



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Departament d'Arquitectura de Computadors

# Tarjetas Gráficas y Aceleradores

## Nvidia Roadmap

Agustín Fernández

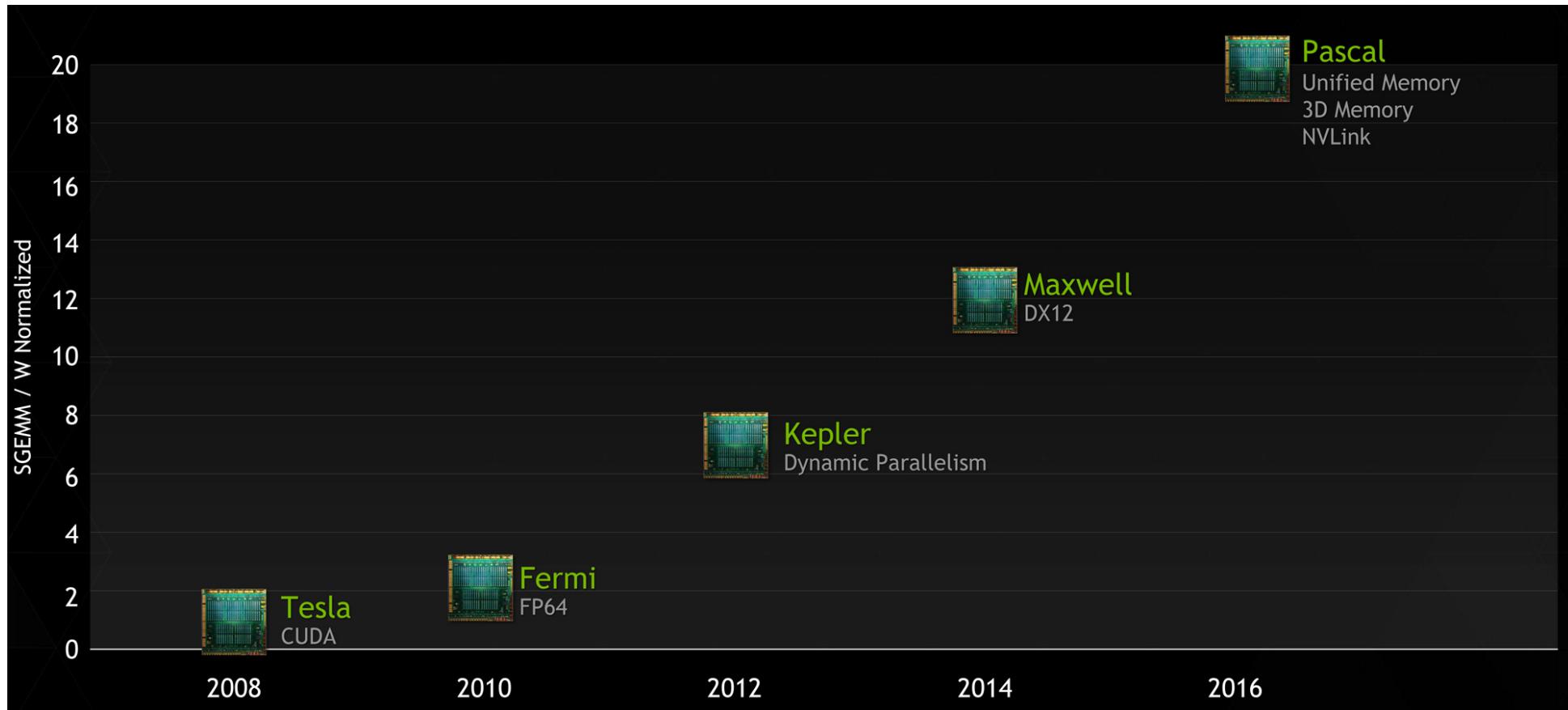
Departament d'Arquitectura de Computadors

Facultat d'Informàtica de Barcelona

Universitat Politècnica de Catalunya



# Nvidia Roadmap



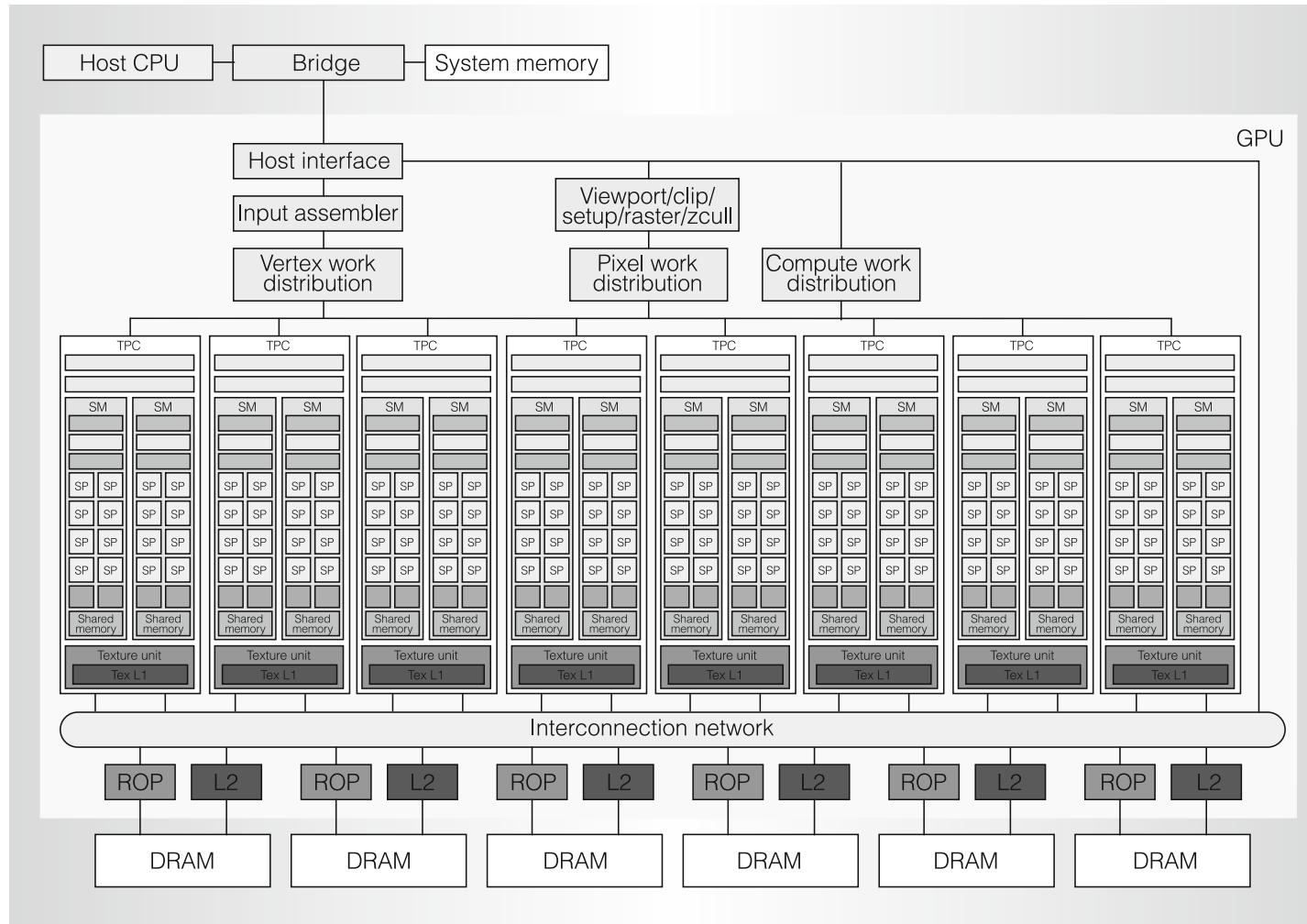
i Rendimiento por W !

[www.nvidia.com](http://www.nvidia.com)

# Arquitectura TESLA

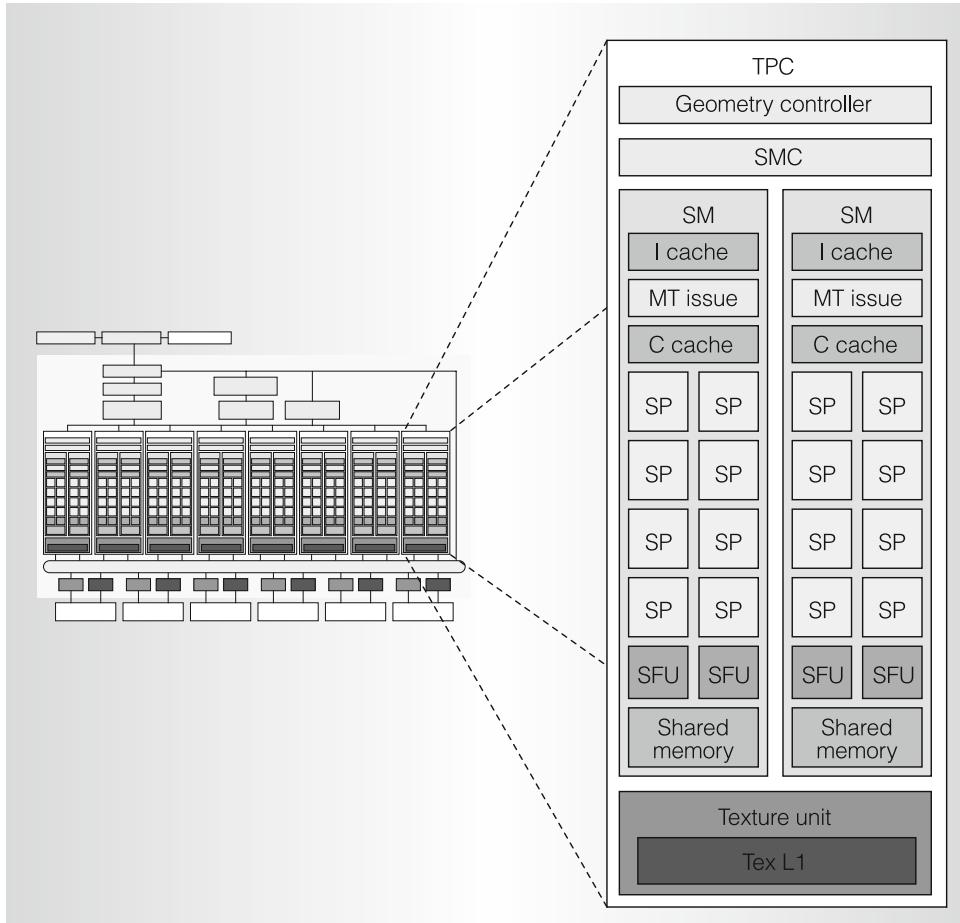
- 1<sup>a</sup> Implementación de NVIDIA de Shaders Unificados
- 2 Generaciones: G80 y GT200 (con sus diferentes variantes)
  - Usada en las familias: GeForce 8, GeForce 9, GeForce 100, GeForce 200 y GeForce 300.
- Stream Processors (SPs)
- Primeras implementaciones de CUDA
- Diferentes frecuencias de funcionamiento (GPU – SPs)
- FMA en Doble Precisión ( $d = a + b*c$ , sin redondear después del \*)
  
- Nuevos métodos de Antialiasing
- Mejoras en el filtrado de texturas

# Arquitectura TESLA



[www.nvidia.com](http://www.nvidia.com)

# Arquitectura TESLA



[www.nvidia.com](http://www.nvidia.com)

**Instrucciones específicas para acceder a cada tipo de memoria:**

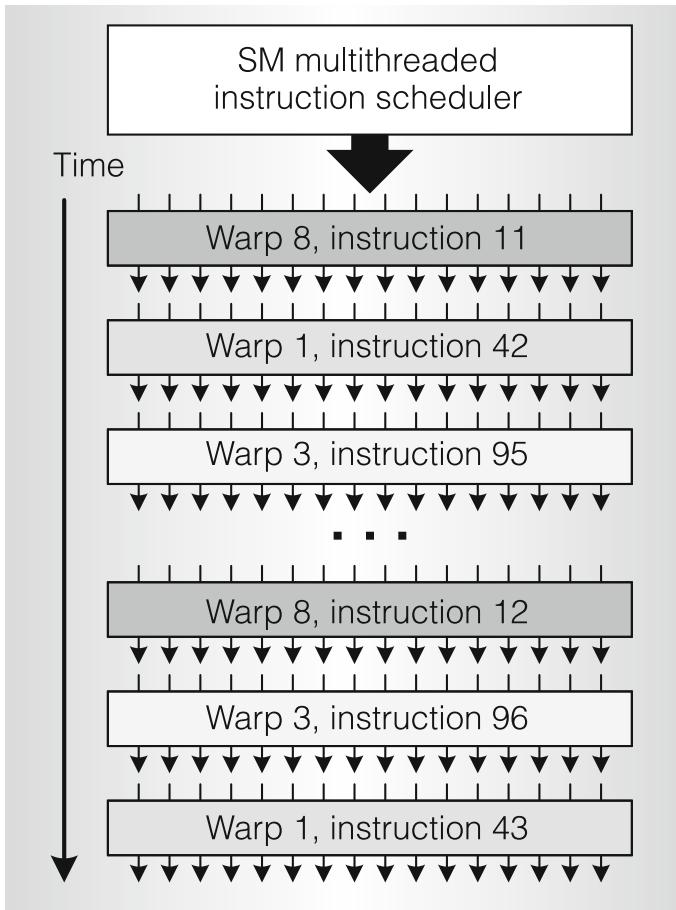
- Local
- Compartida
- Global

**FP “compatible” IEEE 754**

**CUDA funciona por:**

- Shaders Unificados
- Load/store convencional
- Memoria Compartida

# Arquitectura TESLA



Cada SM puede gestionar y ejecutar hasta 768 threads concurrentemente con zero overhead.

La unidad de ejecución es el **WARP (32 threads)**.  
Cada SM mantiene el estado de 24 WARPS.

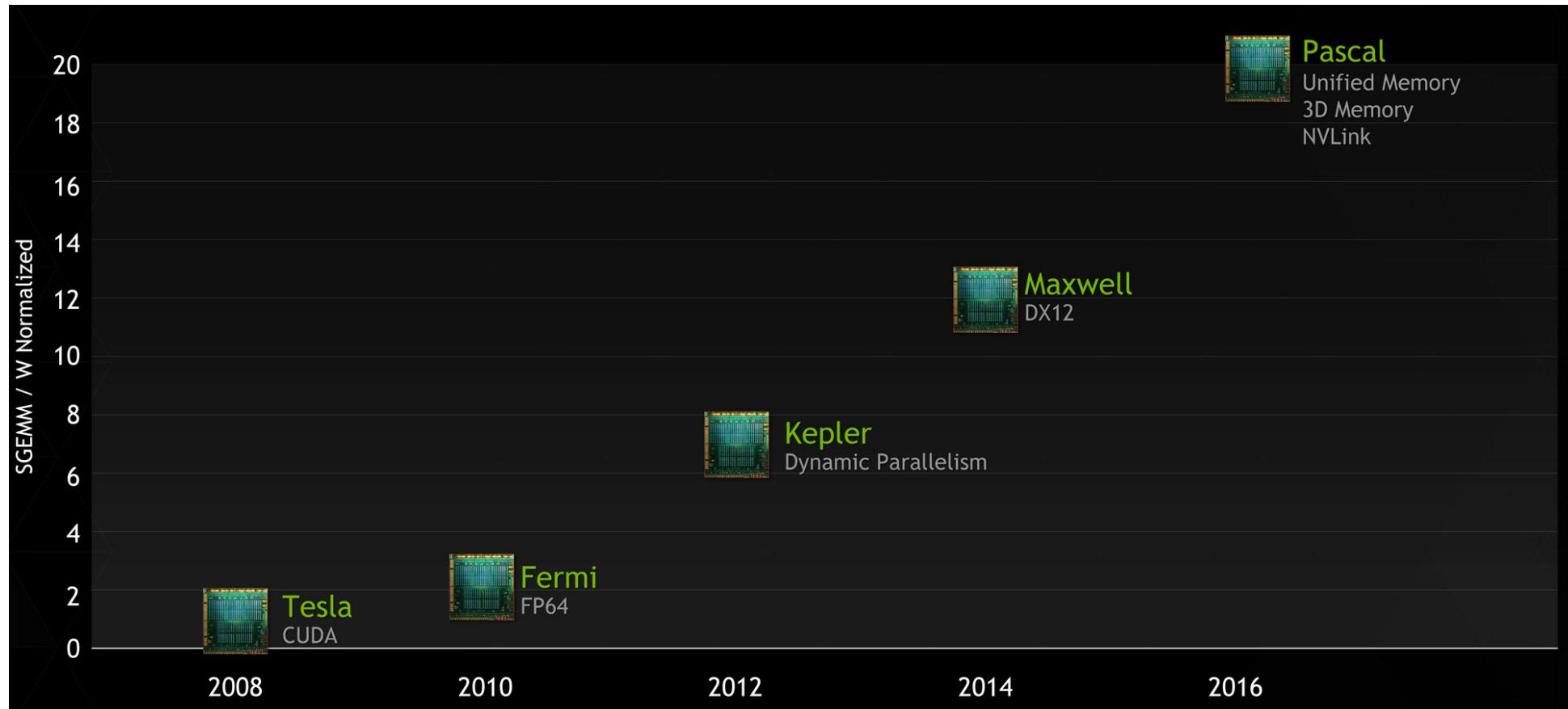
## Thread Scheduling:

- Cuando todos los operandos están disponibles.
- Ejecución en orden (en un warp)
- Políticas: greedy + round robin

# Arquitectura TESLA: 2 generaciones

GPU	G80	GT200
Transistores	681 millones	1.400 millones
CUDA Cores	128	240
Double Precision FP Capability	-	30 FMA ops / clock
Single Precision FP Capability	128 MAD ops/clock	240 MAD ops / clock
Warp schedulers (per SM)	1	1
Special Function Units (SFUs) / SM	2	2
Shared Memory (per SM)	16 KB	16 KB
L1 Cache (per SM)	-	-
L2 Cache (per SM)	-	-
ECC Memory Support	No	No
Concurrent Kernels	No	No
Load/Store Address Width	32 bits	32 bits

# Nvidia Roadmap



i Rendimiento por W !

