ARQUITECTURA DE LAS GPUS Y TIPOS DE MEMORIA

Francisco J. Hernández López

fcoj23@cimat.mx

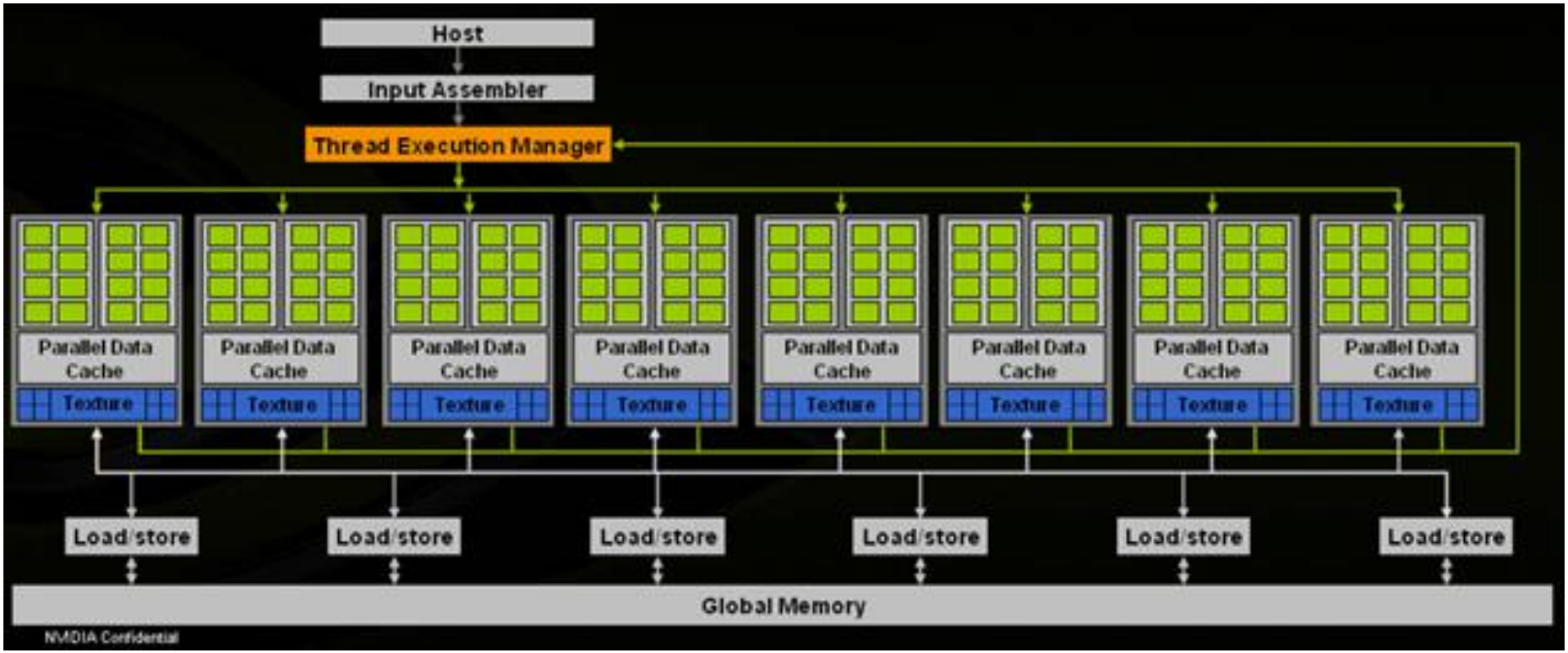
ARQUITECTURA DE LA GPU

- La GPU es un arreglo de unidades de ejecución llamadas *Streaming Multiprocessors* (SM)
- Cada SM consiste de un arreglo de 8 o más *Streaming Processors* (SP) o CUDA Cores
- Se considera que es un **Coprocesador o Acelerador** que debe operar con una CPU
- Los SPs ejecutan el trabajo en un conjunto de hasta 32 unidades (1 warp)
- La ventaja de programar en CUDA, es que a pesar de toda la variabilidad en el hardware, los programas escritos en GPUs viejitas pueden ejecutarse en GPUs nuevas.

Cook, S. (2012). *CUDA programming: a developer's guide to parallel computing with GPUs*. Newnes.

GPU TESLA GEFORCE 8800 GTX CON 16 SM X 8SP

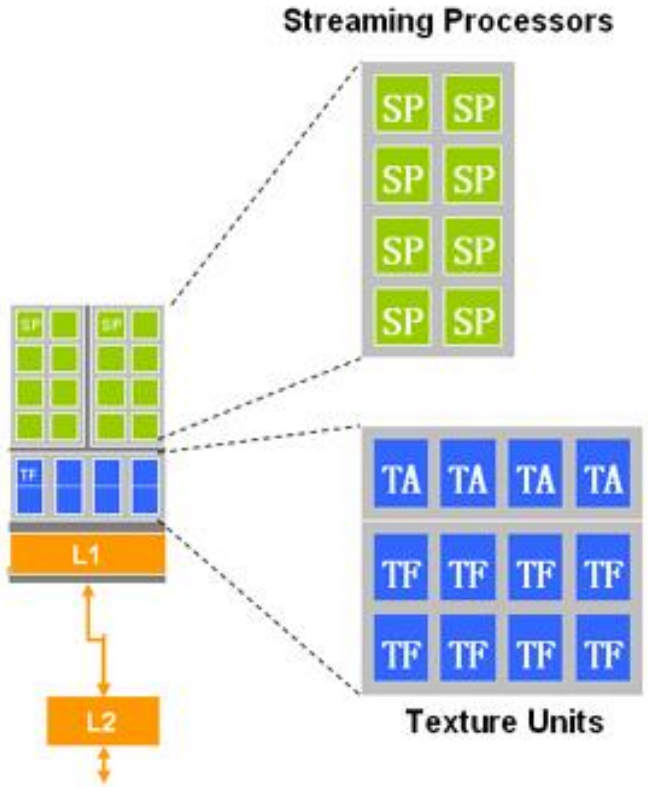
- Hasta 128 CUDA Cores de 1.5 GHz
- Procesadores de punto flotante de precisión simple

