

# CS 380 - GPU and GPGPU Programming

## Lecture 10: GPU Architecture, Pt. 8

Markus Hadwiger, KAUST

# Reading Assignment #5 (until Sep 30)



## Read (required):

- Programming Massively Parallel Processors book (4th edition),  
**Chapter 2** (*Heterogeneous data parallel computing*)
- CUDA NVCC documentation:  
[https://docs.nvidia.com/cuda/pdf/CUDA\\_Compiler\\_Driver\\_NVCC.pdf](https://docs.nvidia.com/cuda/pdf/CUDA_Compiler_Driver_NVCC.pdf)  
Read Chapters 1 – 3; Chapter 6 (GPU Compilation); get an overview of the rest



# Quiz #1

## Organization

- First 30 min of lecture
- No material (book, notes, ...) allowed

## Content of questions

- Lectures (both actual lectures and slides)
- Reading assignments
- Programming assignments (algorithms, methods)
- Solve short practical examples

# GPU Architecture: Real Architectures

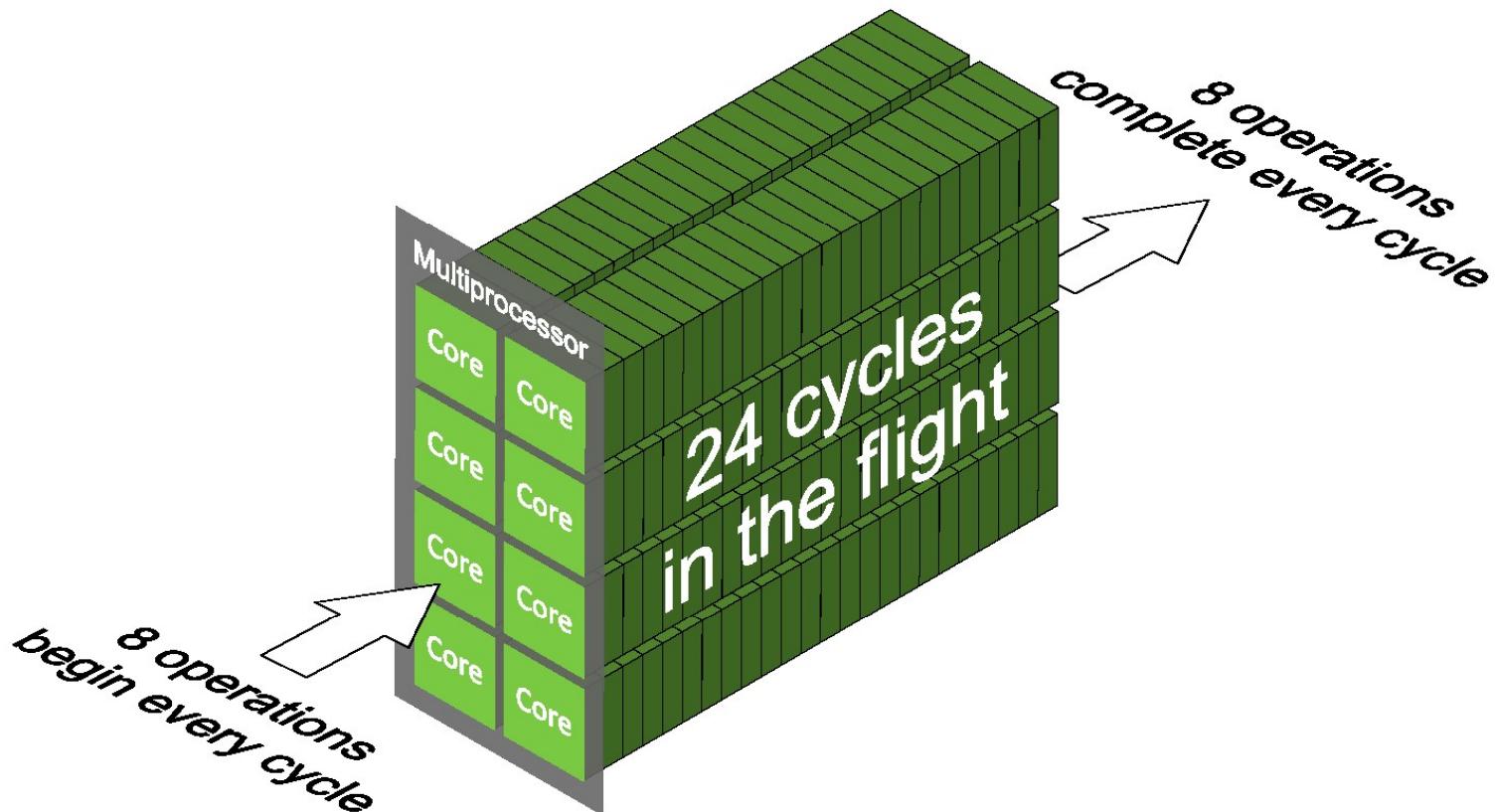
# ALU Instruction Latencies and Instructs. / SM



