

CS 380 - GPU and GPGPU Programming

Lecture 11: GPU Architecture 8

Markus Hadwiger, KAUST

Reading Assignment #6 (until Oct 12)



Read (required):

- Programming Massively Parallel Processors book,
2nd edition: Chapter 5 (*CUDA Memories*)
3rd edition: Chapter 4 (*Memory and Data Locality*)
- Look at the “Tuning Guides” for different architectures in the CUDA SDK

Read (optional):

- This week is NVIDIA GTC (Oct 5 – 9):

<https://www.nvidia.com/en-us/gtc/>

- NVIDIA Ampere (GA101) white paper:

<https://www.nvidia.com/content/dam/en-zz/Solutions/geforce/ampere/pdf/NVIDIA-ampere-GA102-GPU-Architecture-Whitepaper-V1.pdf>

- NVIDIA Ampere (A100) Tensor Core GPU white paper:

<https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf>

NVIDIA Architectures (since first CUDA GPU)



Tesla: 2007-2009

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

Fermi: 2010

- GF100, ... (GTX 480, ...)
- GF110, ... (GTX 580, ...)

Kepler: 2012

- GK104, ... (GTX 680, ...)
- GK110, ... (GTX 780, GTX Titan, ...)

Maxwell: 2015

- GM107, ... (GTX 750Ti, ...)
- GM204, ... (GTX 980, Titan X, ...)

Pascal: 2016

- GP100 (Tesla P100, ...)
- GP10x: x=2,4,6,7,8, ...
(GTX 1080, Titan X *Pascal*...)

Volta: 2017/2018

- GV100, ...
(Tesla V100, Titan V, ...)

Turing: 2018/2019

- TU102, TU104, TU106, TU116, ...
(Titan RTX, RTX 2070, 2080, 2080Ti, ...)

Ampere: 2020

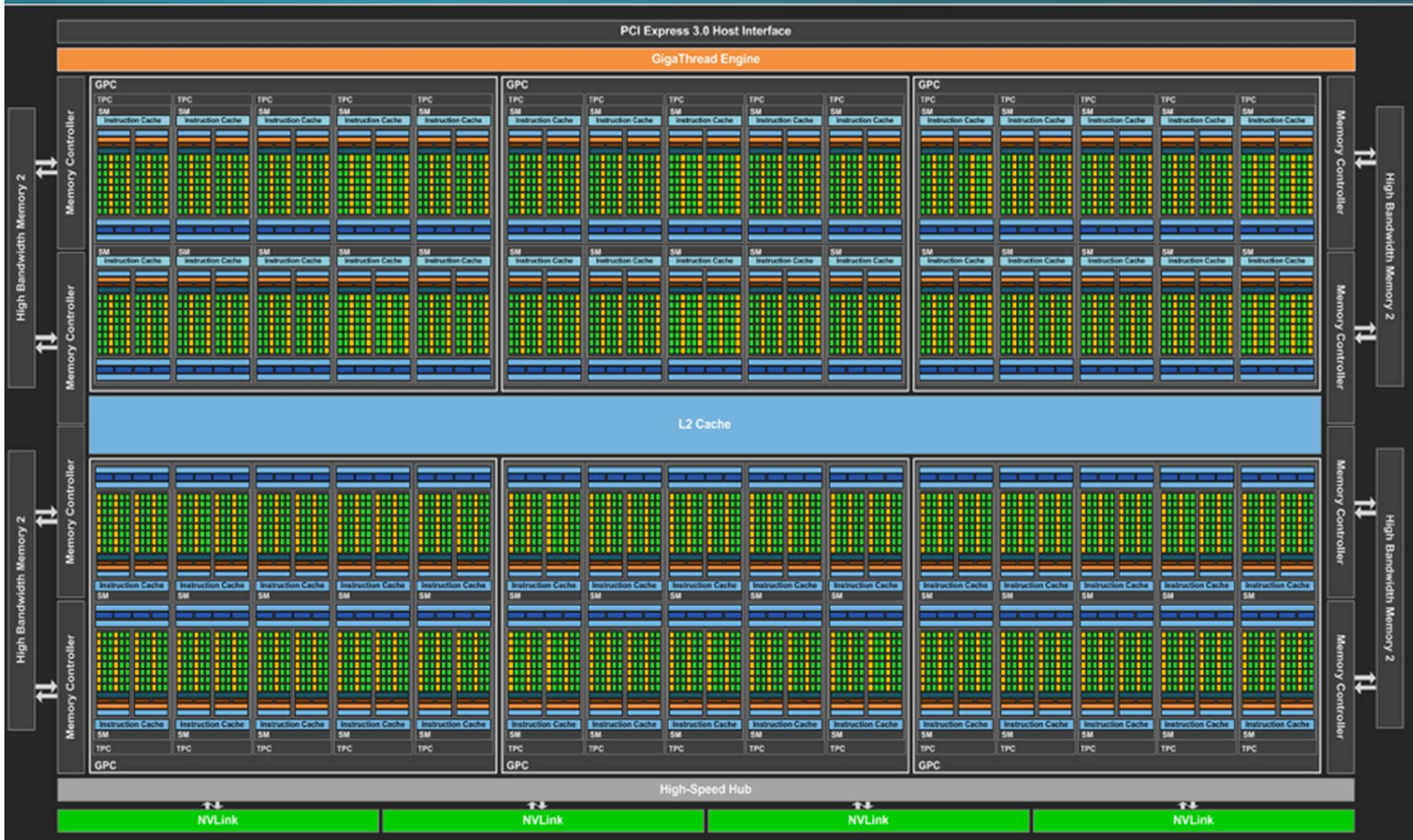
- GA100, GA102, GA104, ...
(A100, RTX 3070, 3080, 3090, ...)



NVIDIA Pascal Architecture 2016

GP100, GP10x: $x=2,4,6,7,8, \dots$
(GTX 1080, Titan X *Pascal*...)