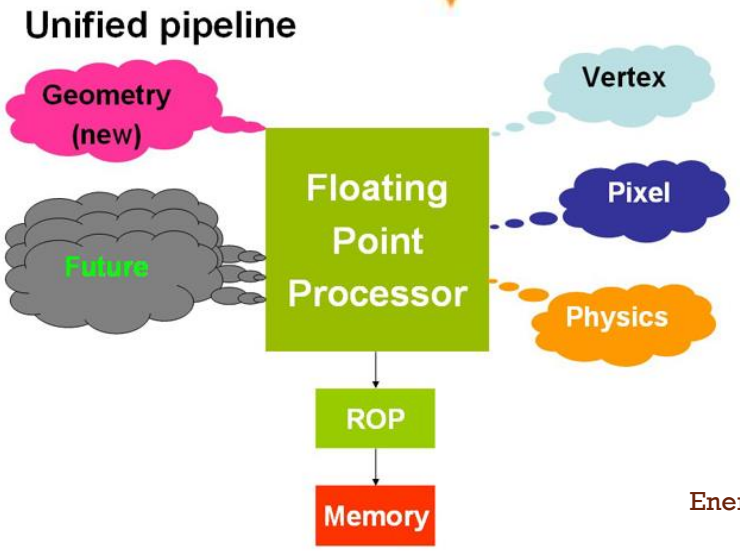


GeForce_8800_GPU_Architecture_Technical_Brief.pdf

SM DEL GPU GEFORCE 8800

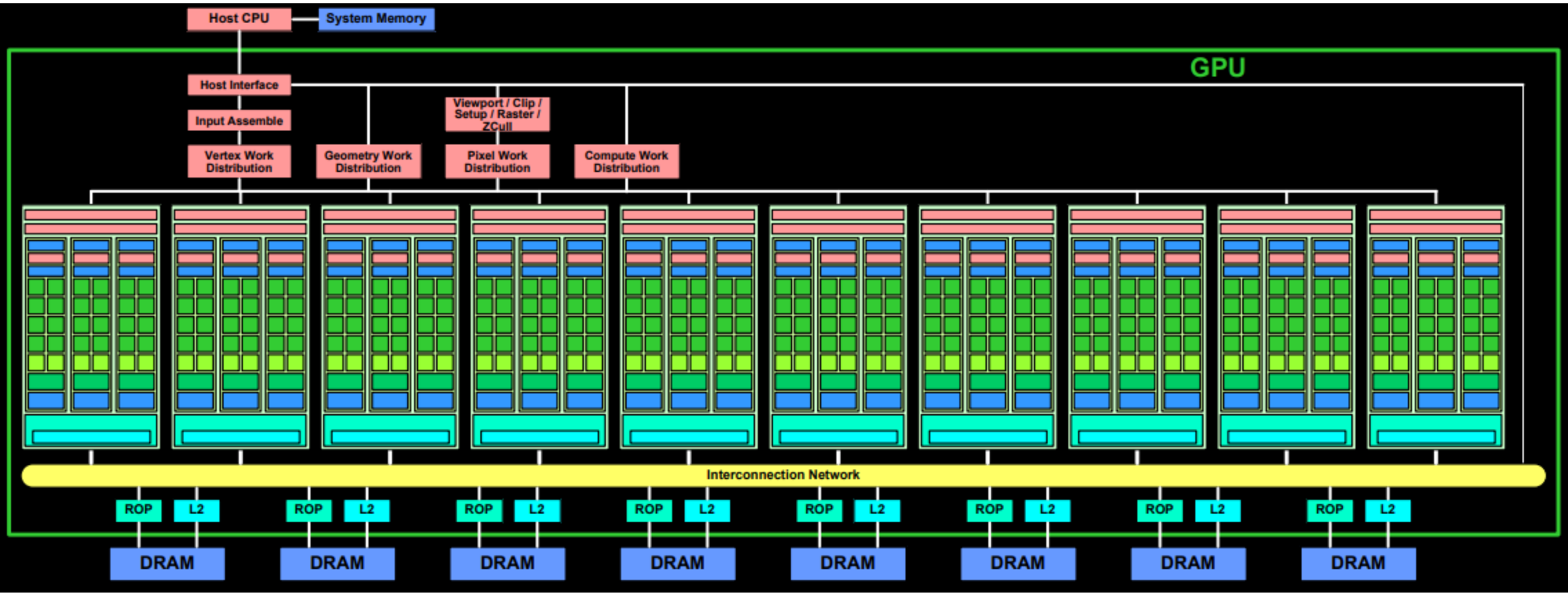
- Comparten un cierto número de unidades de:
  - Filtrado de Texturas (TF)
  - Direccionamiento de Texturas (TA)
  - Caché
- Utilizan el estándar IEEE 754 de precisión en punto flotante de 32-bits





GPU TESLA GTX 280 CON 30 SM X 8 SP

- Tiene 10 TPC (Thread Processing Cluster)
- Esta GPU soporta hasta 32 warps/SM vs 24 warps/SM de la GF 8800

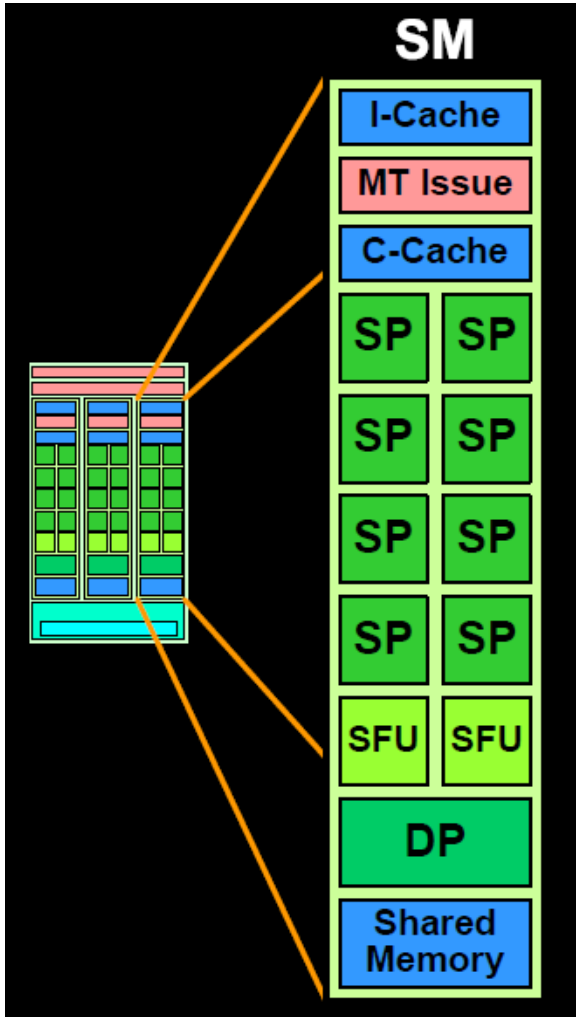


http://download.nvidia.com/developer/cuda/seminar/TDCI_Arch.pdf

1.4 billones de transistores

SM DEL GPU GTX 280

- Ocho SPs
  - Estándar IEEE 754 de precisión en punto flotante de 32-bits
  - Enteros de 32 y 64 bits
- Dos unidades de funciones especiales (SFU)
  - sin, cos, log, exp
- Una unidad de precisión double
- 16KB de memoria compartida



