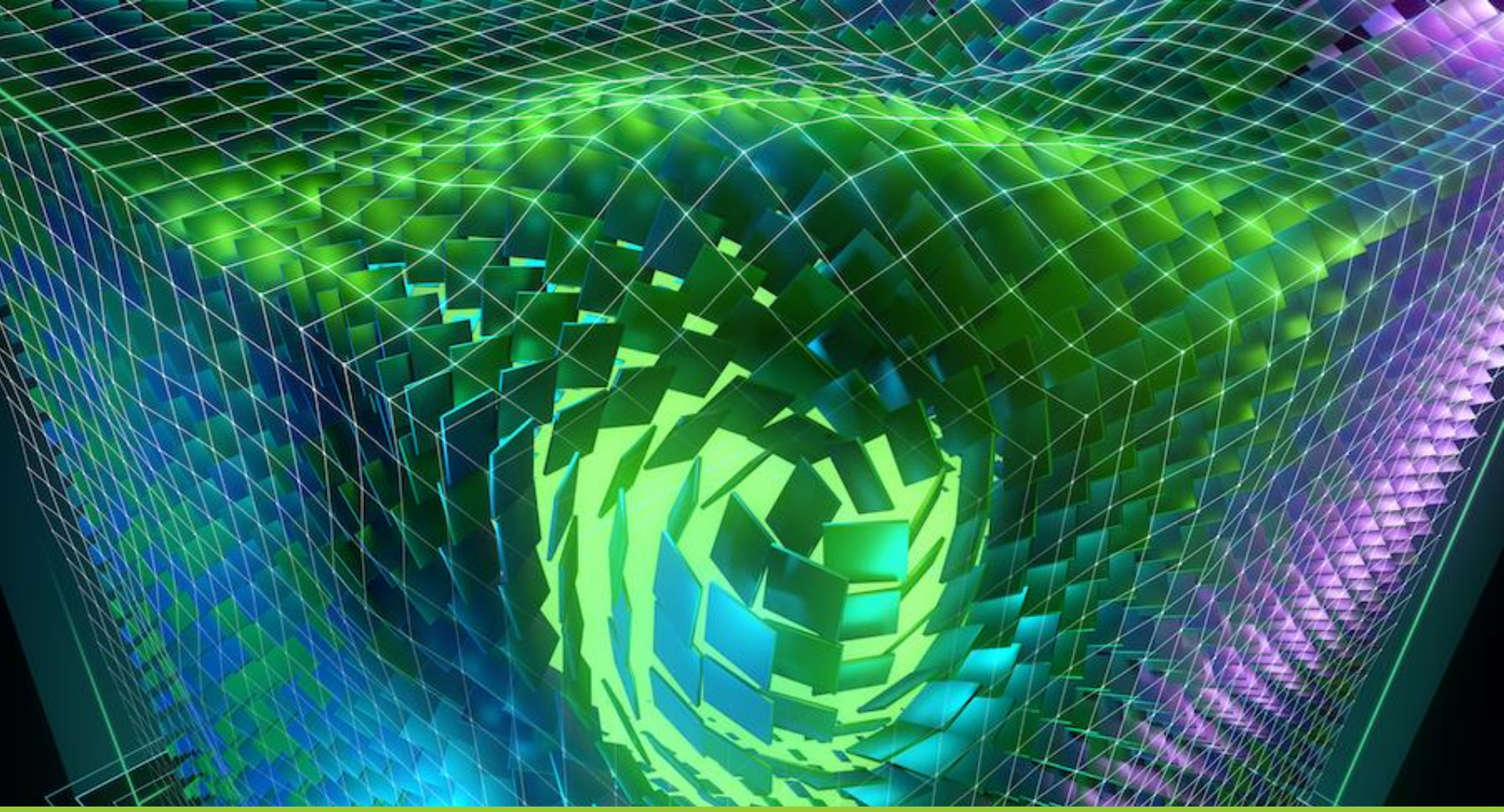


INF351 – Computación de Alto Desempeño

Microarquitecturas y Características

PROF. ÁLVARO SALINAS

Microarquitecturas

Las microarquitecturas corresponden a las distintas organizaciones de un dispositivo respecto a diseños y tecnologías.

