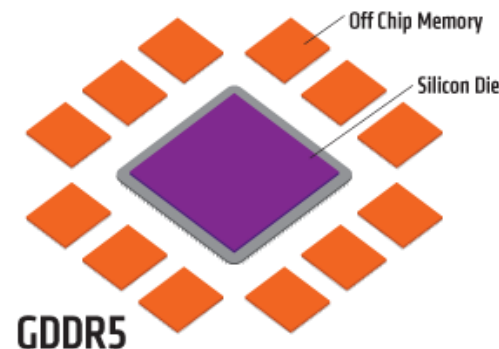
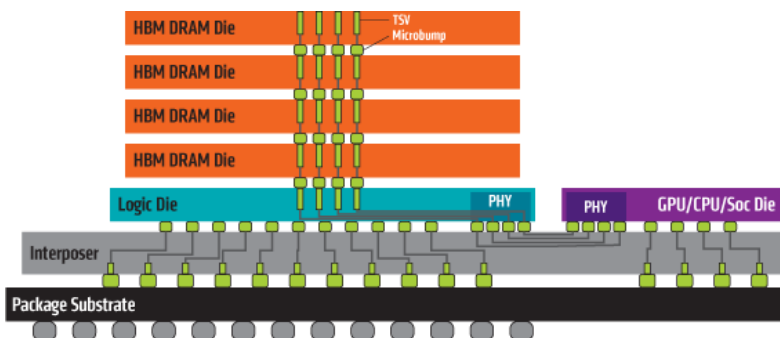
- La adquisición de ATI le permitió a AMD innovar en hardware gráfico → Un resultado son las Unidades de Procesamiento Acelerado (APU).
- Las APUs combinan una CPU y una GPU en el mismo chip, y se basan en la arquitectura GCN
- Pueden ser una buena opción desde la perspectiva del compromiso precio-rendimiento (buenos cocientes de eficiencia energética).
- Se ofrecen para computadoras de escritorio y notebooks.



¿Diferencia con iGPUs de Intel?

Memoria de Alto Ancho de Banda (HBM)

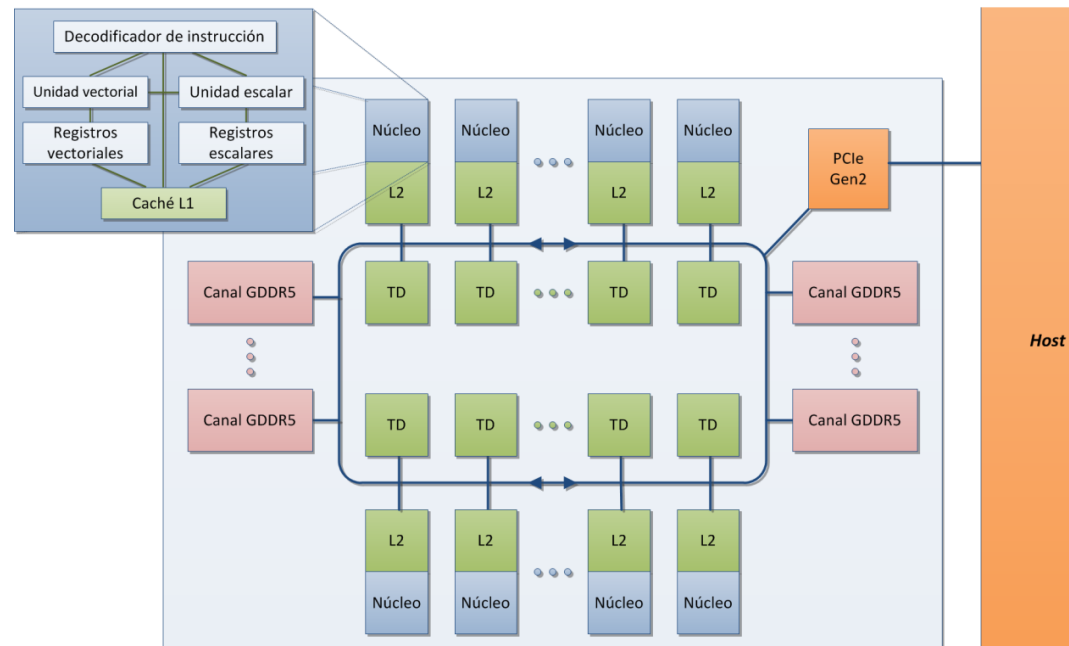
- Una de las innovaciones más importantes de AMD en los últimos años es su tecnología Memoria de Alto Ancho de Banda (HBM)
 - HBM es un nuevo tipo de memoria RAM que organiza los chips de memoria en forma vertical y apilada (*memoria 3D*) y que puede ser aprovechado tanto por GPUs como CPUs.
 - Múltiples mejoras: significativo ahorro de espacio y considerables aumentos en la velocidad de comunicación y en la eficiencia energética
 - Incorporada en las GPUs con nombre clave Fiji de de AMD en 2015. También fue adoptada por NVIDIA para sus placas de la arquitectura Pascal en 2016.