http://download.nvidia.com/developer/cuda/seminar/TDCI_Arch.pdf

GPU FERMI GF100 CON 16 SM X 32 SP

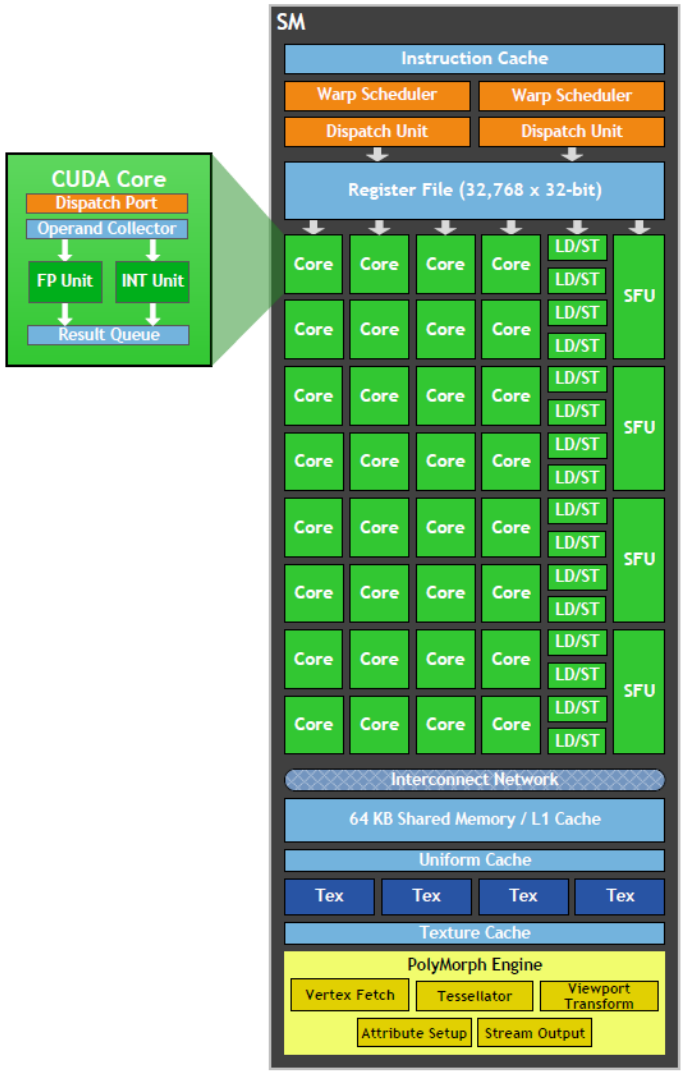
- Tiene 4 GPCs (Graphics Processing Cluster), cada GPC tiene 4 SMs y cada SM tiene 32 SPs.

Wittenbrink, C. M., Kilgariff, E., & Prabhu, A. (2010, August). Fermi gf100 graphics processing unit (gpu). In *Hot Chips* (Vol. 22).

3 billones de transistores

SM DEL GPU FERMI GF100

- 32 SPs, cada uno ejecuta una instrucción (float o int) por ciclo de reloj
- 128KB del archivo de registros, con 32768 registros, cada uno de 32-bits
- 4 SFUs
- 16 unidades de carga y almacenamiento (LD/ST)
- 64KB de caché
- Dos warps se ejecutan de forma concurrente



Wittenbrink, C. M., Kilgariff, E., & Prabhu, A. (2010, August). Fermi gf100 graphics processing unit (gpu). In *Hot Chips* (Vol. 22).

GPU KEPLER GK110 CON 15 SMX X 192 SP

- Los SM ahora se llaman SMX, ya que presentan una estructura interna nueva, agregando nuevas capacidades a la GPU

Kepler GK110 Whitepaper. The Fastest, Most Efficient HPC Architecture Ever Built.

7.1 billones de transistores

SMX DEL GPU KEPLER GK110

- 64 unidades de precision Double (DP Unit)
- 32 unidades de Carga/Almacenamiento (LD/ST)
- 32 unidades de Funciones Especiales (SFU)
- 4 planificadores warp

Kepler GK110 Whitepaper. The Fastest, Most Efficient HPC Architecture Ever Built.

