King Abdullah University of Science and Technology

KAUST

CS 380 - GPU and GPGPU Programming
Lecture 9: GPU Architecture 7

Markus Hadwiger, KAUST

Reading Assignment #5 (until Oct 1)

Read (required):

- Programming Massively Parallel Processors book, Chapter 3 (*Introduction to CUDA*)
- Programming Massively Parallel Processors book, Chapter 4 (*CUDA Threads*) until (including) 4.3

Read (optional):

- NVIDIA Fermi graphics (GF100) and compute white papers:

http://www.nvidia.com/object/IO_86775.html

http://www.nvidia.com/object/IO_86776.html

- NVIDIA Kepler (GK110) white paper:

http://www.nvidia.com/content/PDF/kepler/NVIDIA-Kepler-GK110-Architecture-Whitepaper.pdf

- NVIDIA Maxwell (GM107 and GM204) white papers:

https://international.download.nvidia.com/geforce-com/international/pdfs/GeForce-GTX-750-Ti-Whitepaper.pdf

https://international.download.nvidia.com/geforce-com/international/pdfs/GeForce-GTX-750-Ti-Whitepaper.pdf

Reading Assignment #6 (until Oct 8)

Read (required):

- Programming Massively Parallel Processors book, finish Chapter 4 (CUDA Threads)
- Programming Massively Parallel Processors book, Chapter 5 (CUDA Memories)

Read (optional):

- NVIDIA Pascal (GP100) architecture white paper:

https://images.nvidia.com/content/pdf/tesla/whitepaper/pascal-architecture-whitepaper.pdf

- NVIDIA Volta (GV100) architecture white paper:

https://images.nvidia.com/content/volta-architecture/pdf/Volta-Architecture-Whitepaper-v1.0.pdf

Tesla (G80, GT200) – Fermi Specs

GPU	G80	GT200	Fermi
Transistors	681 million	1.4 billion	3.0 billion
CUDA Cores	128	240	512
Double Precision Floating Point Capability	None	30 FMA ops / clock	256 FMA ops /clock
Single Precision Floating Point Capability	128 MAD ops/clock	240 MAD ops / clock	512 FMA ops /clock
Special Function Units (SFUs) / SM	2	2	4
Warp schedulers (per SM)	1	1	2
Shared Memory (per SM)	16 KB	16 KB	Configurable 48 KB or 16 KB
L1 Cache (per SM)	None	None	Configurable 16 KB or 48 KB
L2 Cache	None	None	768 KB
ECC Memory Support	No	No	Yes
Concurrent Kernels	No	No	Up to 16
Load/Store Address Width	32-bit	32-bit	64-bit

Kepler – Volta Specs

Tesla Product	Tesla K40	Tesla M40	Tesla P100	Tesla V100
GPU	GK180 (Kepler)	GM200 (Maxwell)	GP100 (Pascal)	GV100 (Volta)
SMs	15	24	56	80
TPCs	15	24	28	40
FP32 Cores / SM	192	128	64	64
FP32 Cores / GPU	2880	3072	3584	5120
FP64 Cores / SM	64	4	32	32
FP64 Cores / GPU	960	96	1792	2560
Tensor Cores / SM	NA	NA	NA	8
Tensor Cores / GPU	NA	NA	NA	640
GPU Boost Clock	810/875 MHz	1114 MHz	1480 MHz	1455 MHz
Peak FP32 TFLOP/s*	5.04	6.8	10.6	15
Peak FP64 TFLOP/s*	1.68	.21	5.3	7.5
Peak Tensor Core TFLOP/s*	NA	NA	NA	120
Texture Units	240	192	224	320
Memory Interface	384-bit GDDR5	384-bit GDDR5	4096-bit HBM2	4096-bit HBM2
Memory Size	Up to 12 GB	Up to 24 GB	16 GB	16 GB
L2 Cache Size	1536 KB	3072 KB	4096 KB	6144 KB
Shared Memory Size / SM	16 KB/32 KB/48 KB	96 KB	64 KB	Configurable up to 96 KB
Register File Size / SM	256 KB	256 KB	256 KB	256KB
Register File Size / GPU	3840 KB	6144 KB	14336 KB	20480 KB
TDP	235 Watts	250 Watts	300 Watts	300 Watts
Transistors	7.1 billion	8 billion	15.3 billion	21.1 billion
GPU Die Size	551 mm^2	601 mm^2	610 mm^2	815 mm^2
Manufacturing Process	28 nm	28 nm	16 nm FinFET+	12 nm FFN

CUDA
Compute Capabilities

Compute Capab. – 2.0

- 1024 threads / block
- More threads / SM
- 32K registers / SM
- New synchronization functions

	Compute Capability				
Feature Support (Unlisted features are supported for all compute capabilities)	**1.0**	**1.1**	**1.2**	**1.3**	**2.0**
Integer atomic functions operating on 32-bit words in global memory (Section B.10)	No	yes			
Integer atomic functions operating on 64-bit words in global memory (Section B.10)	No		Yes		
Integer atomic functions operating on 32-bit words in shared memory (Section B.10)					
Warp vote functions (Section B.11)					
Double-precision floating-point numbers	No			Yes	
Floating-point atomic addition operating on 32-bit words in global and shared memory (Section B.10)	No				Yes
__ballot() (Section B.11)					
__threadfence_system() (Section B.5)					
__syncthreads_count(), __syncthreads_and(), __syncthreads_or() (Section B.6)					

	Compute Capability				
Technical Specifications	**1.0**	**1.1**	**1.2**	**1.3**	**2.0**
Maximum x- or y-dimension of a grid of thread blocks	65535				
Maximum number of threads per block	512				1024
Maximum x- or y-dimension of a block	512				1024
Maximum z-dimension of a block	64				
Warp size	32				
Maximum number of resident blocks per multiprocessor	8				
Maximum number of resident warps per multiprocessor	24		32		48
Maximum number of resident threads per multiprocessor	768		1024		1536
Number of 32-bit registers per multiprocessor	8 K		16 K		32 K
Maximum amount of shared memory per multiprocessor	16 KB				48 KB
Number of shared memory banks	16				32
Amount of local memory per thread	16 KB				512 KB
Constant memory size	64 KB				
Cache working set per multiprocessor for constant memory	8 KB				
Cache working set per multiprocessor for texture memory	Device dependent, between 6 KB and 8 KB				
Maximum width for a 1D texture reference bound to a CUDA array	8192				32768
Maximum width for a 1D texture reference bound to linear memory	2^{27}				
Maximum width and height for a 2D texture reference bound to linear memory or a CUDA array	65536 x 32768				65536 x 65536
Maximum width, height, and depth for a 3D texture reference bound to linear memory or a CUDA array	2048 x 2048 x 2048				4096 x 4096 x 4096
Maximum number of instructions per kernel	2 million				

Compute Capabilities 2.0 – 3.5 (Fermi – Kepler)

	FERMI GF100	FERMI GF104	KEPLER GK104	KEPLER GK110
Compute Capability	2.0	2.1	3.0	3.5
Threads / Warp	32	32	32	32
Max Warps / Multiprocessor	48	48	64	64
Max Threads / Multiprocessor	1536	1536	2048	2048
Max Thread Blocks / Multiprocessor	8	8	16	16
32-bit Registers / Multiprocessor	32768	32768	65536	65536
Max Registers / Thread	63	63	63	255
Max Threads / Thread Block	1024	1024	1024	1024
Shared Memory Size Configurations (bytes)	16K 48K	16K 48K	16K 32K 48K	16K 32K 48K
Max X Grid Dimension	2^16-1	2^16-1	2^32-1	2^32-1
Hyper-Q	No	No	No	Yes
Dynamic Parallelism	No	No	No	Yes

Compute Capability of Fermi and Kepler GPUs

Compute Capab. 5.x (Maxwell, Part 1)

Maxwell

- GM107: 5.0
- GM204: 5.2

	Compute Capability					
Technical Specifications	**2.x**	**3.0, 3.2**	**3.5**	**3.7**	**5.0**	**5.2**
Maximum dimensionality of grid of thread blocks	3					
Maximum x-dimension of a grid of thread blocks	65535	$2^{31}-1$				
Maximum y- or z-dimension of a grid of thread blocks	65535					
Maximum dimensionality of thread block	3					
Maximum x- or y-dimension of a block	1024					
Maximum z-dimension of a block	64					
Maximum number of threads per block	1024					
Warp size	32					
Maximum number of resident blocks per multiprocessor	8	16			32	
Maximum number of resident warps per multiprocessor	48	64				
Maximum number of resident threads per multiprocessor	1536	2048				

Compute Capab. 5.x (Maxwell, Part 2)

Maxwell

- GM107: 5.0
- GM204: 5.2

	Compute Capability					
Technical Specifications	**2.x**	**3.0, 3.2**	**3.5**	**3.7**	**5.0**	**5.2**
Number of 32-bit registers per multiprocessor	32 K	64 K		128 K	64 K	
Maximum number of 32-bit registers per thread block	32 K	64 K				
Maximum number of 32-bit registers per thread	63		255			
Maximum amount of shared memory per multiprocessor	48 KB			112 KB	64 KB	96 KB
Maximum amount of shared memory per thread block	48 KB					
Number of shared memory banks	32					
Amount of local memory per thread	512 KB					
Constant memory size	64 KB					
Cache working set per multiprocessor for constant memory	8 KB				10 KB	
Cache working set per multiprocessor for texture memory	12 KB	Between 12 KB and 48 KB				

Compute Capabilities 3.5 – 7.0 (Kepler – Volta)

GPU	Kepler GK180	Maxwell GM200	Pascal GP100	Volta GV100
Compute Capability	3.5	5.2	6.0	7.0
Threads / Warp	32	32	32	32
Max Warps / SM	64	64	64	64
Max Threads / SM	2048	2048	2048	2048
Max Thread Blocks / SM	16	32	32	32
Max 32-bit Registers / SM	65536	65536	65536	65536
Max Registers / Block	65536	32768	65536	65536
Max Registers / Thread	255	255	255	255*
Max Thread Block Size	1024	1024	1024	1024
FP32 Cores / SM	192	128	64	64
# of Registers to FP32 Cores Ratio	341	512	1024	1024
Shared Memory Size / SM	16 KB/32 KB/48 KB	96 KB	64 KB	Configurable up to 96 KB

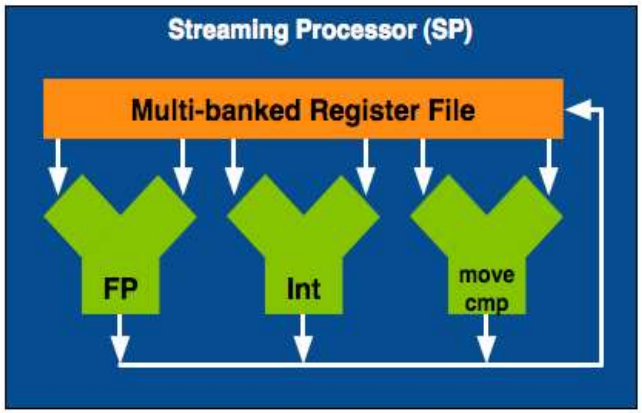
NVIDIA Tesla Architecture

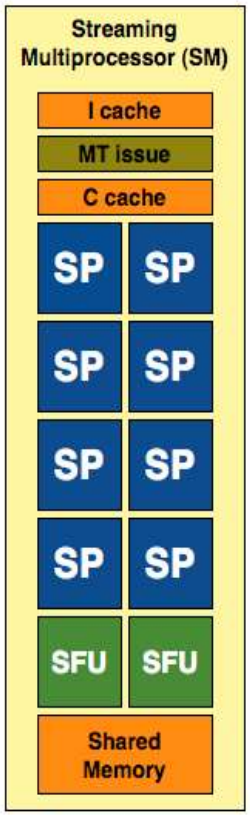
2007-2009

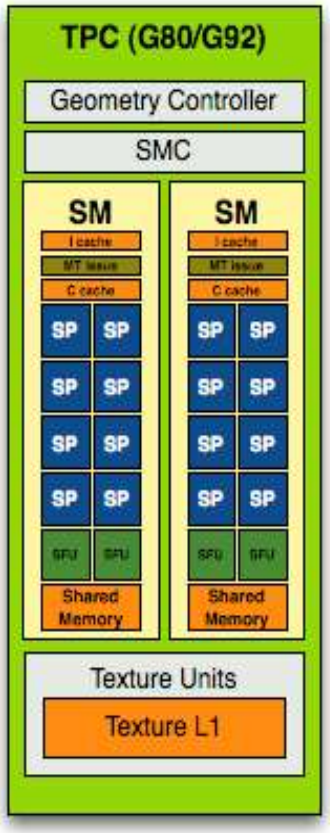
G80, G9x: 2007 (Geforce 8800, ...)
GT200: 2008/2009 (GTX 280, ...)