[www.nvidia.com](http://www.nvidia.com)

# Arquitectura FERMI

## □ Tercera Generación de los Streaming Multiprocessor (SM)

- 32 CUDA cores por SM (4x redimiento de GT200)
- Mejora 8x en FP respecto a GT200
- Puede lanzar 2 warps por SM
- 64 KB de RAM, configurable como Memoria Compartida o L1

## □ Segunda generación del ISA

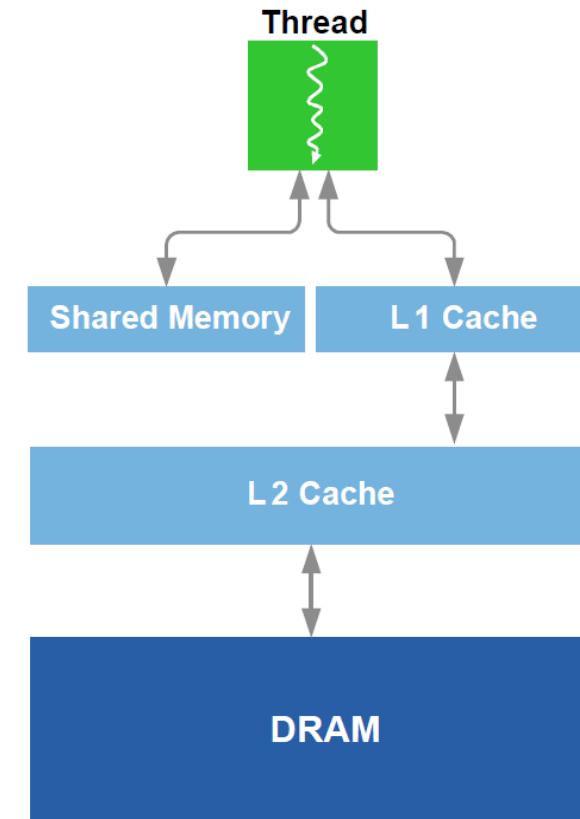
- Espacio de direcciones unificado (Memoria Local, Compartida y Global)
  - ✓ Soporte real C++
- Full IEEE 754 32 y 64 bits
- Direccionamiento de 64 bits

# Arquitectura FERMI

## □ Sistema de Memoria Mejorado

- Camino unificado para los accesos a L1 / Memoria Compartida y L2.
- L1 / Memoria Compartida Configurable:
  - ✓ 16 KB L1 – 48 KB Memoria Compartida
  - ✓ 48 KB L1 – 16 KB Memoria Compartida
- Primera GPU con soporte ECC. Todos los niveles de la jerarquía usan ECC para garantizar la integridad de los datos.
- Mejora 20x instrucciones atómicas

Fermi Memory Hierarchy

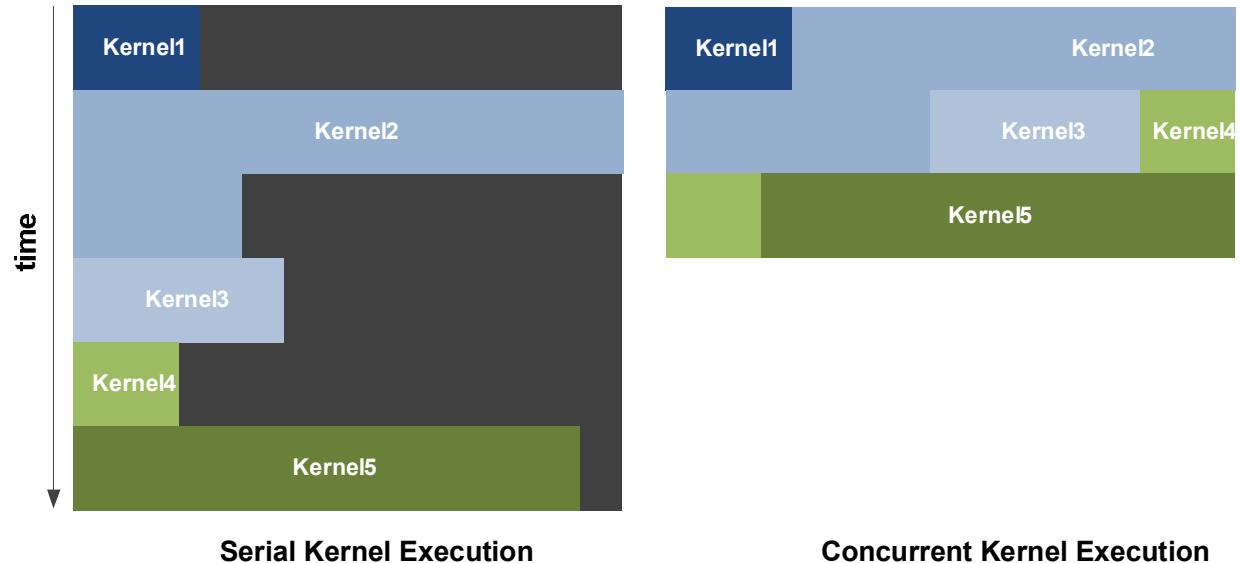


[www.nvidia.com](http://www.nvidia.com)

# Arquitectura FERMI

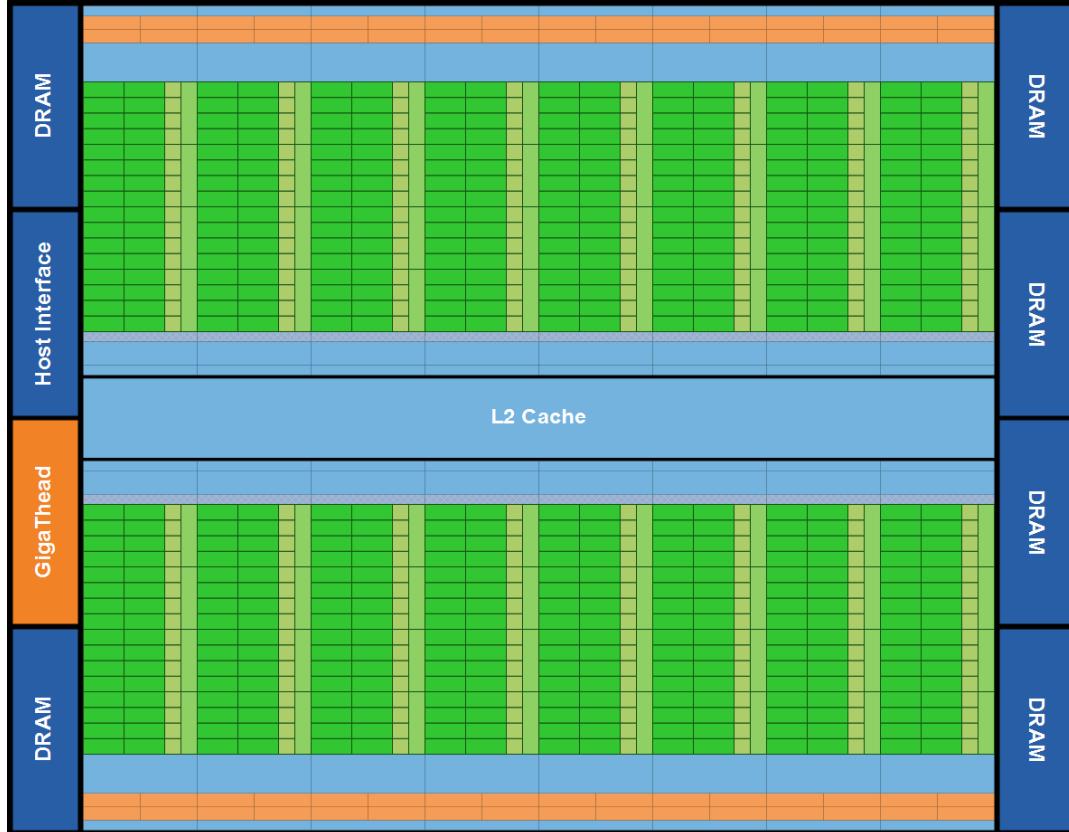
## ☐ NVIDIA GigaThread Engine

- 10x cambio contexto
- Ejecución concurrente de kernels
- Ejecución Blocks desorden
- Permite transferencia dual concurrente (streams CUDA)



[www.nvidia.com](http://www.nvidia.com)

# Arquitectura FERMI



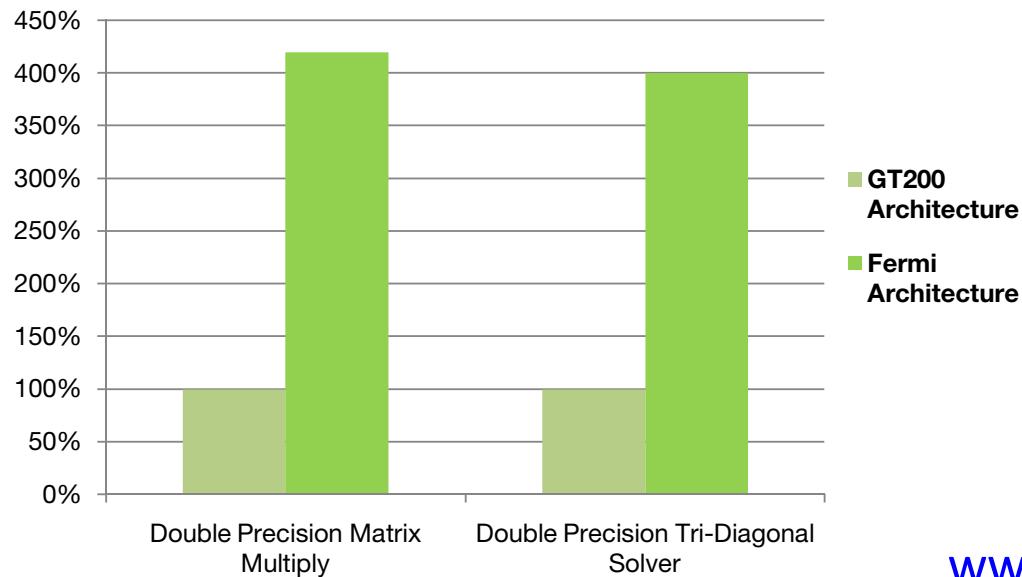
- 16 SMs
- 32 CUDA cores por SM
- 512 CUDA cores
- 6 canales de Memoria GDDR5
- 384 bits interface Memoria

[www.nvidia.com](http://www.nvidia.com)

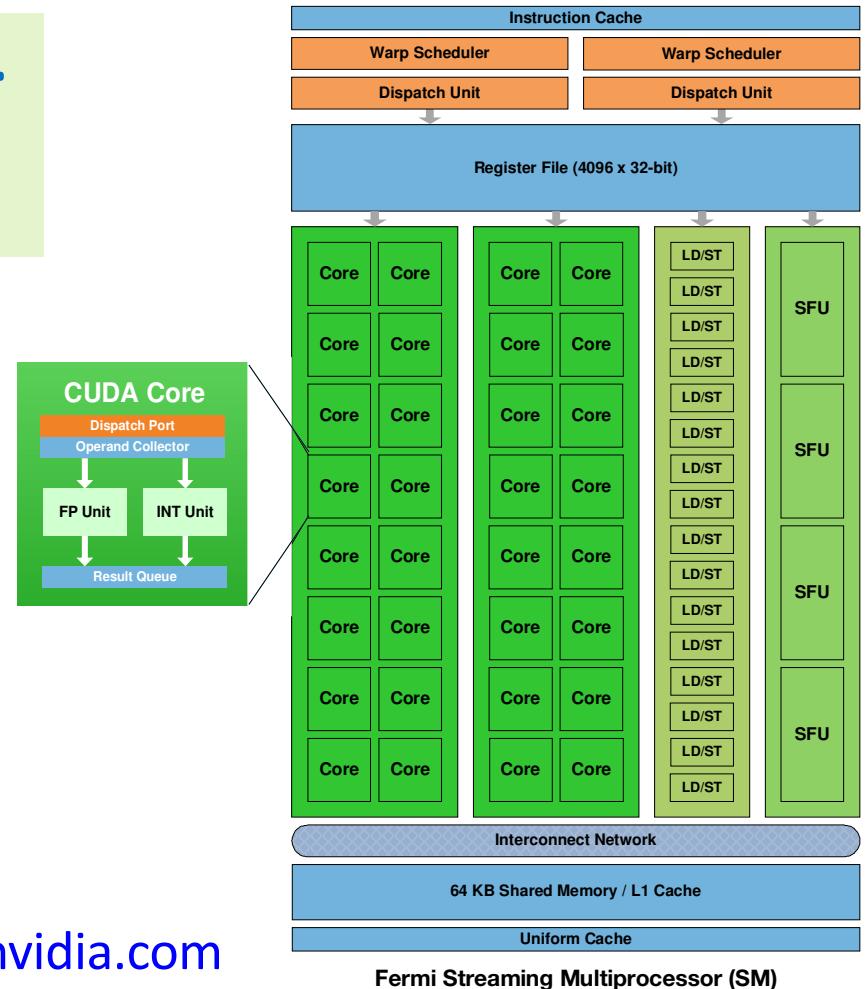
# Arquitectura FERMI

- 32 bits precisión en todas las operaciones.
- Soporte para operandos de 64 bits
- FP Doble Precisión

Double Precision Application Performance



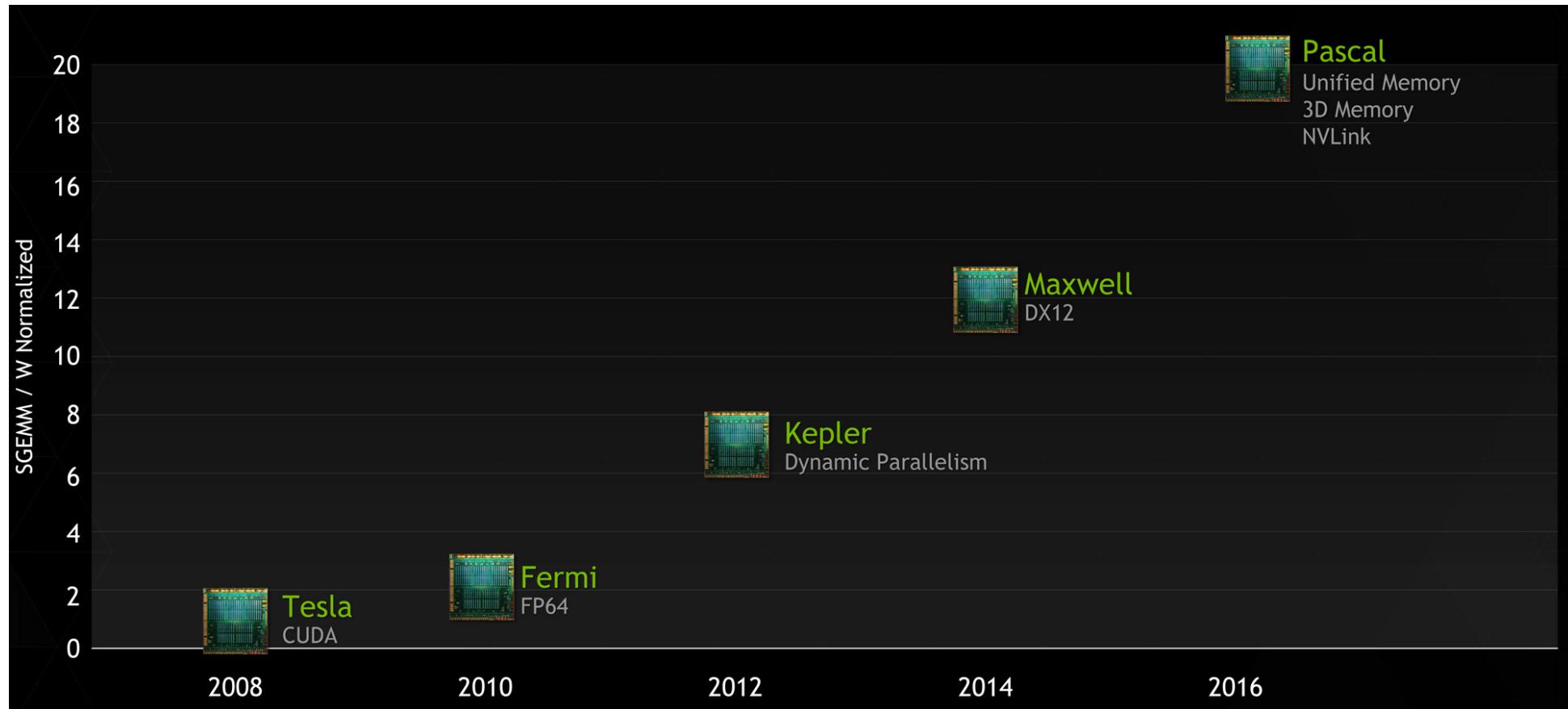
[www.nvidia.com](http://www.nvidia.com)