NVIDIA Pascal Architecture (2016)



NVIDIA Pascal GP100 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

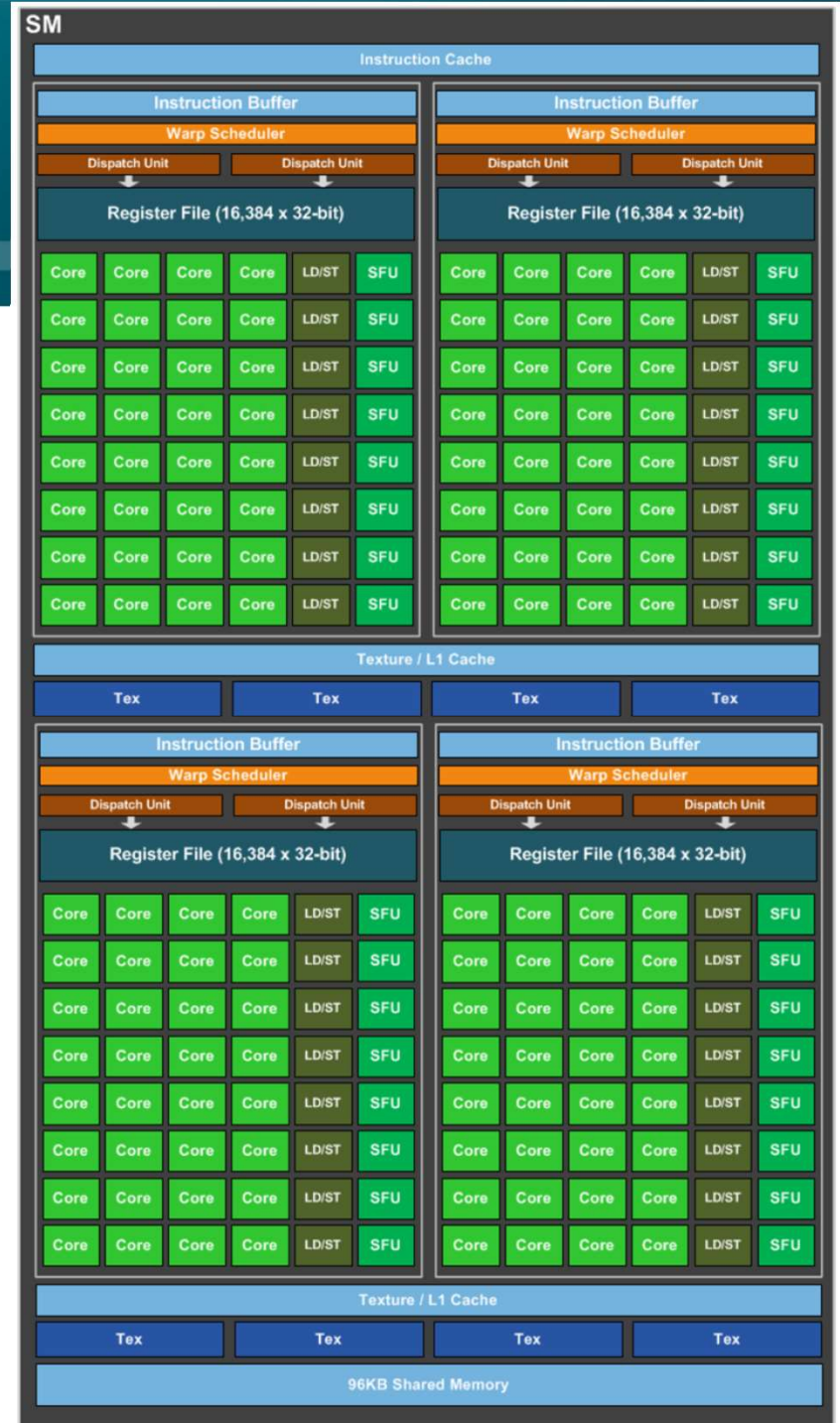
NVIDIA Pascal GP104 SM

Multiprocessor: SM

- 128 CUDA cores

4 partitions inside SM

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



NVIDIA Pascal Architecture (2016)



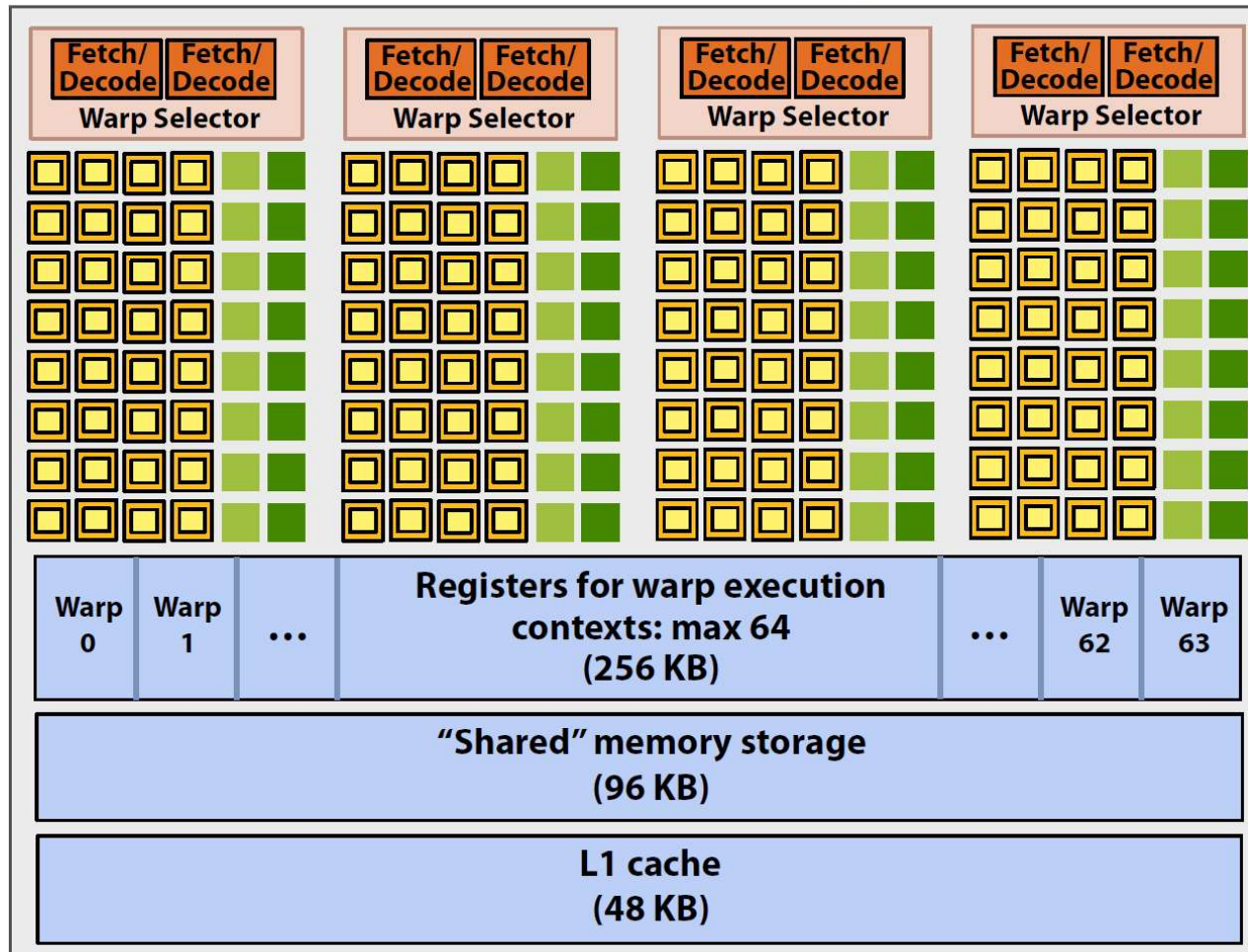
Total chip capacity on Tesla P100 (GP100)