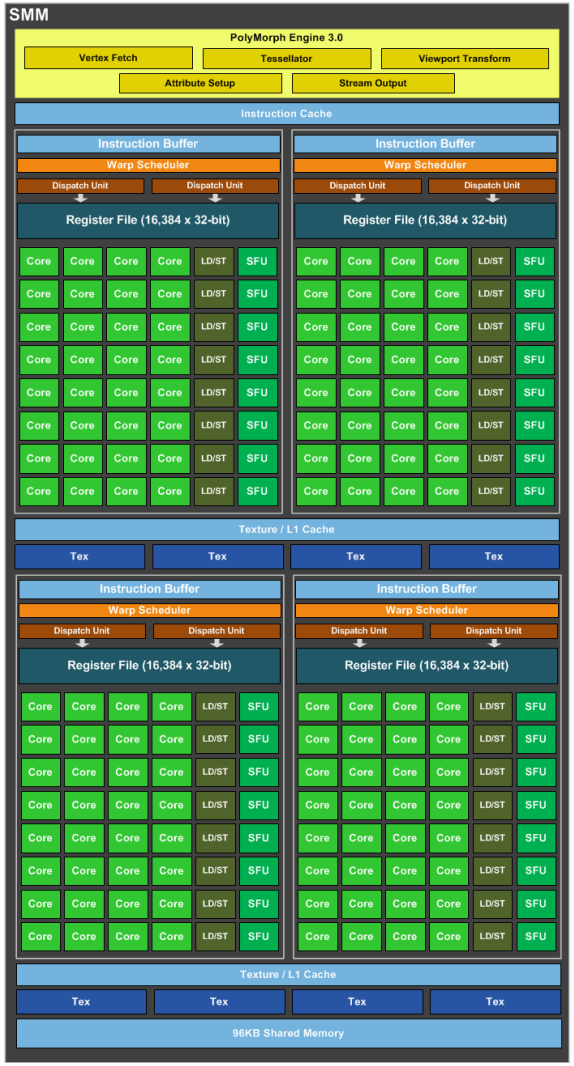
GPU MAXWELL GM204 CON 16 SMM X 128 SP

NVIDIA GeForce GTX 980, Featuring Maxwell Whitepaper

5.2 billones de transistores

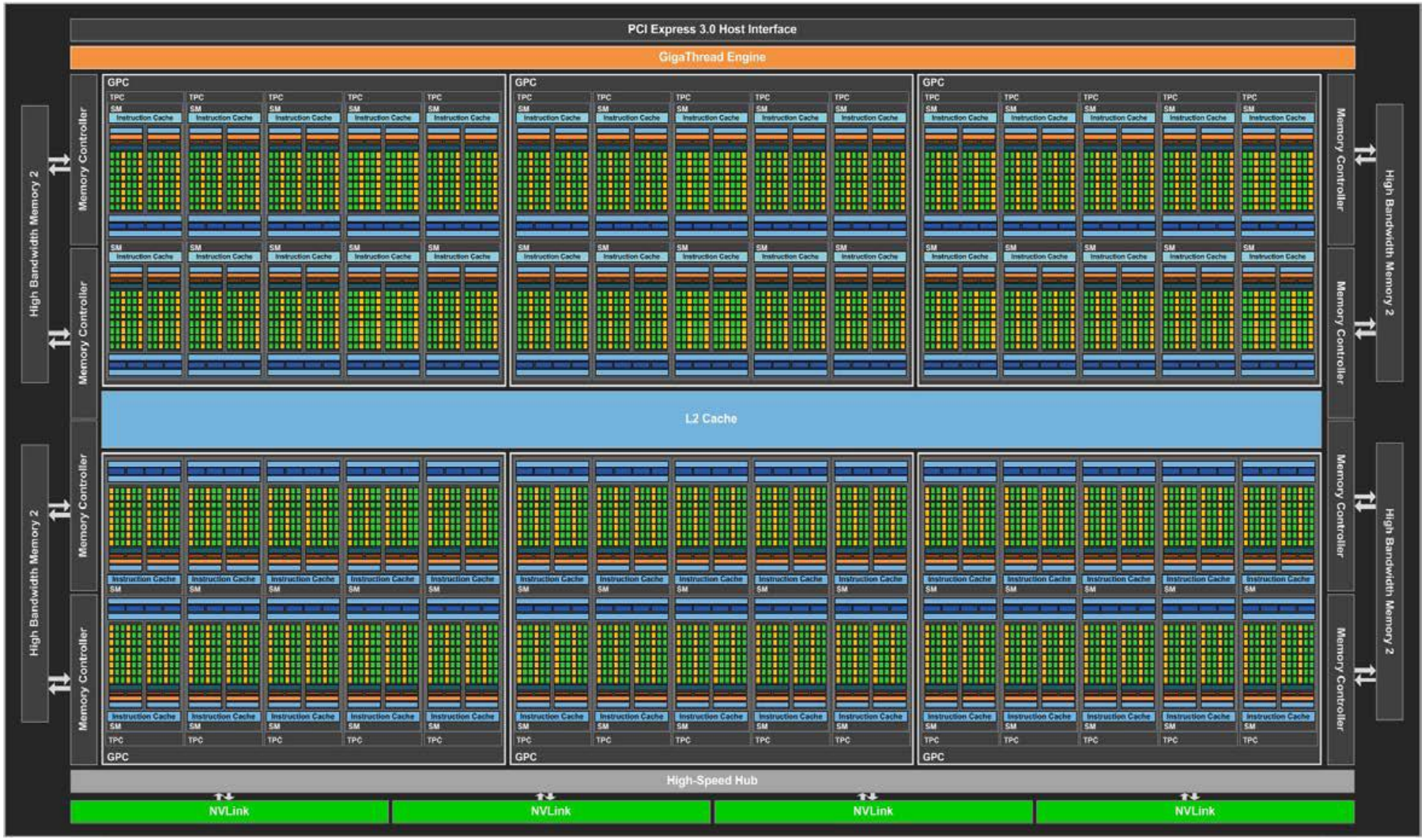
SMM DEL GPU MAXWELL GM204

- Se incrementó la caché L2 a 2MB
- 8 unidades de Textura (cuadritos azules)
- 32 unidades de Funciones Especiales (SFU)
- 32 de Carga y Almacenamiento (LD/ST)
- 4 planificadores warp
- 96KB de memoria compartida
- La caché L1 ahora comparte espacio con la memoria cache de Textura.