En las GPU de NVIDIA capaces de utilizar CUDA se distinguen las siguientes según su orden de lanzamiento:

- Tesla
- Fermi
- Kepler
- Maxwell
- Pascal
- Volta
- Turing

Tesla

Lanzamiento: 2006

Compute capability: 1.x

CUDA cores per SM: 8

Principales cambios respecto a su predecesor:

- Soporte para CUDA
- Unified shader model
- Streaming multiprocessors (SM)

Fermi

Lanzamiento: 2009

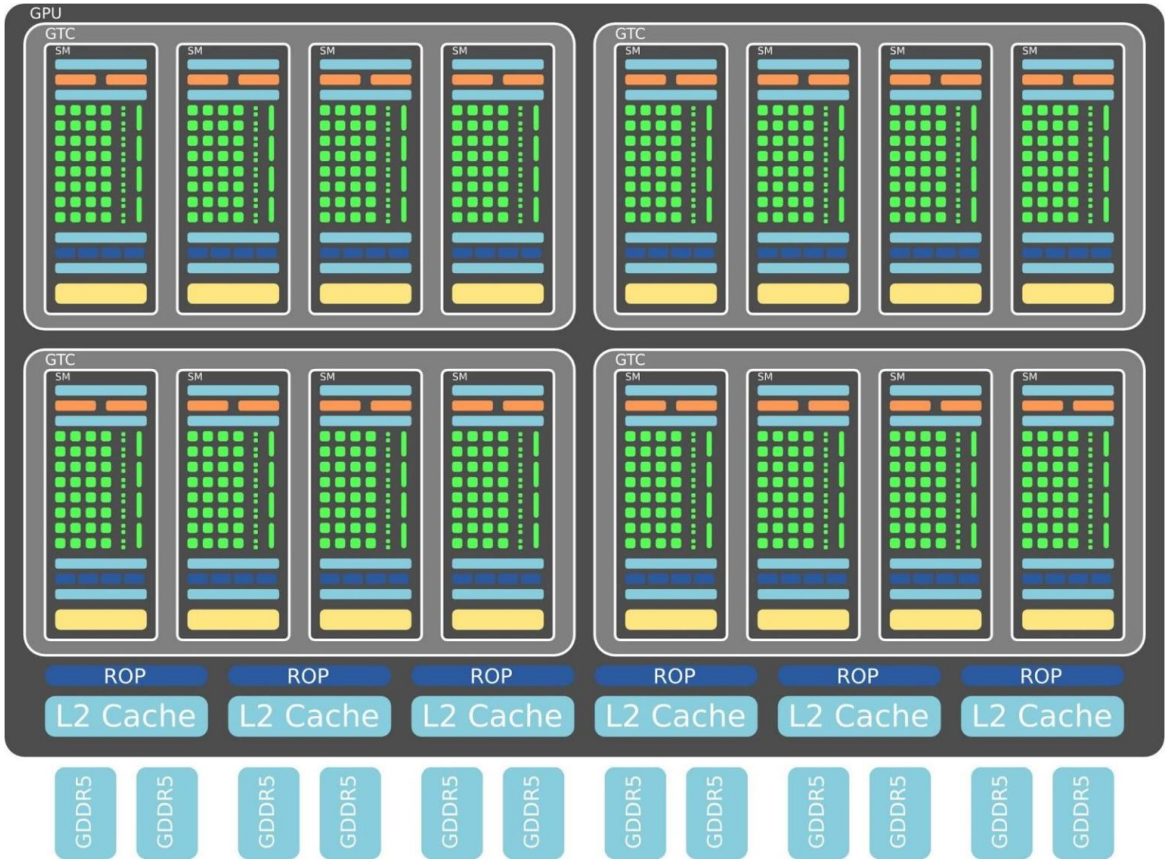