# Arquitectura FERMI

GPU	Tesla G80	Tesla GT200	Fermi
Transistores	681 millones	1.400 millones	3.000 millones
CUDA Cores	128	240	512
Double Precision FP Capability	-	30 FMA ops/clock	256 FMA ops/clock
Single Precision FP Capability	128 MAD ops/clock	240 MAD ops/clock	512 FMA ops/clock
Warp schedulers (per SM)	1	1	2
Special Function Units per SM	2	2	4
Shared Memory (per SM)	16 KB	16 KB	48 KB o 16 KB
L1 Cache (per SM)	-	-	16 KB o 48 KB
L2 Cache (per SM)	-	-	768 KB
ECC Memory Support	No	No	Si
Concurrent Kernels	No	No	Hasta 16
Load/Store Address Width	32 bits	32 bits	64 bits

# Nvidia Roadmap



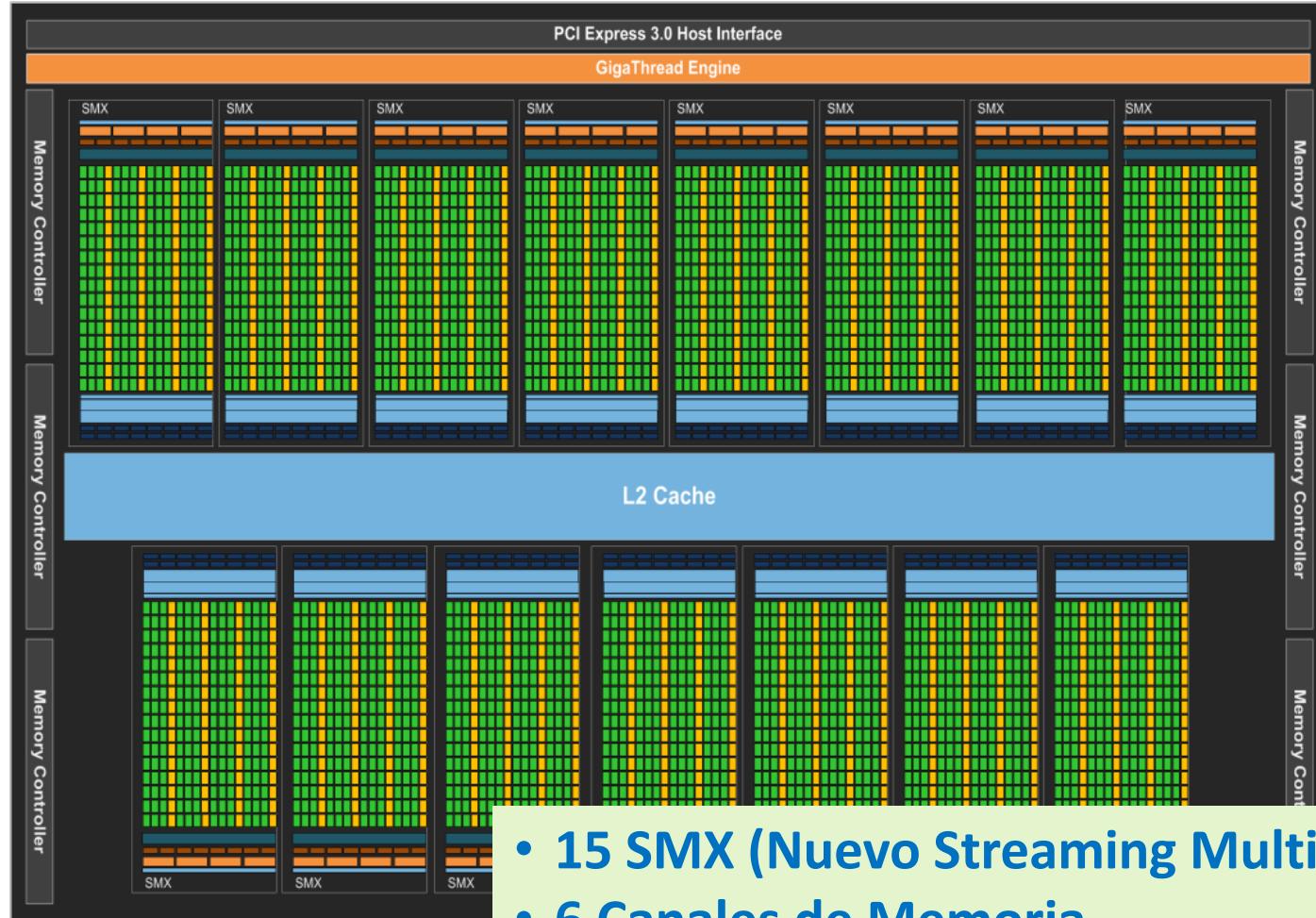
i Rendimiento por W !

[www.nvidia.com](http://www.nvidia.com)

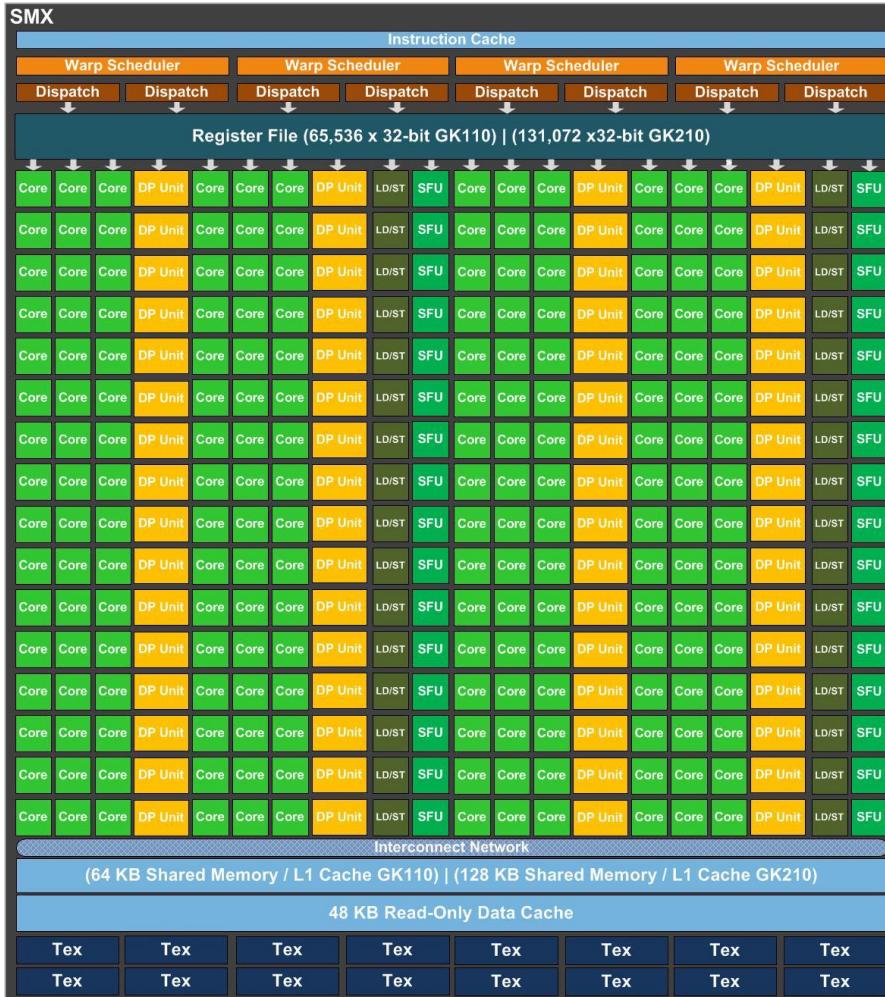
# Arquitectura KEPLER

- Mejor Rendimiento por W
- Más Elementos de Cálculo
- Paralelismo Dinámico
- Hyper-Q
- Grid Management Unit
- NVIDIA GPUDirect

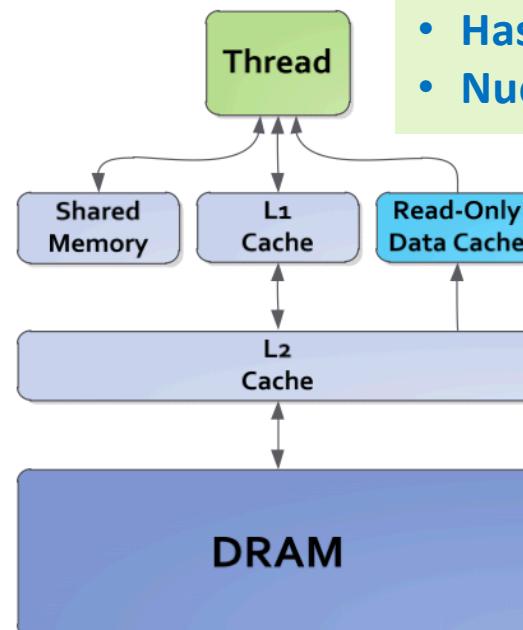
# Arquitectura KEPLER



# Arquitectura KEPLER



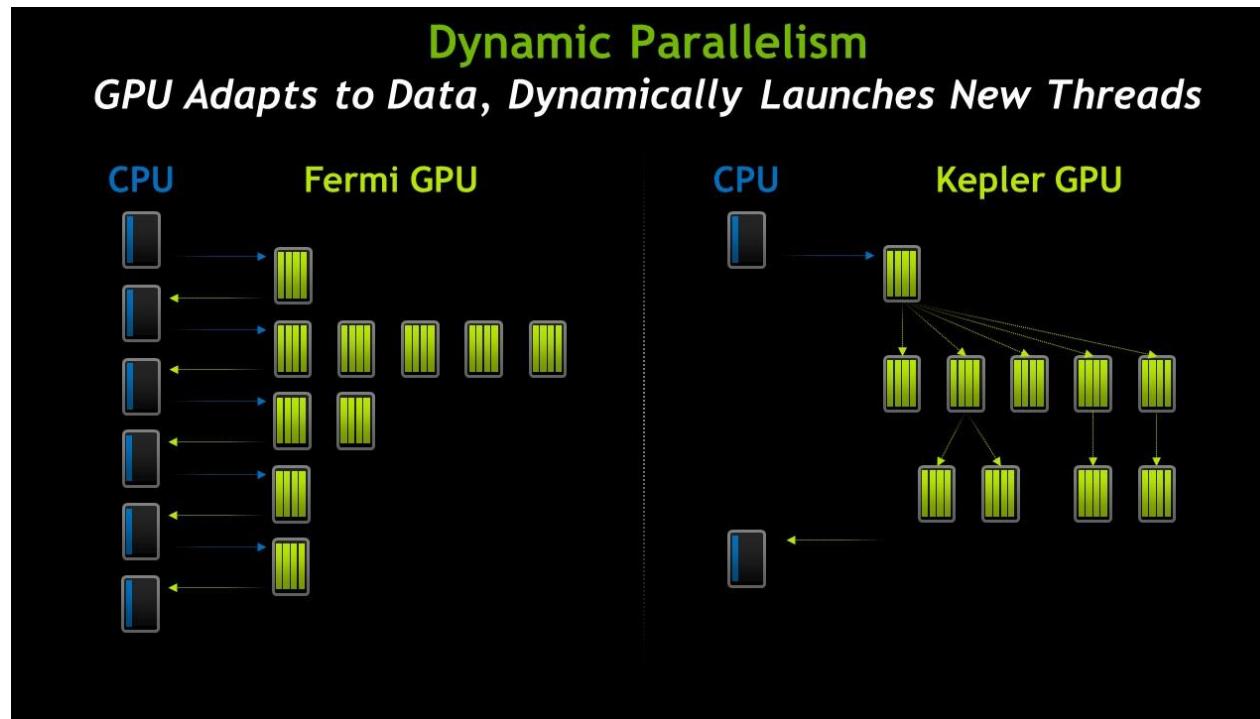
- 192 CUDA cores
- 64 Unidades FP Doble Precisión
- 32 SFU
- 32 Unidades load/store
- 4 warp schedulers
- 2 instrucciones por warp
- Hasta 255 registros por thread
- Nueva cache 48 KB L1 sólo lectura



# Arquitectura KEPLER

## Paralelismo Dinámico

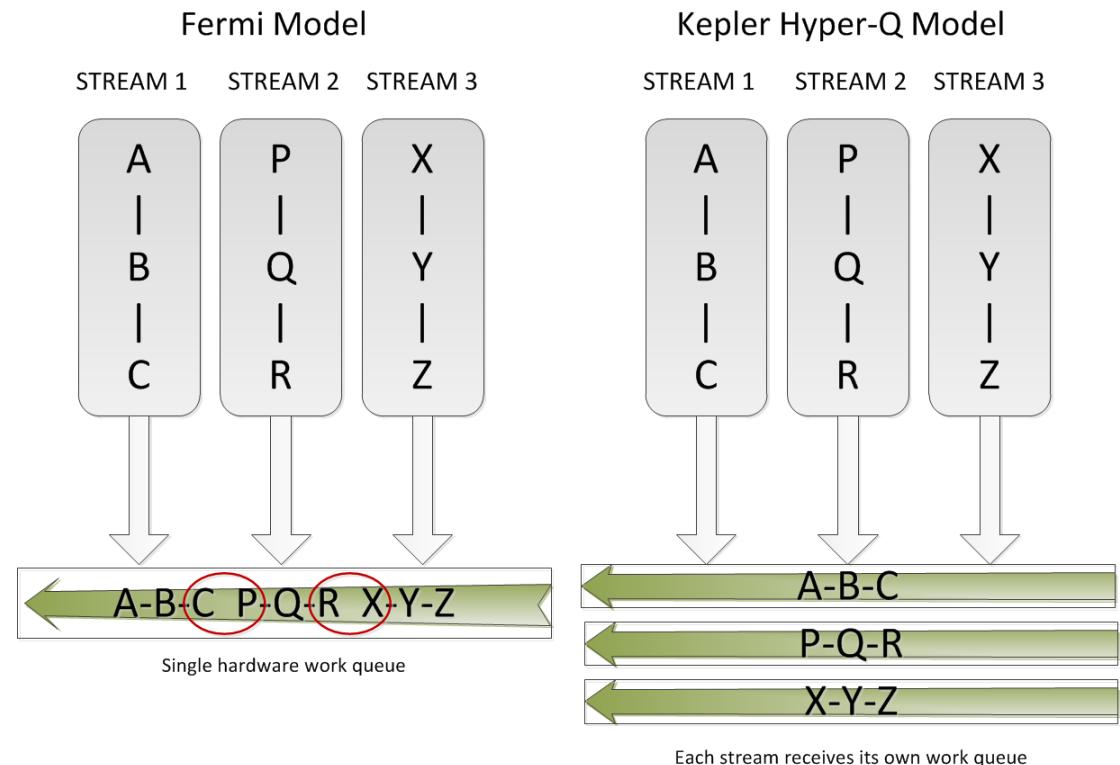
- Se puede lanzar un kernel desde un kernel, sin intervención de la CPU.
- Muy útil para equilibrar la carga