NVIDIA GeForce GTX 980, Featuring Maxwell Whitepaper

GPU PASCAL GP100 CON 60 SM X 64 SP



NVIDIA Tesla P100, Featuring Pascal GP100, the World's Fastest GPU Whitepaper

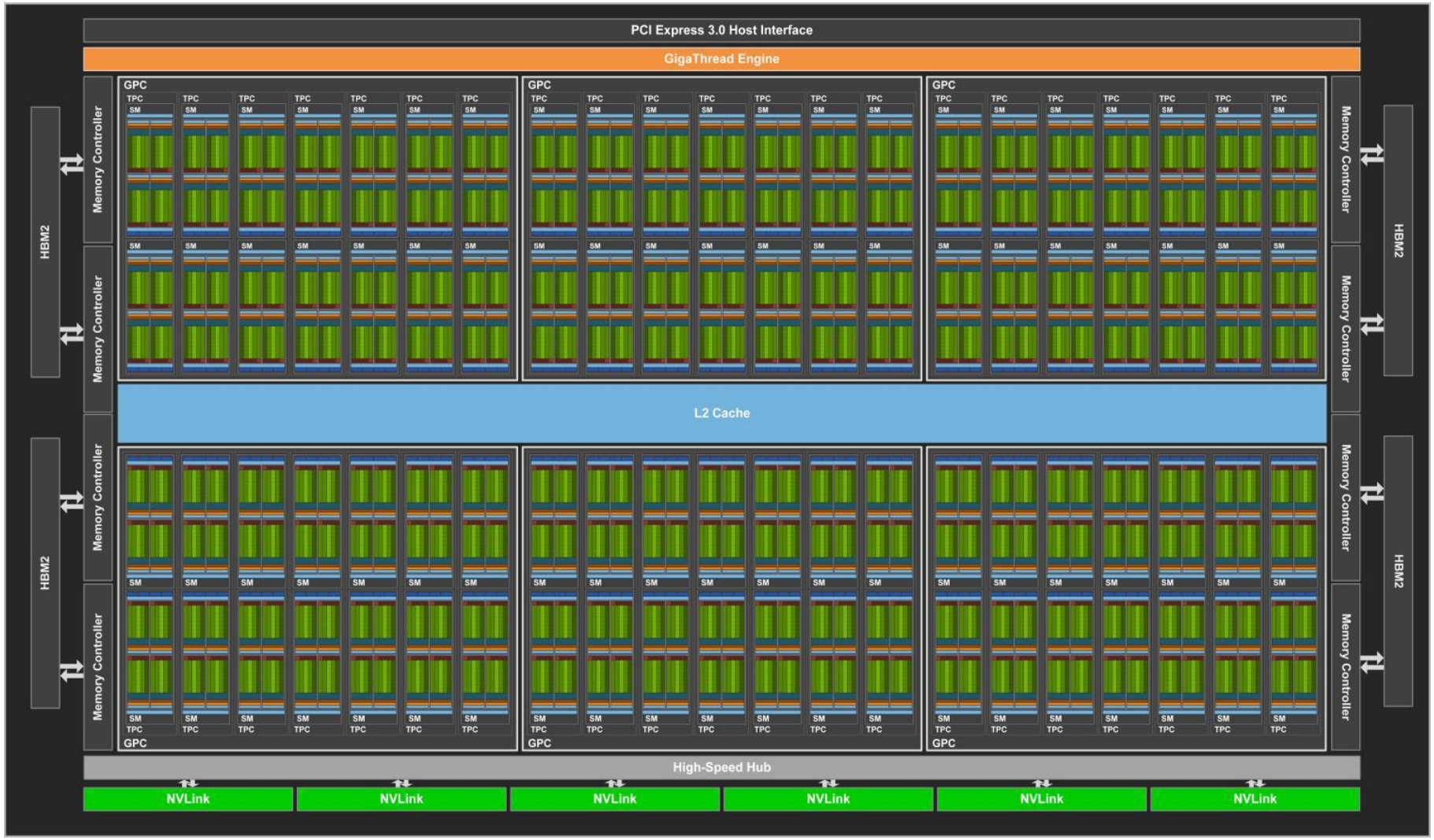
15.3 billones de transistores

SM DEL GPU PASCAL GP100

- Se incrementó la caché L2 a 4MB
- 4 unidades de textura
- 16 SFU
- 16 LD/ST
- 2x32 CUDA Cores de precision simple
- 2x16 CUDA Cores de precision double
- 2 planificadores warp
- 64KB de Mem. Comp.