| CC                                 | 2.0<br>(Fermi)       | 2.1<br>(Fermi)       | 3.x<br>(Kepler)     | 5.x<br>(Maxwell) | 6.0<br>(Pascal) | 6.1/6.2<br>(Pascal) | 7.x<br>(Volta, Turing) | 8.0/8.6<br>(Ampere) | 8.9/9.0<br>(Ada/Hopper) |
|------------------------------------|----------------------|----------------------|---------------------|------------------|-----------------|---------------------|------------------------|---------------------|-------------------------|
| # warp sched.<br>/ SM              | 2                    | 2                    | 4                   | 4                | 2               | 4                   | 4                      | 4                   | 4                       |
| # ALU<br>dispatch<br>/ warp sched. | 1 (over<br>2 clocks) | 2 (over<br>2 clocks) | 2                   | 1                | 1               | 1                   | 1                      | 1                   | 1                       |
| SM busy with<br># warps + inst     | L                    | 2L                   | 8L                  | 4L               | 2L              | 4L                  | 4L                     | 4L                  | 4L                      |
| inst. pipe<br>latency (L)          | 22                   | 22                   | 11                  | 9                | 6               | 6                   | 4                      | 4                   | 4                       |
| <b>SM busy with<br/># warps</b>    | <b>22</b>            | <b>22<br/>+ ILP</b>  | <b>44<br/>+ ILP</b> | <b>36</b>        | <b>12</b>       | <b>24</b>           | <b>16</b>              | <b>16</b>           | <b>16</b>               |

see NVIDIA CUDA C Programming Guides (different versions)  
performance guidelines/multiprocessor level; compute capabilities