# Arquitectura KEPLER

## Hyper-Q

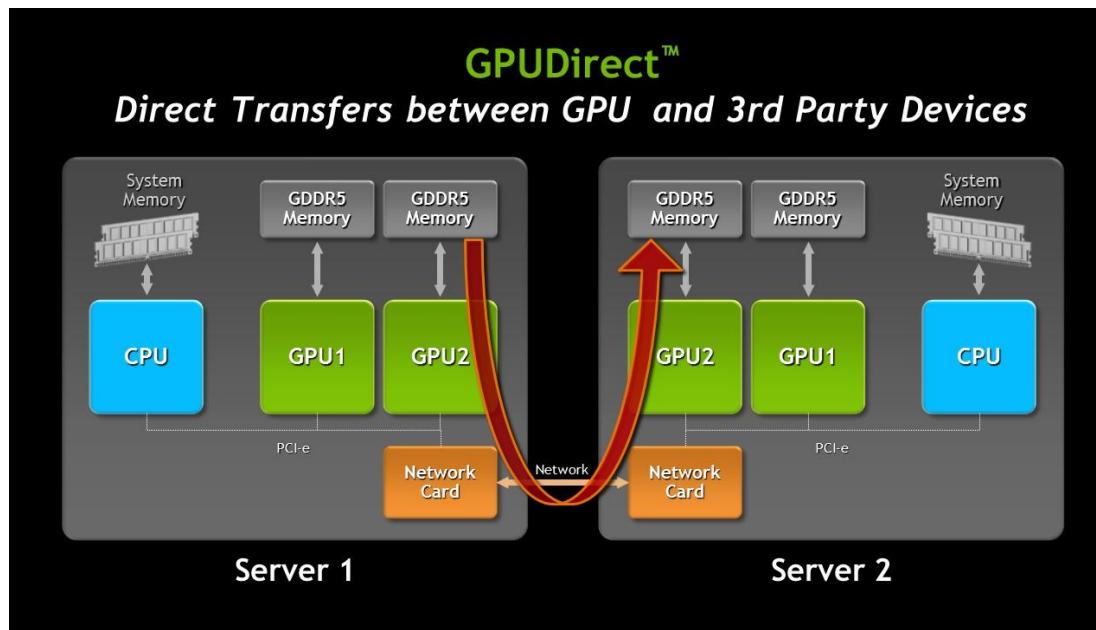
- Fermi permite varios CUDA streams, pero sólo hay 1 cola hardware para todos los streams.
- Kepler dispone de múltiples colas



# Arquitectura KEPLER

## NVIDIA GPUDirect

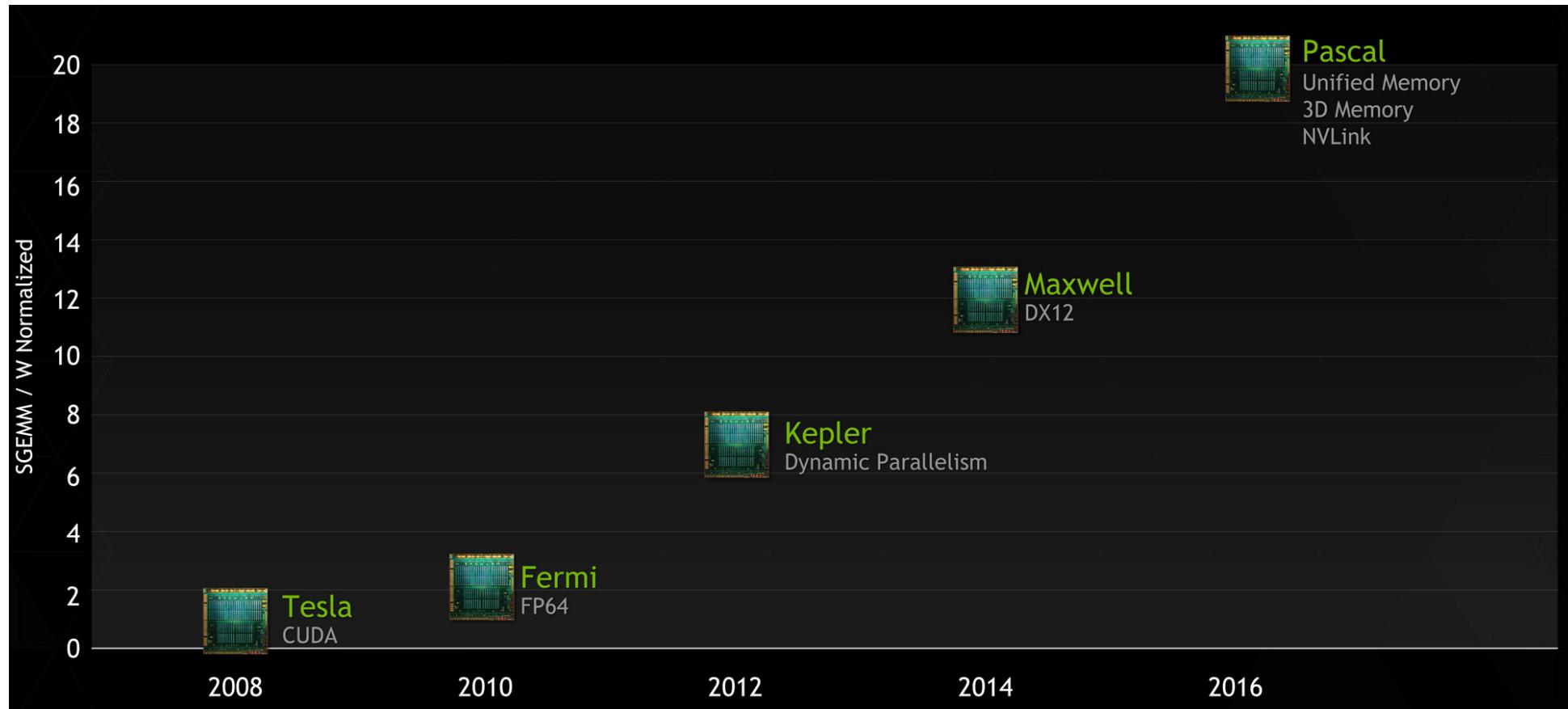
- Soporta RDMA (Remote Direct Memory Access)
- Accesos remotos a la memoria de otra GPU sin intervención de la CPU
- High Performance Computing Servers



# Arquitectura KEPLER

GPU	Fermi GF100	Fermi GF104	Kepler GK104	Kepler GK110	Kepler GK210
Compute Capability	2.0	2.1	3.0	3.5	3.7
Threads / Warp			32		
Max Threads / Block			1.024		
Max Warps / Multiprocessor	48		64		
Max Threads / Multiprocessor	1.536		2.048		
Max Thread Blocks / Multiprocessor	8		16		
32-bit Registers / Multiprocessor	32.768		65.536	131.072	
Max Registers / Thread Block	32.768		65.536		
Max Registers / Thread	63		255		
Max Shared Memory / Multiprocessor		48 KB		112 KB	
Max Shared Memory / Thread Block			48 KB		
Max X Grid Dimension	$2^{16} - 1$		$2^{32} - 1$		
Hyper-Q	No		Yes		
Dynamic Parallelism	No		Yes		

# Nvidia Roadmap



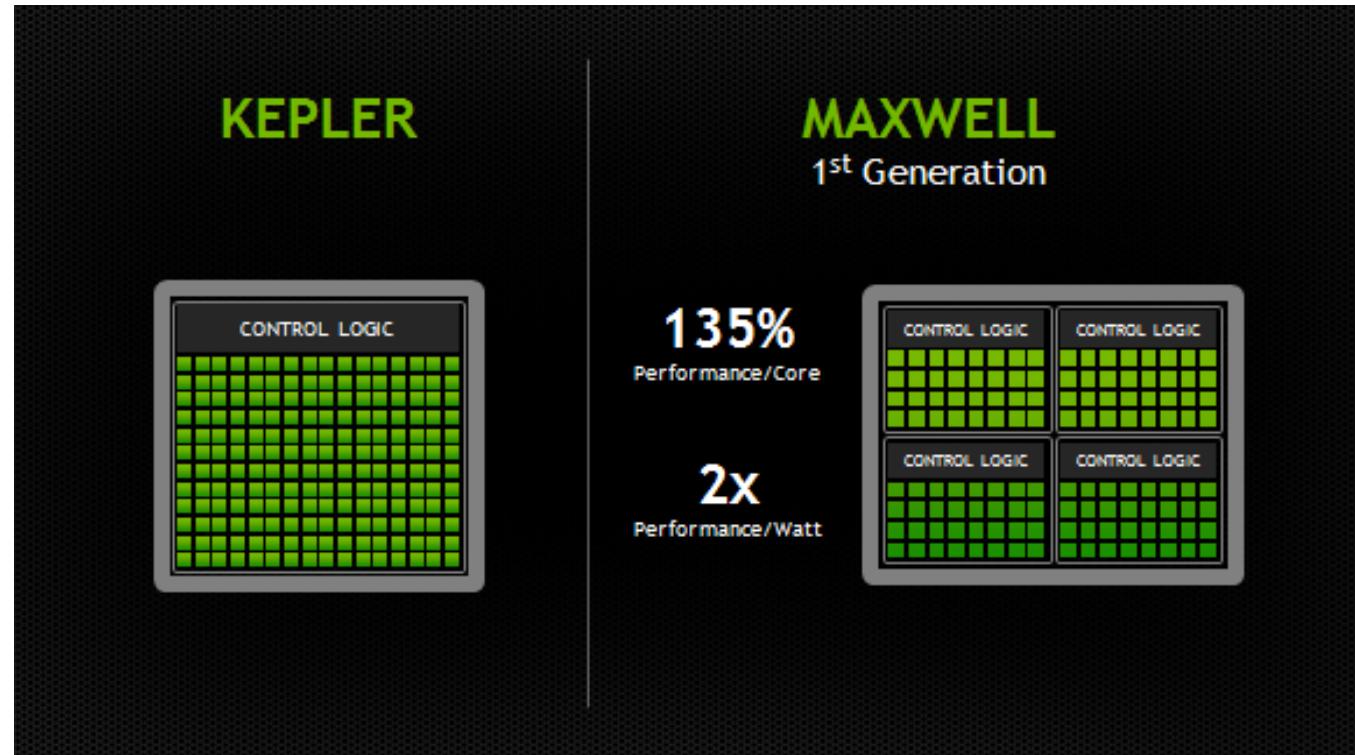
i Rendimiento por W !

[www.nvidia.com](http://www.nvidia.com)

# Arquitectura MAXWELL

## Mejora de los SMM

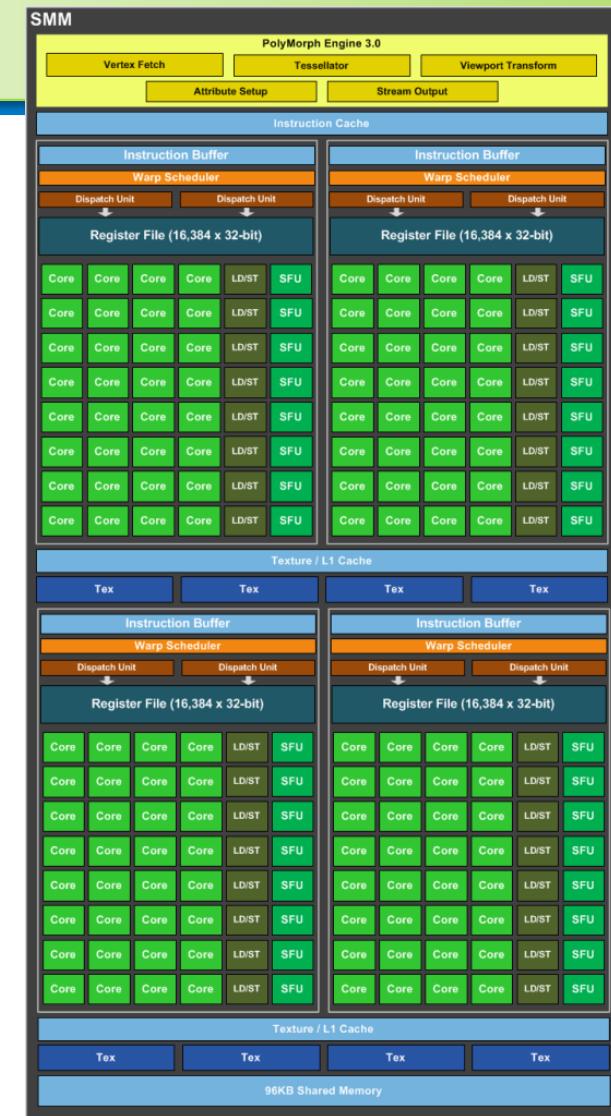
- Disminuye el número de CUDA cores por SM
- Aumenta el número de SM
- Cada SMM se divide en 4 conjuntos



# Arquitectura MAXWELL

## Maxwell Streaming Multiprocessor

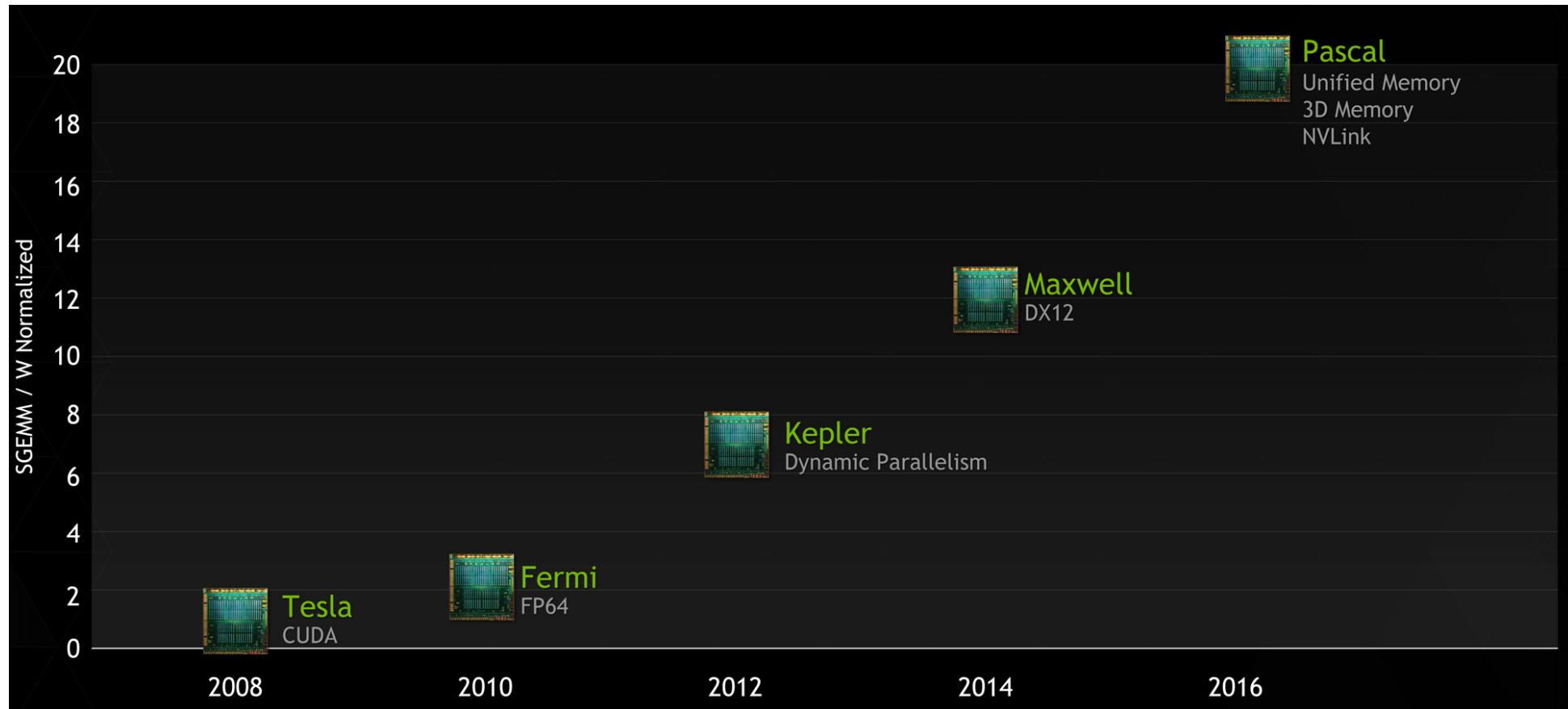
- 4 Conjuntos independientes con 32 CUDA cores, 8 SFU y 8 unidades de load/store.**
- Cada conjunto tiene su propio warp scheduler**
- 64 KB de Memoria Compartida (96 KB)**
- La cache L1 está combinada con la Texture Cache.**
- Se ha mejorado la latencia de las instrucciones**
- Las operaciones atómicas son nativas y en memoria compartida.**



# Arquitectura MAXWELL

GPU	GeForce GTX 680 ( Kepler GK104)	GeForce GTX 980 (Maxwell GM204)
CUDA Cores	1.536	2.048
GFLOPS	3.090	4.612
Compute Capability	3.0	5.2
SMs	8	16
Shared Memory / SM	48 KB	96 KB
Register File Size / SM	256 KB	256 KB
Active Blocks / SM	16	32
Texture Units	128	128
Texel fill-rate	128.8 Gtex/s	144,1 Gtex/s
Memory	2.048 MB	4.096 MB
Memory Bandwidth	192.3 GB/s	224.3 GB/s
L2 Cache	512 KB	2048 KB
TDP	195 W	165 W

# Nvidia Roadmap



i Rendimiento por W !