XEON PHI, GPUS Y PROCESADORES HÍBRIDOS DE INTEL

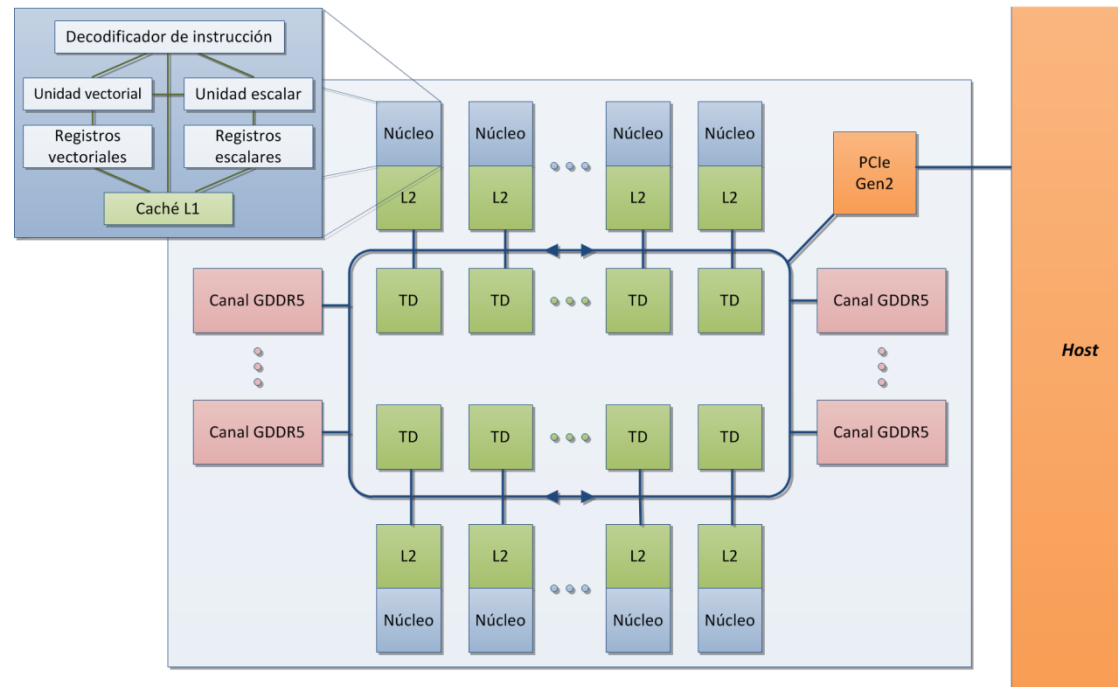
Intel Xeon Phi

- Arquitectura desarrollada por Intel para competir con las GPUs
- Primera generación lanzada en 2013 (Knights Corner, KNC)
 - Coprocesador de hasta 61 núcleos x86 con unidades vectoriales extendidas (512 bits) y SMT (4 hilos hardware por núcleo).
- Caché L1 de 64 Kb + Caché L2 de 512 Kb
- Interconexión en forma de anillo de alta velocidad
- Conexión al host a través del bus PCIe Gen2



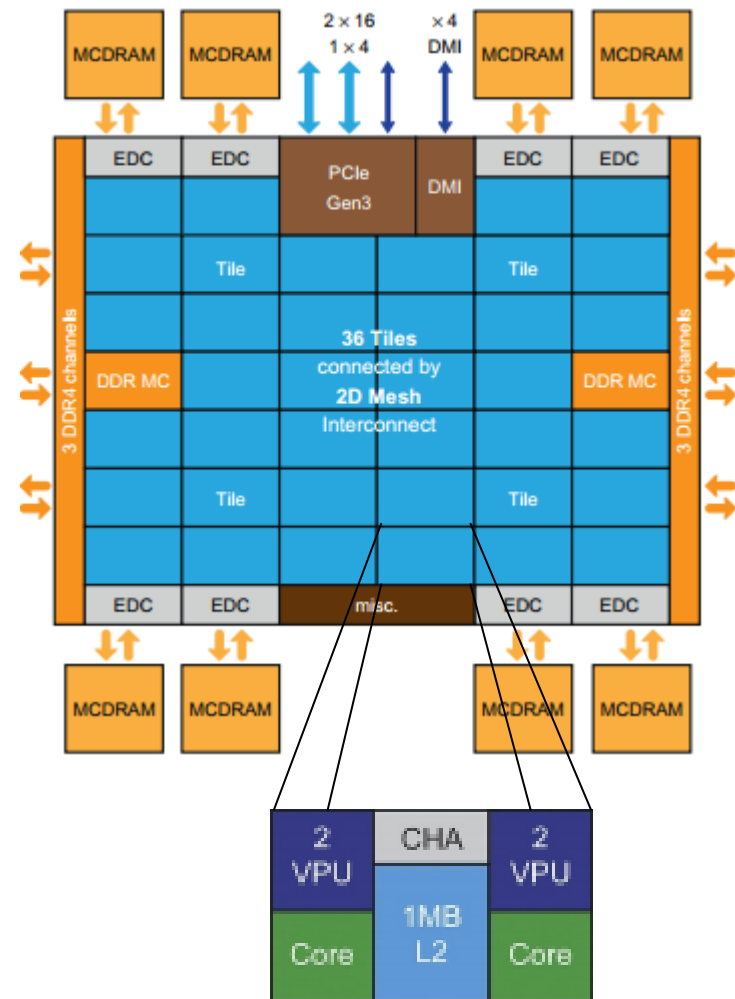
Intel Xeon Phi – Knights Corner

- Ejecuta un sistema operativo propio basado en Linux
- Compatibilidad de código con procesadores Xeon → requiere compilación cruzada
- Dos modos de ejecución:
 - Nativo
 - Offload



Intel Xeon Phi

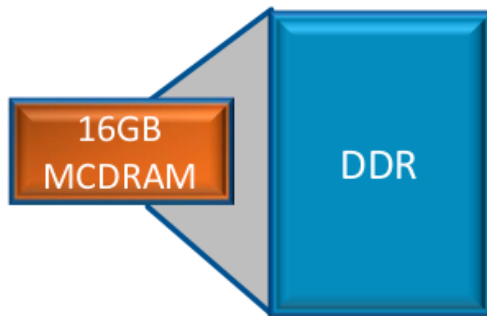
- Segunda generación lanzada en 2015: Knights Landing (KNL)
- Capacidad de operar autónomamente
- Características arquitectónicas
 - Hasta 36 *Tiles* interconectados por malla 2D
 - Cada Tile incluye 2 núcleos:
 - Basados en la micro-arquitectura Intel Atom (fuera de orden, 4 hilos hw por núcleo)
 - 2 unidades vectoriales por núcleo
 - Caché L2 compartida de 1 MB
- Compatibilidad plena con arquitectura x86



Intel Xeon Phi - Knights Landing

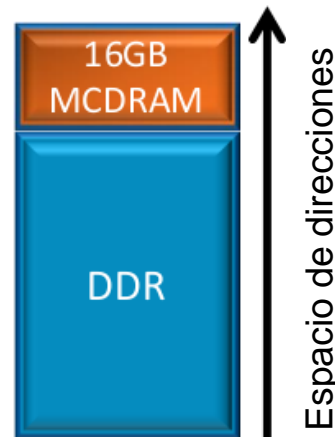
- Incorporación de memoria de alto ancho de banda (HBM) mediante tecnología MCDRAM