Compute capability: 2.x

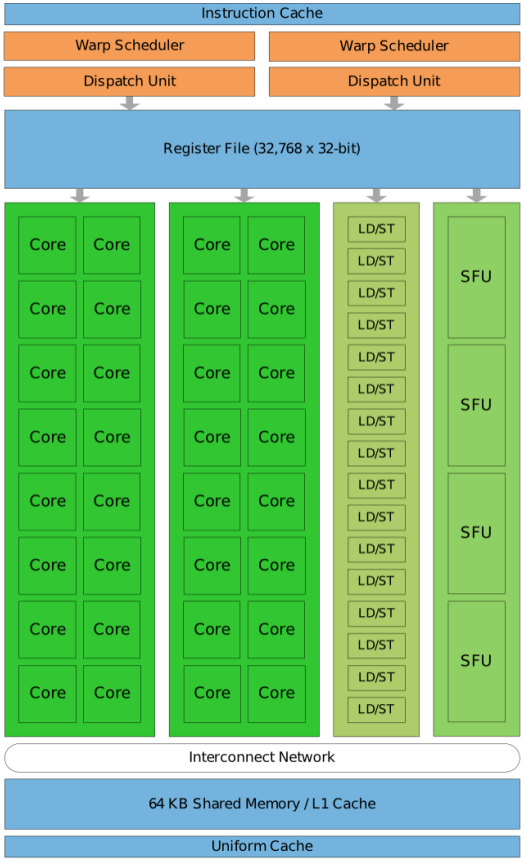
CUDA cores per SM: 32

Principales cambios respecto a su predecesor:

- Más CUDA cores
- GigaThread global scheduler
- Soporte para FP64

Fermi


GPU
GTC
SM
ROP
L2 Cache
GDDR5


Instruction Cache
Warp Scheduler
Dispatch Unit
Register File (32,768 x 32-bit)
Core
LD/ST
SFU
Interconnect Network
64 KB Shared Memory / L1 Cache
Uniform Cache