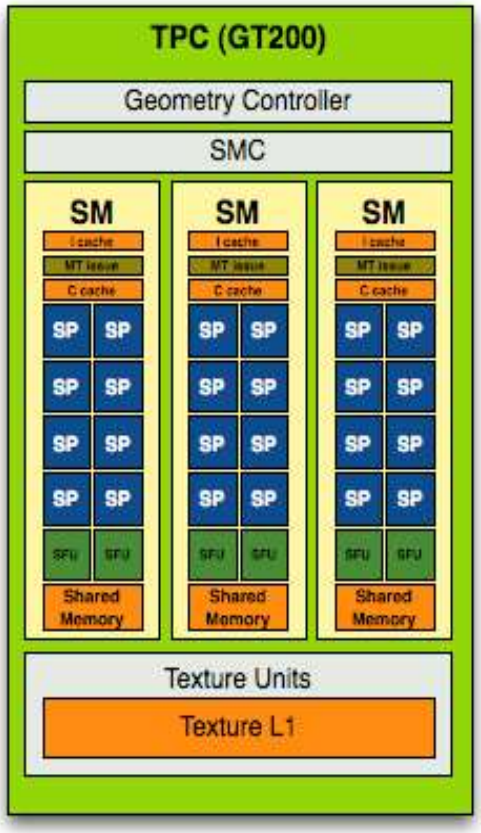
(this is not the Tesla product line!)

NVIDIA Tesla Architecture (not the Tesla product line!), G80: 2007, GT200: 2008/2009









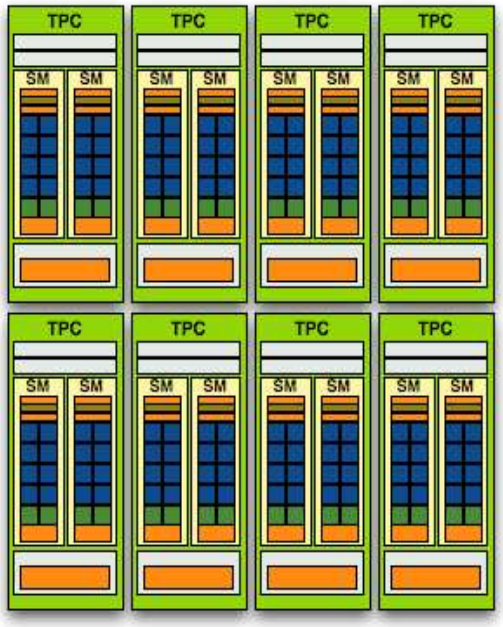
G80: first CUDA GPU!

- Streaming Processor (SP) [nowadays: CUDA core]
- Streaming Multiprocessor (SM)
- Texture/Processing Cluster (TPC)

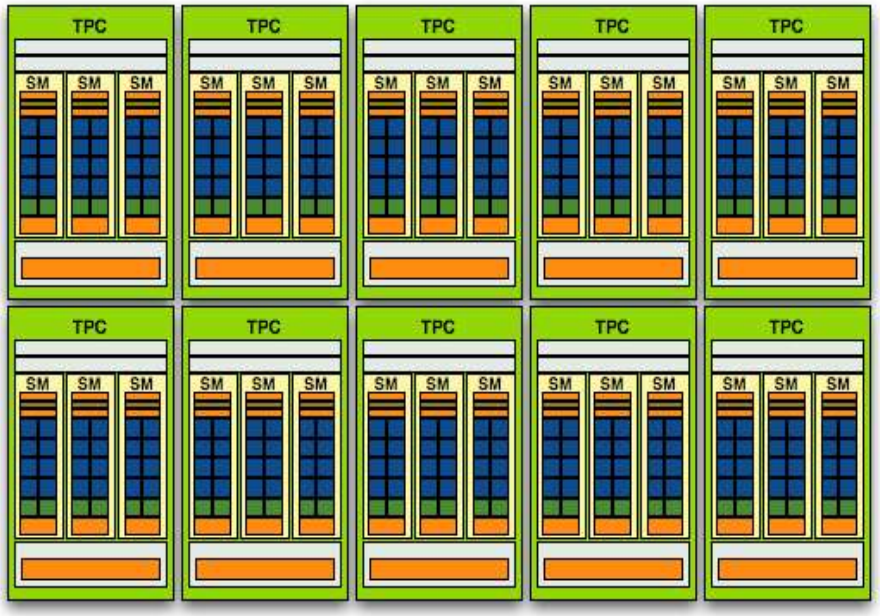
Courtesy AnandTech

NVIDIA Tesla Architecture (not the Tesla product line!), G80: 2007, GT200: 2008/2009

- G80/G92: 8 TPCs * (2 * 8 SPs) = 128 SPs [= CUDA cores]
- GT200: 10 TPCs * (3 * 8 SPs) = 240 SPs [= CUDA cores]
- Arithmetic intensity has increased (ALUs vs. texture units)



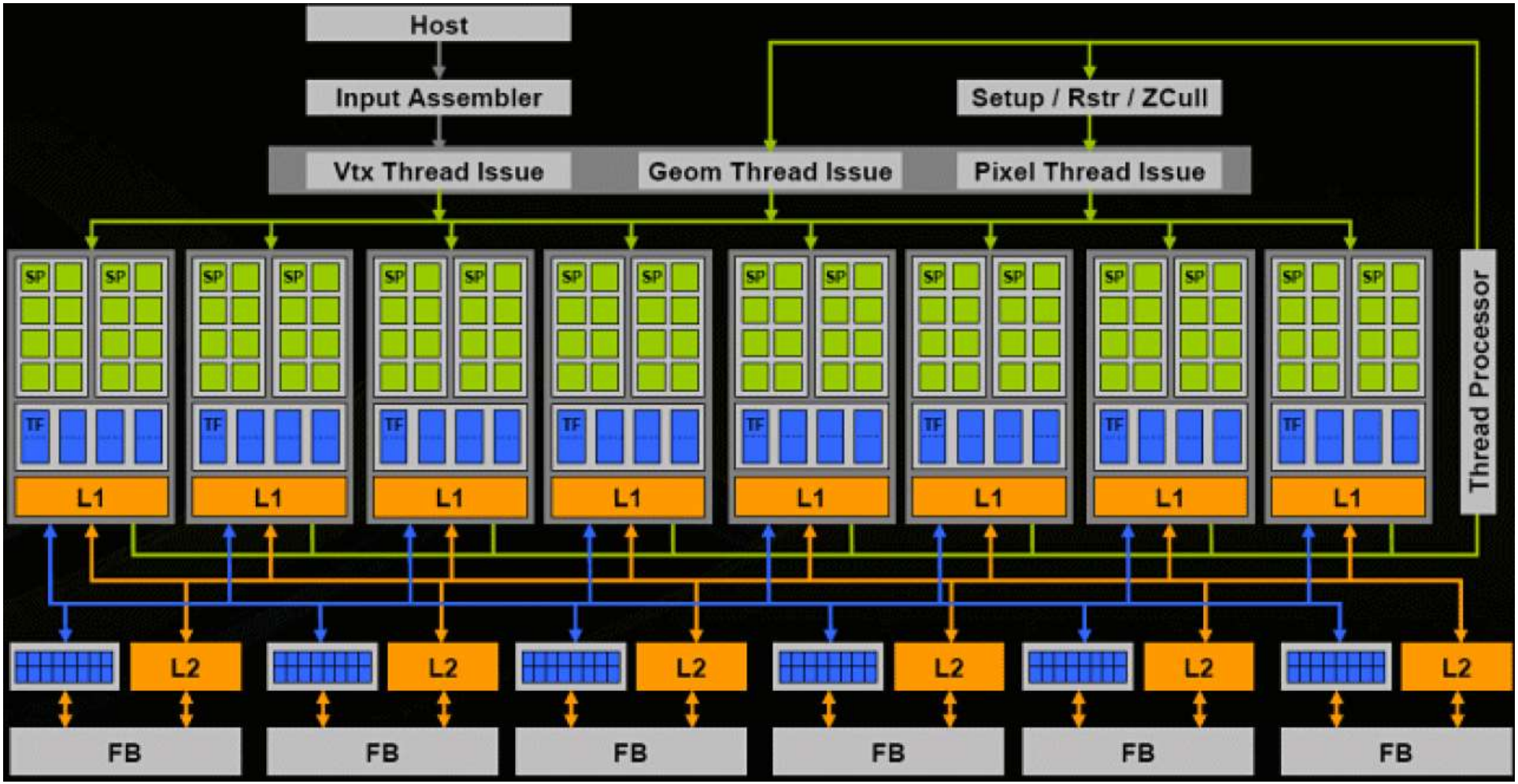
G80 / G92



GT200

Courtesy AnandTech

Example: GeForce 8 (end of 2006 - 2007)



NVIDIA Fermi Architecture

2010

GF100, ... (GTX 480, ...)

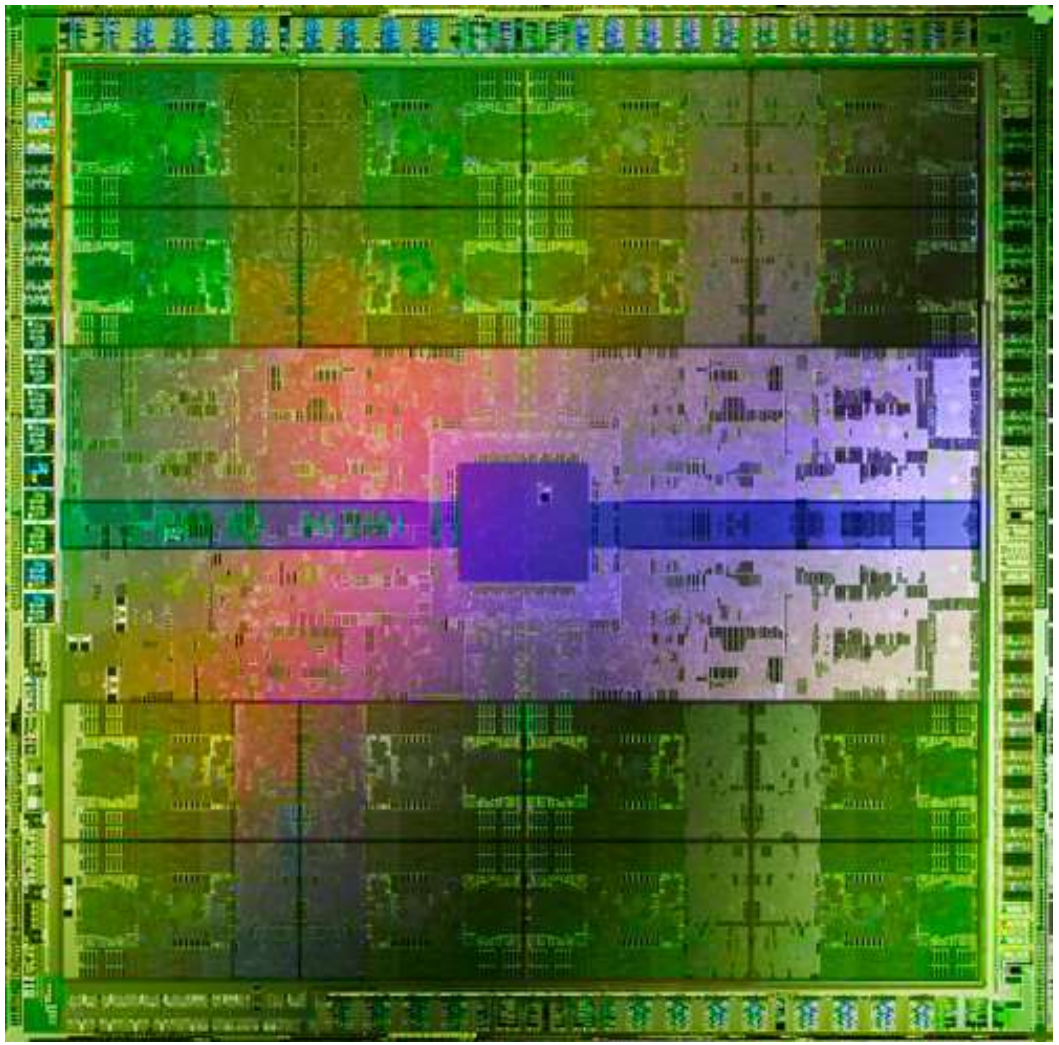
GF110, ... (GTX 580, ...)