Modo cache



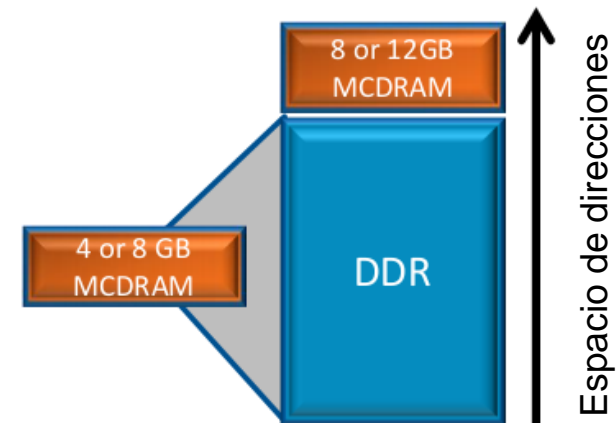
- ✓ Transparente al usuario
- ✓ No requiere cambios en el código fuente
- ✓ Puede sufrir pérdida de rendimiento

Modo *flat*



- ✓ Ofrece el mayor ancho de banda y la más baja latencia
- ✓ Usualmente requiere cambios en el código fuente

Modo híbrido



- ✓ Combinación de las dos anteriores

Intel Xeon Phi

- Tercera generación lanzada en 2017: Knights Mill (KNM)
- Variante de KNL orientada a Deep Learning → Incorpora instrucciones que duplican/cuadriplican rendimiento en FP32/FP16 pero reducen a la mitad en FP64
- Intel discontinuó la línea Xeon Phi en 2019 para orientarse a otros productos



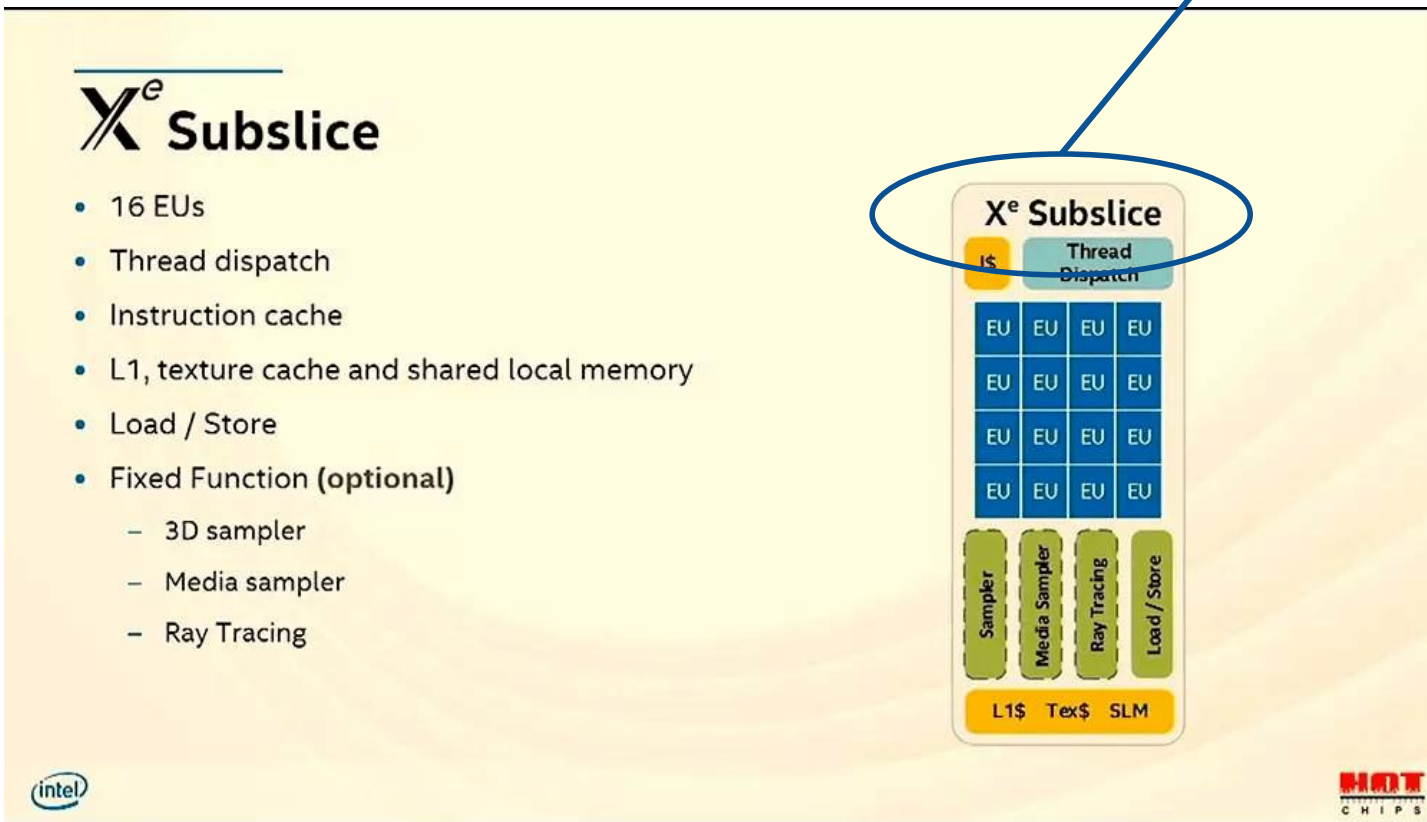
Intel Xe

- La familia de GPUs Xe consiste de un conjunto de microarquitecturas:
 - Xe-LP → Integradas y de bajo consumo
 - Xe-HPG → Gaming (alto rendimiento)
 - Xe-HP → Datacenter (alto rendimiento)
 - Xe-HPC → HPC



Intel Xe

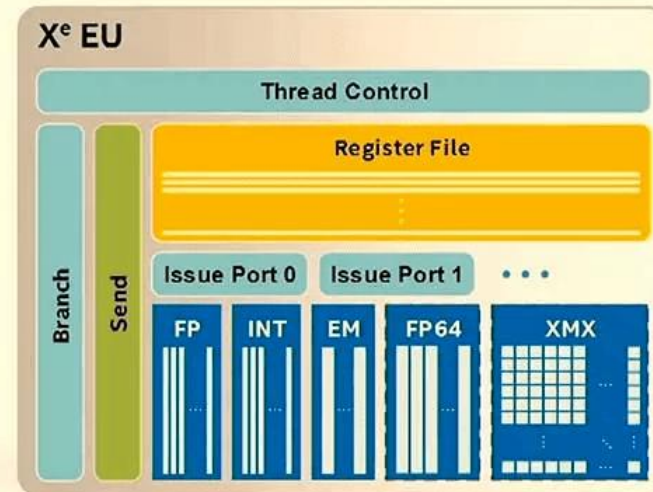
SM (NVIDIA)
CU (AMD)