Kepler

Lanzamiento: 2012

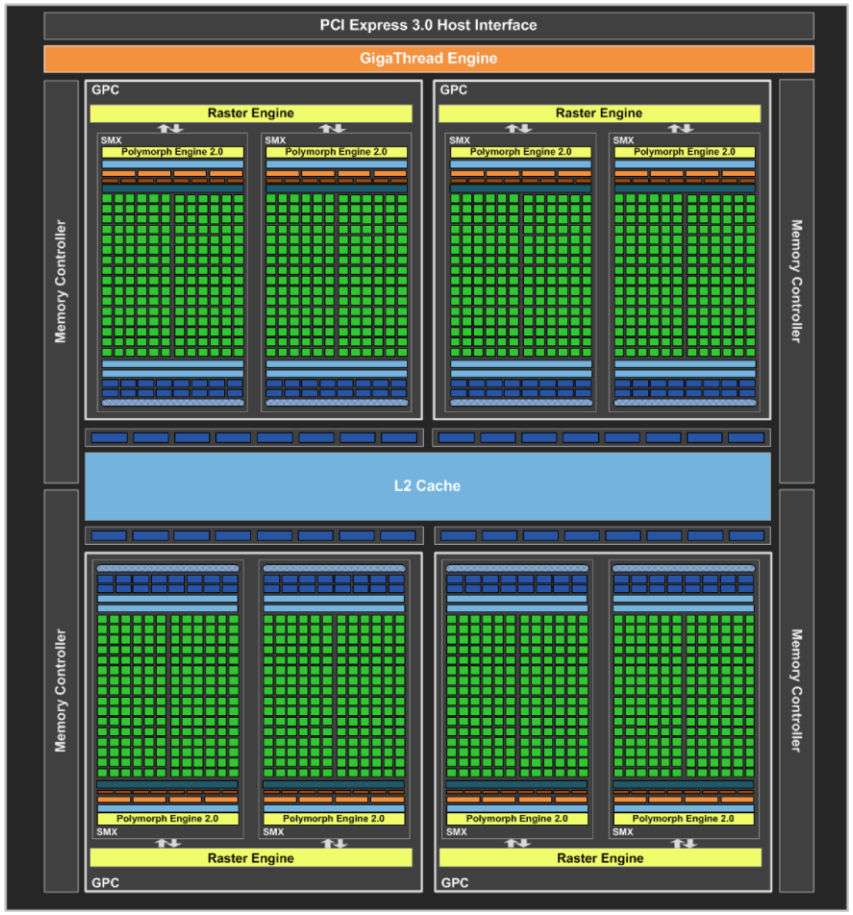
Compute capability: 3.x

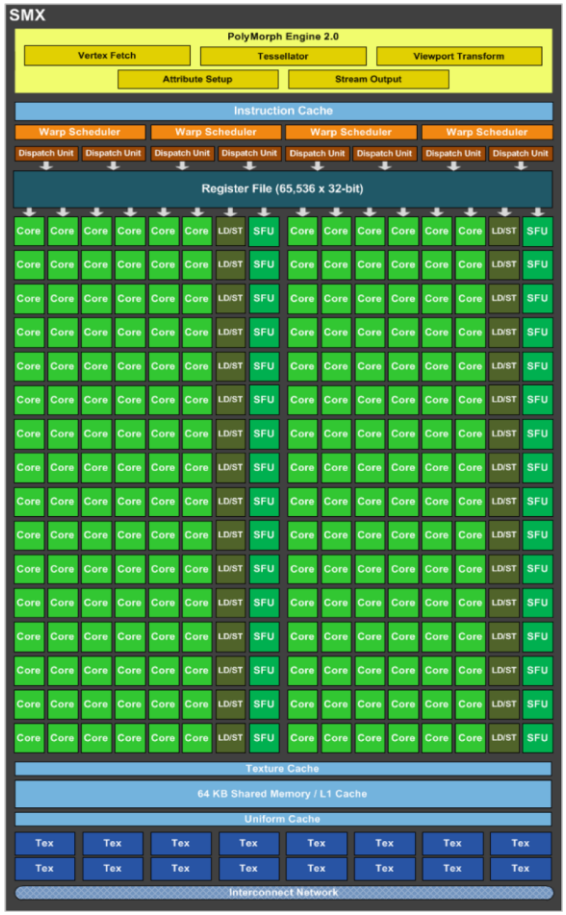
CUDA cores per SM (SMX): 192

Principales cambios respecto a su predecesor:

- Más CUDA cores
- Warps de 32 subprocesos
- SMX son más eficientes en power consumption

Kepler


PCI Express 3.0 Host Interface
GigaThread Engine
GPC
Raster Engine
SMX
Polymorph Engine 2.0
Memory Controller
L2 Cache


SMX
PolyMorph Engine 2.0
Vertex Fetch
Tessellator
Viewport Transform
Attribute Setup
Stream Output
Instruction Cache
Warp Scheduler
Dispatch Unit
Register File (65,536 x 32-bit)
Core
LD/ST
SFU
Texture Cache
64 KB Shared Memory / L1 Cache
Uniform Cache
Tex
Interconnect Network

Kepler


PCI Express 3.0 Host Interface
GigaThread Engine
Memory Controller
Memory Controller
Memory Controller
SMX
L2 Cache
SMX
Instruction Cache
Warp Scheduler
Dispatch
Register File (65,536 x 32-bit)
Core
DP Unit
LD/ST
SFU
Interconnect Network
64 KB Shared Memory / L1 Cache
48 KB Read-Only Data Cache
Tex