- 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

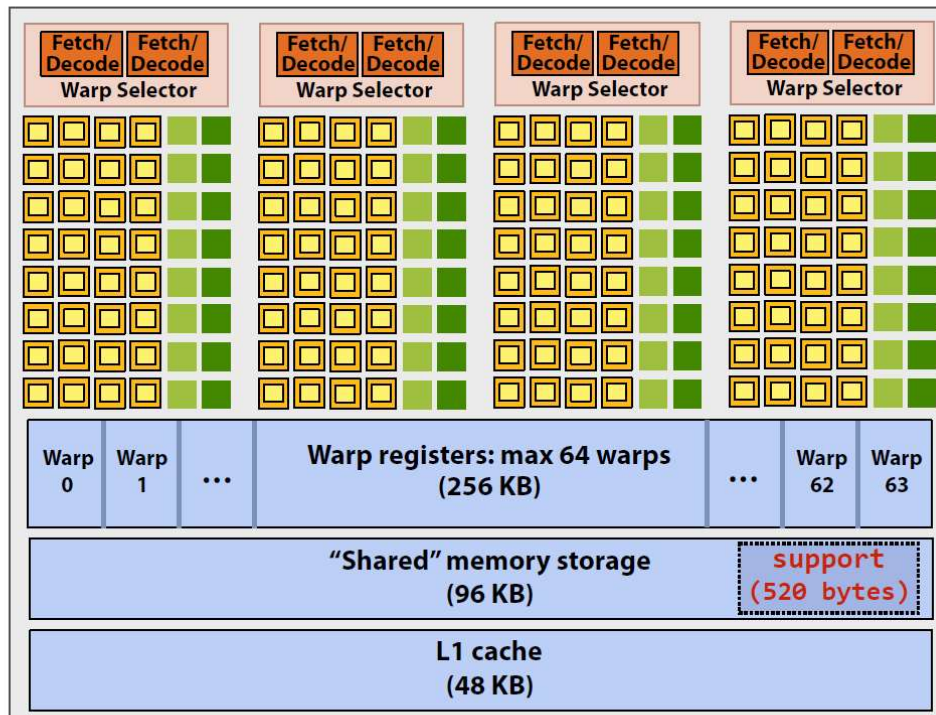


-  = SIMD functional unit, control shared across 32 units (1 MUL-ADD per clock)
-  = load/store
-  = SIMD special function unit (sin, cos, etc.)

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



Tensor Cores



Mixed-precision, fast matrix-matrix multiply and accumulate

$$\mathbf{D} = \begin{pmatrix} \begin{matrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{matrix} & \begin{matrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{matrix} & \begin{matrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{matrix} \end{pmatrix}$$

FP16 or FP32 FP16 FP16 FP16 or FP32

From this, build larger sizes, higher dimensionalities, ...

[+Tensor cores on later architectures add more data types/precisions!]

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

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

Turing (vs. Pascal)



Apart from RT cores, Volta and Turing are very similar
(and both have compute capability 7.x: Volta: 7.0, Turing: 7.5)

GPU Features	GeForce GTX 1080	GeForce RTX 2080	Quadro P5000	Quadro RTX 5000
Architecture	Pascal	Turing	Pascal	Turing
GPCs	4	6	4	6
TPCs	20	23	20	24
SMs	20	46	20	48
CUDA Cores / SM	128	64	128	64
CUDA Cores / GPU	2560	2944	2560	3072
Tensor Cores / SM	NA	8	NA	8
Tensor Cores / GPU	NA	368	NA	384
RT Cores	NA	46	NA	48

TU104

TU104



NVIDIA Turing Architecture

2018/2019

TU102, TU104, TU106, TU116, ...
(RTX 2070, 2080, 2080Ti, Tesla T4, ...)

NVIDIA Turing Architecture (2018/2019)