[www.nvidia.com](http://www.nvidia.com)

# Arquitectura PASCAL

□ Anunciada para 2016

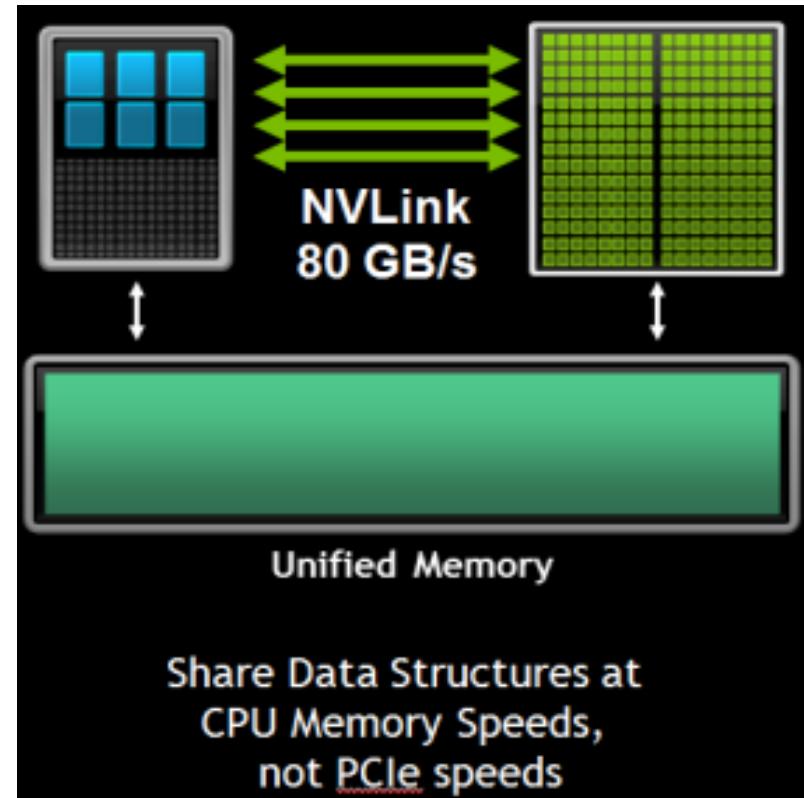
□ Memoria 3D (stacked)

- Aumenta capacidad
- Aumenta ancho de banda
- Disminuye latencia

□ Memoria Unificada CPU - GPU

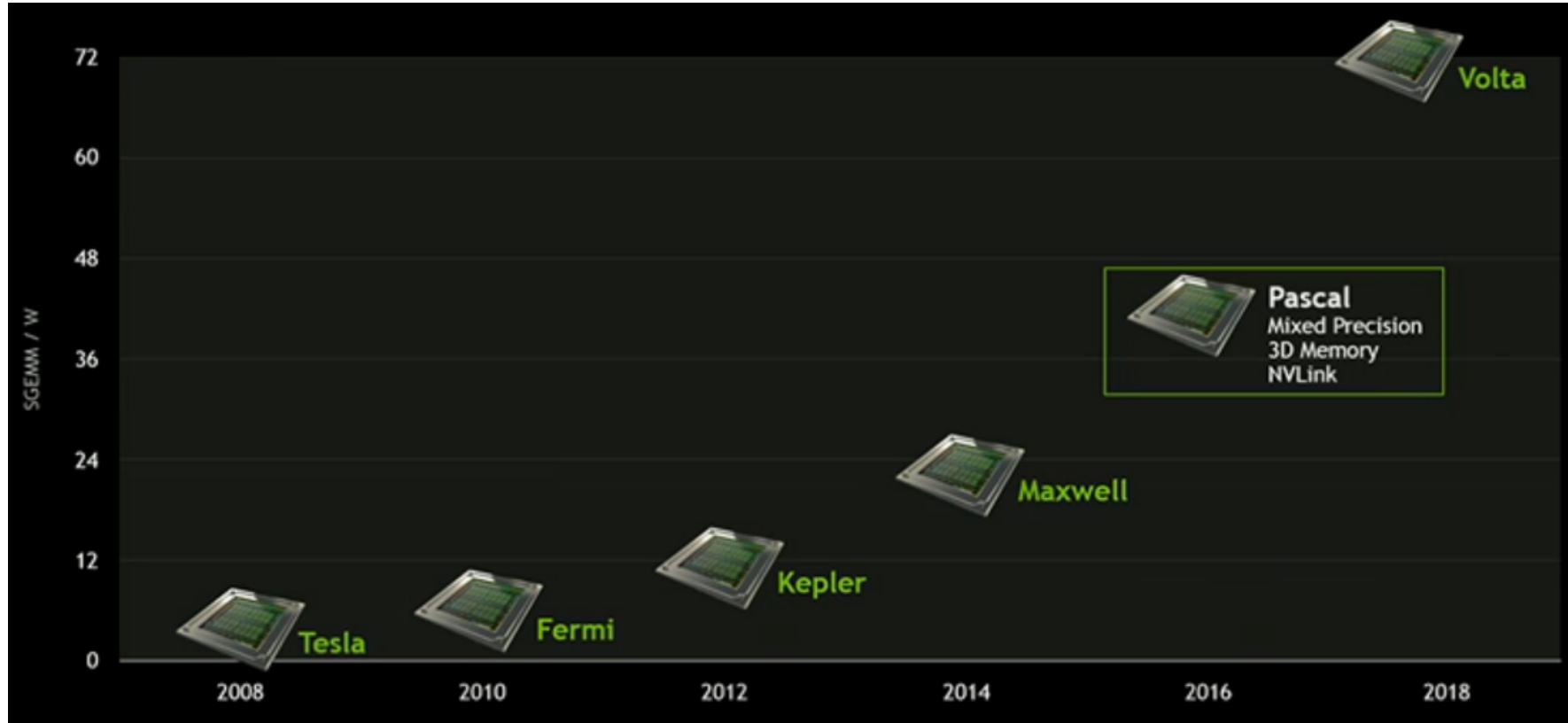
□ NVlink

- Enlace entre CPU/GPU de alta velocidad
- 80GB/s (5x PCIe x16 3.0)



Sólo han usado Memoria 3D en aceleradores (GDDR5X).  
No han implementado la conexión Nvlink entre CPU y GPU.

# Y después de PASCAL?



¡ Ha cambiado el rendimiento !

[www.nvidia.com](http://www.nvidia.com)

No hay información fiable sobre Volta, hasta el 10 de mayo de 2017

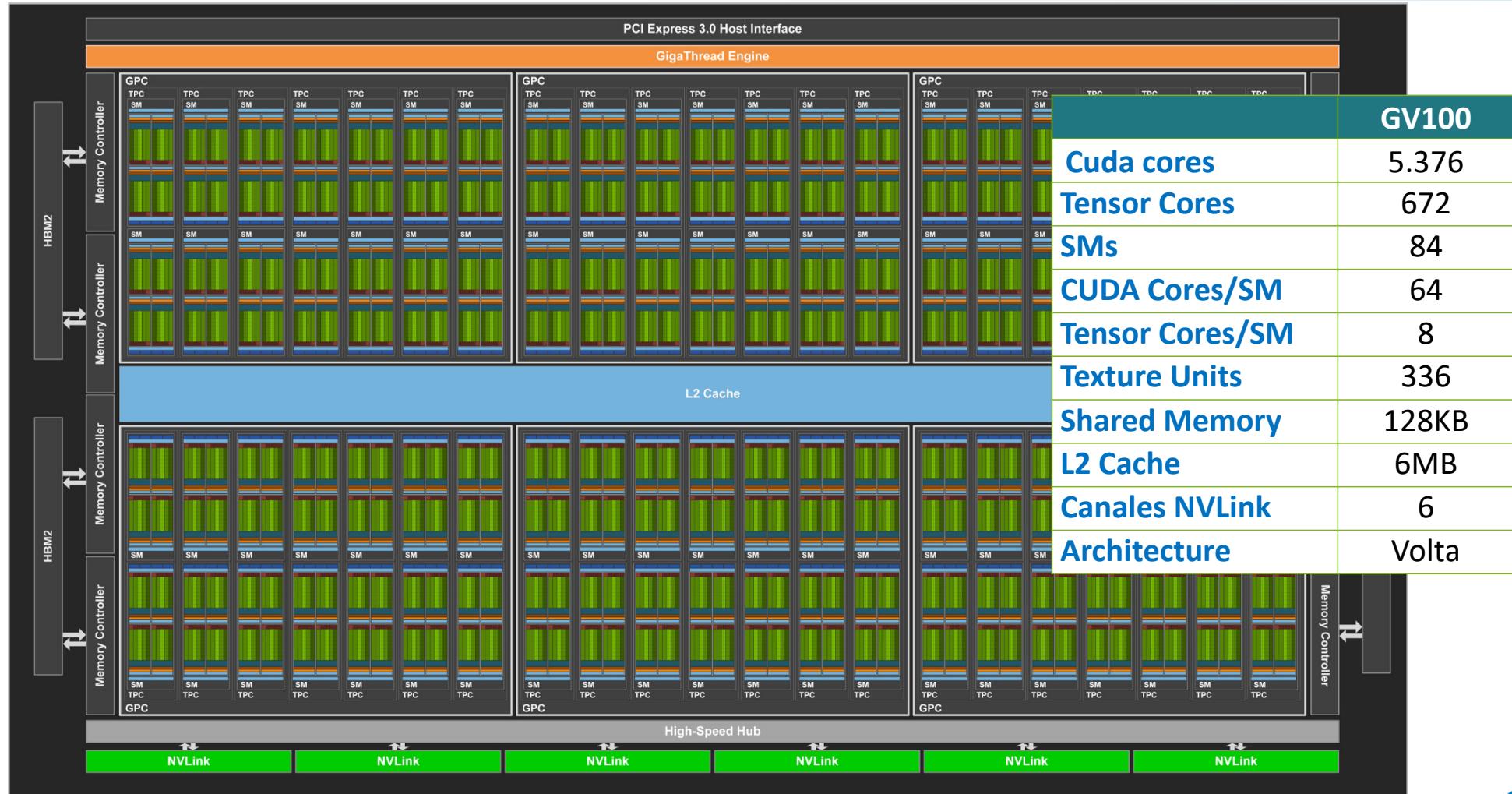
# TESLA comparativa

	Tesla V100	Tesla P100	Tesla K40	Tesla M40
<b>Stream Processors</b>	5.120	3.584	2.880	3.072
<b>Boost Clock(s)</b>	1.455 MHz	1.480 MHz	875 MHz	1.114 MHz
<b>Memory Clock</b>	1.75Gbps HBM2	1.4Gbps HBM2	6Gbps GDDR5	6Gbps GDDR5
<b>Memory Bus Width</b>	4096-bit	4096-bit	384-bit	384-bit
<b>Memory Bandwidth</b>	900 GB/s	720 GB/s	288 GB/s	288 GB/s
<b>VRAM</b>	16GB	16GB	12GB	12GB
<b>Half Precision</b>	30 TFLOPS	21.2 TFLOPS	4.29 TFLOPS	6.8 TFLOPS
<b>Single Precision</b>	15 TFLOPS	10.6 TFLOPS	4.29 TFLOPS	6.8 TFLOPS
<b>Double Precision</b>	7.5 TFLOPS	5.3 TFLOPS	1.43 TFLOPS	213 GFLOPS
<b>GPU</b>	GV100	GP100	GK110B	GM200
<b>Transistor Count</b>	$21 \cdot 10^9$	$15.3 \cdot 10^9$	$7.1 \cdot 10^9$	$8 \cdot 10^9$
<b>TDP</b>	300W	300W	235W	250W
<b>Technology</b>	TSMC 12nm FFN	TSMC 16nm FinFET	TSMC 28nm	TSMC 28nm
<b>Architecture</b>	Volta	Pascal	Kepler	Maxwell 2
<b>Precio (Mar/2020)</b>	11.500 €	6.500 €	3.138 €	6.118 €

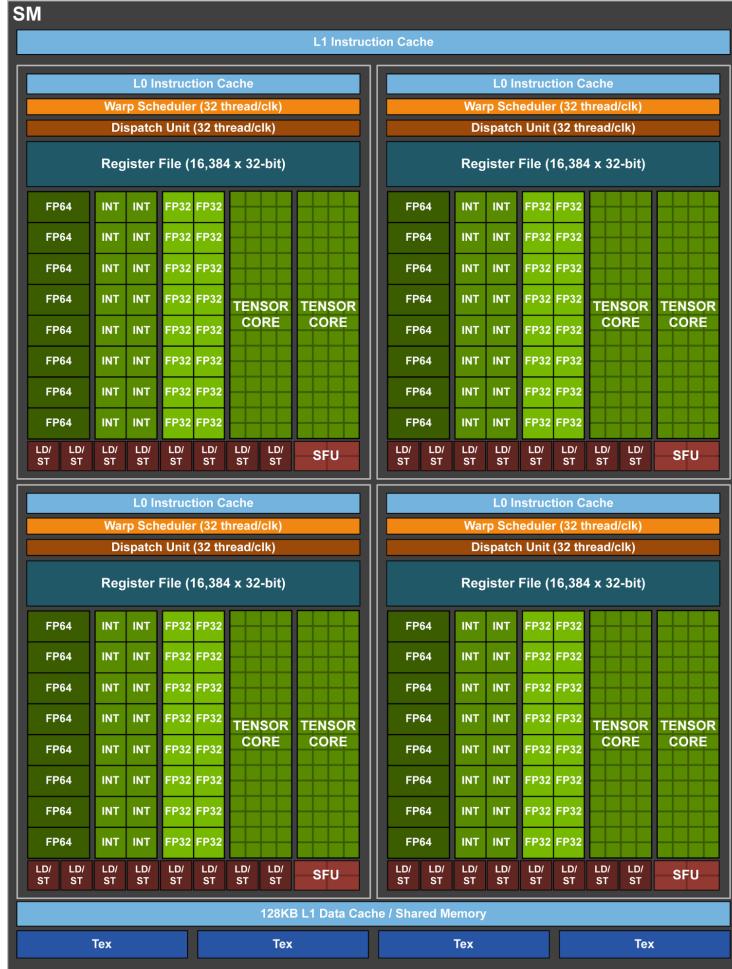
# GPU comparativa

	GV100	GP100	GK110
Cuda cores	5.376	3.840	2.880
Tensor Cores	672	N/A	N/A
SMs	84	60	15
CUDA Cores/SM	64	64	192
Tensor Cores/SM	8	N/A	N/A
Texture Units	336	240	240
Memory (bus width)	HBM2 (4096b)	HBM2 (4096b)	GDDR5 (384b)
Shared Memory	128KB, Configurable	24KB L1, 64KB Shared	48KB
L2 Cache	6MB	4MB	1.5MB
Half Precision	2:1 (Vec2)	2:1 (Vec2)	1:1
Double Precision	1:2	1:2	1:3
Die Size	815 mm <sup>2</sup>	610 mm <sup>2</sup>	552 mm <sup>2</sup>
Transistor Count	21·10 <sup>9</sup>	15.3·10 <sup>9</sup>	7.1·10 <sup>9</sup>
TDP	300W	300W	235W
Manufacturing Process	TSMC 12nm FFN	TSMC 16nm FinFET	TSMC 28nm
Architecture	Volta	Pascal	Kepler

# GV100



# GV100



	GV100
Cuda cores	5.376
Tensor Cores	672
SMs	84
CUDA Cores/SM	64
Tensor Cores/SM	8
Texture Units	336
Shared Memory	128KB
L2 Cache	6MB
Architecture	Volta

¿Tensor cores?

# Tensor cores

- Hardware específico para realizar

$$D = \left( \begin{array}{cccc} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{array} \right) \text{FP16 or FP32} \quad \left( \begin{array}{cccc} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{array} \right) \text{FP16} + \left( \begin{array}{cccc} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{array} \right) \text{FP16 or FP32}$$

- Pensado para *Tensor deep learning operations*
- Los Tensor cores alcanzan los 120 TFLOPS.
- Hasta la aparición de CUDA 9 (2017), no se sabía como se usarían.