NVIDIA Fermi Architecture (2010)

Full size

- 4 GPCs
- 4 SMs each
- 6 64-bit memory controllers (= 384 bit)

NVIDIA Fermi (GF100) Die Photo



Full size

- 4 GPCs
- 4 SMs each

NVIDIA Fermi SM (2010)

Streaming processors now called *CUDA cores*

32 CUDA cores per Fermi streaming multiprocessor (SM)

16 SMs = 512 CUDA cores

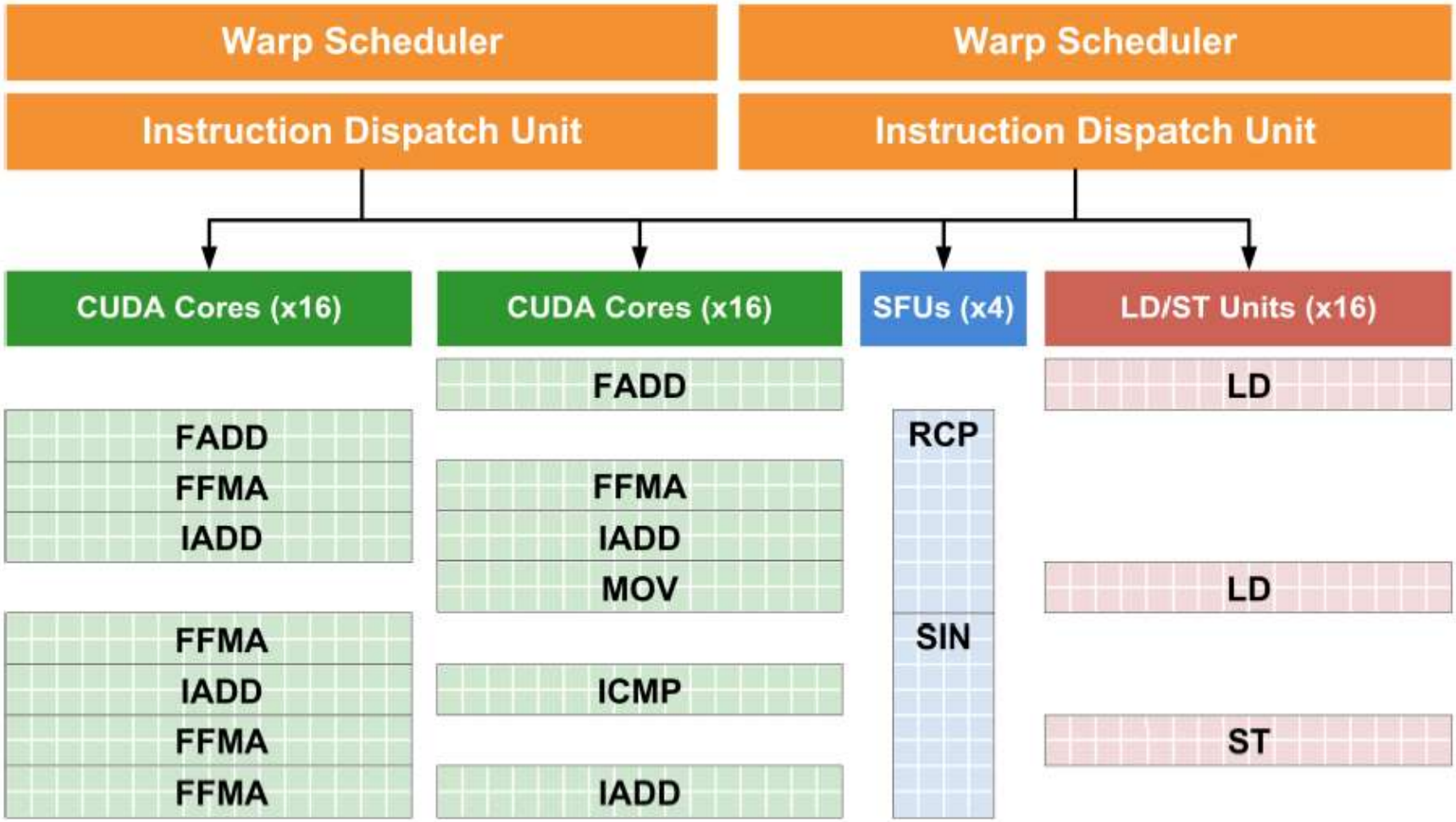
CPU-like cache hierarchy

- L1 cache / shared memory
- L2 cache

Texture units and caches now in SM

(instead of with TPC=multiple SMs in G80/GT200)

Dual Warp Schedulers



Graphics Processor Clusters (GPC)

(instead of TPC on GT200)

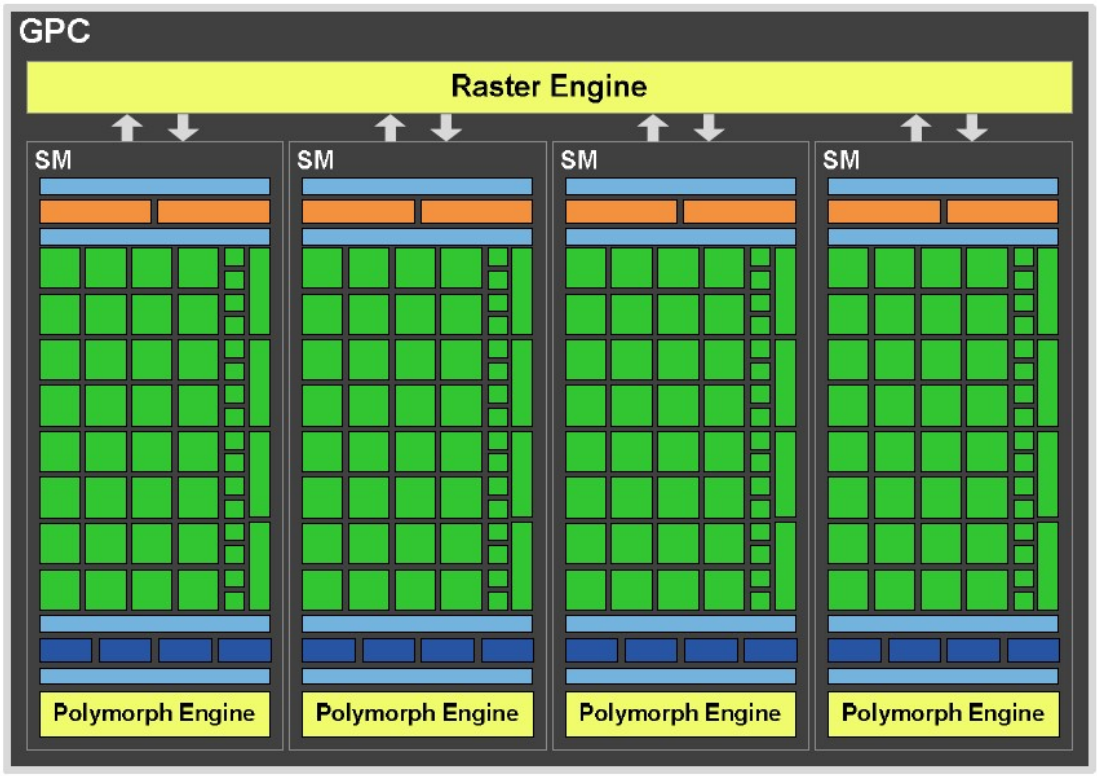
4 SMs

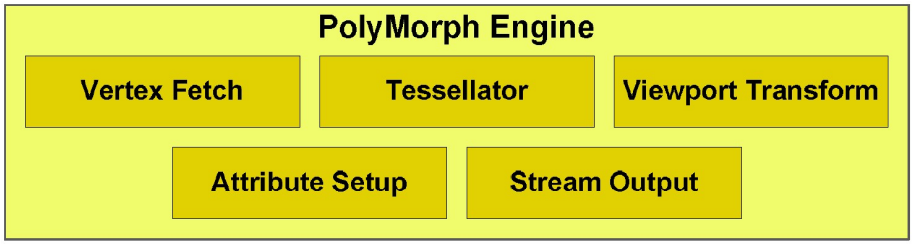
32 CUDA cores / SM

4 SMs / GPC =
128 cores / GPC

Decentralized rasterization and geometry

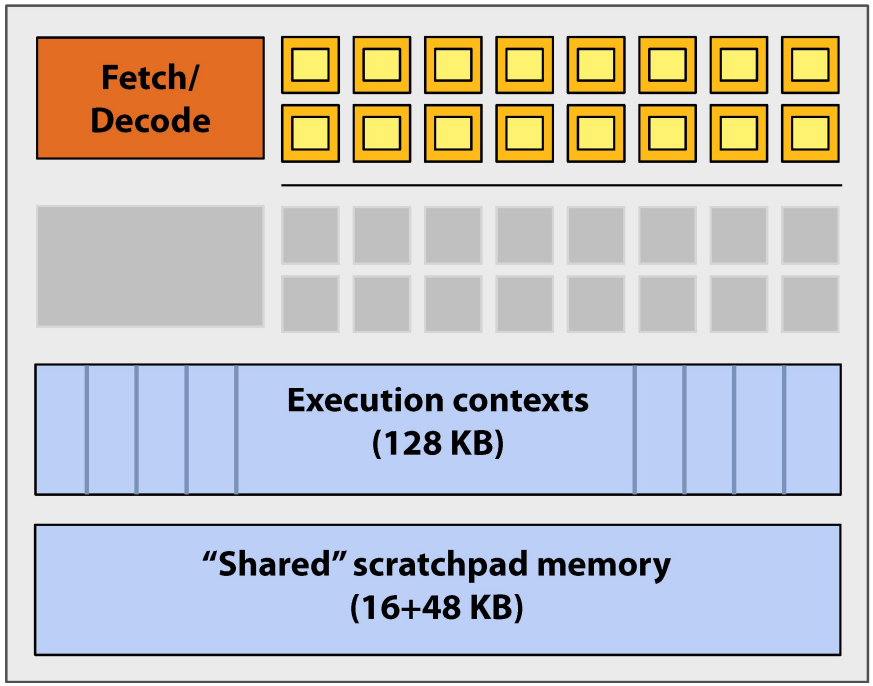
- 4 raster engines
- 16 ”PolyMorph” engines



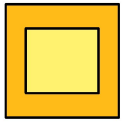


NVIDIA Fermi Architecture (2010)

NVIDIA GeForce GTX 480 "core"



Source: Fermi Compute Architecture Whitepaper
CUDA Programming Guide 3.1, Appendix G