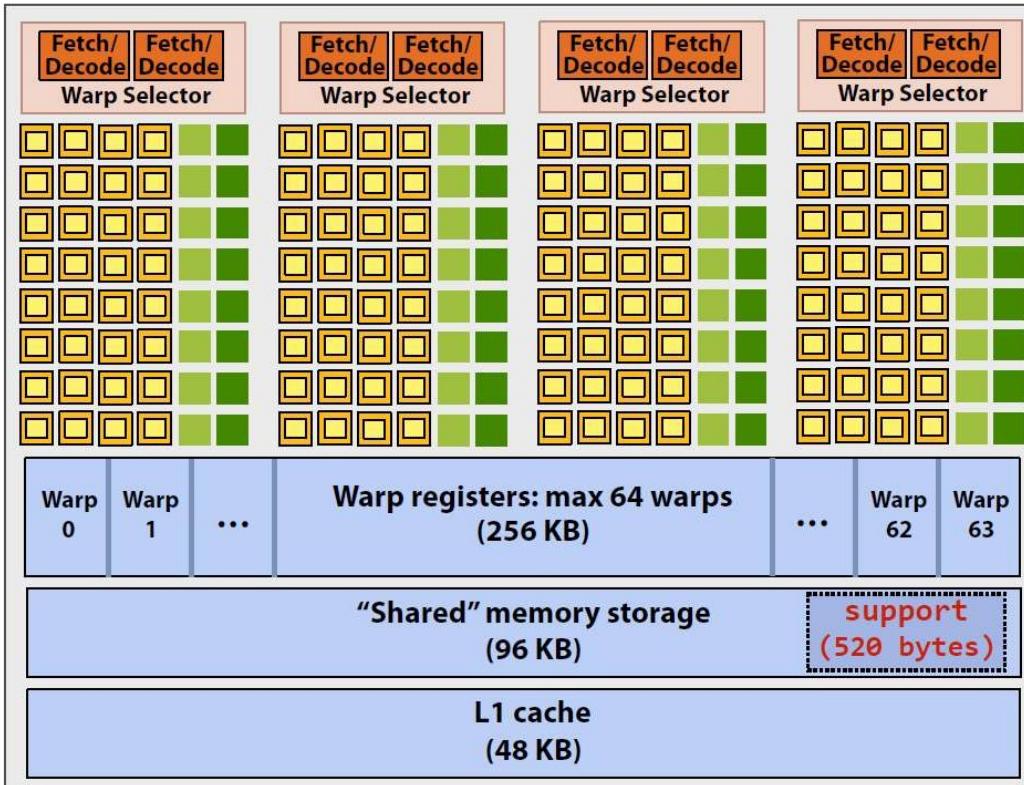
# Use Little's law



Needed parallelism = Latency × Throughput

courtesy of Vasily Volkov

# 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

(compute capability 7.0/7.2)

GV100 (cc 7.0), ... (Titan V, Tesla V100, ...)

GV10B, GV11B (cc 7.2), ... (Tegra Xavier, ...)

# NVIDIA Volta Architecture (2017/2018)





# Instruction Throughput

Instruction throughput numbers in CUDA C Programming Guide (Chapter 8.4)

|   | Compute Capability                        |             |     |     |     |                                   |                       |     |                       |     |     |
|---|---|-------------|-----|-----|-----|-----------------------------------|-----------------------|-----|-----------------------|-----|-----|
|   | 3.5,<br>3.7                               | 5.0,<br>5.2 | 5.3 | 6.0 | 6.1 | 6.2                               | 7.x                   | 8.0 | 8.6                   | 8.9 | 9.0 |
| 16-bit floating-point add, multiply, multiply-add | N/A                                       |             | 256 | 128 | 2   | 256                               | 128                   | 256 | 128                   | 256 |     |
|   |   |             |     |     |     |                                   | 128 for __nv_bfloat16 |     | 128 for __nv_bfloat16 |     |     |
| 32-bit floating-point add, multiply, multiply-add | 192                                       | 128         |     | 64  | 128 |                                   | 64                    |     | 128                   |     |     |
| 64-bit floating-point add, multiply, multiply-add | 64  |             | 4   | 32  | 4   | 32                                | 32                    | 2   | 2                     | 64  |     |
|   | 8 for GeForce GPUs, except for Titan GPUs |             |     |     |     | 2 for compute capability 7.5 GPUs |                       |     |                       |     |     |

# ALU Instruction Latencies and Instructs. / SM



| CC                                 | 2.0<br>(Fermi)       | 2.1<br>(Fermi)       | 3.x<br>(Kepler) | 5.x<br>(Maxwell) | 6.0<br>(Pascal) | 6.1/6.2<br>(Pascal) | 7.x<br>(Volta, Turing) | 8.0/8.6<br>(Ampere) | 8.9/9.0<br>(Ada/Hopper) |
|------------------------------------|----------------------|----------------------|-----------------|------------------|-----------------|---------------------|------------------------|---------------------|-------------------------|
| # warp sched.<br>/ SM              | 2                    | 2                    | 4               | 4                | 2               | 4                   | 4                      | 4                   | 4                       |
| # ALU<br>dispatch<br>/ warp sched. | 1 (over<br>2 clocks) | 2 (over<br>2 clocks) | 2               | 1                | 1               | 1                   | 1                      | 1                   | 1                       |
| SM busy with<br># warps + inst     | L                    | 2L                   | 8L              | 4L               | 2L              | 4L                  | 4L                     | 4L                  | 4L                      |
| inst. pipe<br>latency (L)          | 22                   | 22                   | 11              | 9                | 6               | 6                   | 4                      | 4                   | 4                       |
| SM busy with<br># warps            | 22                   | 22<br>+ ILP          | 44<br>+ ILP     | 36               | 12              | 24                  | 16                     | 16                  | 16                      |

see NVIDIA CUDA C Programming Guides (different versions)  
 performance guidelines/multiprocessor level; compute capabilities

# NVIDIA Volta SM

## Multiprocessor: SM (CC 7.0)

- 64 FP32 + 64 INT32 cores
- 32 FP64 cores
- 32 LD/ST units; 16 SFUs
- 8 tensor cores  
(FP16/FP32 mixed-precision)

## 4 partitions inside SM

- 16 FP32 + 16 INT32 cores each
- 8 FP64 cores each
- 8 LD/ST units; 4 SFUs each
- 2 tensor cores each
- Each has: warp scheduler, dispatch unit, register file