○ Hay que usar librerías

```
#include <mma.h>
using namespace nvcuda;
```

# HBM2 y NVLINK

## □ Utiliza memoria HBM2

- Sólo tiene 4 chips de memoria apilados.
- El estándar define hasta 8 chips apilados, pero nadie los fabricaba
- Ha aumentado la frecuencia respecto a una P100 (de 1,4 Gbps a 1,75 Gbps)

## □ Incorpora NVLink 2

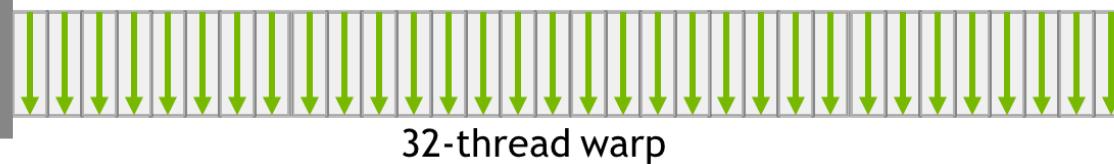
- 6 canales bidireccionales NVLINK a 25 GB/s
- Incluye un protocolo de coherencia entre CPU – GPU
- La versión PCIe no aprovecha NVLink

# Comportamiento de los warps

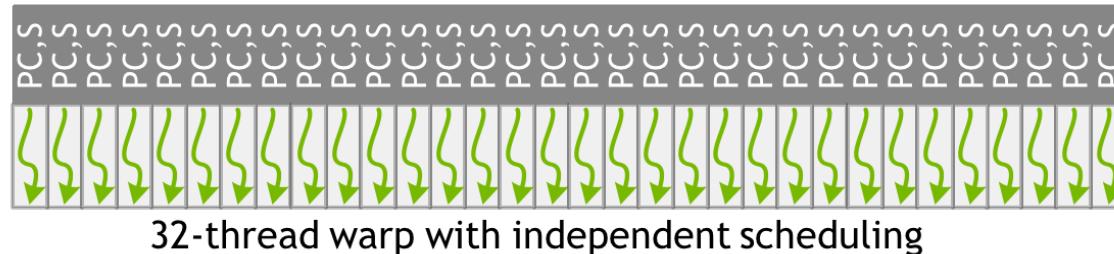
- Hay un cambio en el comportamiento de los warp

Pre-Volta

Program Counter (PC)  
and Stack (S)



Volta

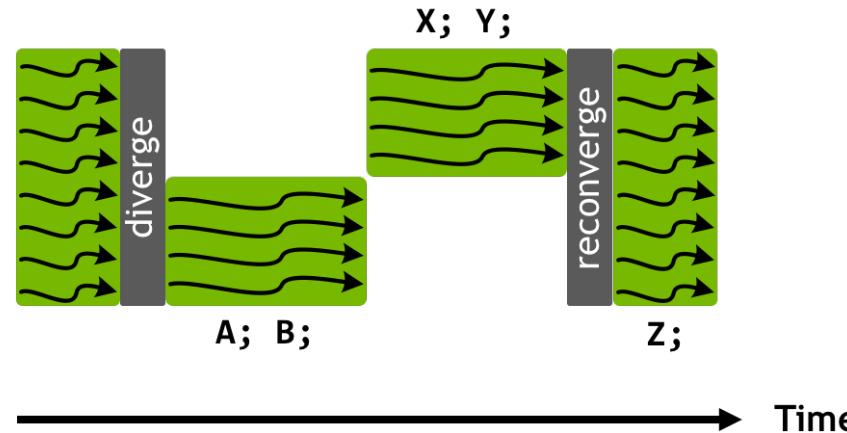


[www.nvidia.com](http://www.nvidia.com)

- Las posibilidades son muchas, la información disponible no es muy alentadora.
- Va a haber cambios en las opciones de sincronización (CUDA 9).

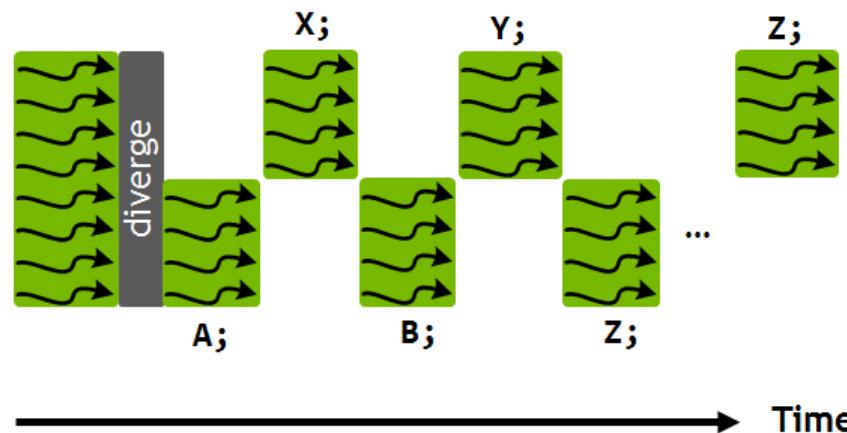
# Comportamiento de los warps

```
if (threadIdx.x < 4) {  
    A;  
    B;  
} else {  
    X;  
    Y;  
}  
Z;
```



Pre Volta

```
if (threadIdx.x < 4) {  
    A;  
    B;  
} else {  
    X;  
    Y;  
}  
Z;
```

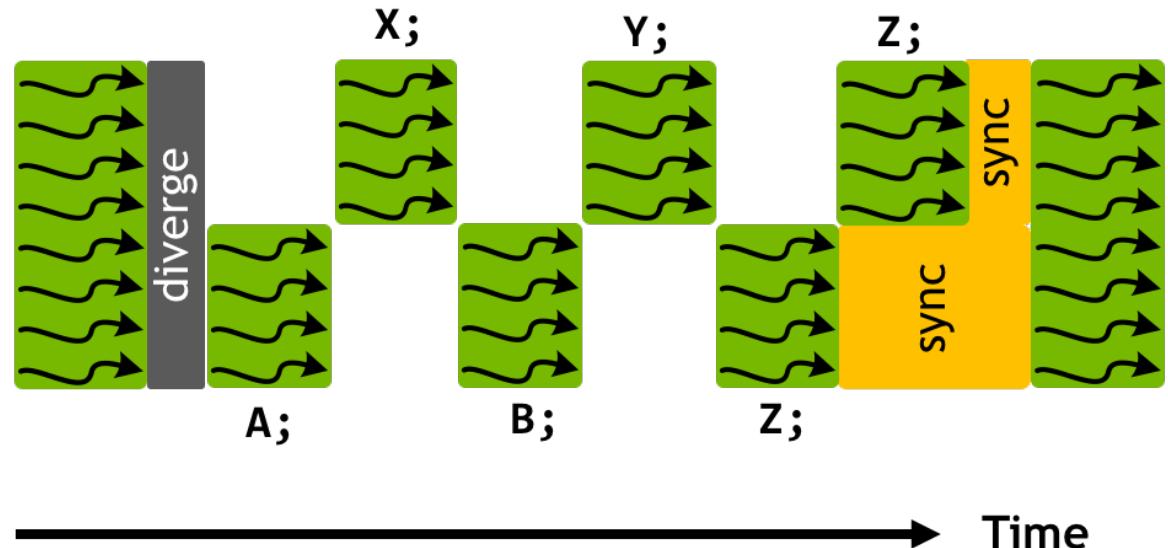


Volta

¿Ventajas?

# Comportamiento de los warps

```
if (threadIdx.x < 4) {  
    A;  
    B;  
} else {  
    X;  
    Y;  
}  
Z;  
__syncwarp()
```



\_\_syncwarp()

# NVIDIA Tesla V100 (for NVLink) (for PCIe)



140 × 78 mm



267 × 111 mm

NVIDIA ha invertido 3.000 M\$ en este proyecto.

[www.nvidia.com](http://www.nvidia.com)

# ¿Qué vende NVIDIA?

## SYSTEM SPECIFICATIONS

GPUs	8X Tesla V100	8X Tesla P100
TFLOPS (GPU FP16)	960	170
GPU Memory	128 GB total system	
CPU	Dual 20-Core Intel Xeon E5-2698 v4 2.2 GHz	
NVIDIA CUDA® Cores	40,960	28,672
NVIDIA Tensor Cores (on V100 based systems)	5,120	N/A
Maximum Power Requirements	3,200 W	
System Memory	512 GB 2,133 MHz DDR4 LRDIMM	
Storage	4X 1.92 TB SSD RAID 0	
Network	Dual 10 GbE, Up to 4 IB EDR	
Software	Ubuntu Linux Host OS See Software Stack for Details	
System Weight	134 lbs	
System Dimensions	866 D x 444 W x 131 H (mm)	
Packing Dimensions	1,180 D x 730 W x 284 H (mm)	
Operating Temperature Range	10-35 °C	



**DGX1, 149.000\$**



**DGX station, 69.000\$**

## SYSTEM SPECIFICATIONS

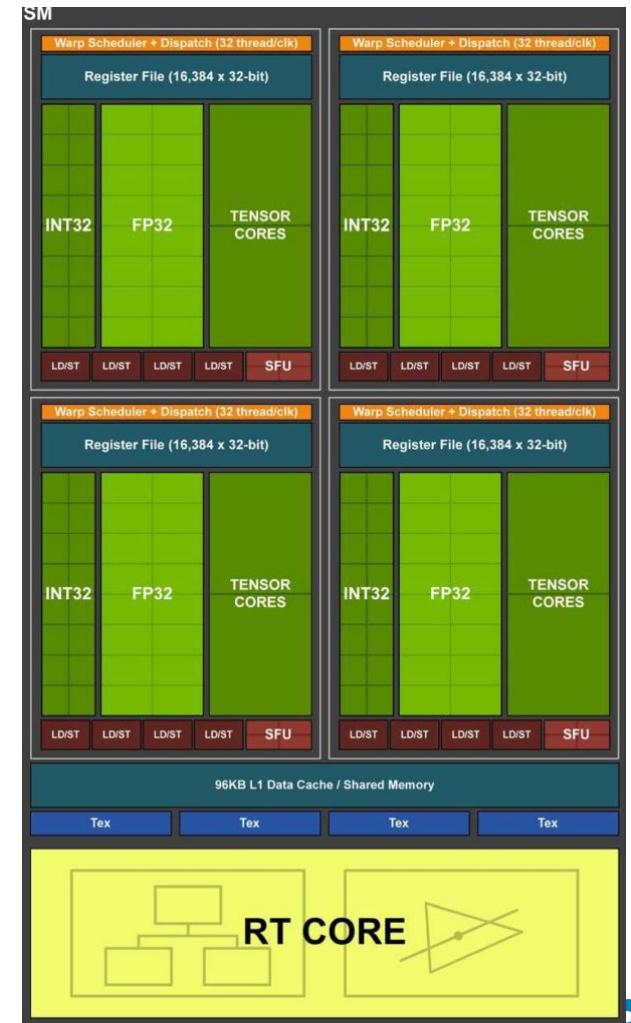
GPUs	4X Tesla V100
TFLOPS (GPU FP16)	480
GPU Memory	64 GB total system
NVIDIA Tensor Cores	2,560
NVIDIA CUDA® Cores	20,480
CPU	Intel Xeon E5-2698 v4 2.2 GHz (20-Core)
System Memory	256 GB LRDIMM DDR4
Storage	Data: 3X 1.92 TB SSD RAID 0 OS: 1X 1.92 TB SSD
Network	Dual 10 Gb LAN
Display	3X DisplayPort, 4K resolution
Acoustics	< 35 dB
System Weight	88 lbs / 40 kg
System Dimensions	518 D x 256 W x 639 H (mm)
Maximum Power Requirements	1,500 W
Operating Temperature Range	10-30 °C
Software	Ubuntu Desktop Linux OS DGX Recommended GPU Driver CUDA Toolkit

# Y ahora llega Turing

- Hardware específico para Ray Tracing en tiempo real
- Aplicaciones Deep Learning orientadas a gráficos
- Memoria de Alto rendimiento GDDR6
- Segunda generación del NVIDIA NVLink
- USB-C and VirtualLink
- La implementación de la GPU TU102 incluye:
  - 4,608 CUDA Cores
  - 72 RT Cores
  - 576 Tensor Cores
  - 288 texture units
  - 12 32-bit GDDR6 memory controllers (384-bits) (616 GB/s)

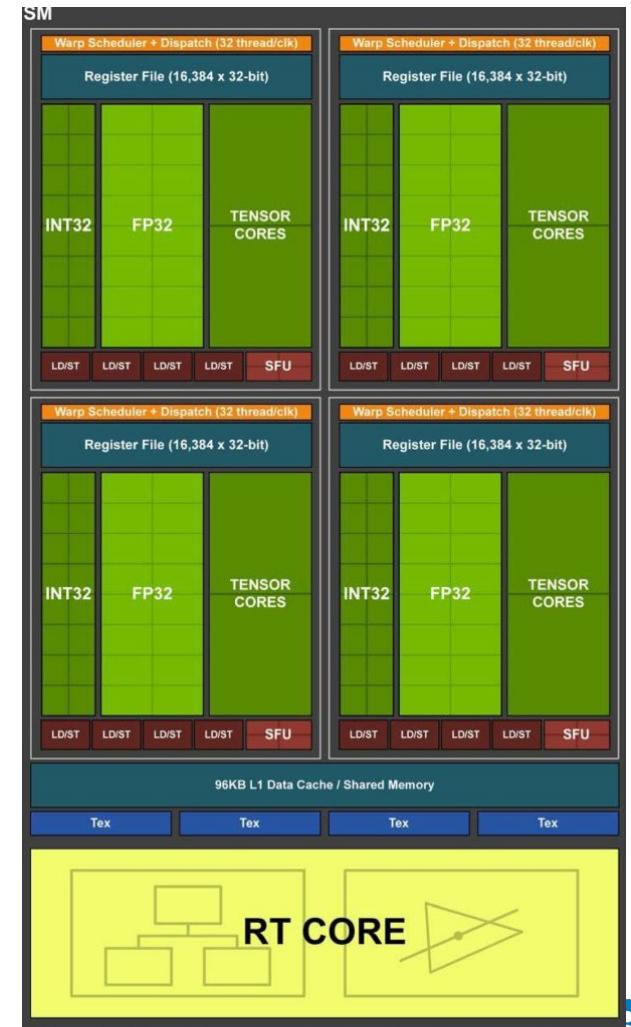
# El Nuevo Streaming Multiprocessor

- El Data path de Turing permite ejecutar simultáneamente instrucciones con enteros con instrucciones en coma flotante.
- Nuevos Tensor Cores, añaden la capacidad de trabajar con enteros para inferencia.
- Deep Learning Super Sampling (DLSS) soportada por los Tensor cores. Utilizando menos samples que en las técnicas tradicionales permite generar imágenes de alta calidad, aprovechando redes neuronales previamente entrenadas.



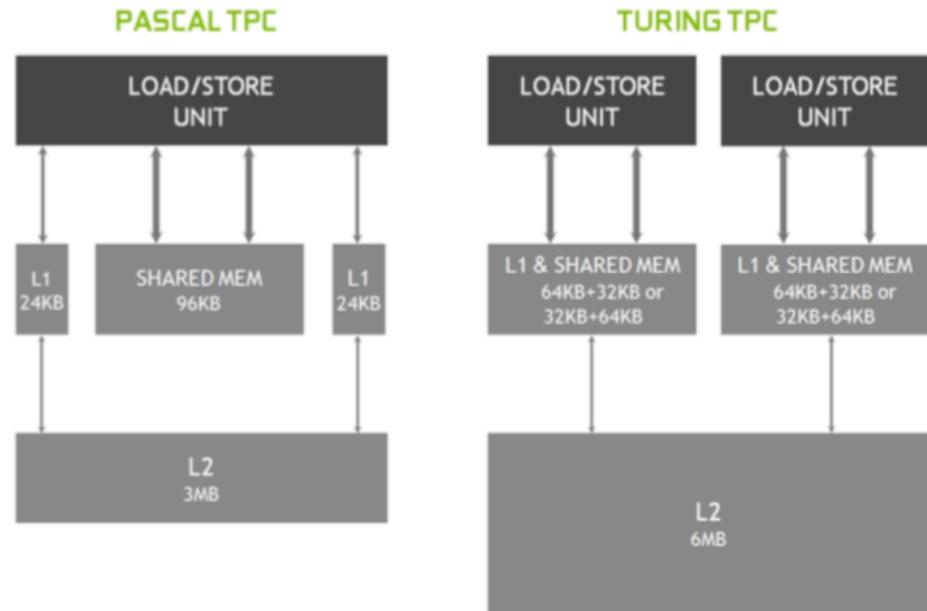
# El Nuevo Streaming Multiprocessor

- Ray Tracing en tiempo real. Dispone de hardware específico para renderizar juegos 3D y modelos profesionales con sombras precisas, reflexiones y refracciones.
- El camino con memoria ha sido rediseñado para unificar shared memory, cache de texturas y memoria cache L1. Dobra el tamaño de L1 y dobla el ancho de banda.