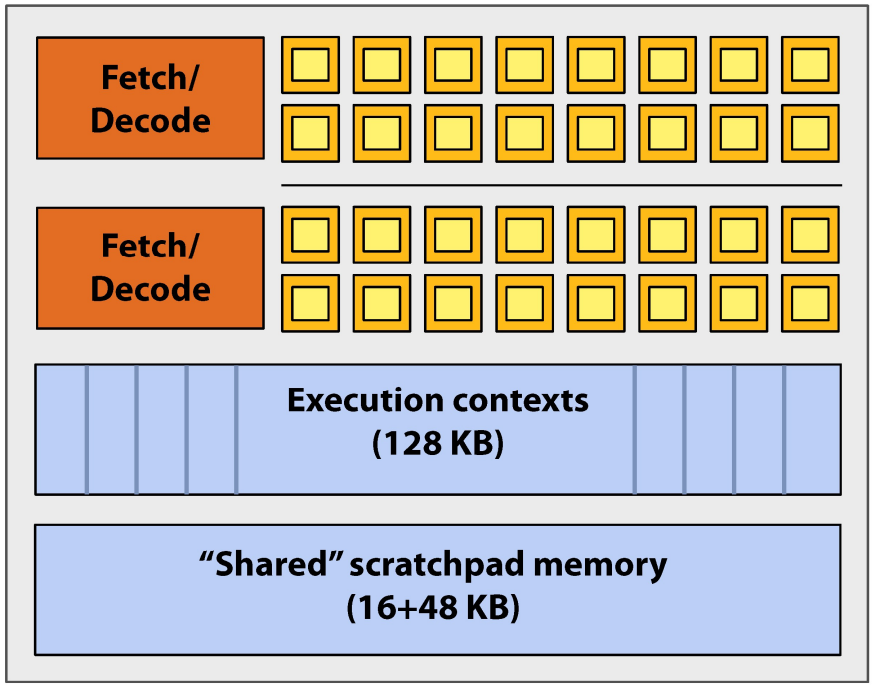


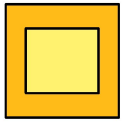
= SIMD function unit,
control shared across 16 units
(1 MUL-ADD per clock)

- Groups of 32 fragments share an instruction stream
- Up to 48 groups are simultaneously interleaved
- Up to 1536 individual contexts can be stored

NVIDIA Fermi Architecture (2010)

NVIDIA GeForce GTX 480 "core"





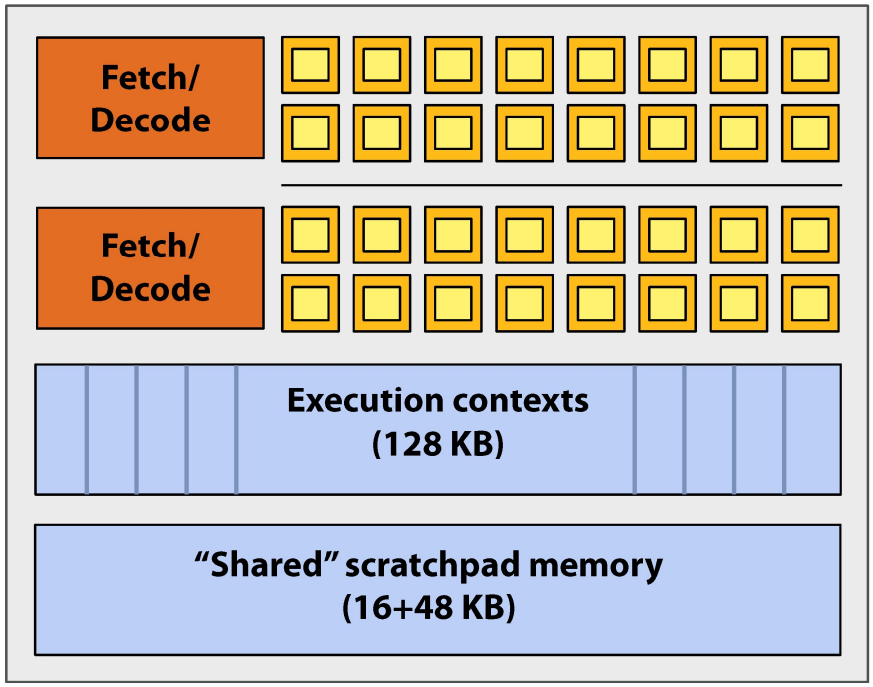
= SIMD function unit, control shared across 16 units (1 MUL-ADD per clock)

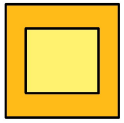
- The core contains 32 functional units
- Two groups are selected each clock (decode, fetch, and execute two instruction streams in parallel)

Source: Fermi Compute Architecture Whitepaper
CUDA Programming Guide 3.1, Appendix G

NVIDIA Fermi Architecture (2010)

NVIDIA GeForce GTX 480 "SM"





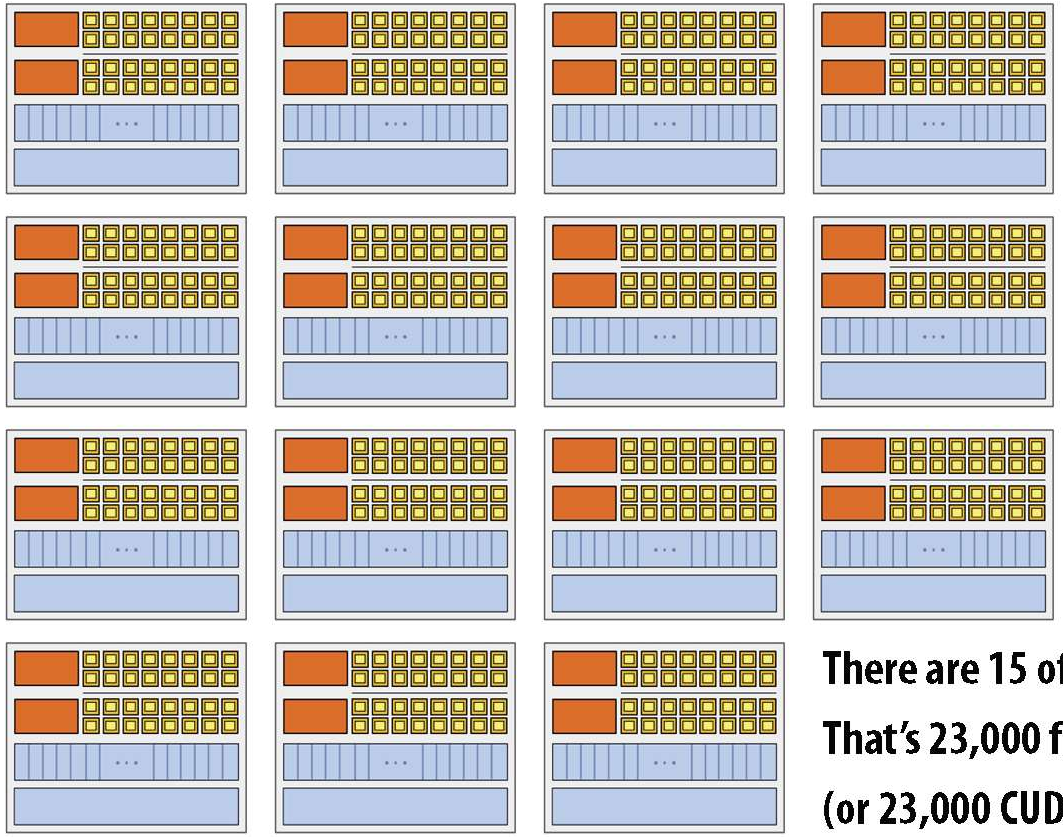
= CUDA core
(1 MUL-ADD per clock)

- The SM contains 32 CUDA cores
- Two warps are selected each clock (decode, fetch, and execute two warps in parallel)
- Up to 48 warps are interleaved, totaling 1536 CUDA threads

Source: Fermi Compute Architecture Whitepaper
CUDA Programming Guide 3.1, Appendix G

NVIDIA Fermi Architecture (2010)

NVIDIA GeForce GTX 480



There are 15 of these things on the GTX 480:

That's 23,000 fragments!

(or 23,000 CUDA threads!)

NVIDIA Kepler Architecture

2012

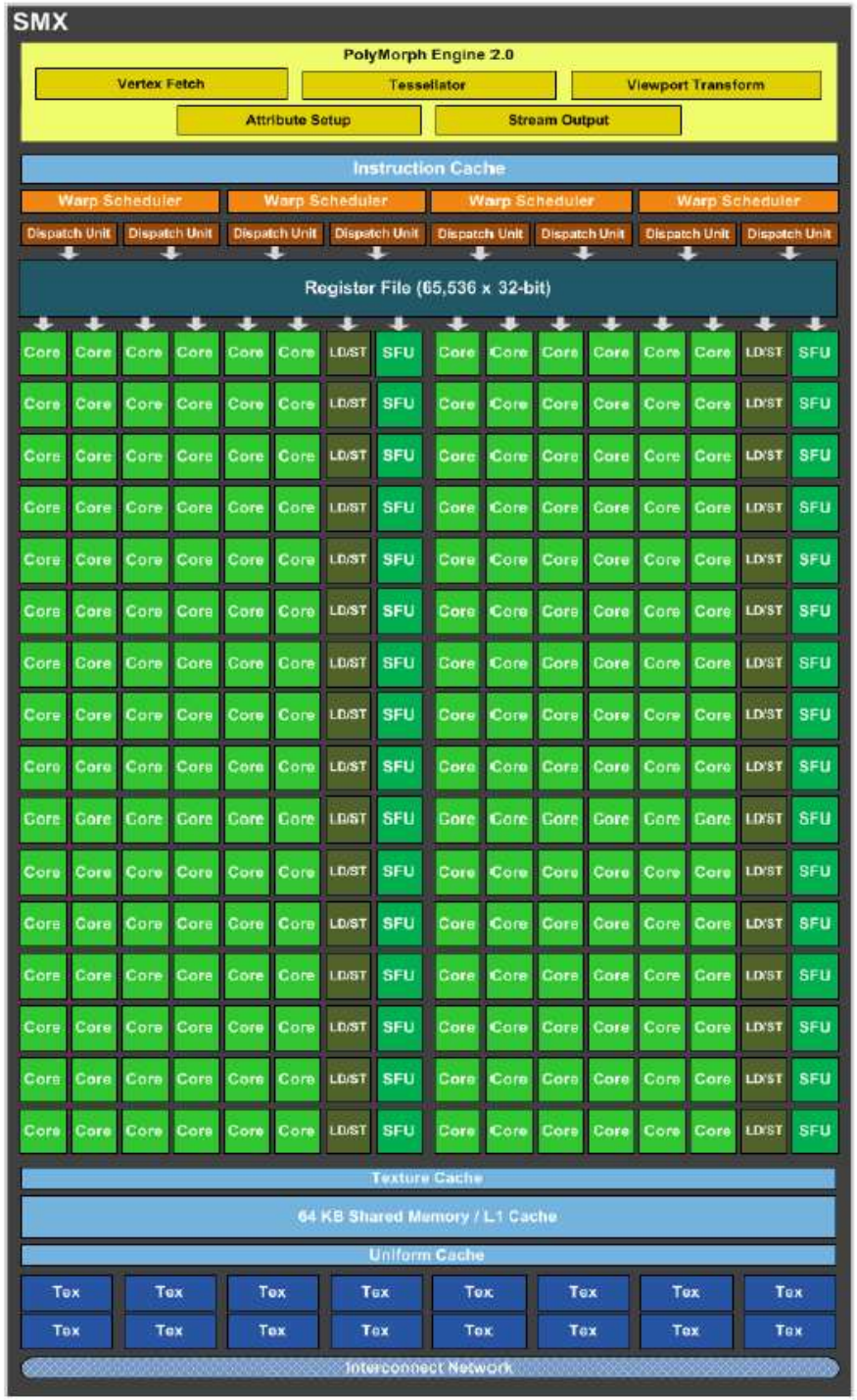
GK104, ... (GTX 680, ...)

GK110, ... (GTX 780, GTX Titan, ...)