# Tensor Cores

Mixed-precision, fast matrix-matrix multiply and accumulate

$$\mathbf{D} = \left( \begin{array}{cccc} \mathbf{A}_{0,0} & \mathbf{A}_{0,1} & \mathbf{A}_{0,2} & \mathbf{A}_{0,3} \\ \mathbf{A}_{1,0} & \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & \mathbf{A}_{1,3} \\ \mathbf{A}_{2,0} & \mathbf{A}_{2,1} & \mathbf{A}_{2,2} & \mathbf{A}_{2,3} \\ \mathbf{A}_{3,0} & \mathbf{A}_{3,1} & \mathbf{A}_{3,2} & \mathbf{A}_{3,3} \end{array} \right) \text{FP16 or FP32} \times \left( \begin{array}{cccc} \mathbf{B}_{0,0} & \mathbf{B}_{0,1} & \mathbf{B}_{0,2} & \mathbf{B}_{0,3} \\ \mathbf{B}_{1,0} & \mathbf{B}_{1,1} & \mathbf{B}_{1,2} & \mathbf{B}_{1,3} \\ \mathbf{B}_{2,0} & \mathbf{B}_{2,1} & \mathbf{B}_{2,2} & \mathbf{B}_{2,3} \\ \mathbf{B}_{3,0} & \mathbf{B}_{3,1} & \mathbf{B}_{3,2} & \mathbf{B}_{3,3} \end{array} \right) \text{FP16} + \left( \begin{array}{cccc} \mathbf{C}_{0,0} & \mathbf{C}_{0,1} & \mathbf{C}_{0,2} & \mathbf{C}_{0,3} \\ \mathbf{C}_{1,0} & \mathbf{C}_{1,1} & \mathbf{C}_{1,2} & \mathbf{C}_{1,3} \\ \mathbf{C}_{2,0} & \mathbf{C}_{2,1} & \mathbf{C}_{2,2} & \mathbf{C}_{2,3} \\ \mathbf{C}_{3,0} & \mathbf{C}_{3,1} & \mathbf{C}_{3,2} & \mathbf{C}_{3,3} \end{array} \right) \text{FP16 or FP32}$$

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<br>(Maxwell)  | GP100<br>(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 <sup>*</sup>           | 5.04                 | 6.8                 | 10.6                | 15                          |
| Peak FP64 TFLOP/s <sup>*</sup>           | 1.68                 | .21                 | 5.3                 | 7.5                         |
| Peak Tensor Core<br>TFLOP/s <sup>*</sup> | NA                   | NA                  | NA                  | 120                         |
| Texture Units                            | 240                  | 192                 | 224                 | 320                         |
| Memory Interface                         | 384-bit GDDR5        | 384-bit GDDR5       | 4096-bit<br>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 /<br>SM               | 16 KB/32 KB/48<br>KB | 96 KB               | 64 KB               | Configurable up to 96<br>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 <sup>2</sup>  | 601 mm <sup>2</sup> | 610 mm <sup>2</sup> | 815 mm <sup>2</sup>         |
| Manufacturing Process                    | 28 nm                | 28 nm               | 16 nm<br>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

(compute capability 7.5)

TU102, TU104, TU106, TU116, ... (cc 7.5)  
(Titan RTX, RTX 2070, 2080, 2080Ti, Tesla T4, ...)