TU 102

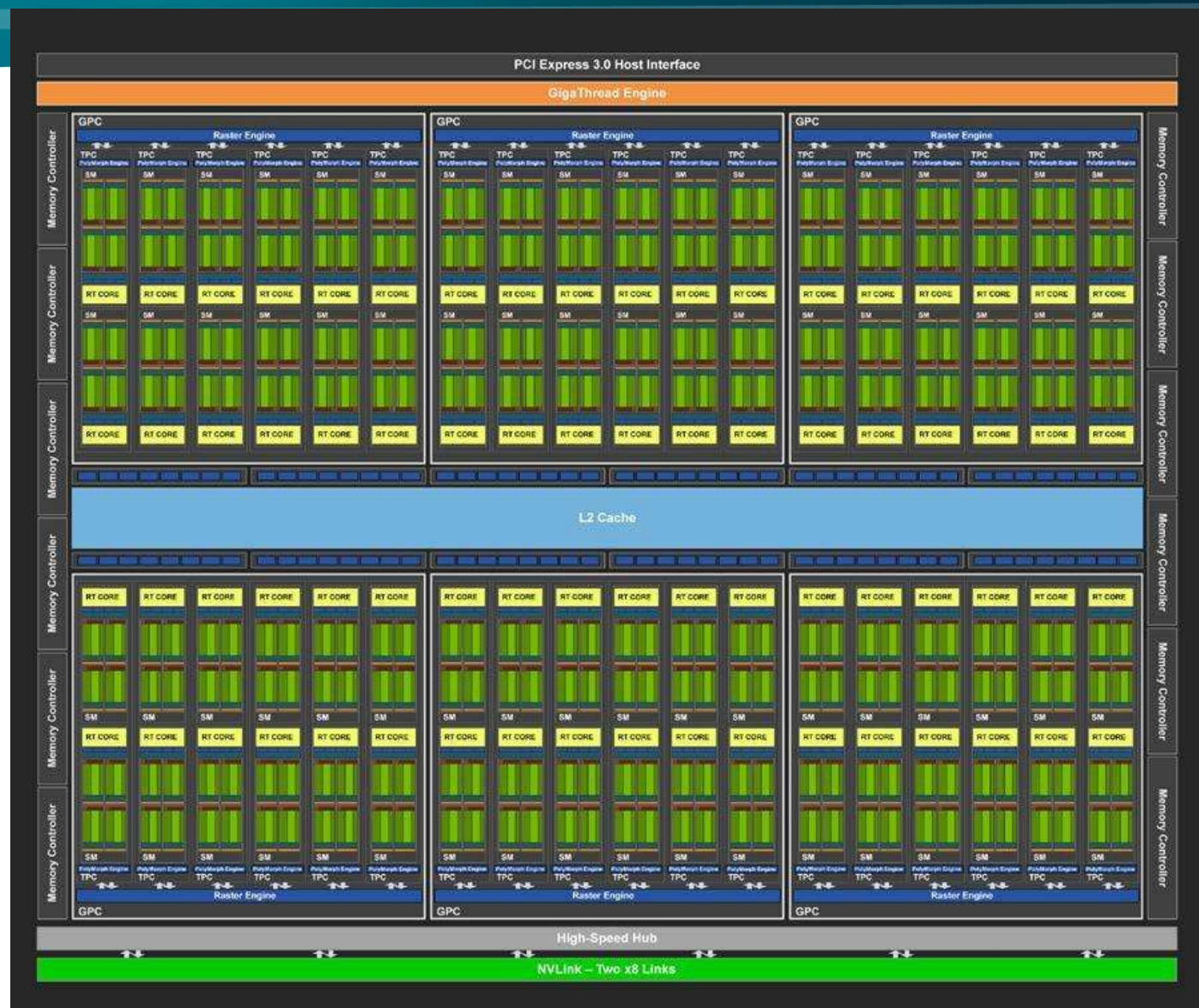
(Geforce:

RTX 2080 Ti,

Quadro:

RTX 6000,

RTX 8000, ...)



NVIDIA Turing Architecture (2018/2019)



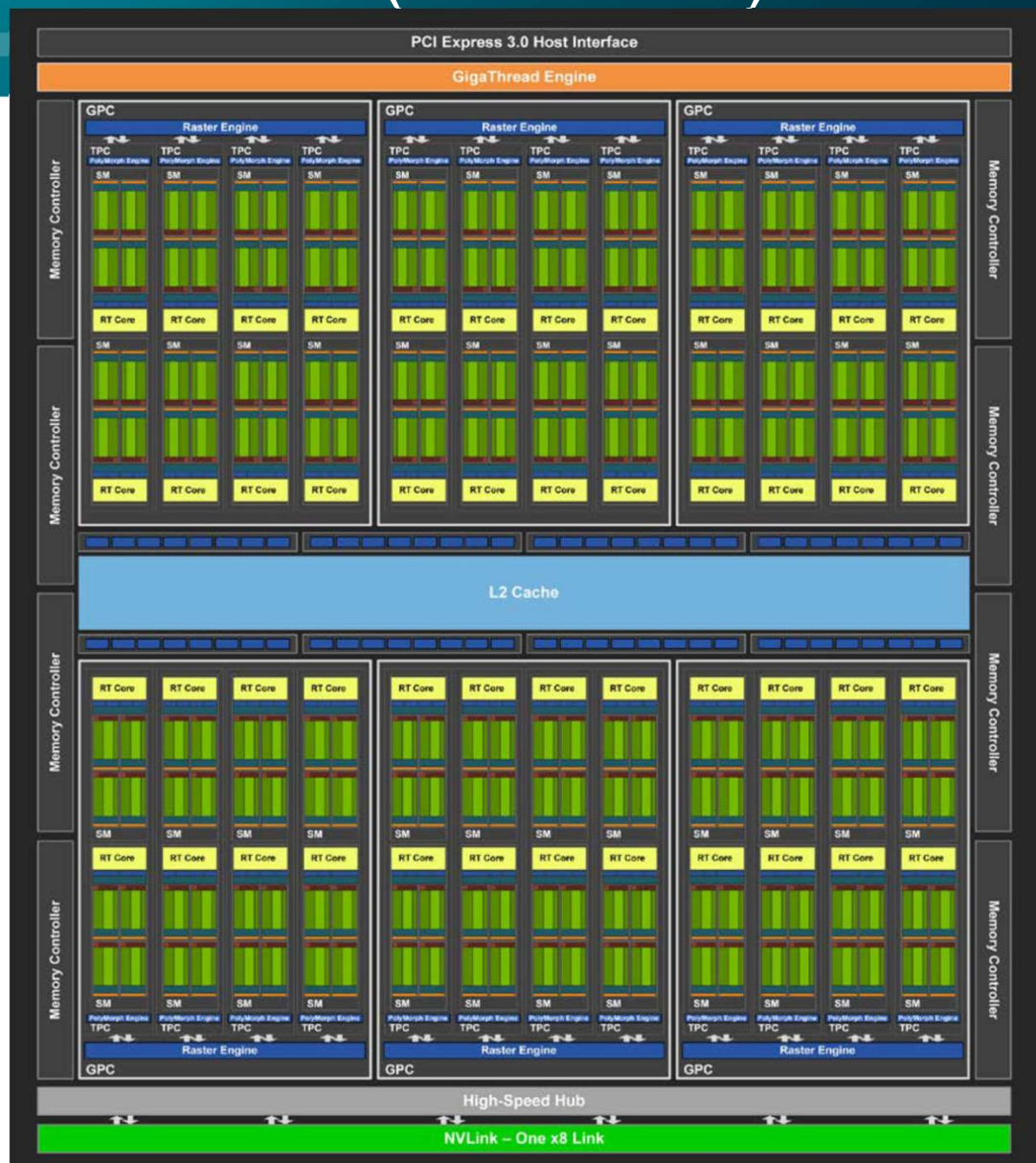
TU 104

(Geforce:

RTX 2080,

Quadro:

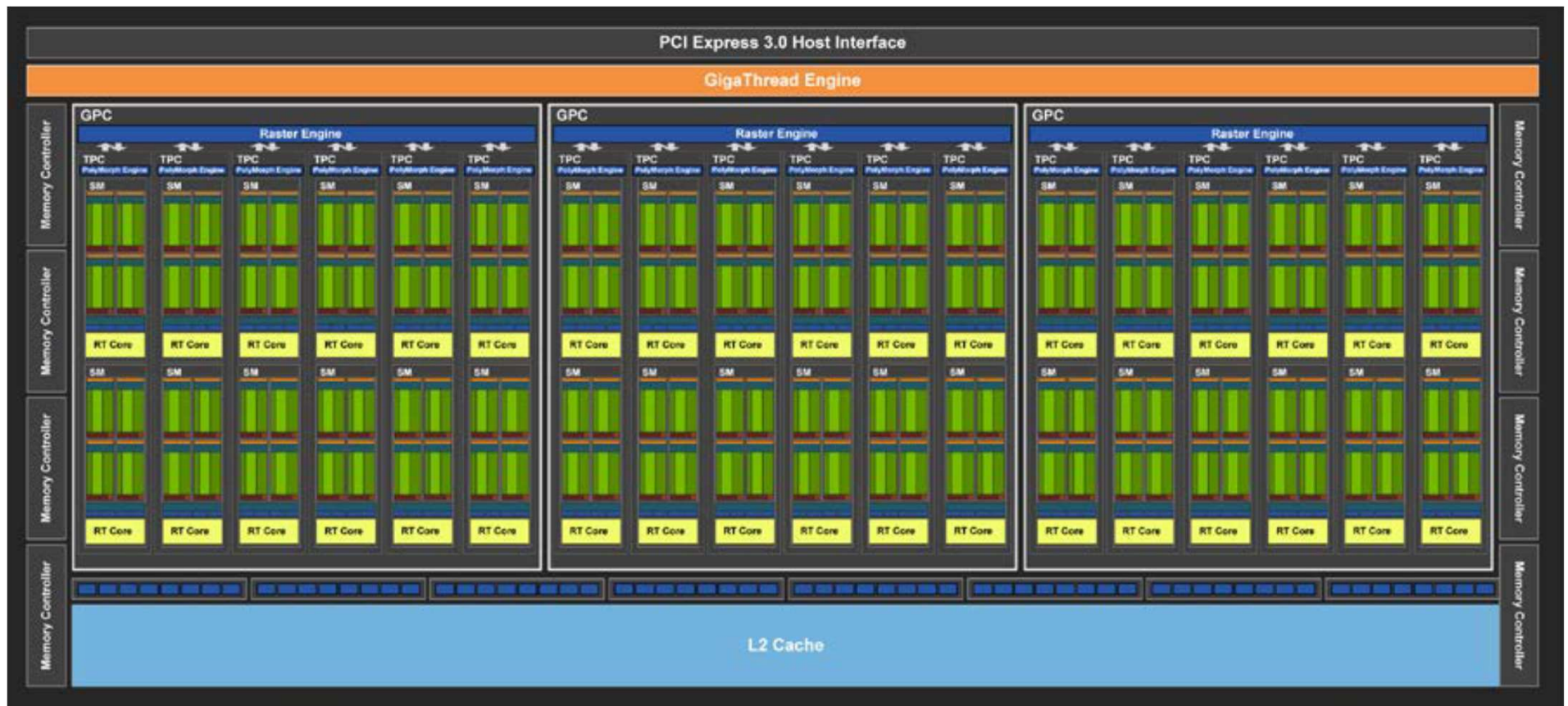
RTX 5000, ...)



NVIDIA Turing Architecture (2018/2019)



TU 106 (Geforce RTX 2070, ...)