Intel Xe

X^e Execution Unit

- Thread control
- Register file
- Branch
- Send
- Multiple issue ports
- Configurable mapping of vector pipes
 - Floating Point
 - Integer
 - Extended Math
 - FP64 (optional)
 - Matrix Extension (XMN) (optional)



Intel Xe

- Ponte Vecchio → GPU de la serie Xe-HPC
 - Integradas a la supercomputadora Aurora del Laboratorio Argonne, EEUU
 - Se comercializan como Intel Data Center GPU Max 1100/1550



Precision	Intel Ponte Vecchio	NVIDIA Hopper H100
FP32	52 TFLOP/s	60 TFLOP/s
FP64	52 TFLOP/s	30 TFLOP/s
XMV Float TF32	419 TFLOP/s	NA
XMV BF16	839 TFLOP/s	2,000 TFLOP/s
XMV FP16	839 TFLOP/s	2,000 TFLOP/s
XMV INT8	1,628 TFLOP/s	4,000 TFLOP/s

Procesadores híbridos de Intel

- Arquitectura presentada en 2021 que combina dos tipos de procesadores en el mismo chip

Performance-cores (P-cores):

- Physically larger, high-performance cores designed for raw speed while maintaining efficiency.
- Tuned for high turbo frequencies and high IPC (instructions per cycle).
- Ideal for crunching through the heavy single-threaded work demanded by many game engines.
- Capable of hyper-threading, which means running two software threads at once.



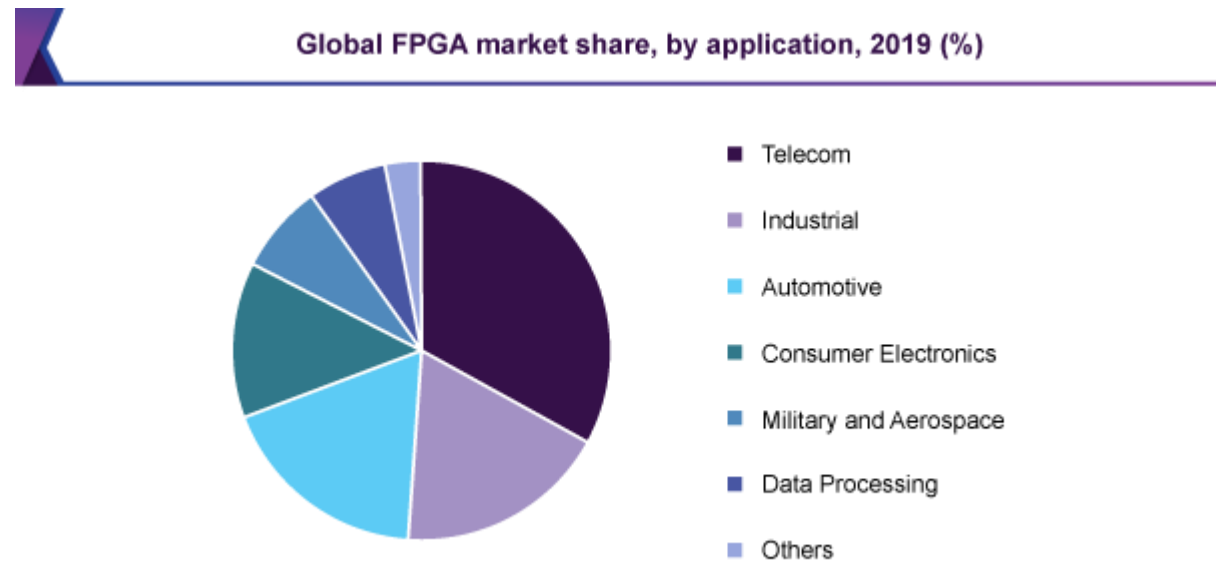
Efficient-cores (E-cores):

- Physically smaller, with multiple E-cores fitting into the physical space of one P-core.
- Designed to maximize CPU efficiency, measured as performance-per-watt.
- Ideal for scalable, multi-threaded performance. They work in concert with P-cores to accelerate core-hungry tasks (like when rendering video, for example).
- Optimized to run background tasks efficiently. Smaller tasks can be offloaded to E-cores — for example, handling Discord or antivirus software — leaving P-cores free to drive gaming performance.
- Capable of running a single software thread.

FPGAs y ASICs

FPGAs

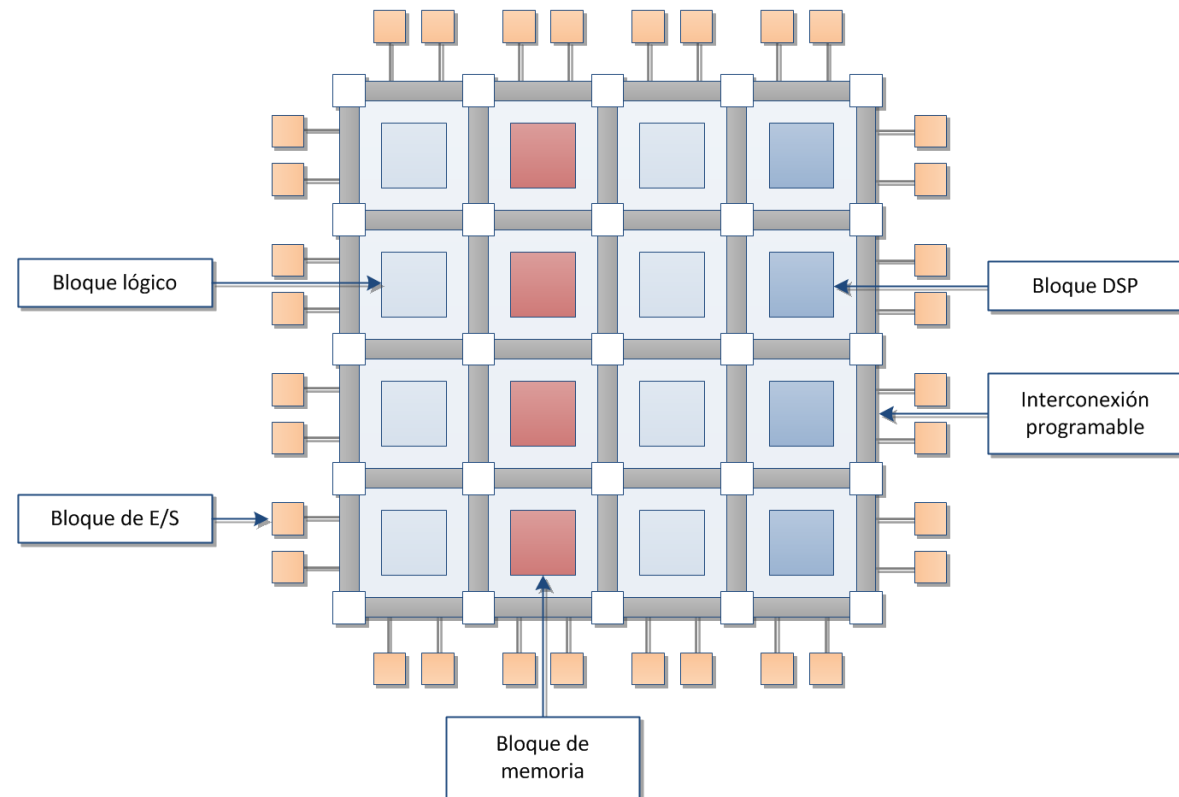
- Las FPGA se crearon en el año 1984 por los co-fundadores de Xilinx, Ross Freeman y Bernard Vonderschmitt → *Idea innovadora*
- Desde entonces, las FPGAs han evolucionado significativamente incorporando características como la adopción de estándares de E/S de alta velocidad, mejoras en la compatibilidad con las CPUs y un continuo aumento en la cantidad de recursos de las placas.
- Inicialmente usadas para procesamiento digital de señales, hoy su uso se ha extendido (datacenter, IA, criptografía, etc)