# NVIDIA Turing Architecture (2018/2019)



TU 102

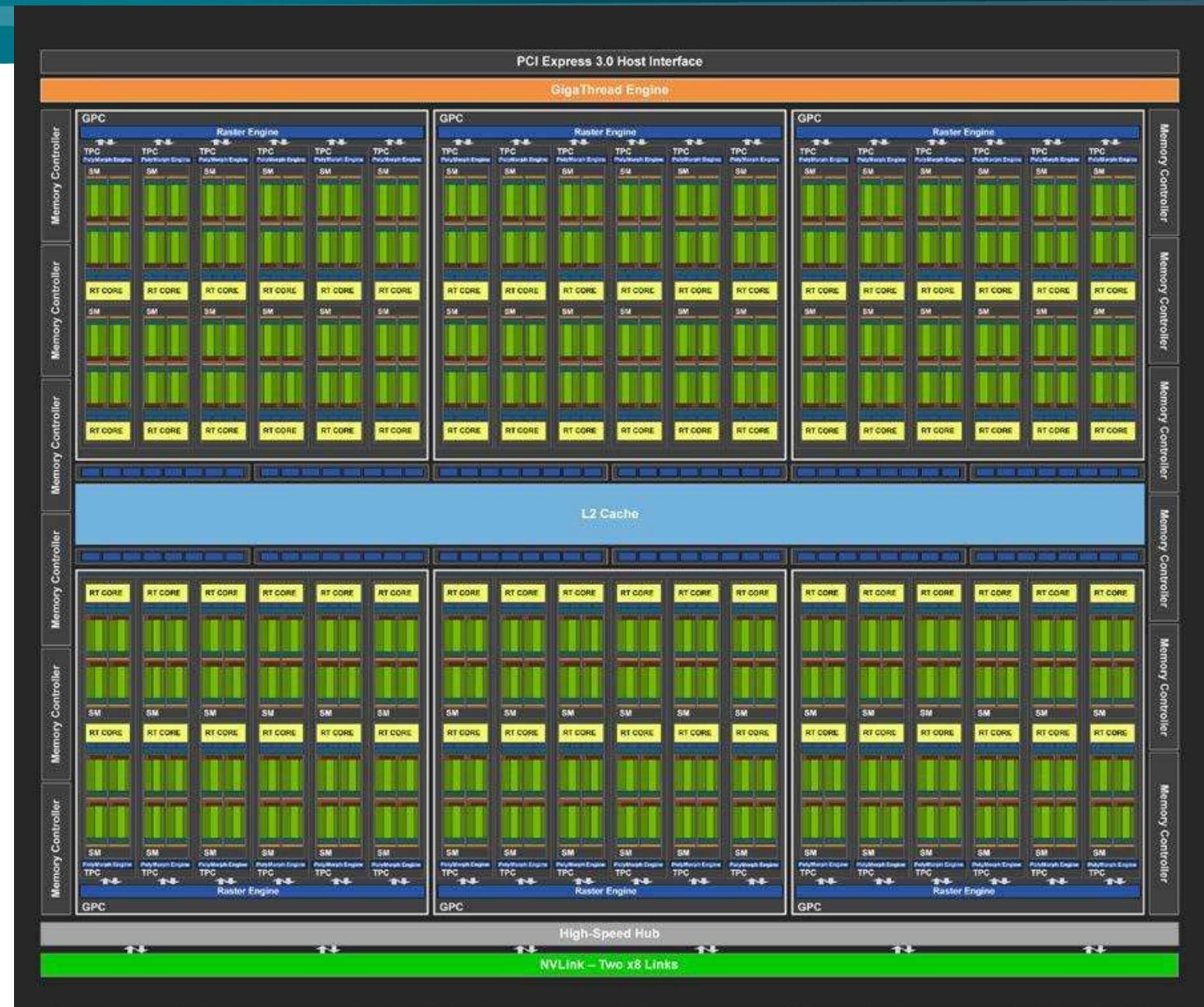
(Geforce:

RTX 2080 Ti,

Quadro:

RTX 6000,

RTX 8000, ...)