NVIDIA Kepler Architecture (2012)

GK104 SMX

- 192 CUDA cores (192 = 6 * 32)
- 32 LD/ST units
- 32 SFUs
- 16 texture units



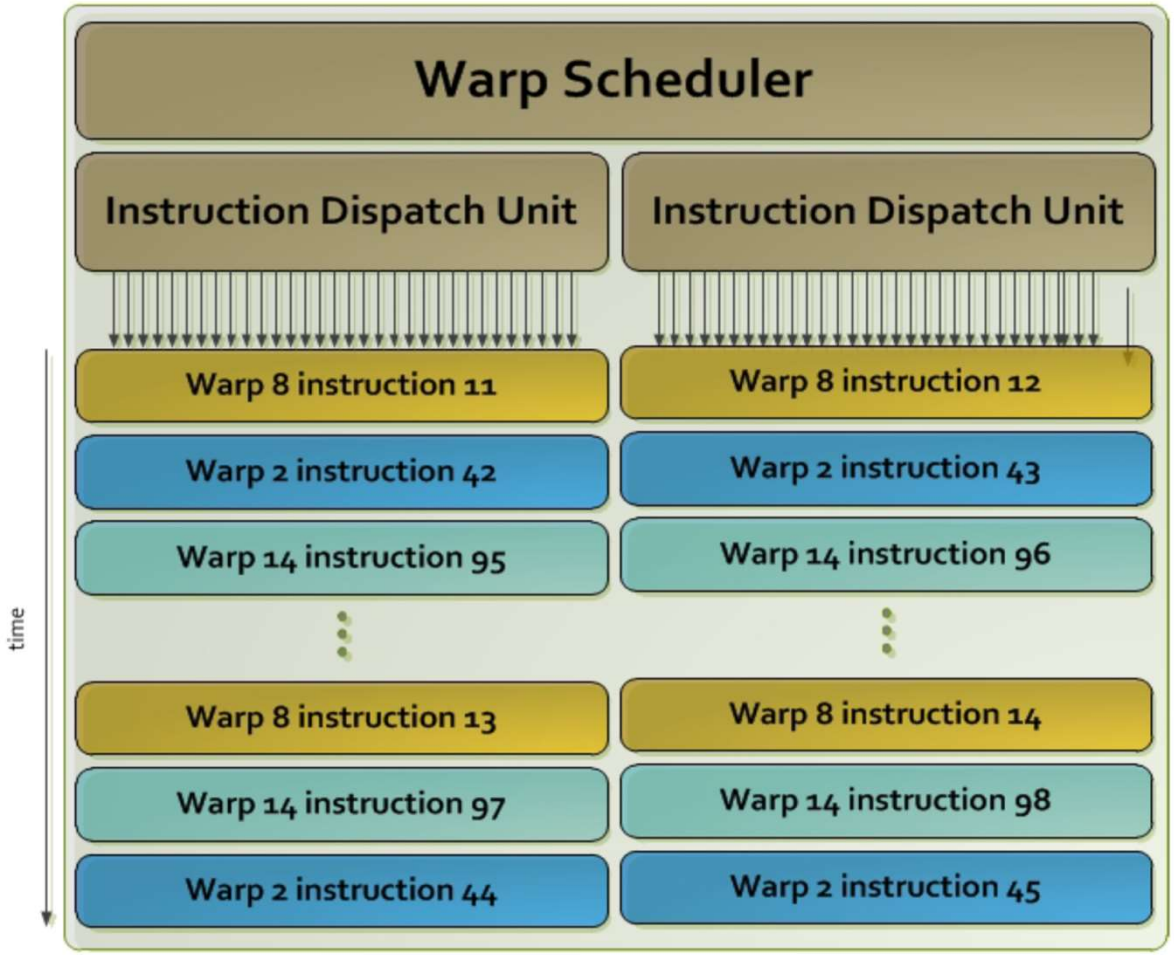
GK110 SMX

- 192 CUDA cores (192 = 6 * 32)
- 64 DP units
- 32 LD/ST units
- 32 SFUs
- 16 texture units

New read-only data cache (48KB)

Two Dispatch Units per Warp Scheduler

Instruction level parallelism

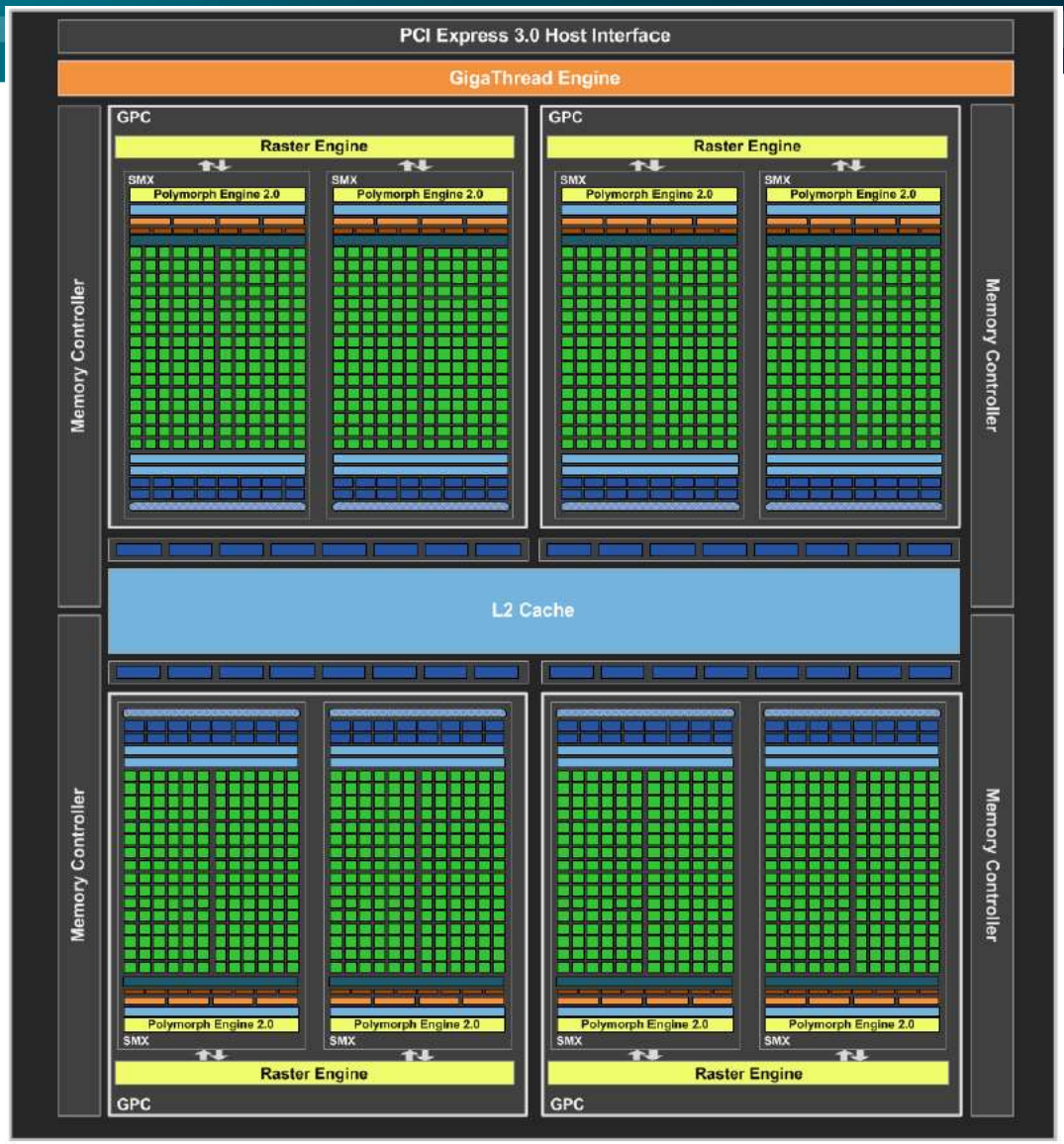


NVIDIA Kepler Architecture (2012)

Three different versions

- Compute capability 3.0 (GK104)
  - Geforce GTX 680, …
  - Quadro K5000
  - Tesla K10
- Compute capability 3.5 (GK110)
  - Geforce GTX 780 / Titan / Titan Black
  - Quadro K6000
  - Tesla K20, Tesla K40
- Compute capability 3.7 (GK210)
  - Tesla K80
  - Came out much later (~end of 2014)

NVIDIA Kepler / GK104 Structure



Full size

- 4 GPCs
- 2 SMXs each

= 8 SMXs,
1536 CUDA cores

NVIDIA Kepler / GK110 Structure (1)

Full size

- 15 SMXs (Titan Black; Titan: 14)
- 2880 CUDA cores (Titan Black; Titan: 2688)
- 5 GPCs of 3 SMXs each

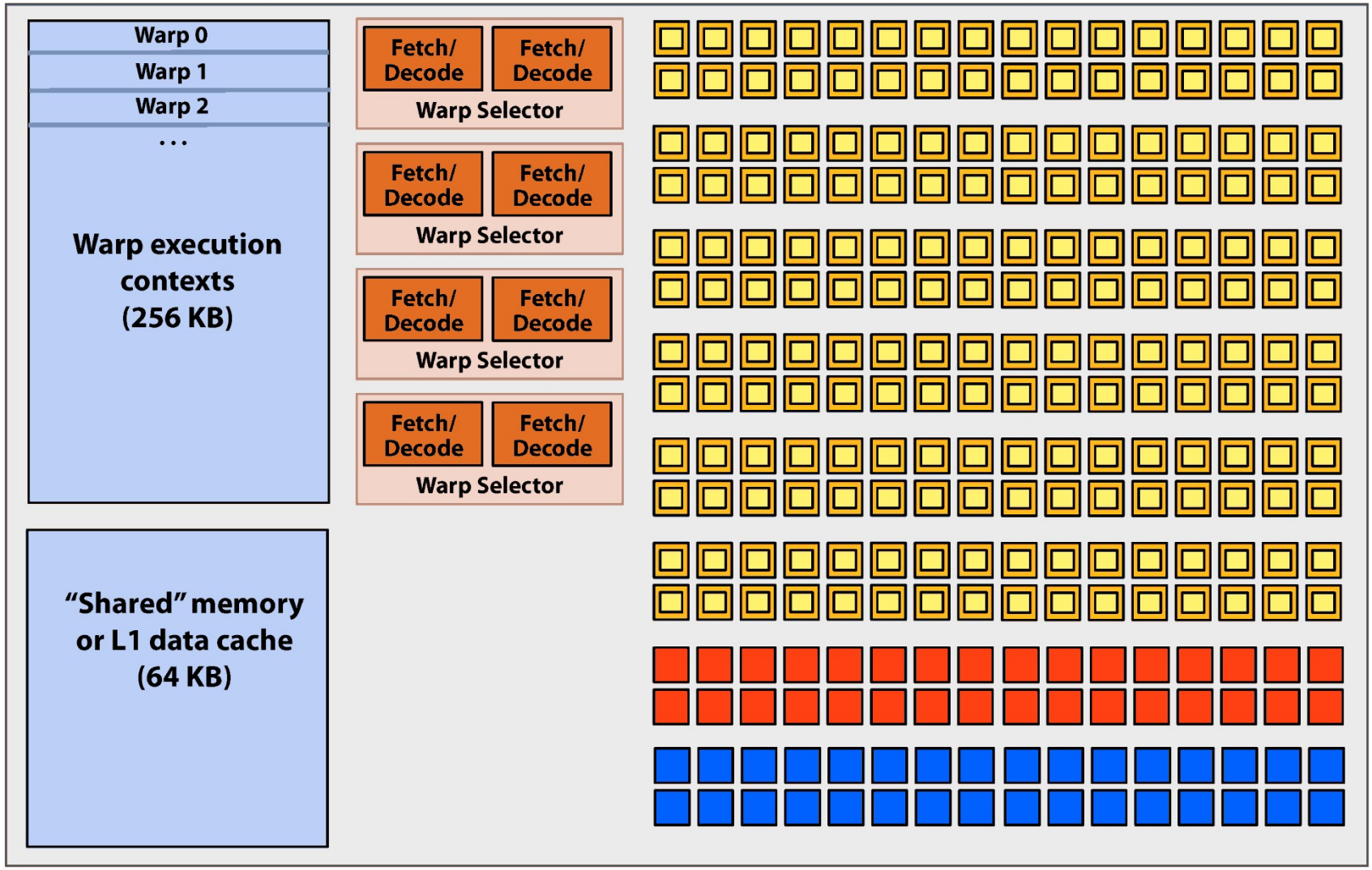
NVIDIA Kepler / GK110 Structure (2)

Titan (not Black)