# Cambios en la Jerarquía de memoria

2x L1 Bandwidth  
Lower L1 Hit Latency  
Up to 2.7x L1 Capacity  
2x L2 Capacity



## ❑ GDDR6

Es la memoria más rápida de la industria

Parece que han abandonado las HBM2

## ❑ Usa Técnicas de Compresión de datos al escribir en el frame buffer. Reduce el tráfico con memoria un 50%

# Nuevas conexiones con el exterior

## □ USB-C y VIRTUALLINK

- USB-C es el nuevo estándar que permite mover imagen (display), datos y power por un mismo conector.
- Virtuallink está diseñado para los nuevos dispositivos de realidad virtual conectados a través del USB-C

## □ Nueva versión del NVLINK

- La nueva TU102 incluye 2 NVLINKx8, cada uno de ellos capaz de mover 25 GB/s en cada dirección. En total tienen un ancho de banda de 100 GB/s
- En modo multiGPU, utiliza este conector en vez del SLI (también lo tiene). Mejoras espectaculares.

# Tecnología NVIDIA NGX

- No funciona en GPUs anteriores. Funciona con CUDA 10, DirectX y Vulkan.
- **Deep Learning Super-Sampling (DLSS)**
  - Nvidia lleva mucho tiempo trabajando con capacidades de IA relacionadas con la reconstrucción de imágenes, y ha encontrado la forma de explotar esto en los videojuegos.
  - Usa una red previamente entrenada con numerosos ejemplos de super alta calidad.
  - Permite generar un frame con resolución HD y luego escalarlo a 4K. Alta calidad con un coste razonable.
  - Permite aplicar técnicas de antialiasing con una calidad muy elevada a un coste razonable (usando los Tensor cores).

# Tecnología NVIDIA NGX

## INPAINTING

Permite eliminar partes de una imagen y generar una alternativa realista de forma totalmente automática. [\[link vídeo\]](#)



## AI SUPER REZ

Incrementa la resolución de una imagen o video por 2x, 4x, u 8x.

# Tecnología NVIDIA NGX

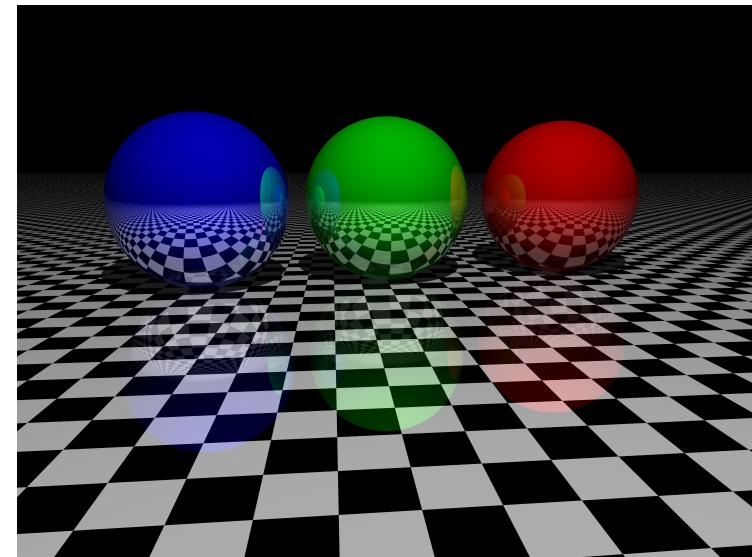
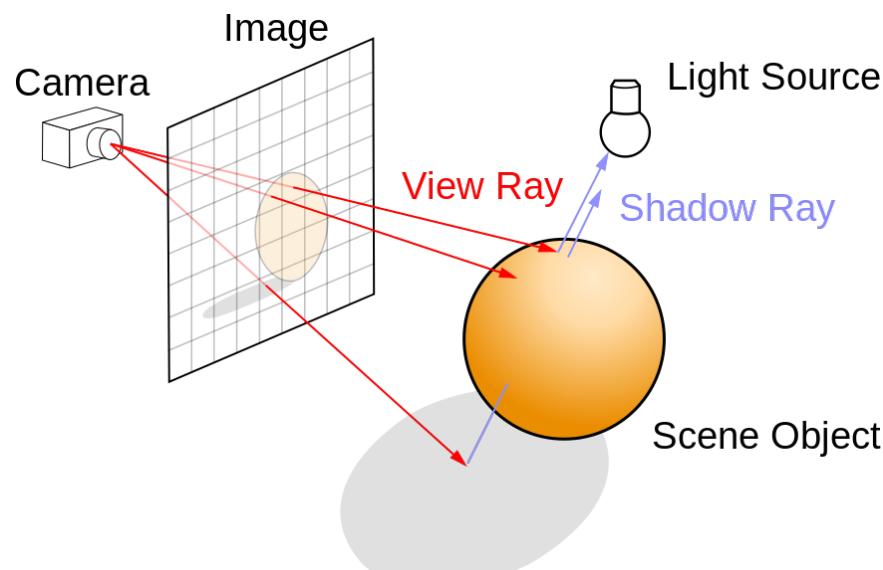
## □ AI SLOW-MO

Inserta imágenes interpoladas en un video para añadir imágenes a cámara lenta. Analiza cada frame para identificar características, objetos, movientos de cámara y generar los frames intermedios. Por ejemplo, convierte un video de 30 fps en un video de 240 fps [\[link vídeo\]](#).



# Algoritmos de renderización: Ray Tracing

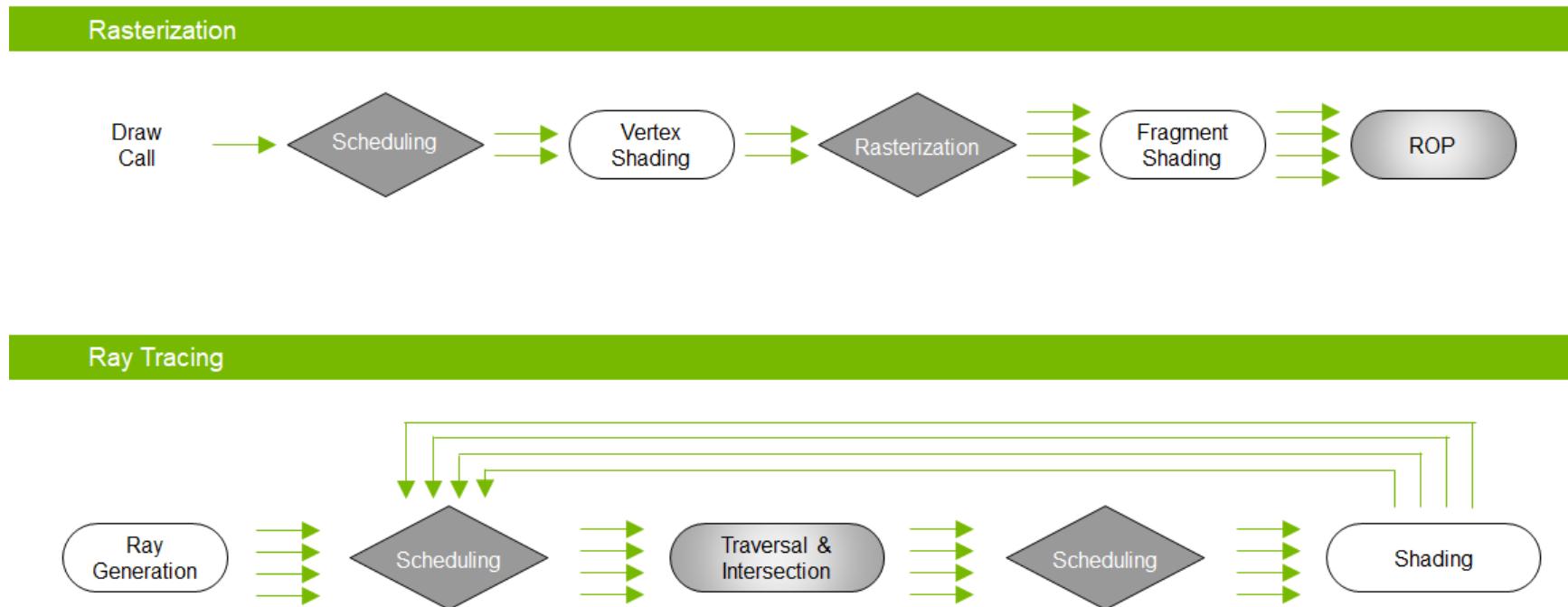
- Para cada píxel  $(x,y)$  se calcula la interacción entre los objetos y las luces:
  - RayTracing, Radiosity, Photon Map
  - El cálculo de cada píxel es muy costoso: Iluminación indirecta (sombras, reflexiones indirectas de los mismos objetos, ...).
  - **Usan algoritmos iterativos / recursivos. Poco adecuados para juegos interactivos**



**PROBLEMA: ¿Cuándo intersecta un rayo con un triángulo?**

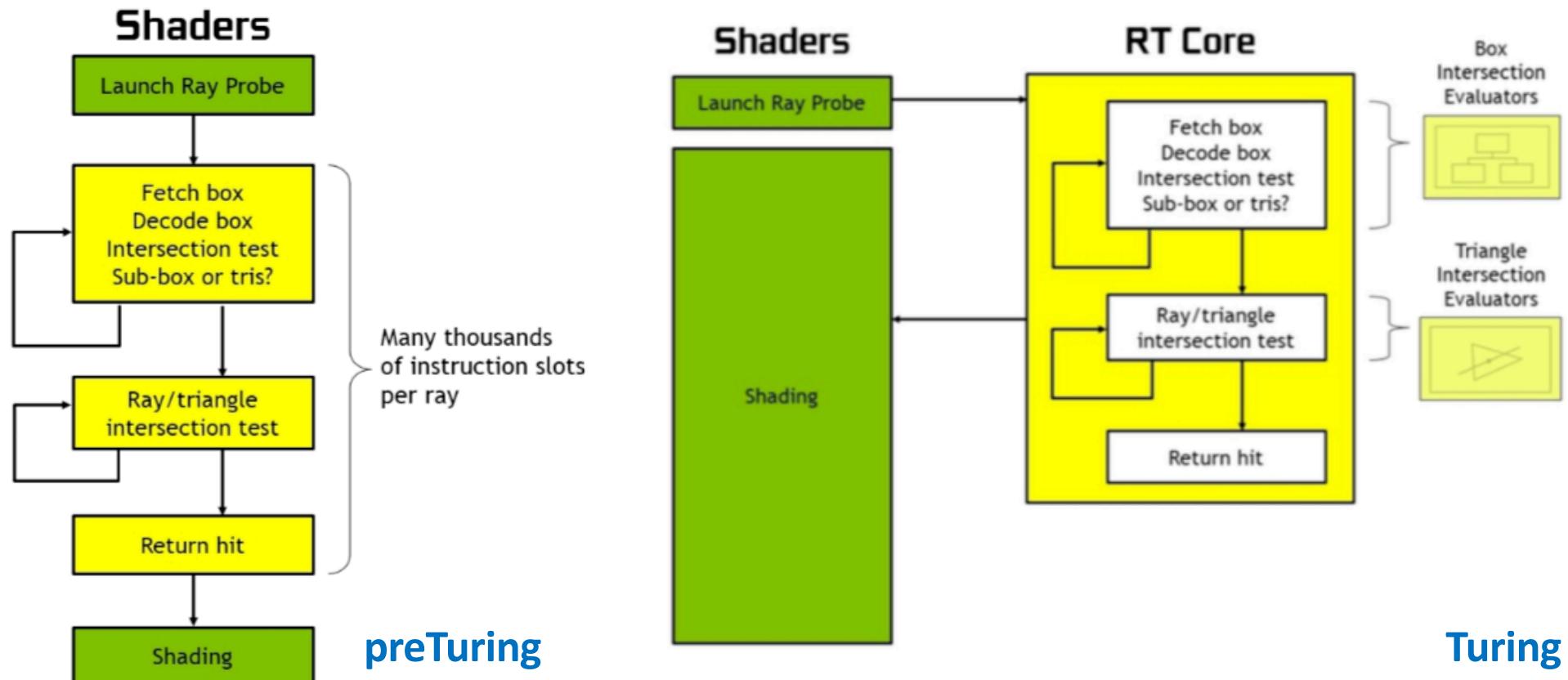
# Ray Tracing en Turing

- Es una combinación de hardware y el NVIDIA RTX software.
- Hybrid Rendering



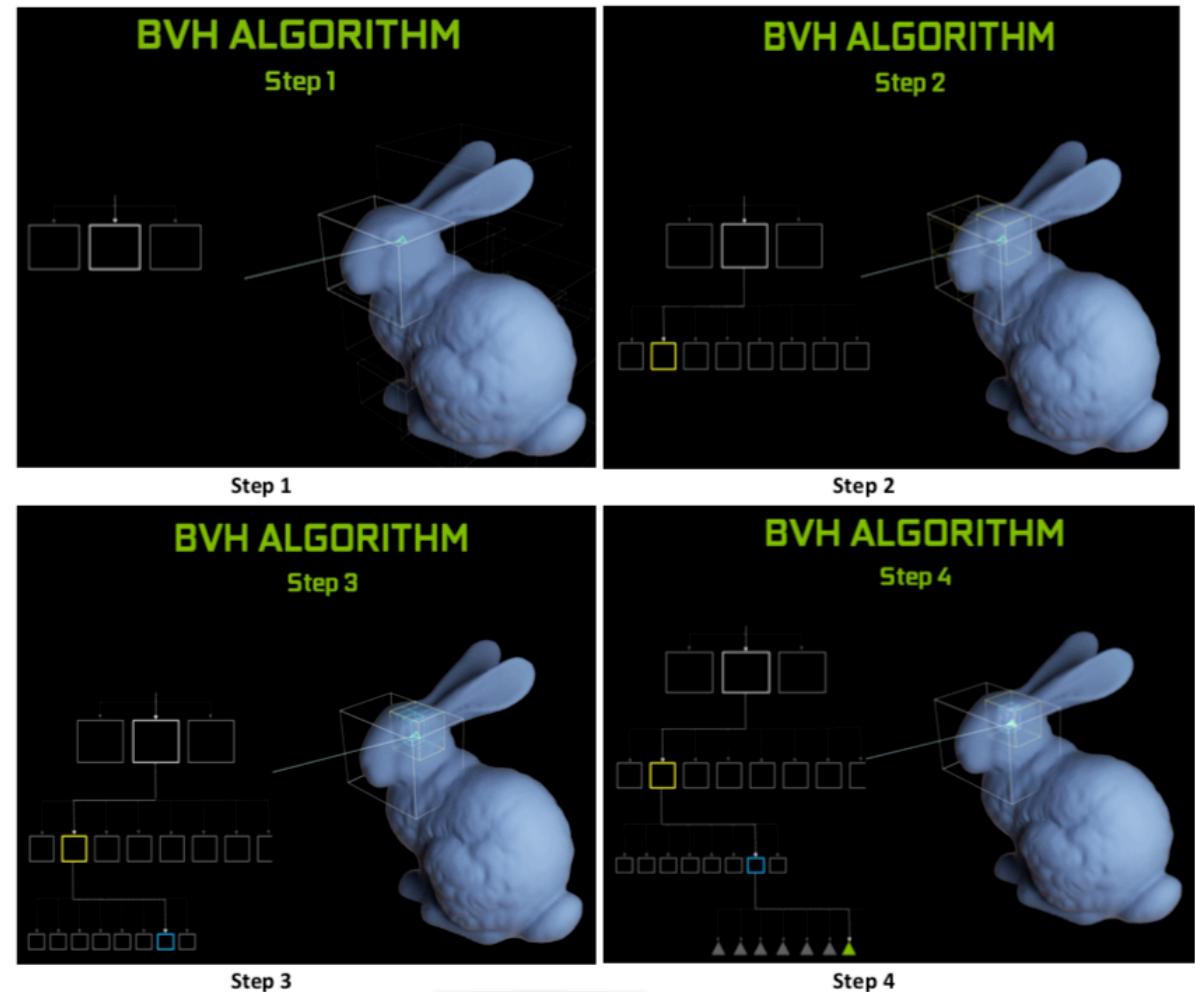
- [\[Video 1\]](#) [\[Video 2\]](#)

# Ray Tracing en una GPU



# Ray Tracing en una GPU cualquiera

- En los RT Cores aceleran el proceso utilizando **Bounding Volume Hierarchy** (BVH).
- No generamos cientos de rayos por pixel. De hecho sólo se generan unos pocos usando los RT cores en combinación con técnicas para filtrar el ruido de la imagen, produciendo imágenes excelentes. Usan algoritmos de IA parecidos a los que hemos comentado



# Ray Tracing en Turing

Resumiendo, Ray Tracing en tiempo real funciona en Turing por:

- Hybrid rendering

Reduce la cantidad de rayos necesario para generar la escena.