# NVIDIA Turing Architecture (2018/2019)



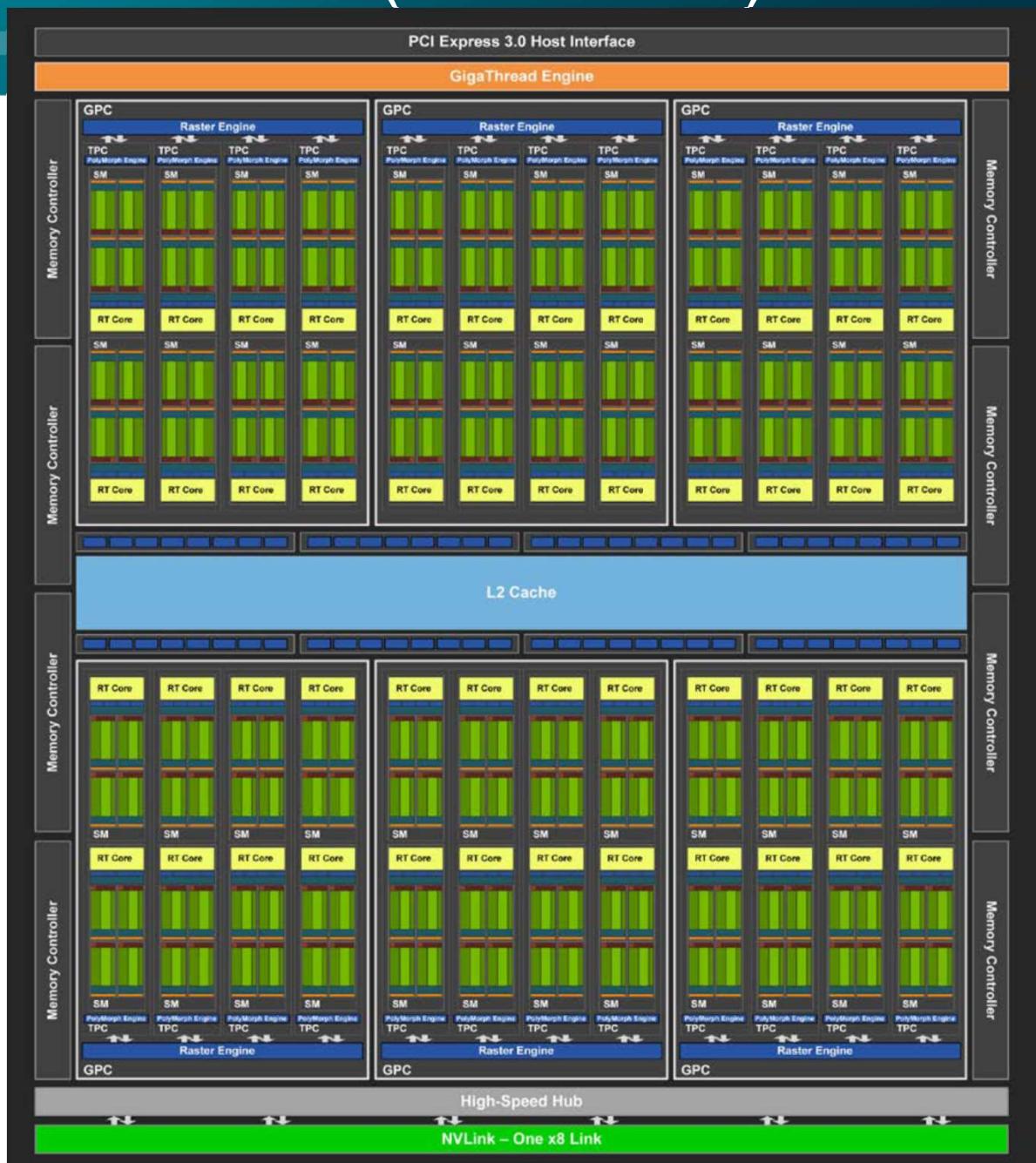
TU 104

(Geforce:

RTX 2080,

Quadro:

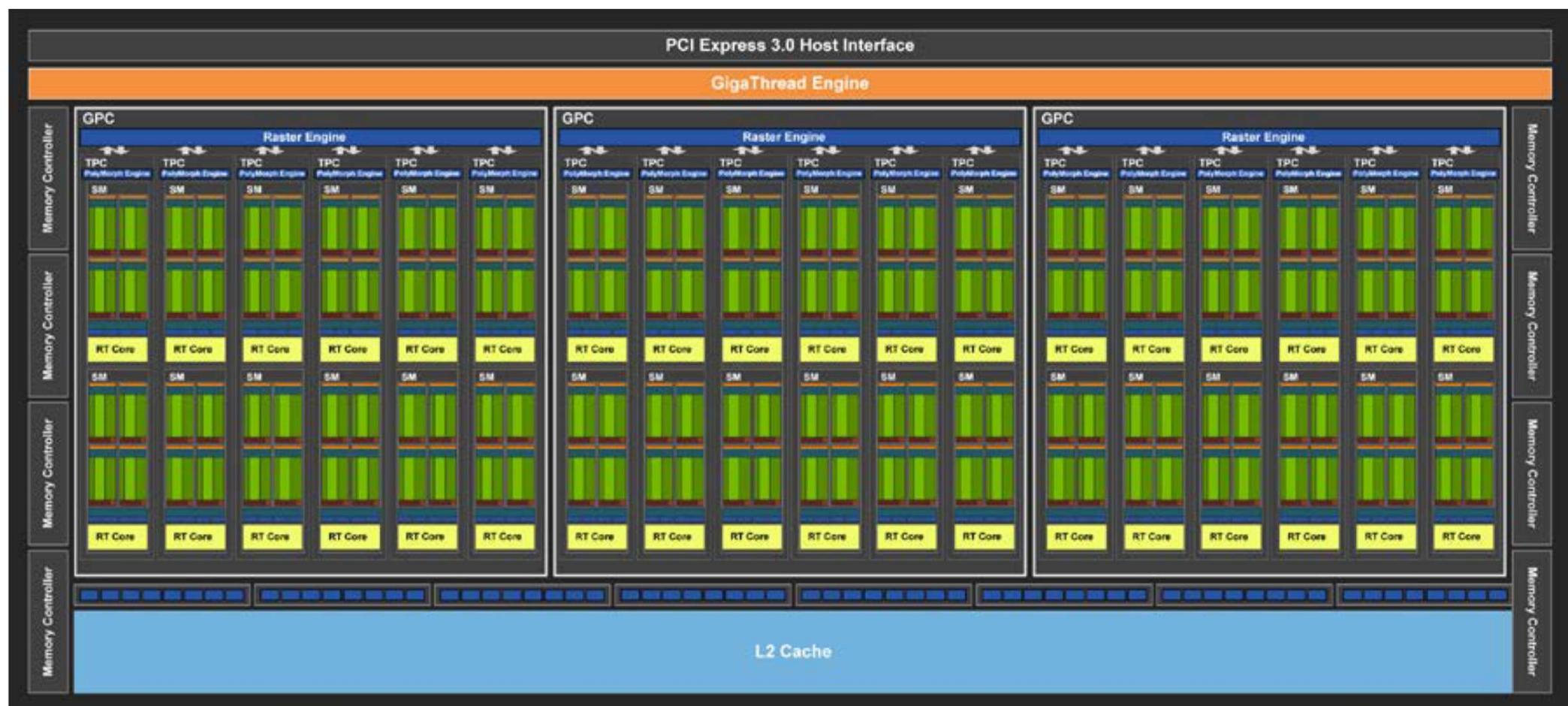
RTX 5000, ...)