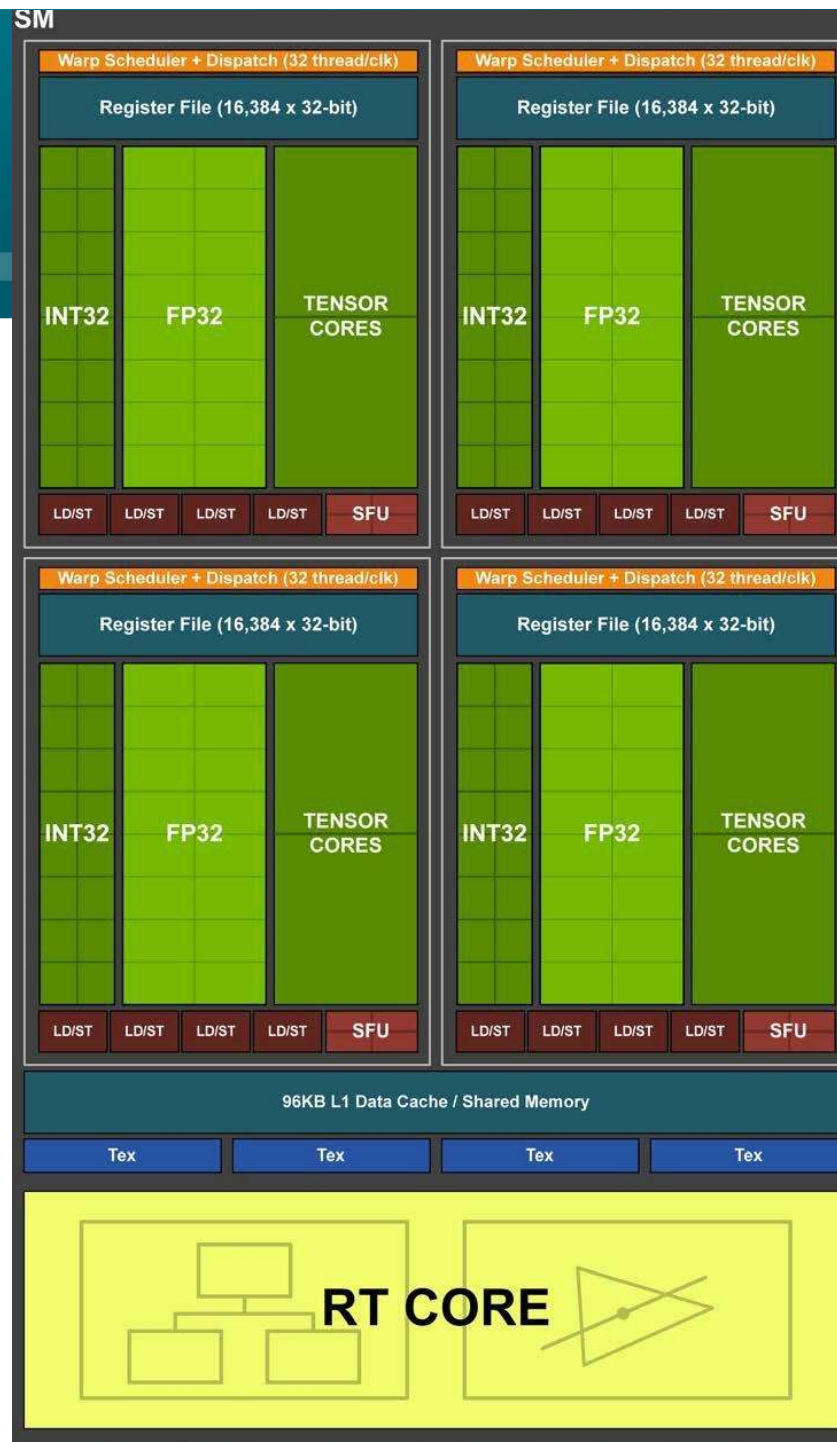
NVIDIA Turing SM

Multiprocessor: SM

- 64 FP32 + INT32 cores
- 2 (!) FP64 cores
- 8 Turing tensor cores (FP16/32, INT4/8 mixed-precision)
- 1 RT (ray tracing) core

4 partitions inside SM

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





NVIDIA Ampere Architecture 2020

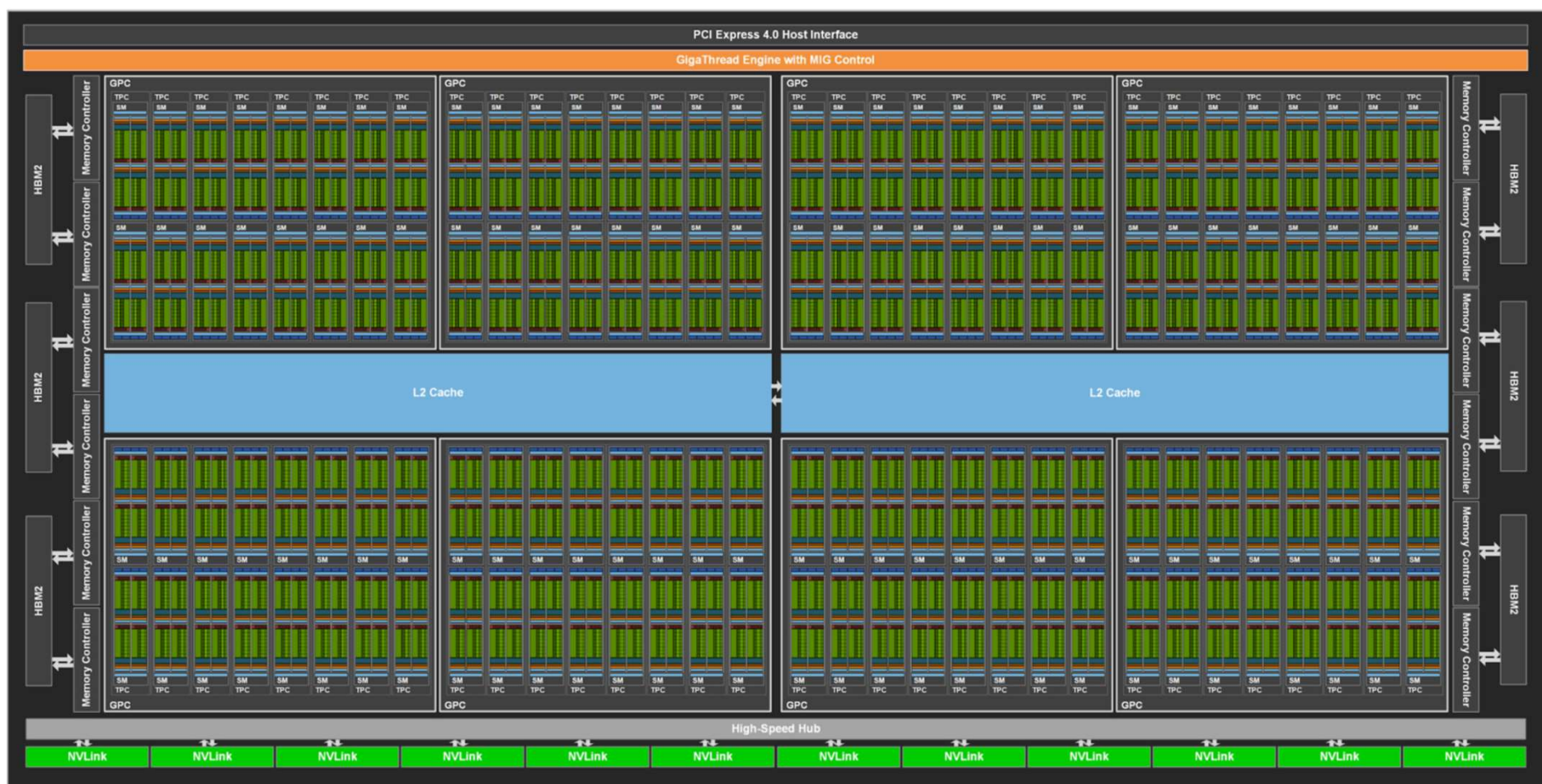
GA100, GA102, GA104, ...
(A100, RTX 3070, RTX 3080, RTX 3090, ...)