- Denoising algorithms

Los filtros eliminan las posibles inconsistencias, artefactos, etc de la imagen.

- BVH algorithm

Usado para calcular la intersección de los triangulos con los rayos de forma eficiente.

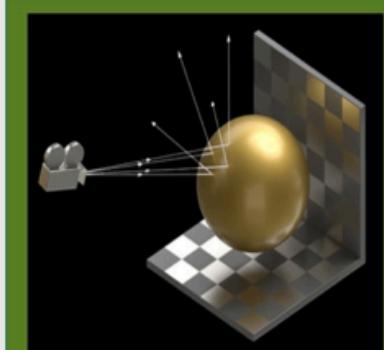
- RT Cores

Es el elemento hardware que permite acelerar el calculo hasta en 10x respecto a las mejores GPUs del mercado.

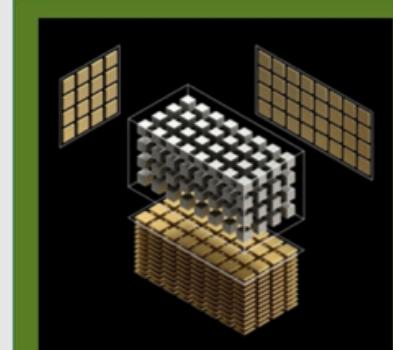
# Ampere GA10x Architecture, un paso de gigante



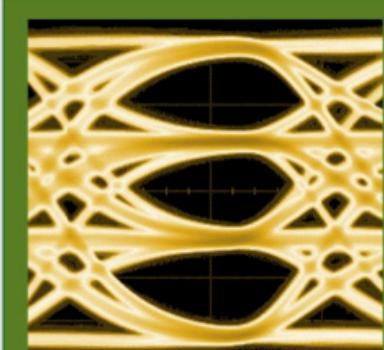
New SM



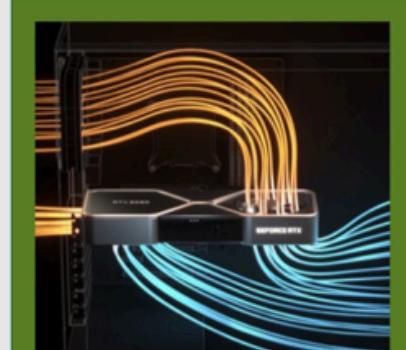
2<sup>nd</sup> Gen  
RT Core



3<sup>rd</sup> Gen  
Tensor Core



GDDR6X



System Design

NVIDIA AMPERE GA102 GPU ARCHITECTURE. *Second-Generation RTX*

- Mejor rendimiento y aprovechamiento energético que Turing.
- GeForce RTX 3090 es la GPU más potente construida por NVIDIA, diseñada para juegos 8K HDR y supuestamente puede ejecutar muchos juegos a 8K@60fps

# Ampere GA10x Architecture, un paso de gigante

## ☐SMs completamente rediseñados

- 2x FP32
- 3<sup>a</sup> gen Tensor Cores
  - ✓ Sparsity
- 2<sup>a</sup> gen RT Cores
  - ✓ Motion Blur
- Nuevas configuraciones de L1 y Shared Memory
  - ✓ 128KB + 0KB, 120KB + 8 KB, 112KB + 16 KB, 96KB + 32KB, 64KB + 64 KB, 28KB + 100KB

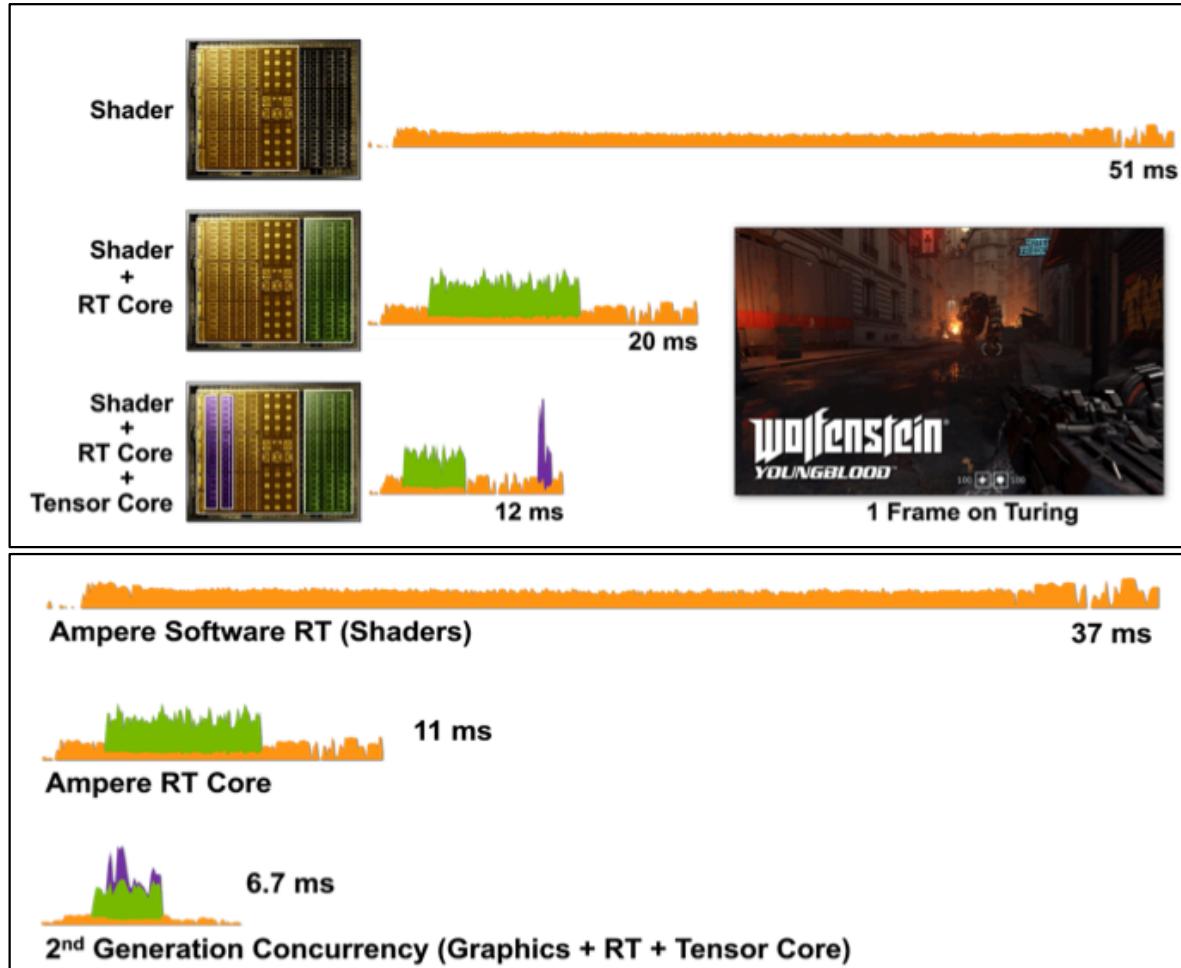


# Ampere, una arquitectura orientada a RayTracing



- Arquitectura orientada a RayTracing: Marbles at night [\[link vídeo\]](#)
- Presentación de Ampere por Jen-Hsun Huang (CEO NVIDIA) [\[link vídeo\]](#)

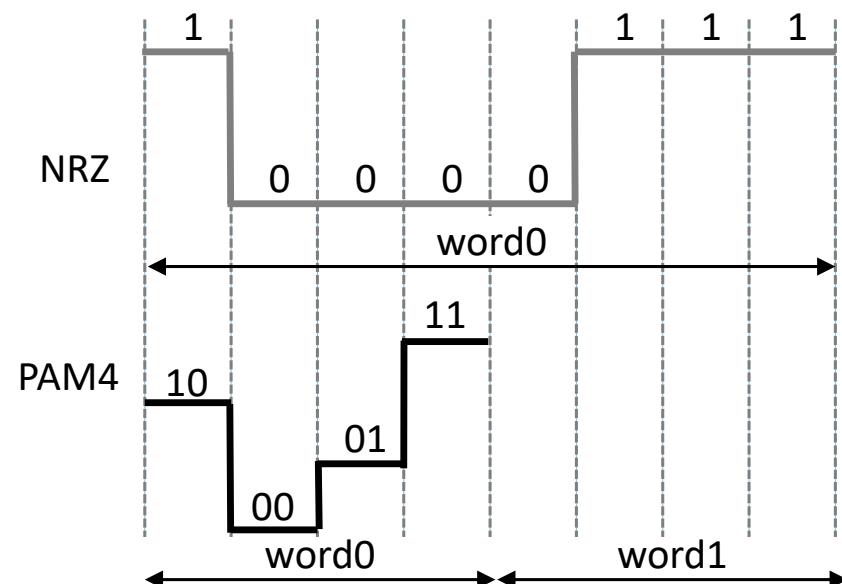
# Ampere, una arquitectura orientada a RayTracing



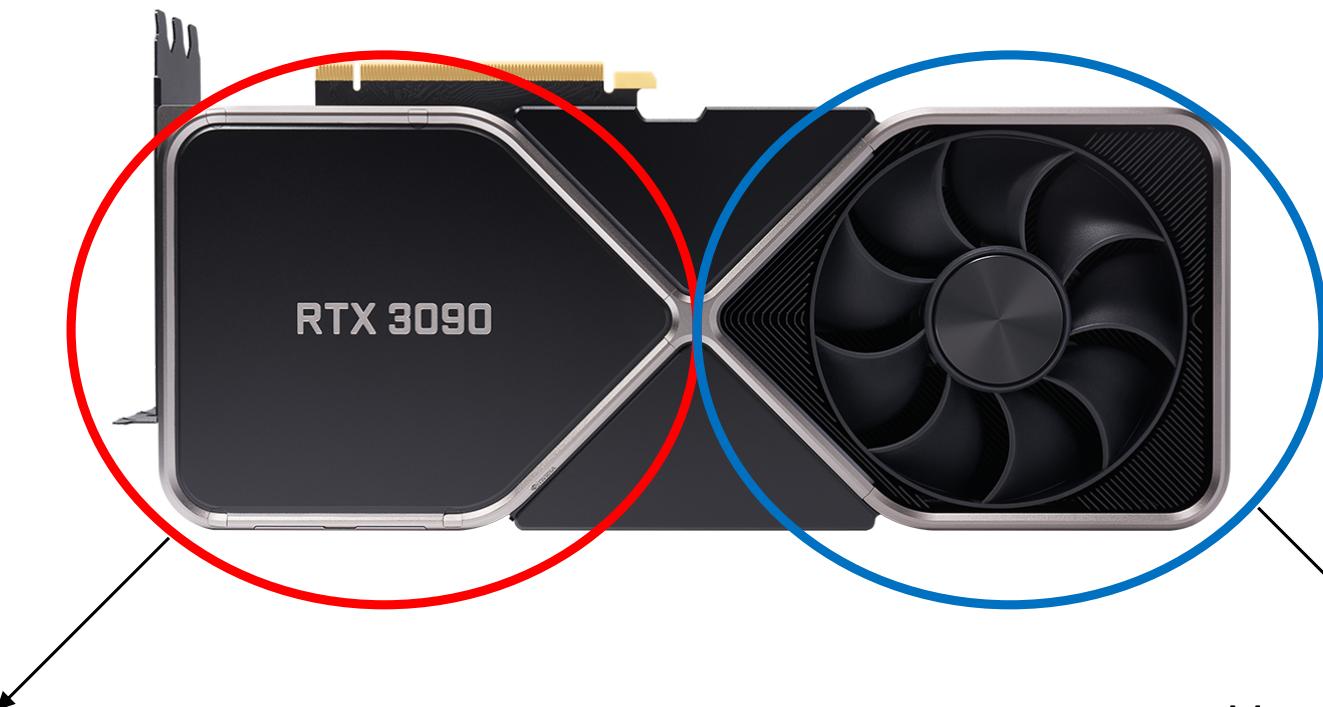
NVIDIA AMPERE GA102 GPU ARCHITECTURE. Second-Generation RTX

# GDDR6X

- Desarrollada por Micron y NVIDIA
- En una GDDR convencional se transmite 1 bit en cada flanco de reloj, codificación NRZ (2 niveles de voltaje)
- En una GDDR6X se transmiten 2 bits (00, 01, 10, 11) en cada flanco de reloj, codificación PAM4 (4 niveles de voltaje)
  - La información se mueve codificada
- Permite trabajar a menor frecuencia:
  - Manteniendo/mejorando el ancho de banda
  - Menor calor generado
  - Menor consumo



# NVIDIA RTX 3090, Sistema refrigeración revolucionario



GPU, memòria, ... + ventilador  
Refrigera la placa y saca el **aire caliente** fuera de la cabina

Propaganda NVIDIA

- 3 veces más silenciosa
- 20 grados menos que una RTX 2000

Ventilador, ayuda a que entre aire en la cabina

# Comparativa GeForce RTX 2080 vs RTX 3080

	GeForce RTX 2080	GeForce RTX 3080
<b>GPU Codename</b>	TU104	GA102
<b>GPU Architecture</b>	NVIDIA Turing	NVIDIA Ampere
<b>GPCs</b>	6	6
<b>TPCx</b>	23	34
<b>SMs</b>	46	68
<b>CUDA Cores /SM</b>	64	128
<b>CUDA Cores / GPU</b>	2944	8704
<b>Tensor Cores /SM</b>	8 (2nd gen)	4 (3rd gen)
<b>Tensor Cores /GPU</b>	368 (2nd gen)	272 (3rd gen)
<b>RT Cores</b>	46 (1st gen)	68 (2nd gen)
<b>GPU Boost Clock</b>	1800 MHz	1710 MHz
<b>Peak FP32</b>	10,6 TFLOPs	29,8 TFLOPs
<b>Peak FP16</b>	21,2 TFLOPs	29,8 TFLOPs
<b>Peak INT32</b>	10,6 TOPs	14,9 TOPs
<b>Peak FP16 Tensor</b>	84,8 TFLOPs	119/238 TFLOPs
<b>Peak INT8 Tensor</b>	169,6 TOPs	238/476 TOPs

	GeForce RTX 2080	GeForce RTX 3080
<b>Memory size</b>	8GB	10GB
<b>Memory type</b>	GDDR6	GDDR6X
<b>Memory interface</b>	256 bits	320 bits
<b>Memory data rate</b>	14 Gbps	19 Gbps
<b>Memory bandwidth</b>	448 GB/s	760 GB/s
<b>ROPs</b>	64	96
<b>Pixel fill rate</b>	115,2 Gpix/s	164,2 Gpix/s
<b>TUs</b>	184	272
<b>Texel fill rate</b>	331,2 Gtex/s	465 Gtex/s
<b>L1 / Shared Memory</b>	4416 KB	8704 KB
<b>L2</b>	4096 KB	5120 KB
<b>Register file size</b>	11776 KB	17408 KB
<b>TGP</b>	225 W	320 W
<b>Transistor count</b>	$13,6 \times 10^9$	$28,3 \times 10^9$
<b>Die size</b>	545 mm <sup>2</sup>	628,4 mm <sup>2</sup>
<b>Manufacturing</b>	12nm	8nm

# Bibliografía & Documentación

- [www.nvidia.com](http://www.nvidia.com)
- Erik Lindholm, John Nickolls, Stuart Oberman and John Montrym  
“NVIDIA TESLA: A Unified Graphics and Computer Architecture”  
IEEE Micro 2008
- NVIDIA’s Next Generation CUDA Compute Architecture: Fermi  
Whitepaper NVIDIA, 2009
- NVIDIA’s Next Generation CUDA Compute Architecture: Kepler GK110/210  
Whitepaper NVIDIA, 2014
- NVIDIA Turing GPU Architecture  
Whitepaper NVIDIA, 2018
- NVIDIA Ampere GA102 GPU Architecture,  
NVIDIA, 2021



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Departament d'Arquitectura de Computadors

# Tarjetas Gráficas y Aceleradores

## Nvidia Roadmap

Agustín Fernández

Departament d'Arquitectura de Computadors

Facultat d'Informàtica de Barcelona

Universitat Politècnica de Catalunya