- 14 SMXs
- 2688 CUDA cores
- 5 GPCs with 3 SMXs or 2 SMXs each

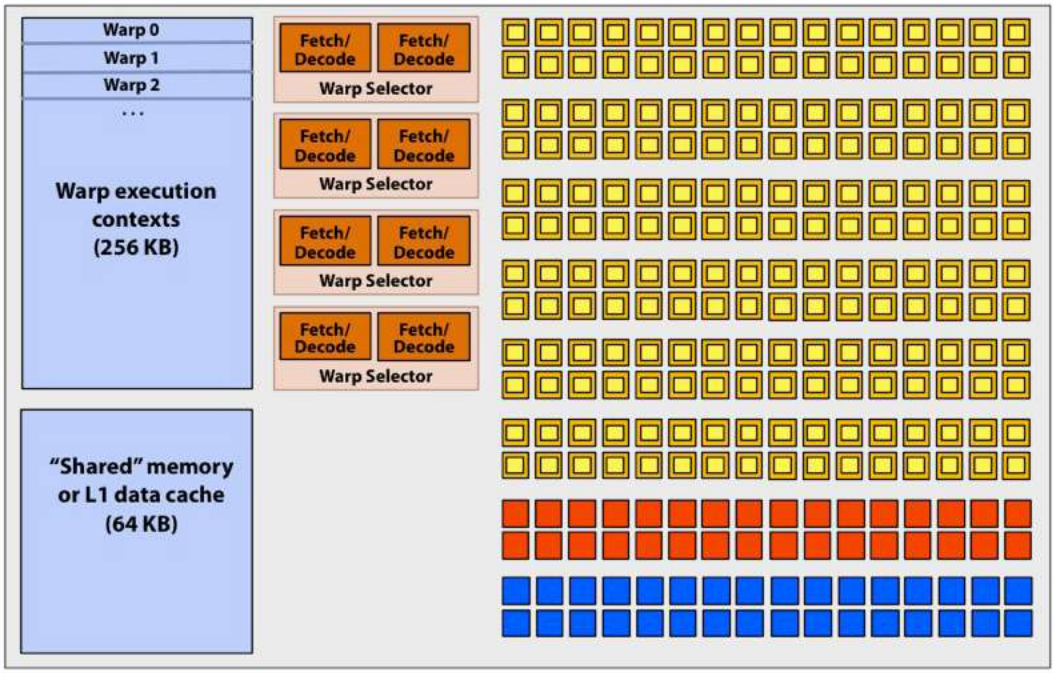
Bonus slides: NVIDIA GTX 680 (2012)

NVIDIA Kepler GK104 architecture SMX unit (one "core")



Bonus slides: NVIDIA GTX 680 (2012)

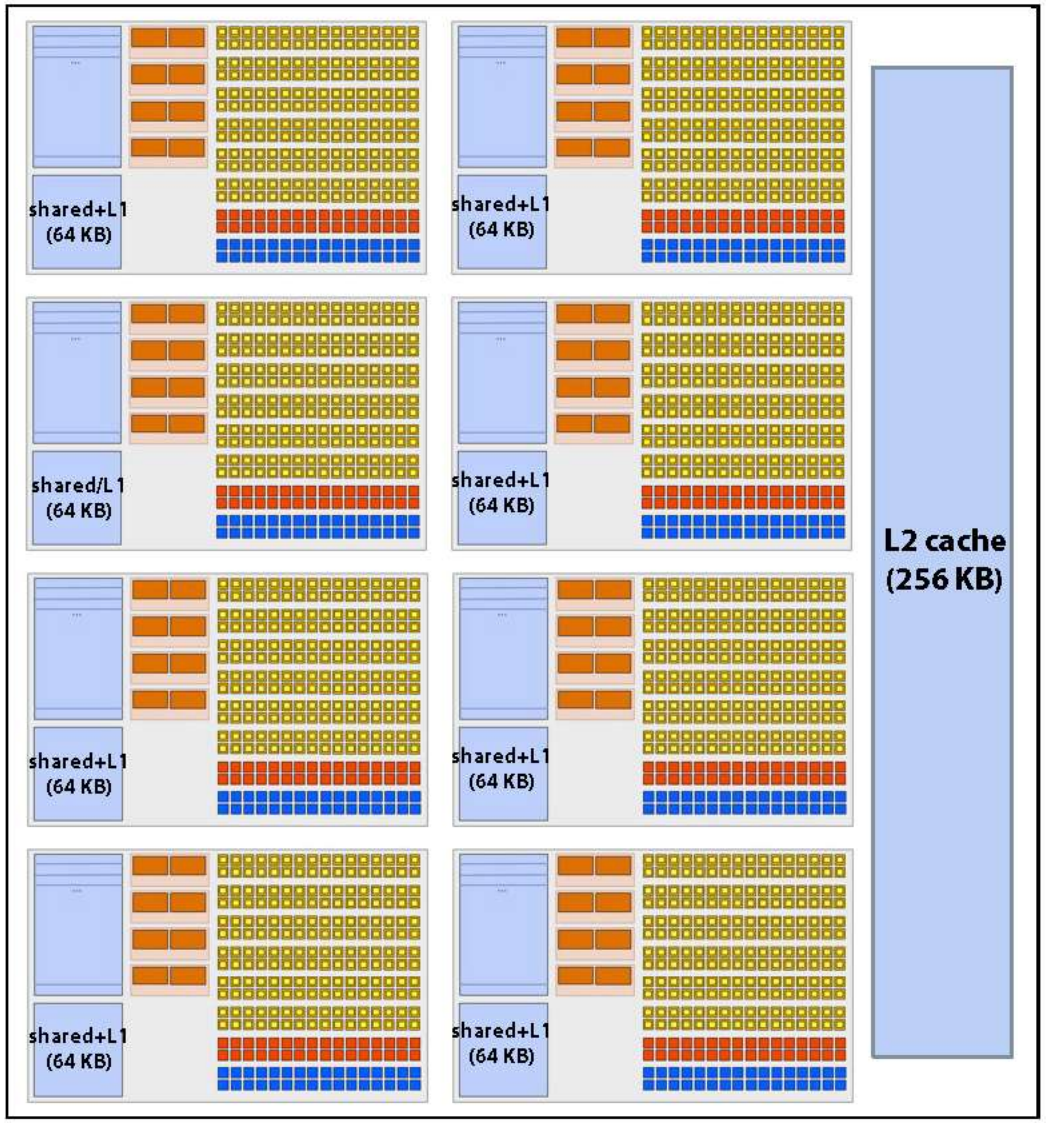
NVIDIA Kepler GK104 architecture SMX unit (one "core")



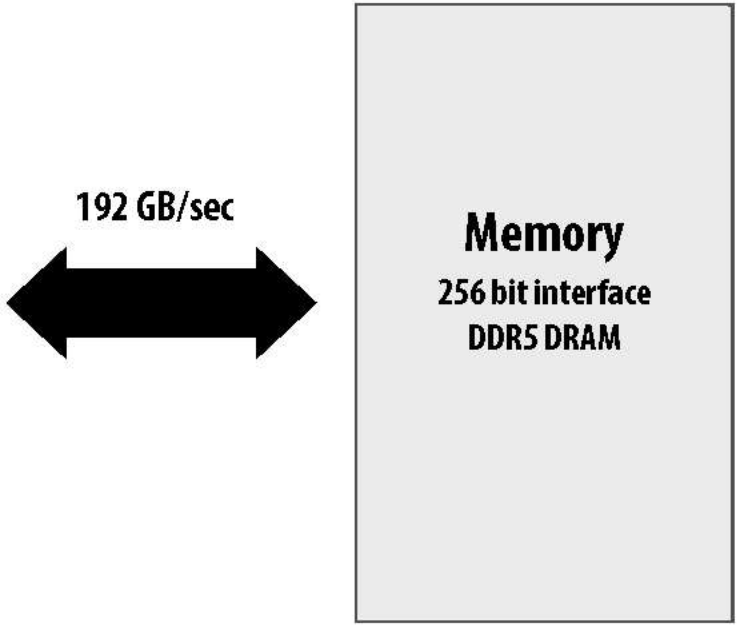
- SMX core resource limits:
  - Maximum warp execution contexts: 64 (2,048 total CUDA threads)
  - Maximum thread blocks: 16
- SMX core operation each clock:
  - Select up to four runnable warps from up to 64 resident on core (thread-level parallelism)
  - Select up to two runnable instructions per warp (instruction-level parallelism)
  - Execute instructions on available groups of SIMD ALUs, special-function ALUs, or LD/ST units

Bonus slides: NVIDIA GTX 680 (2012)

NVIDIA Kepler GK104 architecture



- 1 GHz clock
- Eight SMX cores per chip
- 8 x 192 = 1,536 SIMD mul-add ALUs
  = 3 TFLOPs
- Up to 512 interleaved warps per chip
  (16,384 CUDA threads/chip)
- TDP: 195 watts



NVIDIA Maxwell Architecture

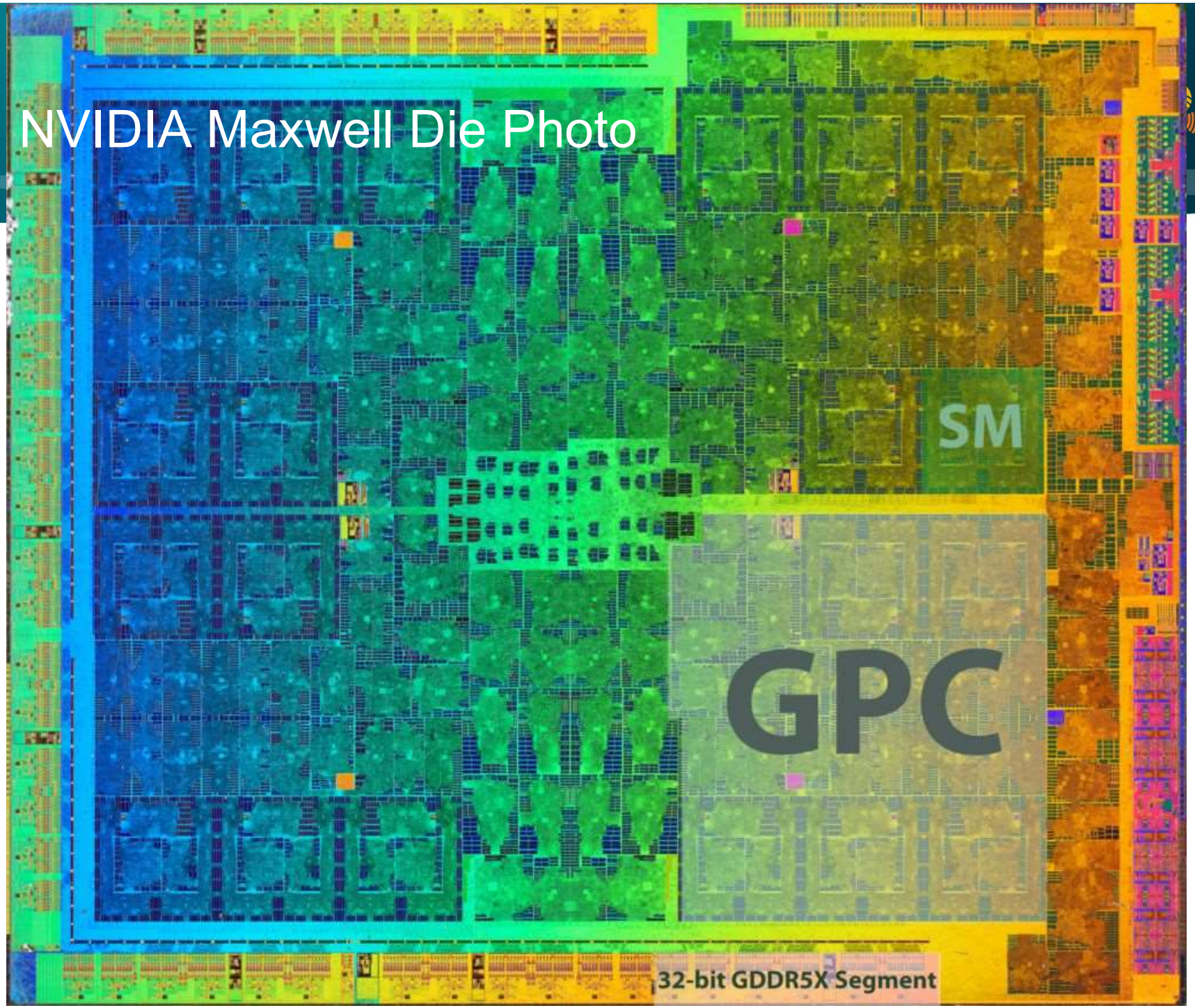
2015

GM107, ... (GTX 750Ti, ...)

GM204, ... (GTX 980, Titan X, ...)