Maxwell

Lanzamiento: 2014

Compute capability: 5.x

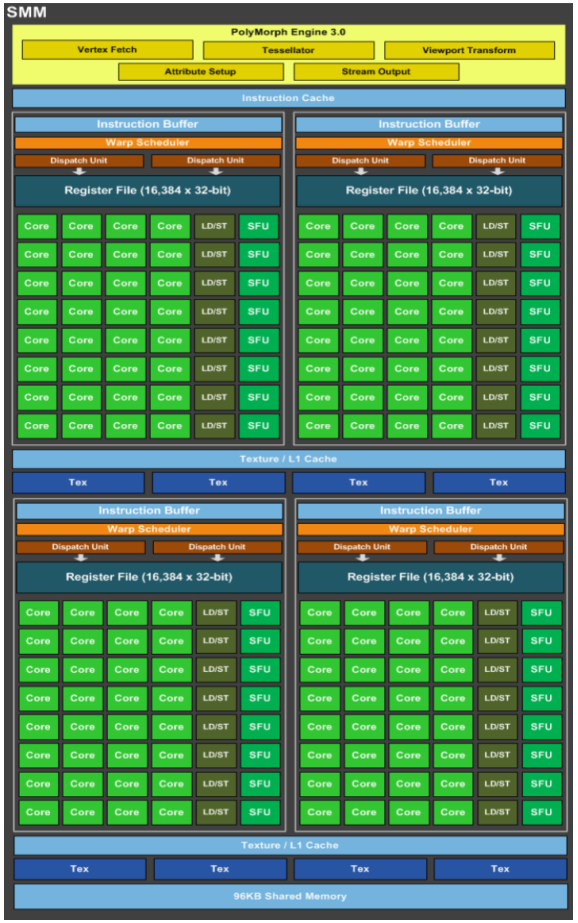
CUDA cores per SM (SMM): 128

Principales cambios respecto a su predecesor:

- Gran aumento de memoria L2 cache
- SMM sacrifican un poco de performance por mucha eficiencia

Maxwell


PCI Express 3.0 Host Interface
GigaThread Engine
GPC
Raster Engine
SMM
Memory Controller
L2 Cache


SMM
PolyMorph Engine 3.0
Vertex Fetch
Tessellator
Viewport Transform
Attribute Setup
Stream Output
Instruction Cache
Instruction Buffer
Warp Scheduler
Dispatch Unit
Register File (16,384 x 32-bit)
Core
LD/ST
SFU
Texture / L1 Cache
Tex
96KB Shared Memory