# NVIDIA Turing Architecture (2018/2019)



TU 106 (Geforce RTX 2070, ...)





# Instruction Throughput

Instruction throughput numbers in CUDA C Programming Guide (Chapter 8.4)

|   | Compute Capability                        |             |     |     |     |                                   |                       |     |                       |     |     |
|---|---|-------------|-----|-----|-----|-----------------------------------|-----------------------|-----|-----------------------|-----|-----|
|   | 3.5,<br>3.7                               | 5.0,<br>5.2 | 5.3 | 6.0 | 6.1 | 6.2                               | 7.x                   | 8.0 | 8.6                   | 8.9 | 9.0 |
| 16-bit floating-point add, multiply, multiply-add | N/A                                       |             | 256 | 128 | 2   | 256                               | 128                   | 256 | 128                   | 256 |     |
|   |   |             |     |     |     |                                   | 128 for __nv_bfloat16 |     | 128 for __nv_bfloat16 |     |     |
| 32-bit floating-point add, multiply, multiply-add | 192                                       | 128         |     | 64  | 128 |                                   | 64                    |     | 128                   |     |     |
| 64-bit floating-point add, multiply, multiply-add | 64  |             | 4   | 32  | 4   | 32                                | 32                    | 2   | 2                     | 64  |     |
|   | 8 for GeForce GPUs, except for Titan GPUs |             |     |     |     | 2 for compute capability 7.5 GPUs |                       |     |                       |     |     |

# ALU Instruction Latencies and Instructs. / SM



| CC                                 | 2.0<br>(Fermi)       | 2.1<br>(Fermi)       | 3.x<br>(Kepler) | 5.x<br>(Maxwell) | 6.0<br>(Pascal) | 6.1/6.2<br>(Pascal) | 7.x<br>(Volta, Turing) | 8.0/8.6<br>(Ampere) | 8.9/9.0<br>(Ada/Hopper) |
|------------------------------------|----------------------|----------------------|-----------------|------------------|-----------------|---------------------|------------------------|---------------------|-------------------------|
| # warp sched.<br>/ SM              | 2                    | 2                    | 4               | 4                | 2               | 4                   | 4                      | 4                   | 4                       |
| # ALU<br>dispatch<br>/ warp sched. | 1 (over<br>2 clocks) | 2 (over<br>2 clocks) | 2               | 1                | 1               | 1                   | 1                      | 1                   | 1                       |
| SM busy with<br># warps + inst     | L                    | 2L                   | 8L              | 4L               | 2L              | 4L                  | 4L                     | 4L                  | 4L                      |
| inst. pipe<br>latency (L)          | 22                   | 22                   | 11              | 9                | 6               | 6                   | 4                      | 4                   | 4                       |
| SM busy with<br># warps            | 22                   | 22<br>+ ILP          | 44<br>+ ILP     | 36               | 12              | 24                  | 16                     | 16                  | 16                      |

see NVIDIA CUDA C Programming Guides (different versions)  
 performance guidelines/multiprocessor level; compute capabilities