Source: www.grandviewresearch.com

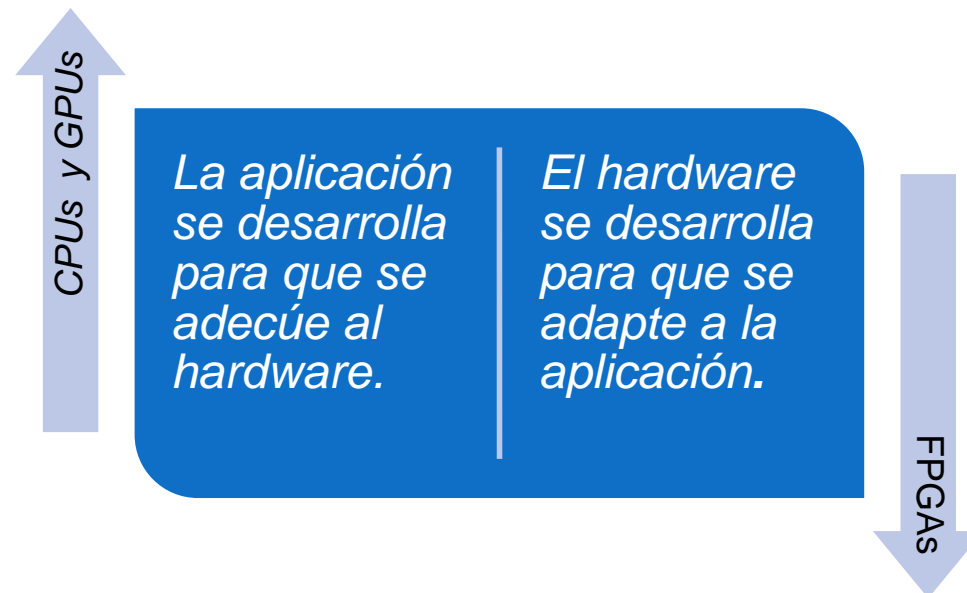
FPGAs

- Una FPGA consiste de circuitos integrados reconfigurables compuestos por interconexiones programables que unen bloques lógicos programables, bloques de memoria embebidos y bloques DSPs → *Hardware programmable*
- La comunicación con el exterior se realiza a través de los bloques de E/S, los cuales se organizan en forma de anillo alrededor de la circunferencia del dispositivo.



FPGAs

- CPUs y GPUs presentan topologías y rutas de datos estáticas para procesar las instrucciones de los programas → los recursos de las FPGAs puede ser configurados e interconectados para crear pipelines de instrucciones a medida en los cuales procesar los datos
- Por ejemplo, si un algoritmo sólo necesita realizar cierto tipo de aritmética con enteros, ¿vale la pena destinar recursos para punto flotante? ¿para otro tipo de operaciones que no sean las de interés?



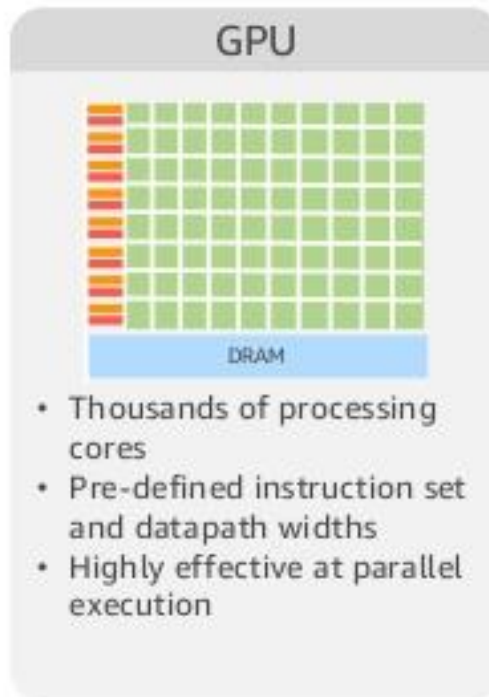
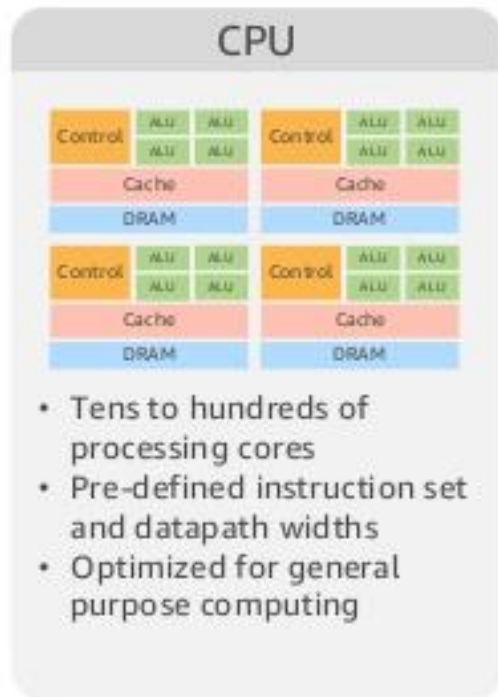
FPGAs