NVIDIA Tesla P100, Featuring Pascal GP100, the World's Fastest GPU Whitepaper

GPU VOLTA GV100 CON 84 SM X 64 SP



NVIDIA TESLA V100 GPU ARCHITECTURE THE WORLD'S MOST ADVANCED DATA CENTER GPU Whitepaper

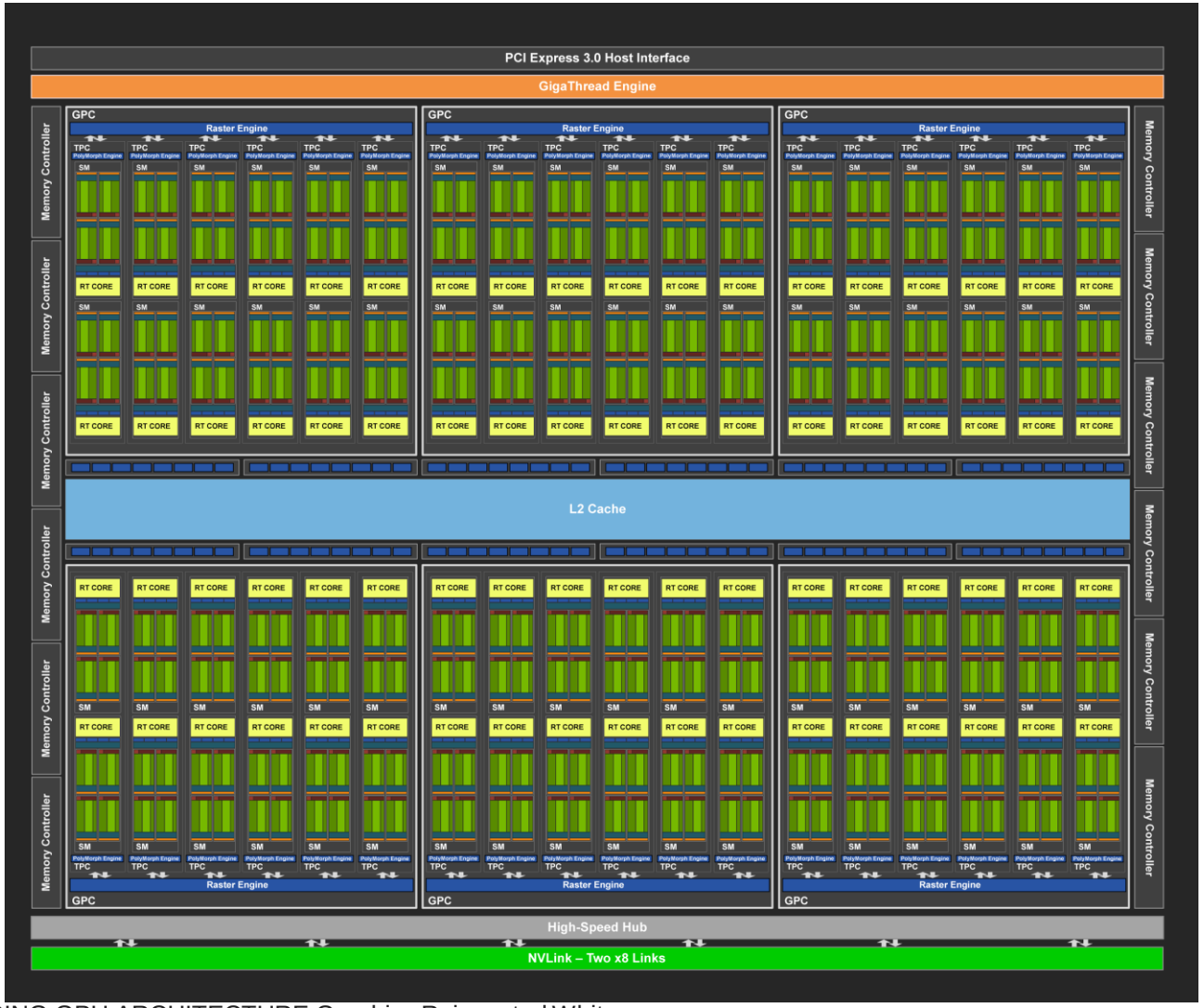
21.1 billones de transistores

SM DEL GPU VOLTA GV100

- Se incrementó la caché L2 a 6MB
- 4 unidades de textura
- 16 SFU, 32 LD/ST
- 64 FP32 cores
- 64 INT32 cores
- 32 FP64 cores
- 8 Tensor cores (672 en toda la GPU)
- Archivo de registros de 64 KB
- Caché L0 (mayor eficiencia que los buffers de instrucciones)
- 4 planificadores warp
- 128KB de mem. caché L1 y textura

NVIDIA TESLA V100 GPU ARCHITECTURE THE WORLD'S MOST ADVANCED DATA CENTER GPU Whitepaper

GPU TURING TU102 CON 72 SM X 64 SP

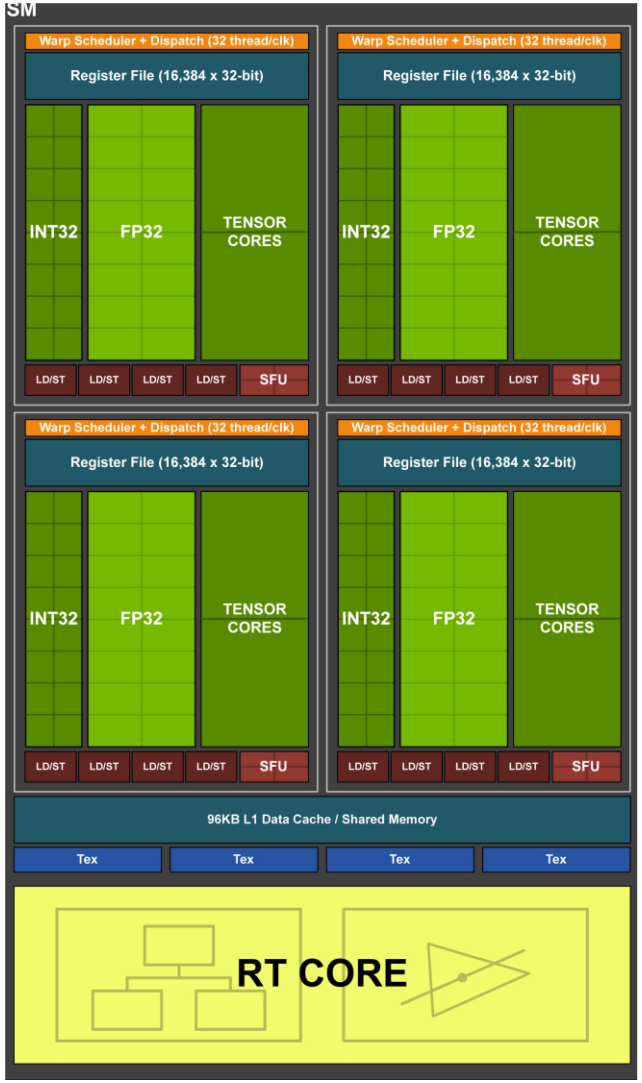


NVIDIA TURING GPU ARCHITECTURE Graphics Reinvented Whitepaper

18.6 billones de transistores

SM DEL GPU TURING TU102

- Caché L2 de 6MB
- 4 unidades de textura
- 16 SFU, 32 LD/ST
- 64 FP32 cores
- 64 INT32 cores
- 2 FP64 cores
- 8 Tensor cores
- 1 RT core (72 RT cores en toda la GPU)
- Archivo de registros de 64 KB
- Caché L0
- 4 planificadores warp
- 96KB de mem. caché L1 y textura