NVIDIA Maxwell Architecture (2015)

NVIDIA Maxwell Die Photo



Maxwell (GM) Architecture

Multiprocessor: SMM

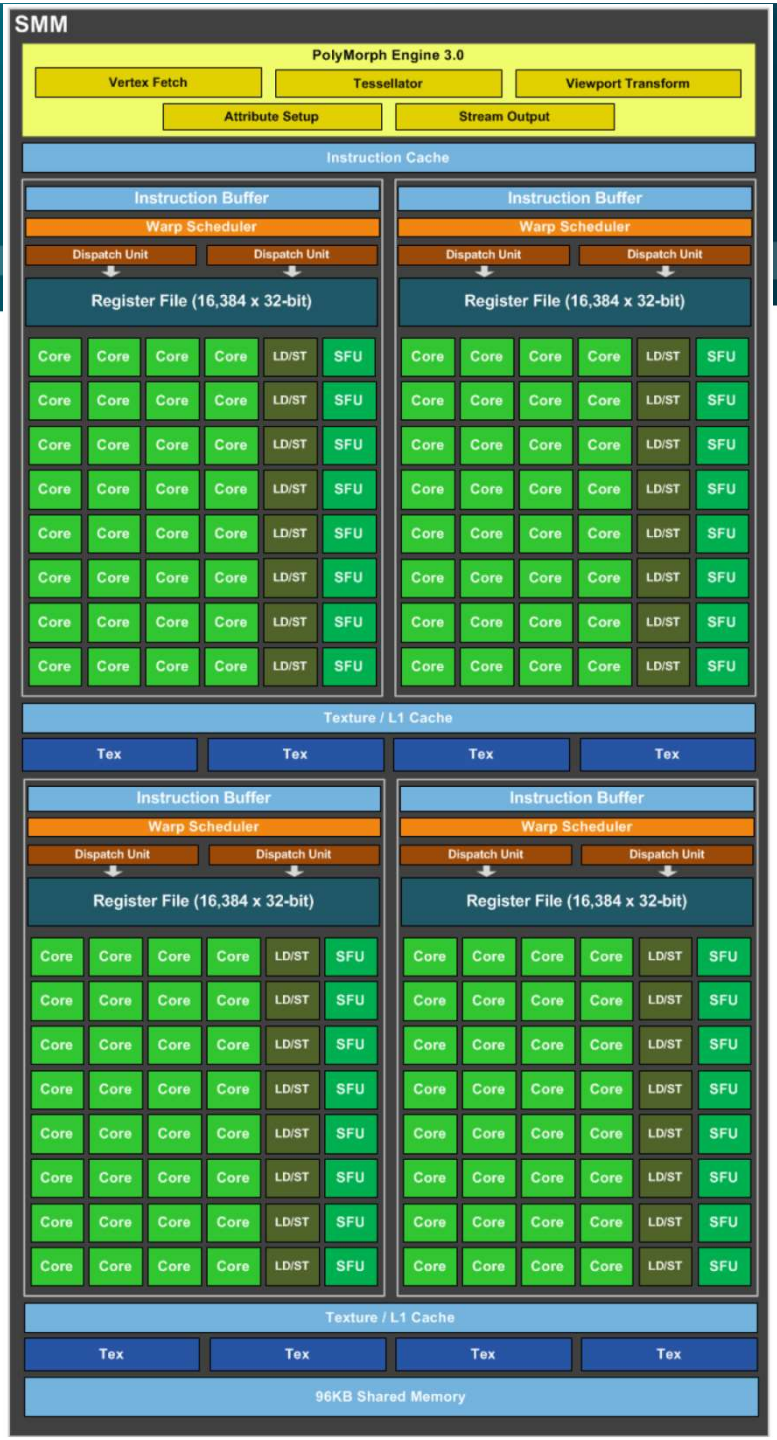
- 128 CUDA cores
- 4 DP units

4 partitions inside SMM

- 32 CUDA cores each
- 8 LD/ST units each
- Each has its own warp scheduler, two dispatch units, register file

Shared memory and L1 cache now separate!

- L1 cache shares with texture cache
- Shared memory is its own space



Maxwell (GM) Architecture

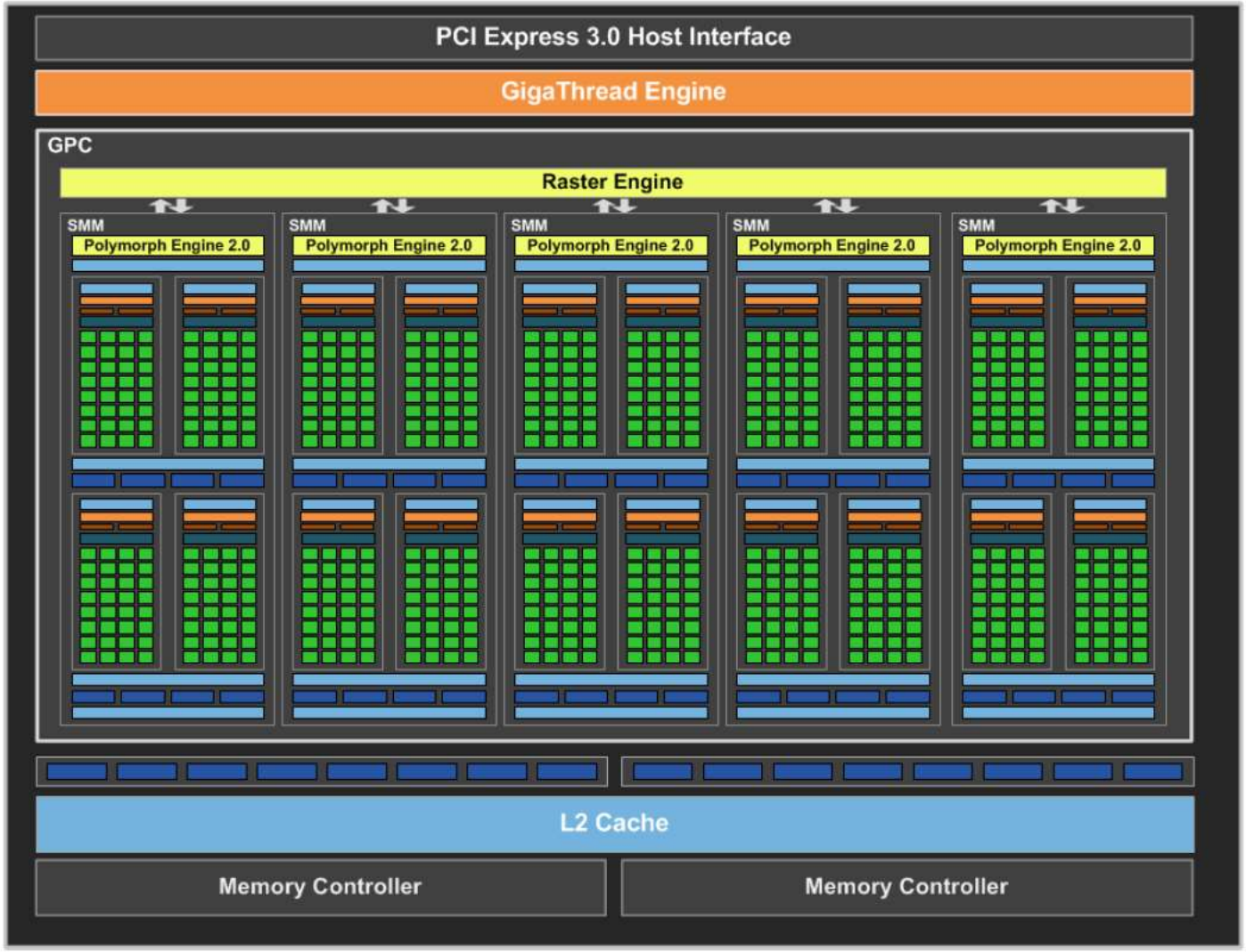
First gen.

GM107

(GTX 750Ti)

5 SMMs

(640 CUDA cores in total)



Maxwell (GM) Architecture

Second gen.

GM204

(GTX 980)

16 SMMs

(2048 CUDA cores in total)

4 GPCs of 4 SMMs

Maxwell (GM) vs. Kepler (GK) Architecture

GK107 vs. GM107

GPU	GK107 (Kepler)	GM107 (Maxwell)
CUDA Cores	384	640
Base Clock	1058 MHz	1020 MHz
GPU Boost Clock	N/A	1085 MHz
GFLOPs	812.5	1305.6
Texture Units	32	40
Texel fill-rate	33.9 Gigatexels/sec	40.8 Gigatexels/sec
Memory Clock	5000 MHz	5400 MHz
Memory Bandwidth	80 GB/sec	86.4 GB/sec
ROPs	16	16
L2 Cache Size	256KB	2048KB
TDP	64W	60W
Transistors	1.3 Billion	1.87 Billion
Die Size	118 mm^2	148 mm^2
Manufacturing Process	28-nm	28-nm

Maxwell (GM) vs. Kepler (GK) Architecture

GK107 vs. GM204

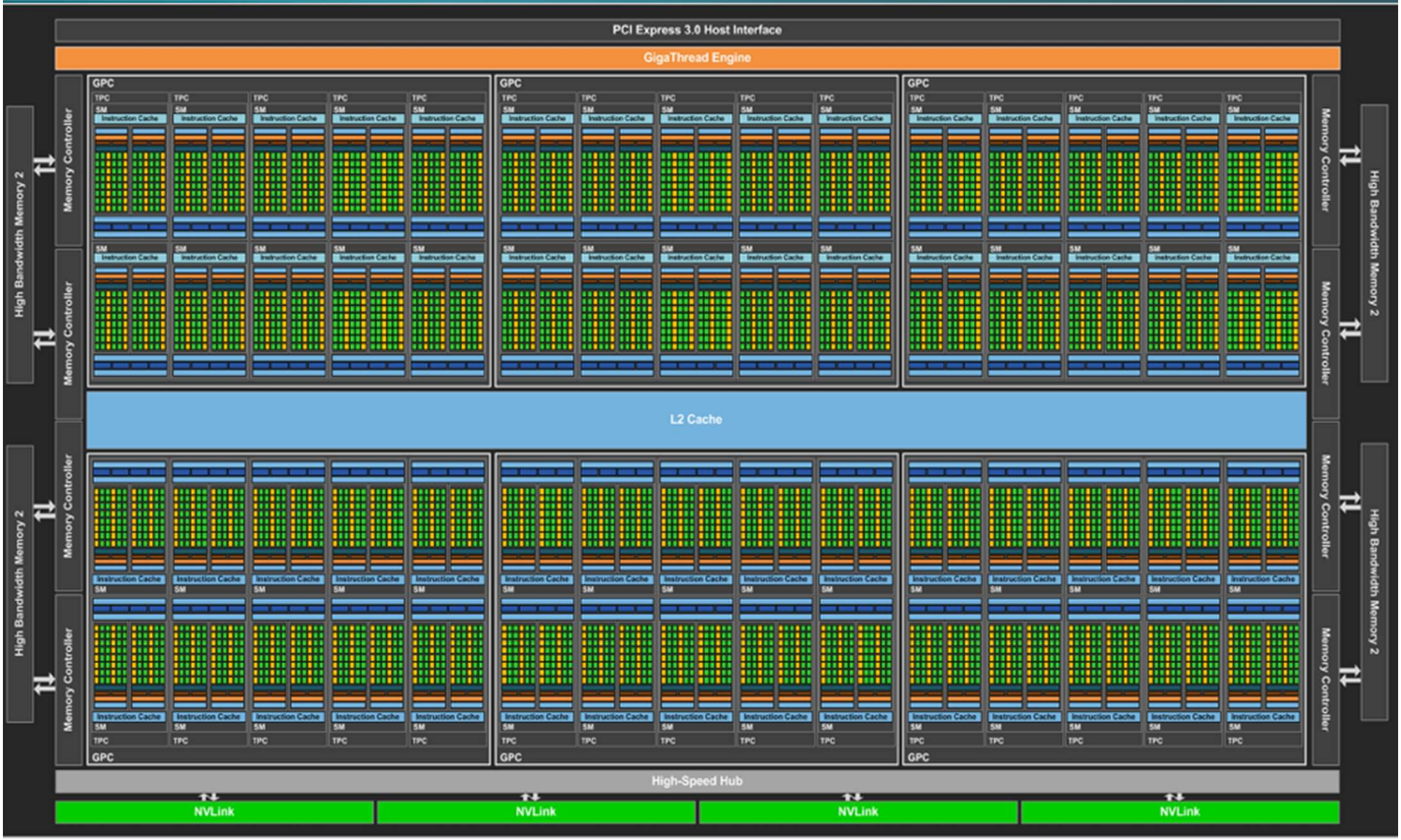
GPU	GeForce GTX 680 (Kepler)	GeForce GTX 980 (Maxwell)
SMs	8	16
CUDA Cores	1536	2048
Base Clock	1006 MHz	1126 MHz
GPU Boost Clock	1058 MHz	1216 MHz
GFLOPs	3090	4612[1]
Texture Units	128	128
Texel fill-rate	128.8 Gigatexels/sec	144.1 Gigatexels/sec
Memory Clock	6000 MHz	7000 MHz
Memory Bandwidth	192 GB/sec	224 GB/sec
ROPs	32	64
L2 Cache Size	512KB	2048KB
TDP	195 Watts	165 Watts
Transistors	3.54 billion	5.2 billion
Die Size	294 mm^2	398 mm^2
Manufacturing Process	28-nm	28-nm

NVIDIA Pascal Architecture

2016

GP100, ... (GTX 1080, Titan X *Pascal*...)