Pascal

Lanzamiento: 2016

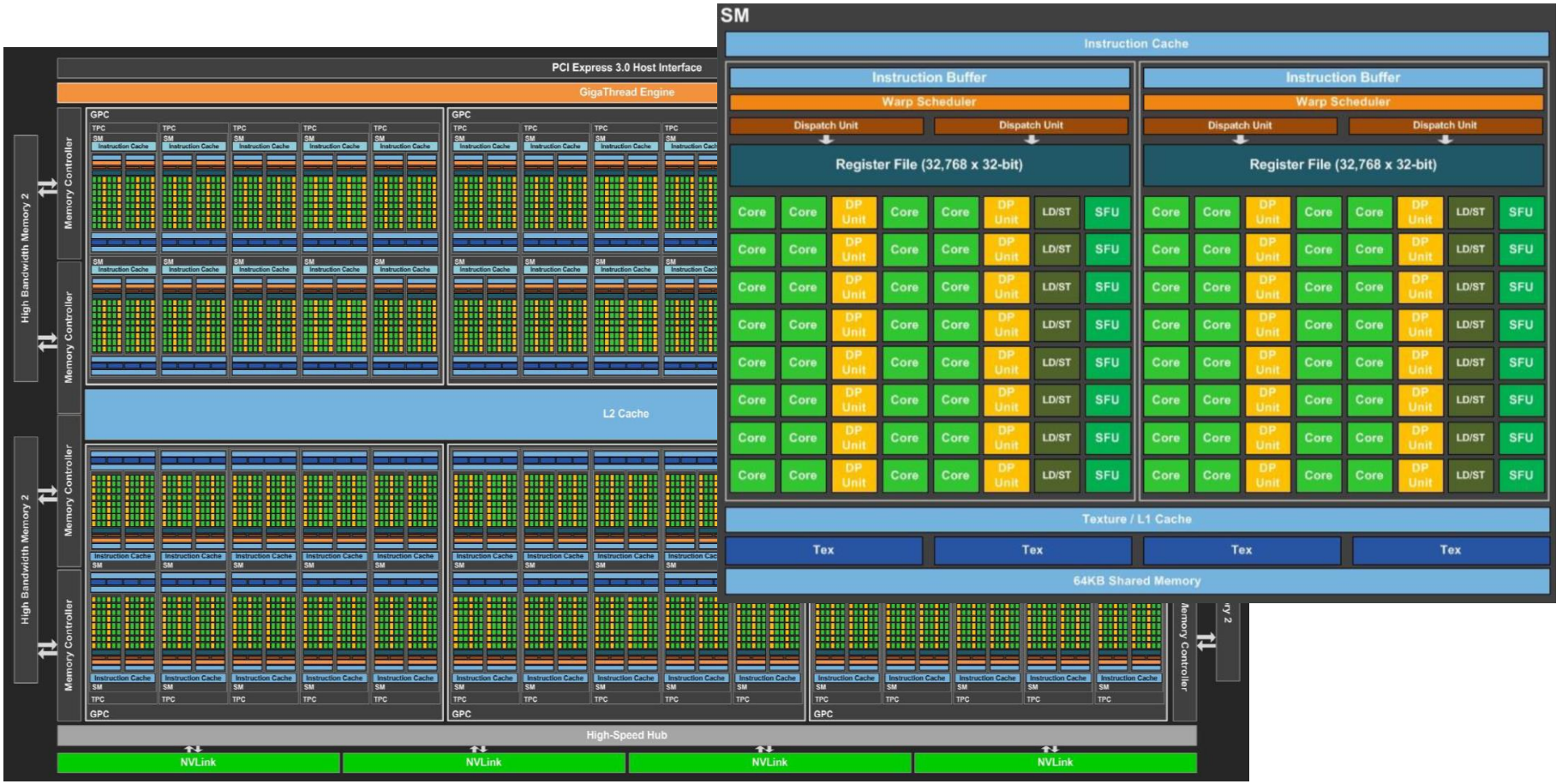
Compute capability: 6.x

CUDA cores per SM: 64

Principales cambios respecto a su predecesor:

- High Bandwith Memory 2
- Operaciones de FP16 se realizan el doble de rápido que operaciones de FP32, y éstas a su vez, el doble de rápido que las de FP64
- Más registros por CUDA core y más Shared Memory
- Dynamic scheduler

Pascal


SM
Instruction Cache
Instruction Buffer
Warp Scheduler
Dispatch Unit
Register File (32,768 x 32-bit)
Core
DP Unit
LD/ST
SFU
Texture / L1 Cache
Tex
64KB Shared Memory
PCI Express 3.0 Host Interface
GigaThread Engine
GPC
TPC
SM
Instruction Cache
Memory Controller
High Bandwidth Memory 2
L2 Cache
High-Speed Hub
NVLink

Volta

Lanzamiento: 2017

Compute capability: 7.x

CUDA cores per SM: 64

Principales cambios respecto a su predecesor:

- Tensor cores
- NVLink 2.0 (solo Tesla)

Volta


PCI Express 3.0 Host Interface
GigaThread Engine
HBM2
Memory Controller
GPC
TPC
SM
L2 Cache
High-Speed Hub
NVLink
SM
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
FP64
INT
FP32
TENSOR CORE
LD/ST
SFU
128KB L1 Data Cache / Shared Memory
Tex

Turing

Lanzamiento: 2018

Compute capability: 7.5

CUDA cores per SM: 64

Principales cambios respecto a su predecesor:

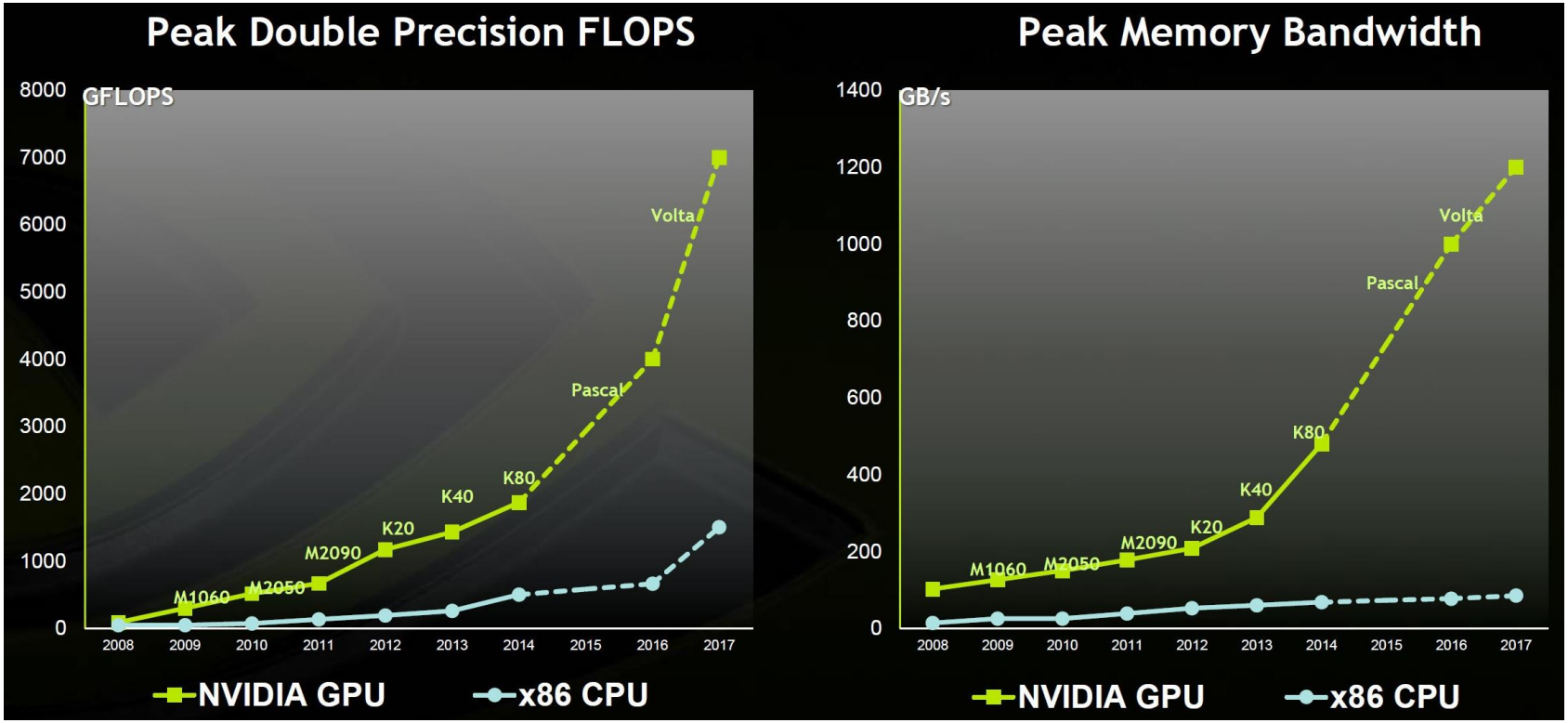
- Ray-Tracing (RT) cores

Turing


PCI Express 3.0 Host Interface
GigaThread Engine
GPC
Raster Engine
TPC
SM
RT CORE
L2 Cache
Memory Controller
High-Speed Hub
NVLink – Two x8 Links


SM
Register File (16,384 x 32-bit)
INT32
FP32
TENSOR CORES
SFU
96KB L1 Data Cache / Shared Memory
Tex
RT CORE

Evolución


Peak Double Precision FLOPS
GFLOPS
8000
7000
6000
5000
4000
3000
2000
1000
0
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
M1060
M2050
M2090
K20
K40
K80
Pascal
Volta
NVIDIA GPU
x86 CPU
Peak Memory Bandwidth
GB/s
1400
1200
1000
800
600
400
200
0
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
M1060
M2050
M2090
K20
K40
K80
Pascal
Volta
NVIDIA GPU
x86 CPU

Diferencias según compute capability

Para saber cuáles son las características soportadas por cada compute capability, así también como las especificaciones técnicas de cada generación, consulte el siguiente link:

https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#features-and-technical-specifications

Para consultar la compute capability de su dispositivo, acceda al siguiente link:

https://en.wikipedia.org/wiki/CUDA#GPUs_supported