NVIDIA TURING GPU ARCHITECTURE Graphics Reinvented Whitepaper

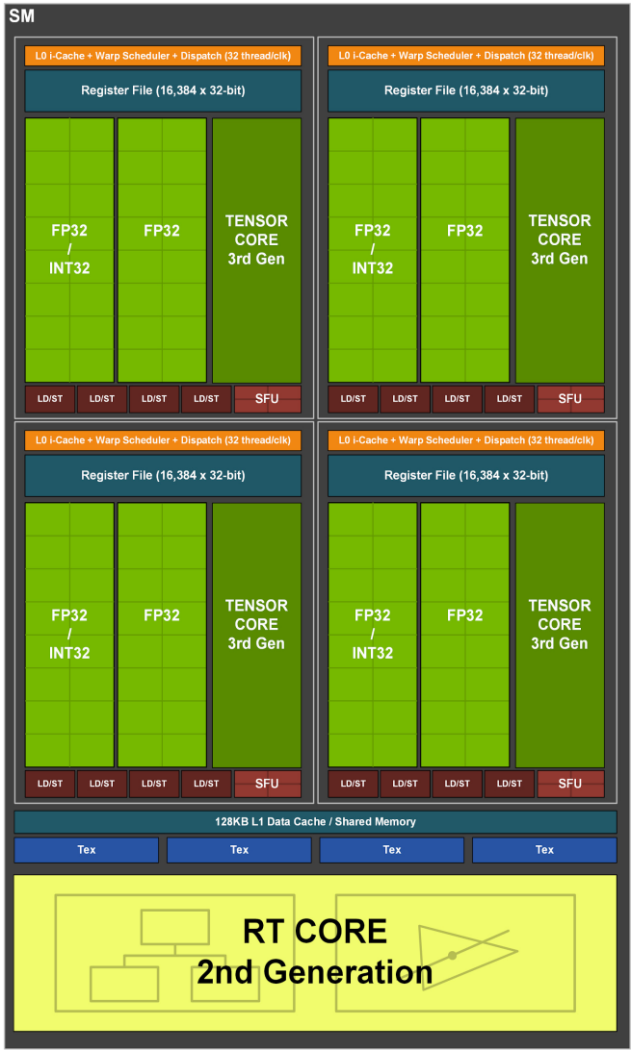
GPU AMPERE GA102 CON 84 SM X 128 SP

NVIDIA AMPERE GA102 GPU ARCHITECTURE THE ULTIMATE PLAY Whitepaper

28.3 billones de transistores

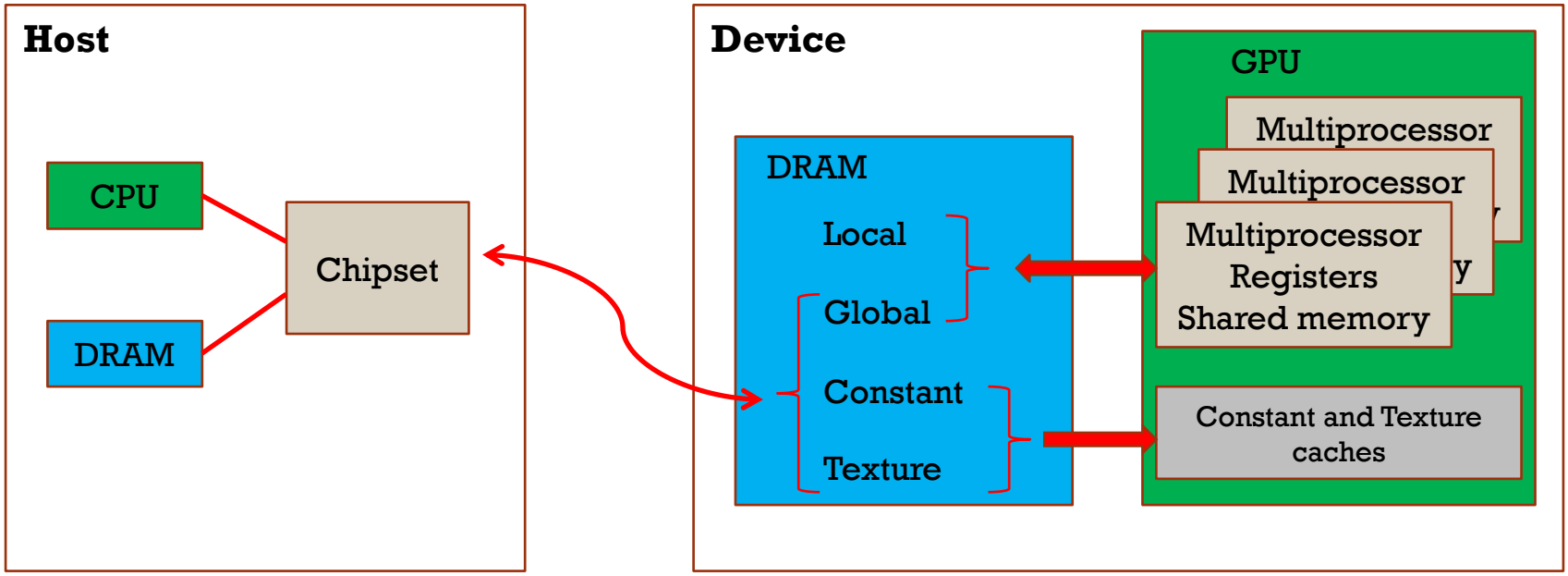
SM DEL GPU AMPERE GA102

- caché L2 de 6MB
- 4 unidades de Textura
- 16 SFU, 32 LD/ST
- 128 FP32 cores
- 64 INT32 cores
- 2 FP64 cores
- 4 Tensor cores (3ra. generación)
- 1 RT core
- Archivo de registros de 64 KB
- Caché L0
- 4 planificadores warp
- 128KB de mem. caché L1 y textura



NVIDIA AMPERE GA102 GPU ARCHITECTURE THE ULTIMATE PLAY Whitepaper

MODELO DE LA MEMORIA EN CUDA

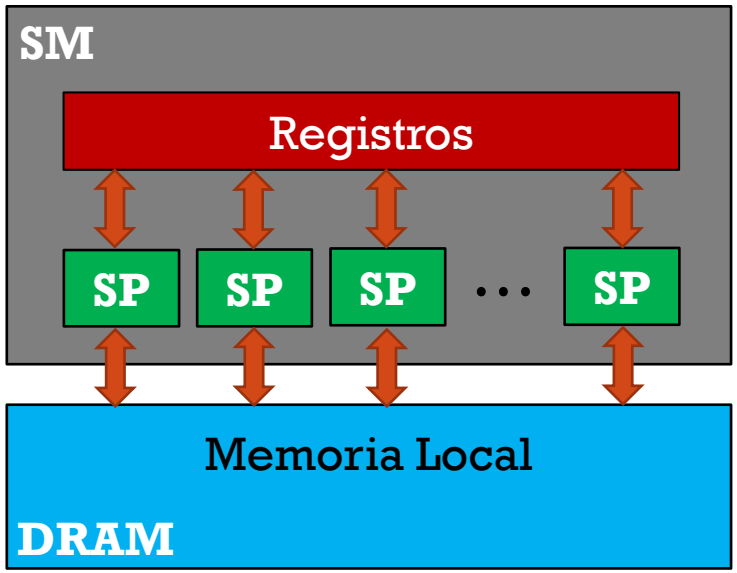


Type	Location	Access	Scope	Lifetime
Register	On-chip	R/W	Thread	Thread
Local	Off-chip	R/W	Thread	Thread
Shared	On-chip	R/W	Block	Block
Global	Off-chip	R/W	Grid	Controlled by host
Constant	Off-chip	R	Grid	Controlled by host
Texture	Off-chip	R	Grid	Controlled by host

Barlas, G. (2014). *Multicore and GPU Programming: An integrated approach*. Elsevier.