NVIDIA Pascal Architecture (2016)



NVIDIA Pascal SM

Multiprocessor: SM

- 64 CUDA cores
- 32 DP units

2 partitions inside SM

- 32 CUDA cores each; 16 DP units each; 8 LD/ST units each
- Each has its own warp scheduler, two dispatch units, register file

NVIDIA Pascal Architecture (2016)

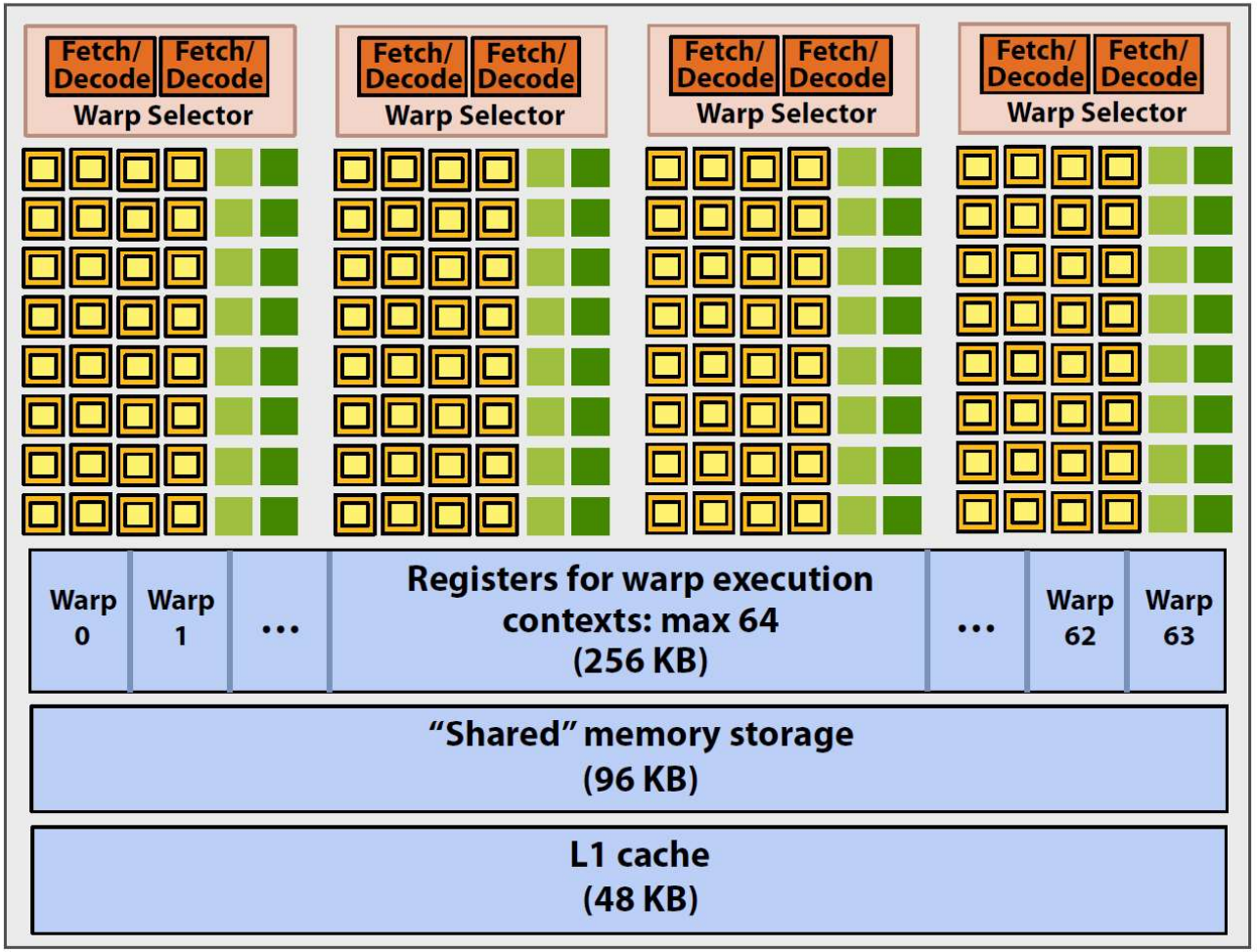
Total chip capacity on Tesla P100

- 56 SMs
  - 64 CUDA cores / SM = 3,584 CUDA cores in total
  - 32 DP units / SM = 1,792 DP units in total
- 28 TPCs (2 SMs per TPC)
- 6 GPCs

Maximum capacity would be 60 SMs and 30 TPCs

NVIDIA GTX 1080 (2016)

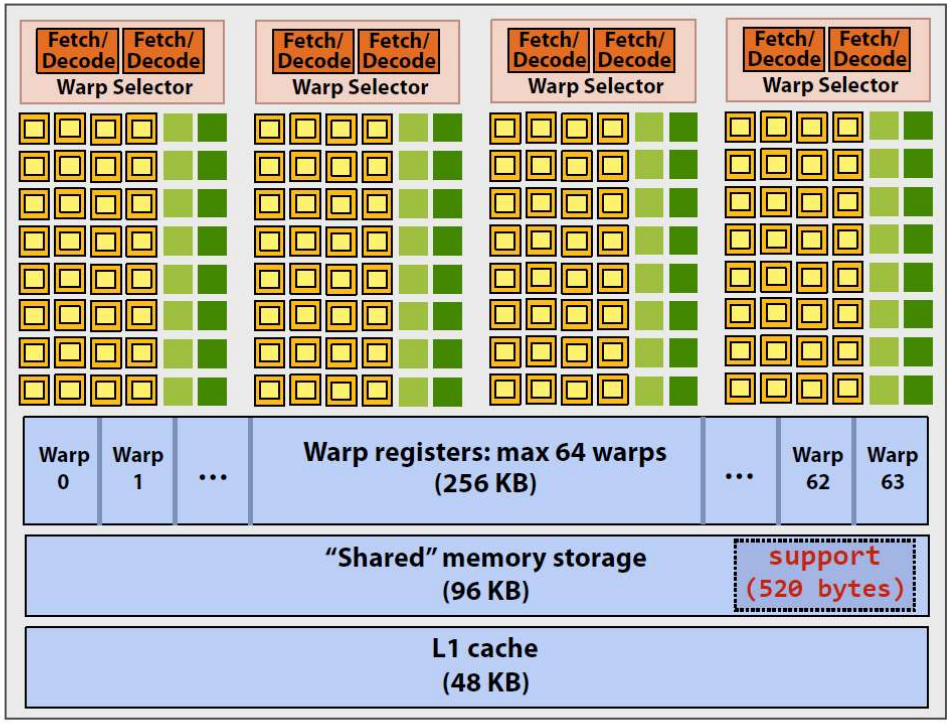
This is one NVIDIA Pascal GP104 streaming multi-processor (SM) unit



SM resource limits:

- Max warp execution contexts: 64 (2,048 total CUDA threads)
- 96 KB of shared memory

Running a single thread block on a SM "core"




```
#define THREADS_PER_BLK 128

__global__ void convolve(int N, float* input,
                         float* output)
{
   __shared__ float support[THREADS_PER_BLK+2];
   int index = blockIdx.x * blockDim.x +
               threadIdx.x;

   support[threadIdx.x] = input[index];
   if (threadIdx.x < 2) {
      support[THREADS_PER_BLK+threadIdx.x]
        = input[index+THREADS_PER_BLK];
   }

   __syncthreads();

   float result = 0.0f;  // thread-local
   for (int i=0; i<3; i++)
     result += support[threadIdx.x + i];

   output[index] = result;
}
```

Recall, CUDA kernels execute as SPMD programs

On NVIDIA GPUs groups of 32 CUDA threads share an instruction stream. These groups called "warps".

A `convolve` thread block is executed by 4 warps (4 warps x 32 threads/warp = 128 CUDA threads per block)

(Warps are an important GPU implementation detail, but not a CUDA abstraction!)

SM core operation each clock:

- **Select up to four runnable warps from 64 resident on SM core (thread-level parallelism)**
- **Select up to two runnable instructions per warp (instruction-level parallelism) ***

NVIDIA Volta Architecture

2017/2018

NVIDIA Volta Architecture (2017/2018)

NVIDIA Volta SM

Multiprocessor: SM

- 64 FP32 + INT32 cores
- 32 FP64 cores
- 8 tensor cores (FP16/FP32 mixed-precision)

4 partitions inside SM

- 16 FP32 + INT32 cores each
- 8 FP64 cores each
- 8 LD/ST units each
- 2 tensor cores each
- Each has: warp scheduler, dispatch unit, register file

NVIDIA Volta Architecture (2017/2018)

Total chip capacity on Tesla V100 (GV100 architecture)

- 80 SMs
  - 64 FP32 cores / SM = 5,120 FP32 cores in total
  - 64 INT32 cores / SM = 5,120 INT32 cores in total
  - 32 FP64 cores / SM = 2,560 FP64 cores in total
  - 4 FP16/FP32 mixed-prec. tensor cores = 650 tensor cores in total
- 40 TPCs (2 SMs per TPC)
- 6 GPCs

Maximum capacity would be 84 SMs and 42 TPCs

Kepler – Volta Specs

(repeated)

Tesla Product	Tesla K40	Tesla M40	Tesla P100	Tesla V100
GPU	GK180 (Kepler)	GM200 (Maxwell)	GP100 (Pascal)	GV100 (Volta)
SMs	15	24	56	80
TPCs	15	24	28	40
FP32 Cores / SM	192	128	64	64
FP32 Cores / GPU	2880	3072	3584	5120
FP64 Cores / SM	64	4	32	32
FP64 Cores / GPU	960	96	1792	2560
Tensor Cores / SM	NA	NA	NA	8
Tensor Cores / GPU	NA	NA	NA	640
GPU Boost Clock	810/875 MHz	1114 MHz	1480 MHz	1455 MHz
Peak FP32 TFLOP/s*	5.04	6.8	10.6	15
Peak FP64 TFLOP/s*	1.68	.21	5.3	7.5
Peak Tensor Core TFLOP/s*	NA	NA	NA	120
Texture Units	240	192	224	320
Memory Interface	384-bit GDDR5	384-bit GDDR5	4096-bit HBM2	4096-bit HBM2
Memory Size	Up to 12 GB	Up to 24 GB	16 GB	16 GB
L2 Cache Size	1536 KB	3072 KB	4096 KB	6144 KB
Shared Memory Size / SM	16 KB/32 KB/48 KB	96 KB	64 KB	Configurable up to 96 KB
Register File Size / SM	256 KB	256 KB	256 KB	256KB
Register File Size / GPU	3840 KB	6144 KB	14336 KB	20480 KB
TDP	235 Watts	250 Watts	300 Watts	300 Watts
Transistors	7.1 billion	8 billion	15.3 billion	21.1 billion
GPU Die Size	551 mm²	601 mm²	610 mm²	815 mm²
Manufacturing Process	28 nm	28 nm	16 nm FinFET+	12 nm FFN

Thank you.