- Si bien tanto la frecuencia del reloj como el pico de rendimiento suelen ser más bajos que los correspondientes a las CPUs y a las GPUs, la capacidad de configurar el hardware para que se adapte al problema específico a resolver le da la posibilidad de obtener mejores rendimientos.
- Además, como no hay desperdicio de recursos de silicio, en general son más eficientes desde el punto de vista energético

Dispositivo	GFLOPS (SP)	Watt (TDP)	GFLOPS/Watt
Intel® Xeon® ER-4669 v3 (18 cores, 36 ht, 2.1-2.9 Ghz)	604.8	135	4.48
NVIDIA Tesla K40 (2880 núcleos CUDA, 745 Mhz)	4290	300	8.05
Xilinx Virtex7 XCTV200T (2160 bloques DSP, ~740 Mhz)	1636	~40	~41

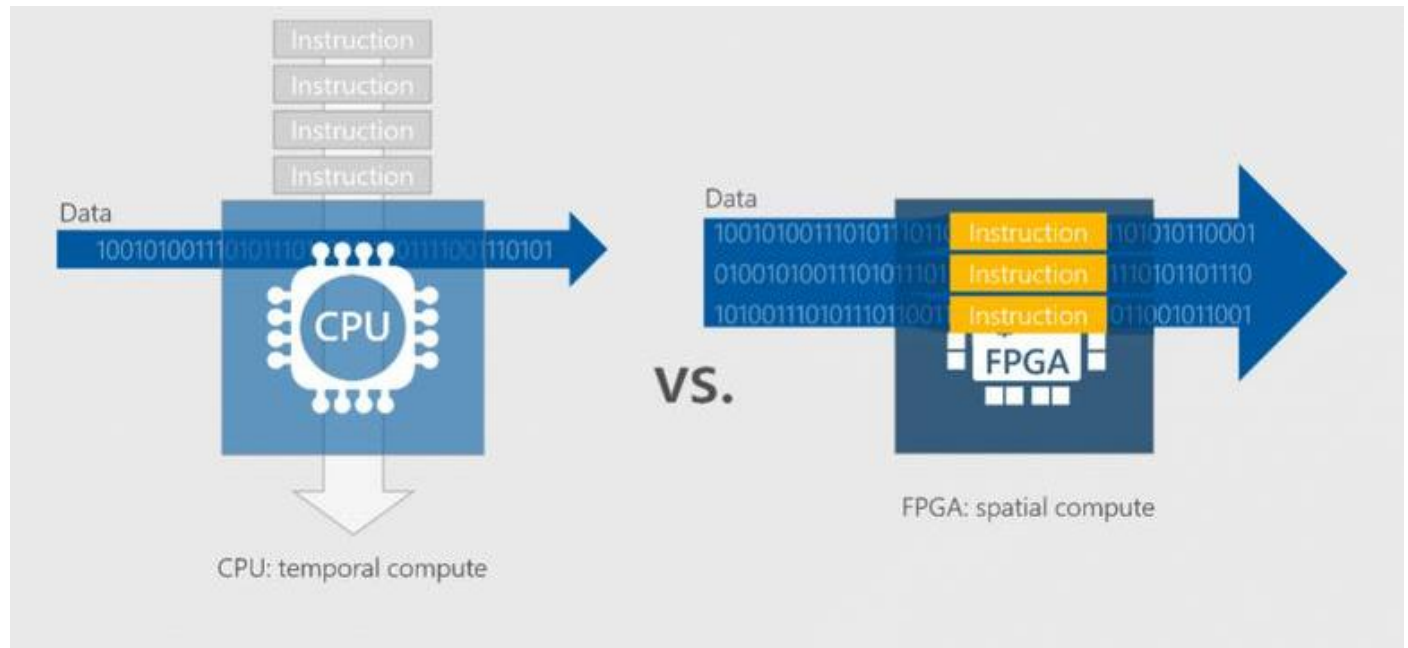
FPGAs

PARALLEL PROCESSING IN GPU AND FPGA



FPGAs

- CPU vs FPGA

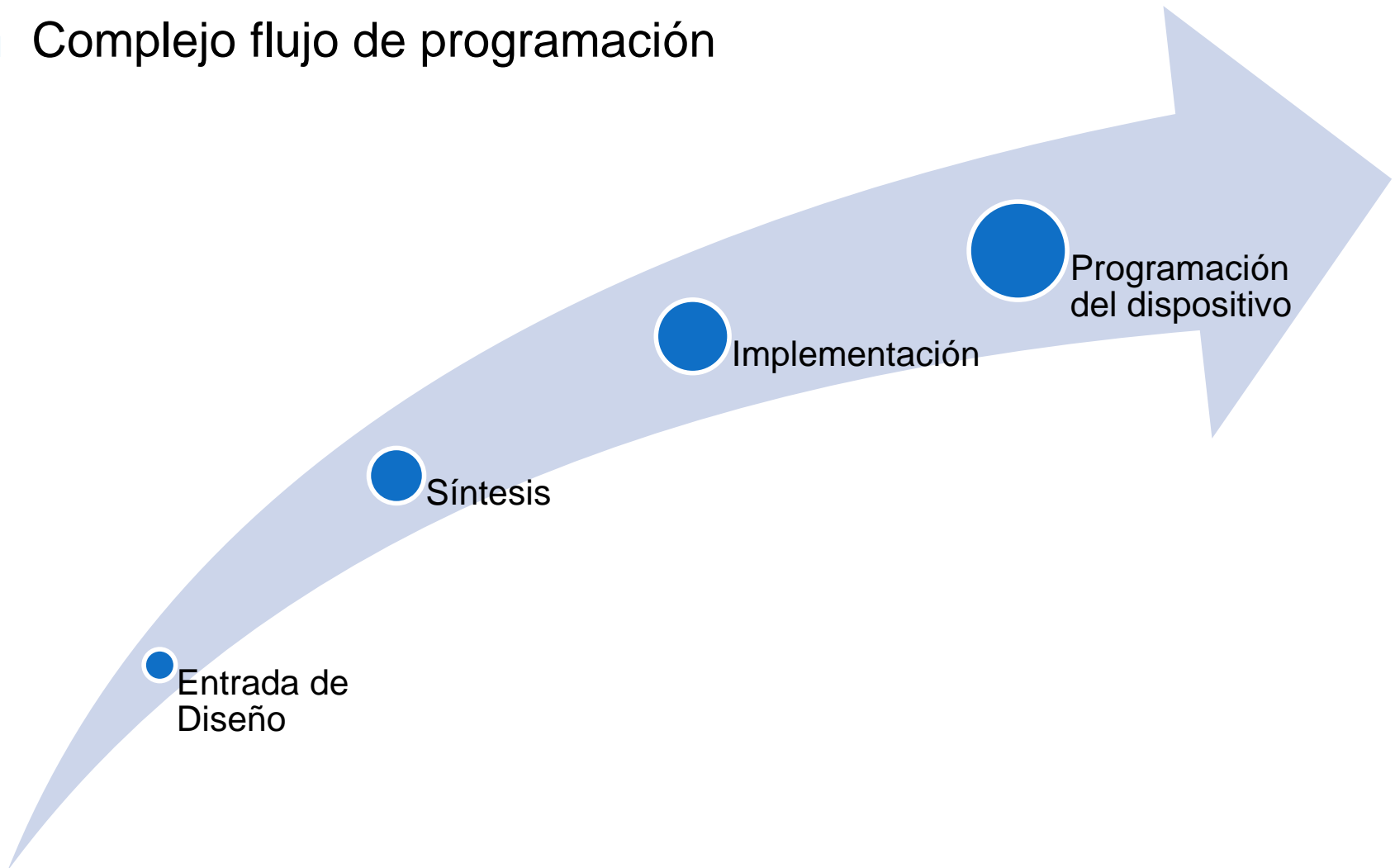


✗ No siempre es conveniente emplear FPGAs. Entre otras características, para obtener alto rendimiento se requiere:

- ✗ Operaciones aritméticas simples (punto fijo → mejor)
- ✗ Amplio paralelismo de datos
- ✗ Estructuras de control regulares y sencillas

FPGAs

- Complejo flujo de programación



FPGAs

- Opciones para Entrada de diseño: CAD vs HDL vs HLS
 - Los diseños generados a partir de herramientas CAD son más fáciles leer y comprender, aunque solo suelen funcionar con proyectos pequeños.
 - Para diseños complejos, la opción tradicional es el uso de Lenguajes de Descripción de Hardware (HDL), un enfoque basado en código.