NVIDIA Ampere GA100 Architecture (2020)



GA 100 (A100 Tensor Core GPU)

Full GPU: 128 SMs (in 8 GPCs/64 TPCs)



NVIDIA GA100 SM

Multiprocessor: SM

- 64 FP32 + 64 INT32 cores
- 32 FP64 cores
- 4 3rd gen tensor cores
- 1 2nd gen RT (ray tracing) core

4 partitions inside SM

- 16 FP32 + 16 INT32 cores
- 8 FP64 cores
- 8 LD/ST units each
- 1 3rd gen tensor core each
- Each has: warp scheduler, dispatch unit, 16K register file



NVIDIA Ampere GA10x Architecture (2020)



GA 102 (RTX 3070, 3080, 3090)

Full GPU: 84 SMs (in 7 GPCs/42 TPCs)



NVIDIA GA10x SM

Multiprocessor: SM

- 128 (64+64) FP32 + 64 INT32 cores
- 2 (!) FP64 cores
- 4 3rd gen tensor cores
- 1 2nd gen RT (ray tracing) core

4 partitions inside SM

- 16+16 FP32 + 16 INT32 cores
- 4 LD/ST units each
- 1 3rd gen tensor core each
- Each has: warp scheduler, dispatch unit, 16K register file



Turing vs. Ampere GA102



Graphics Card	GeForce RTX 2080 Founders Edition	GeForce RTX 2080 Super Founders Edition	GeForce RTX 3080 10 GB Founders Edition
GPU Codename	TU104	TU104	GA102
GPU Architecture	NVIDIA Turing	NVIDIA Turing	NVIDIA Ampere
GPCs	6	6	6
TPCs	23	24	34
SMs	46	48	68
CUDA Cores / SM	64	64	128
CUDA Cores / GPU	2944	3072	8704
Tensor Cores / SM	8 (2nd Gen)	8 (2nd Gen)	4 (3rd Gen)
Tensor Cores / GPU	368	384 (2nd Gen)	272 (3rd Gen)
RT Cores	46 (1st Gen)	48 (1st Gen)	68 (2nd Gen)
GPU Boost Clock (MHz)	1800	1815	1710
Peak FP32 TFLOPS (non-Tensor) ¹	10.6	11.2	29.8
Peak FP16 TFLOPS (non-Tensor) ¹	21.2	22.3	29.8
Peak BF16 TFLOPS (non-Tensor) ¹	NA	NA	29.8
Peak INT32 TOPS (non-Tensor) ^{1,3}	10.6	11.2	14.9

Turing vs. Ampere GA102



Peak FP16 Tensor TFLOPS with FP16 Accumulate¹	84.8	89.2	119/238 ²
Peak FP16 Tensor TFLOPS with FP32 Accumulate¹	42.4	44.6	59.5/119 ²
Peak BF16 Tensor TFLOPS with FP32 Accumulate¹	NA	NA	59.5/119 ²
Peak TF32 Tensor TFLOPS¹	NA	NA	29.8/59.5 ²
Peak INT8 Tensor TOPS¹	169.6	178.4	238/476 ²
Peak INT4 Tensor TOPS¹	339.1	356.8	476/952 ²
Frame Buffer Memory Size and Type	8192 MB GDDR6	8192 MB GDDR6	10240 MB GDDR6X
Memory Interface	256-bit	256-bit	320-bit
Memory Clock (Data Rate)	14 Gbps	15.5 Gbps	19 Gbps
Memory Bandwidth	448 GB/sec	496 GB/sec	760 GB/sec
ROPs	64	64	96
Pixel Fill-rate (Gigapixels/sec)	115.2	116.2	164.2
Texture Units	184	192	272
Texel Fill-rate (Gigatexels/sec)	331.2	348.5	465
L1 Data Cache/Shared Memory	4416 KB	4608 KB	8704 KB

Turing vs. Ampere GA102



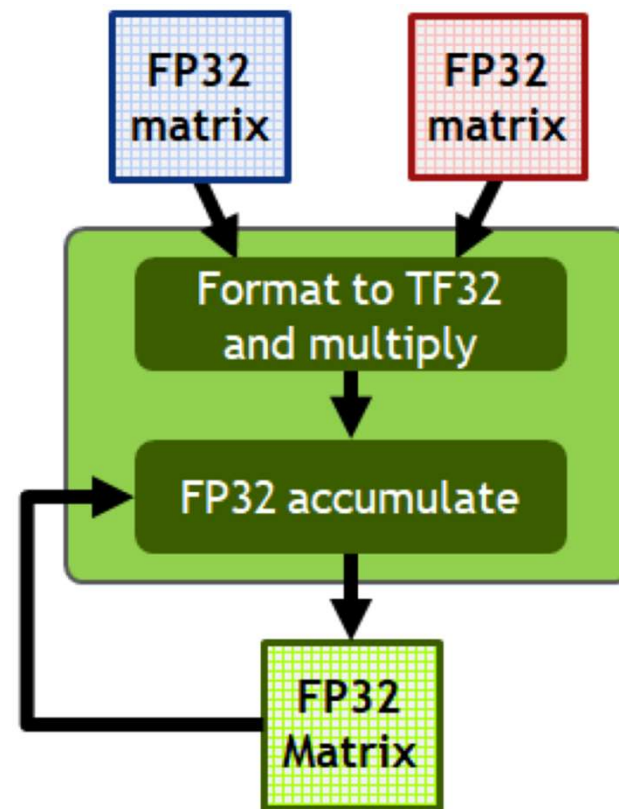
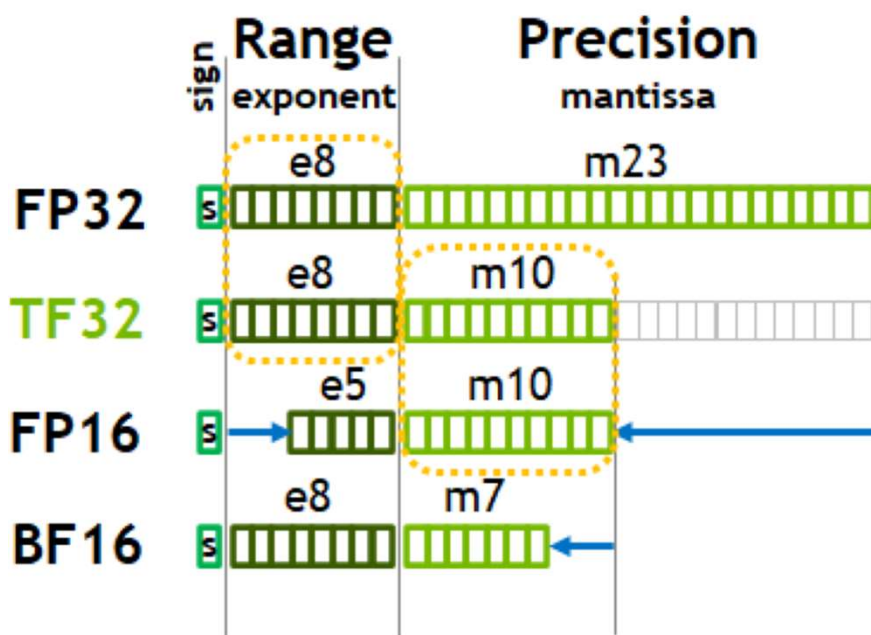
L2 Cache Size	4096 KB	4096 KB	5120 KB
Register File Size	11776 KB	12288 KB	17408 KB
TGP (Total Graphics Power)	225 W	250 W	320W
Transistor Count	13.6 Billion	13.6 Billion	28.3 Billion
Die Size	545 mm ²	545 mm ²	628.4 mm ²
Manufacturing Process	TSMC 12 nm FFN (FinFET NVIDIA)	TSMC 12 nm FFN (FinFET NVIDIA)	Samsung 8 nm 8N NVIDIA Custom Process