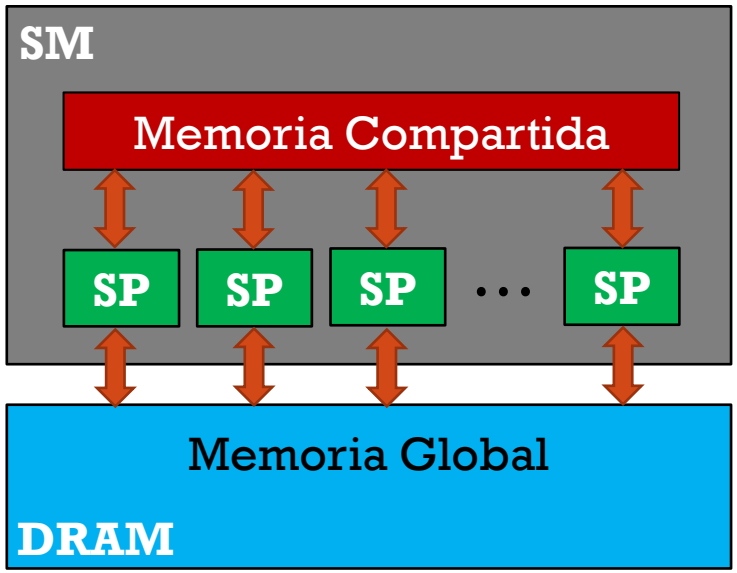
MEMORIA LOCAL Y REGISTROS

- Usados para mantener las variables automáticas
- La capacidad de la GPU determina el número máximo de registros que se pueden utilizar por hilo
- Si se excede este número máximo, las variables locales se almacenan en una pila en la memoria local (que no está en el chip), por lo que el rendimiento puede disminuir
- Tamaño de los registros:
  - 16KB en GPUs de capacidad 1.x
  - 32KB en GPUs de capacidad 2.0
  - 64KB en GPUs de capacidad >= 3.x



MEMORIA COMPARTIDA

- Se encuentra en el Chip
- Mantiene los datos que son usados con mucha frecuencia
- Se puede usar para intercambiar datos entre los SPs del mismo bloque de hilos asignado por el SM
- Tamaños: 16KB, 32KB, 48KB y 96KB (Maxwell)



- Podemos reservar memoria compartida de dos formas:
  - Estática: __**shared**__ type var[tam]
  - Dinámica: **extern** __**shared**__ type var[ ], usando el tercer parámetro de la llamada a la función Kernel

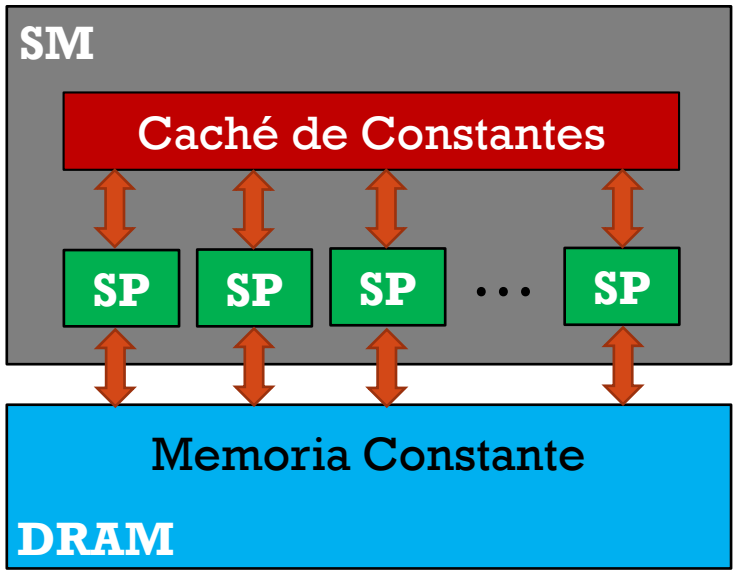
```
NameFunc <<< Dg, Db, Ns >>> (parametro);
```

MEMORIA CONSTANTE

- Es solamente de lectura, de la memoria global, va a una parte de la memoria caché destinada para Constantes
- Tamaño: 64KB en total y 8KB por SM en la memoria caché
- Declarar constantes:

  __constant__ type var1;

  __constant__ type var2[tam];

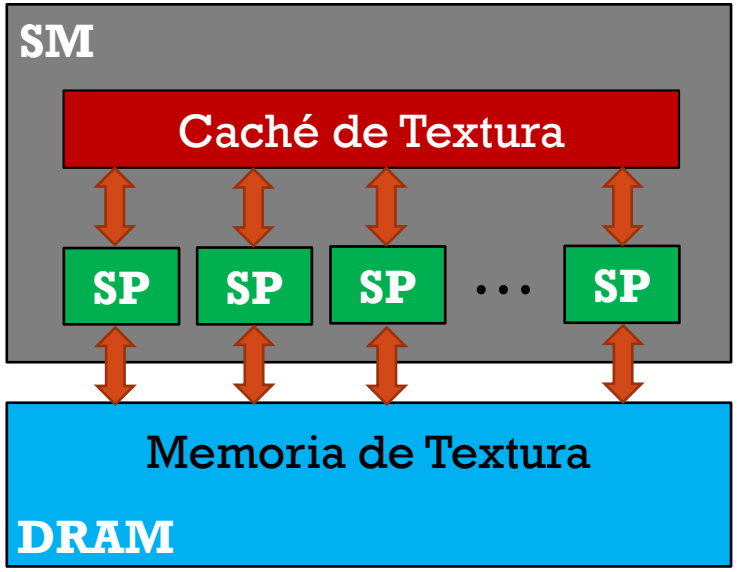


- Pasar variables del host a la memoria constante de la GPU

```
cudaMemcpyToSymbol(const char  *symbol, const void *src, size_t count, size_t offset, enum cudaMemcpyKind kind)
```

MEMORIA DE TEXTURA

- Es solamente de lectura, de la memoria global, va a una parte de la memoria caché destinada para Textura
- Puede realizar interpolaciones de punto flotante como parte del proceso de lectura a nivel hardware
- 16 unidades de textura y 12KB por SM en Kepler




```
Device 0: "GeForce GTX 980M"
  CUDA Driver Version / Runtime Version          10.1 / 8.0
  CUDA Capability Major/Minor version number:    5.2
  Total amount of global memory:                 4096 MBytes (4294967296 bytes)
  (12) Multiprocessors, (128) CUDA Cores/MP:     1536 CUDA Cores
  GPU Max Clock rate:                            1127 MHz (1.13 GHz)
  Memory Clock rate:                             2505 Mhz
  Memory Bus Width:                              256-bit
  L2 Cache Size:                                 2097152 bytes
  Maximum Texture Dimension Size (x,y,z)         1D=(65536), 2D=(65536, 65536),
3D=(4096, 4096, 4096)
  Maximum Layered 1D Texture Size, (num) layers  1D=(16384), 2048 layers
  Maximum Layered 2D Texture Size, (num) layers  2D=(16384, 16384), 2048 layers
```

GRACIAS POR SU ATENCIÓN

Francisco J. Hernández-López

fcoj23@cimat.mx

WebPage:

www.cimat.mx/~fcoj23