 - En los últimos años, se han desarrollado alternativas de alto nivel a los HDLs, llamadas Lenguajes de Alto Nivel (HLL) o también Síntesis de Alto Nivel (HLS)

FPGAs

- Lenguaje de Descripción de Hardware (HDL) → es un lenguaje de programación especializado que se utiliza para definir la estructura, diseño y operación de circuitos electrónicos, especialmente los digitales.
 - Verilog y VHDL son las opciones más populares de HDL → Aunque fueron desarrollados bajo principios similares, su sintaxis es diferente siendo VHDL más verboso y propenso a errores
-
- | | |
|---|-------------------------------|
| ✗ Tediosos y propensos a errores | ✗ No son portables |
| ✗ Complejidad adicional por noción explícita del tiempo | ✗ Baja productividad |
| | ✗ Dificultan el mantenimiento |

FPGAs

```

1 module MISR(dat_in,reset,clk,dat_out);
2 input [3:0]dat_in;
3 input reset,clk;
4 output [3:0]dat_out;
5 reg [3:0]dat_out;
6 reg [3:0]misr_tempreg;
7
8 always@(posedge clk)
9 begin
10     if (reset == 1)
11         dat_out <= 4'b0000;
12
13     //
14     ///////////////////////////////////////////////////
15 module TOP(A,B,C,CHECK,OUT);
16
17 input A,B,C;
18 output CHECK,OUT;
19 wire F,FB,FS;
20
21 mj3 M1(.A(A),.B(B),.C(C),.F(F));
22 invmj3 M2(.A(A),.B(B),.C(C),.FB(FB));
23
24 assign CHECK = F ^ FB;
25
26
27 mj33 M3(.A(A),.B(B),.C(C),.FS(FS));
28
29 assign OUT = (CHECK) ? F : FS;
30
31
32
33 endmodule
34
35
36
37
38