1. Peak rates are based on GPU Boost Clock.
2. Effective TOPS / TFLOPS using the new Sparsity Feature
3. TOPS = IMAD-based integer math

Tensor Cores: Many Mixed Precision Options



New in Ampere: TF32, BF16, FP64



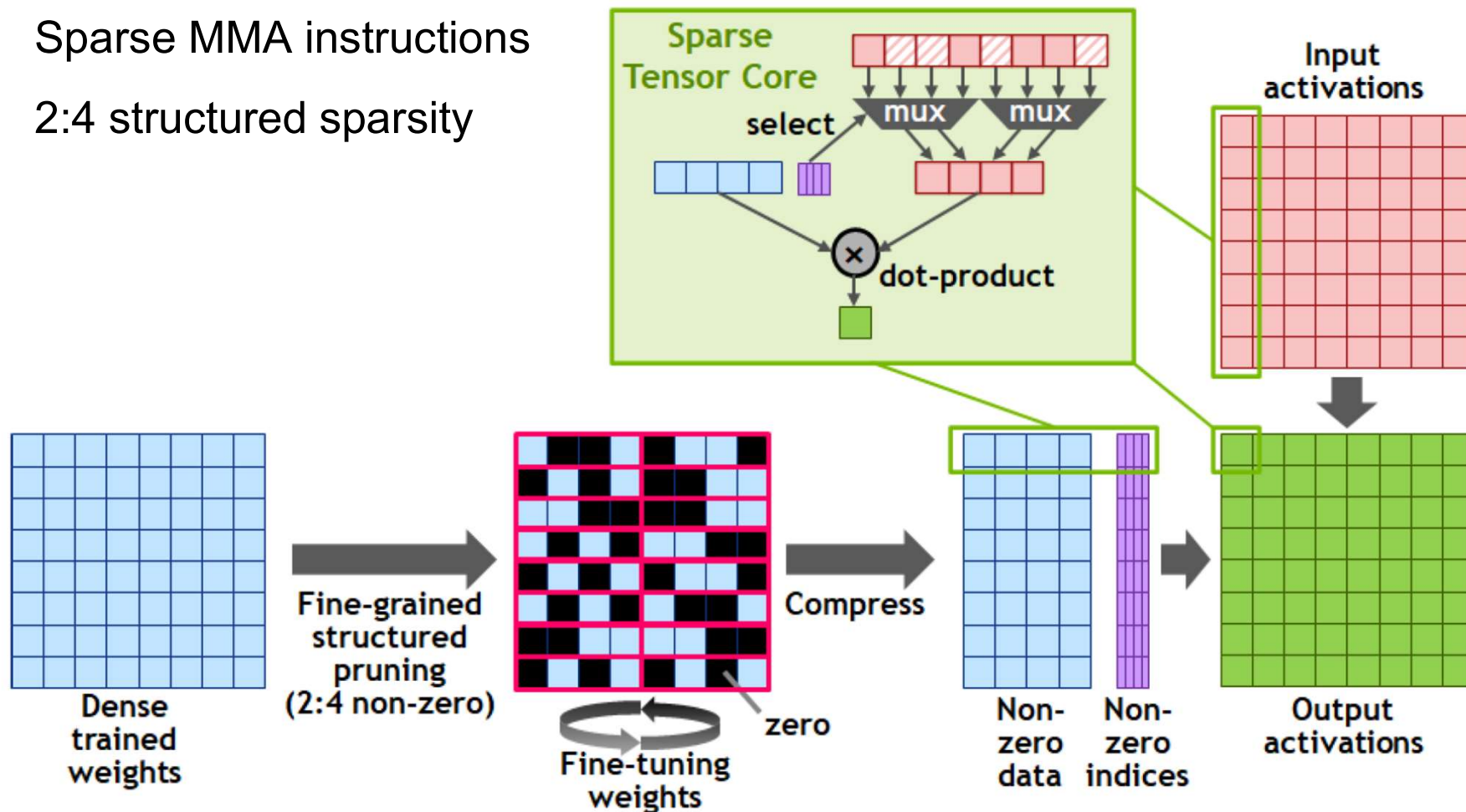
plus FP64 (new in Ampere; GA100 only)

plus INT4/INT8/binary data types (already introduced in Turing)

Tensor Cores: Sparsity Support



Sparse MMA instructions
2:4 structured sparsity





CUDA

Compute Capabilities

Compute Capab. – 2.0

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

Feature Support (Unlisted features are supported for all compute capabilities)	Compute Capability				
	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
<code>_ballot()</code> (Section B.11)					
<code>_threadfence_system()</code> (Section B.5)					
<code>_syncthreads_count()</code> , <code>_syncthreads_and()</code> , <code>_syncthreads_or()</code> (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	16K	16K	16K
	48K	48K	32K	32K
			48K	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

Compute Capabilities – 8.0 (Ampere)



GPU Features	NVIDIA Tesla P100	NVIDIA Tesla V100	NVIDIA A100
GPU Codename	GP100	GV100	GA100
GPU Architecture	NVIDIA Pascal	NVIDIA Volta	NVIDIA Ampere
Compute Capability	6.0	7.0	8.0
Threads / Warp	32	32	32
Max Warps / SM	64	64	64
Max Threads / SM	2048	2048	2048
Max Thread Blocks / SM	32	32	32
Max 32-bit Registers / SM	65536	65536	65536
Max Registers / Block	65536	65536	65536
Max Registers / Thread	255	255	255
Max Thread Block Size	1024	1024	1024
FP32 Cores / SM	64	64	64
Ratio of SM Registers to FP32 Cores	1024	1024	1024
Shared Memory Size / SM	64 KB	Configurable up to 96 KB	Configurable up to 164 KB

Thank you.