```

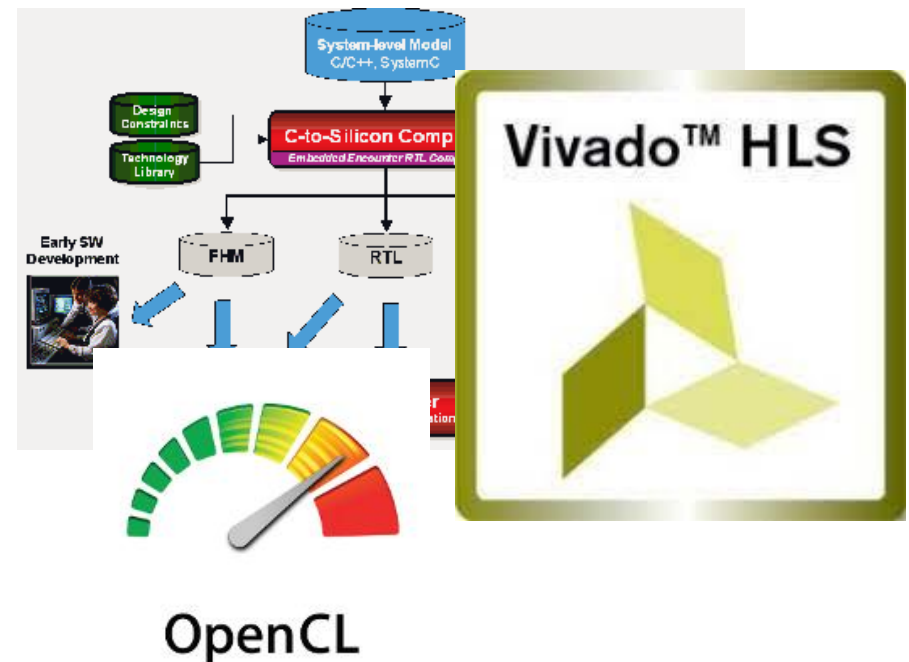
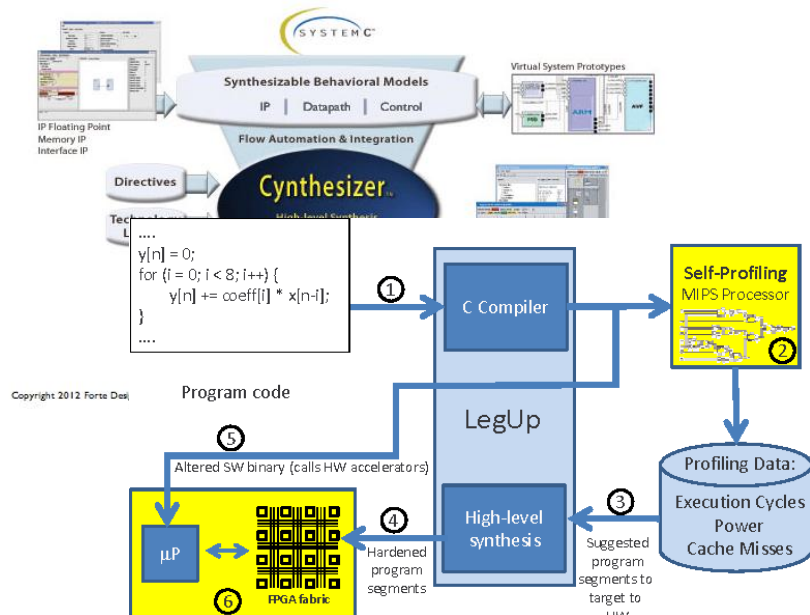
```

module PTIO #(parameter N Bits = 4)
//
//Student Name: XYZ
//Student Number: 12345678
//Design Name : up_counter
// File Name : up_counter.v
// Function : 8 bit Up counter
//
module up_counter (
out , // Output of the counter
enable , // enable for counter
clk , // clock Input
reset // reset Input
);
//-----Define Ports-----
//-----Output Ports-----
output [7:0] out;
//-----Input Ports-----
input enable, clk, reset;
//-----Internal Variables-----
reg [7:0] out;
//-----Code Starts Here-----
always @(posedge clk)
if (reset)
begin
out <= 8'b0;
end
else if (enable)
begin
out <= out + 1;
end
endmodule

```

FPGAs

- En la década del 2000, varios fabricantes de FPGAs empezaron a ofrecer herramientas para síntesis de alto nivel (HLS)
- Bajo este enfoque, los programadores desarrollan el código usando lenguajes de alto nivel (HLL), como C, C++ o SystemC → la herramienta es responsable de generar el código HDL equivalente.



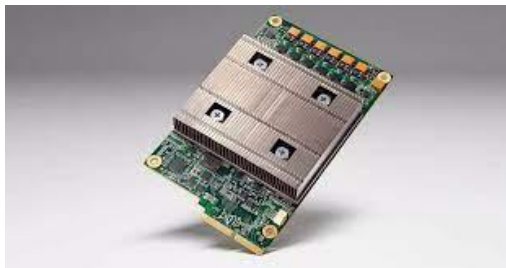
FPGAs

- El uso de HLS permite que los desarrolladores trabajen a un nivel de abstracción más alto y produzcan diseños de hardware reusables, sin requerir una amplia experiencia en el tema.
- Este enfoque ha permitido reducir los tiempo de desarrollo en la industria (*time-to-market*)
 - ✔ Disminuye costo de programación
 - ✔ Aumenta productividad
 - ✔ Brinda portabilidad (OpenCL)
 - ✔ Facilita mantenimiento
- No hay tendencias claras respecto al rendimiento y uso de recursos

Más allá del lenguaje utilizado, el proceso de sintetización puede tomar horas y no hay garantía de que tenga éxito!

ASICs

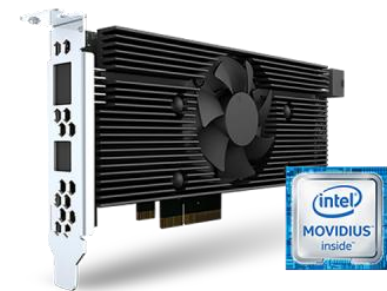
- Un Circuito Integrado de Aplicación Específica (ASIC) hace referencia a un chip destinado a cumplir el propósito para el cual fue diseñado y que no puede ser reprogramado/modificado para realizar otra función
- Los ASICs están presentes en muchos de los dispositivos que nos rodean habitualmente: autos, televisores, celulares, heladeras, entre otros.



Google TPU



Avalon Miner



Intel VPU

FPGAs y ASICs



Resumen de aceleradores

- Si bien cada dispositivo cuenta con sus características propias, existen diversos aspectos que son comunes a todos ellos:
 - Complejidad de programación
 - Técnicas de programación y optimización específicas, múltiples niveles de paralelismo, balance de carga, diversidad de modelos de lenguajes/modelos de programación, ausencia de estándar (consolidado)
 - Memoria del dispositivo separada
 - Al estar separada de la memoria del host, su administración es clave para obtener alto rendimiento
 - Recientemente se han desarrollado modelos de memoria unificada
 - Patrón de acceso a la memoria
 - Requieren de patrones de acceso específicos para obtener el mejor rendimiento, que no siempre coinciden con los de las CPU

Resumen de aceleradores

- Si bien cada dispositivo cuenta con sus características propias, existen diversos aspectos que son comunes a todos ellos
 - Many-cores
 - Gran cantidad de núcleos pequeños
 - Multi-hilado
 - Aplicado de diferentes maneras, buscan ocultar la latencia de la memoria
 - Vectorización (SIMD)
 - Cada uno con su enfoque, comparten el “espíritu”
 - Ejecución de instrucciones en orden
 - La lógica de control está simplificada
 - Memorias caché más pequeñas
 - La mayor parte de los recursos se destina a unidades funcionales
 - Cobra mayor importancia explotar localidad de datos

Bibliografía usada para esta clase

- M. Giles and I. Reguly, Trends in high-performance computing for engineering calculations, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 372, no. 2022, p. 20130319, 2014.
- M. Vestias and H. Neto, Trends of CPU, GPU and FPGA for high-performance computing, in Field Programmable Logic and Applications (FPL), 2014 24th International Conference on, Sept 2014, pp. 1-6.
- E. Rucci, Evaluación de Rendimiento y Eficiencia Energética de Sistemas Heterogéneos para Bioinformática, Tesis Doctoral, UNLP, 2016.