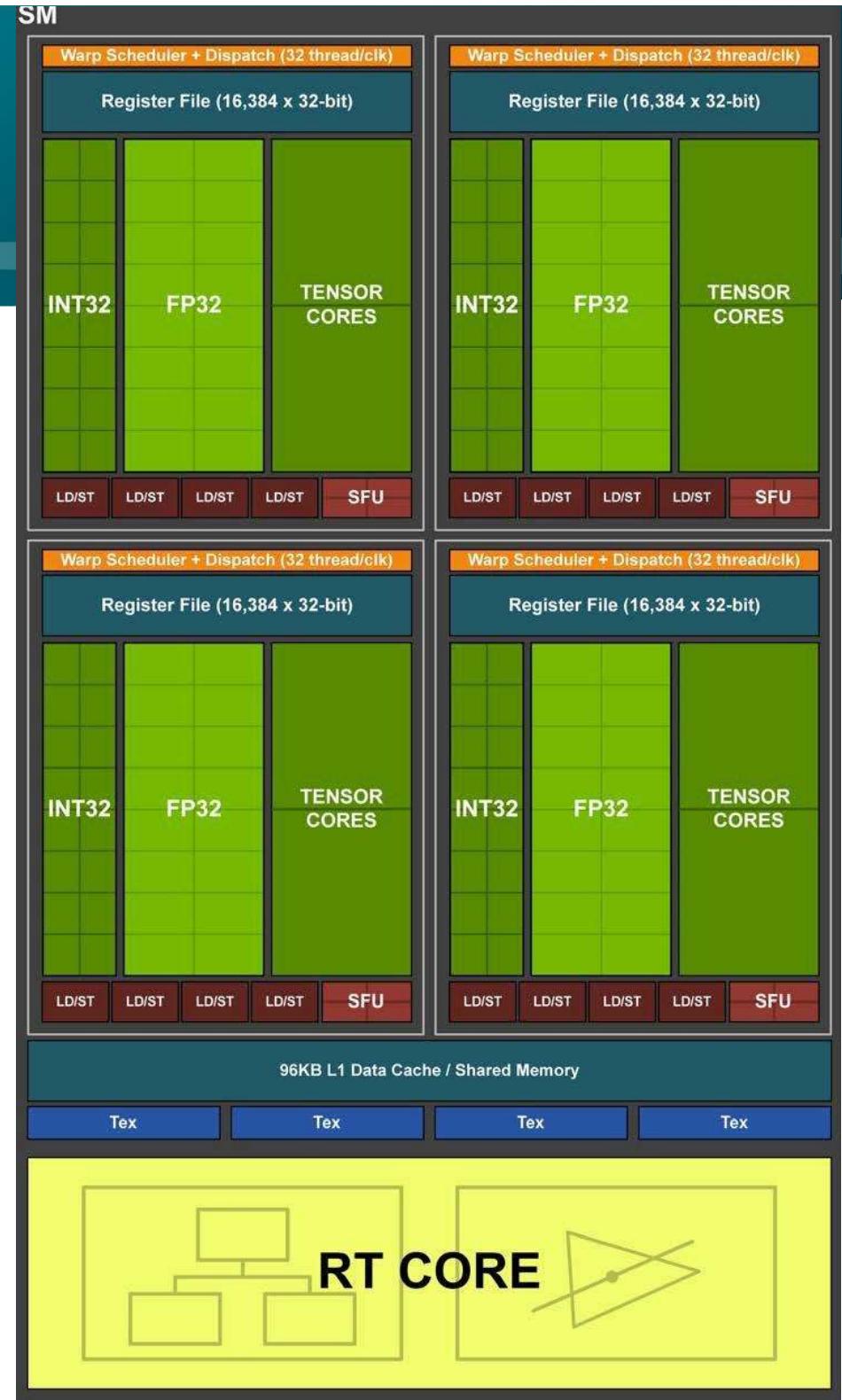
# NVIDIA Turing SM

## Multiprocessor: SM (CC 7.5)

- 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; 4 SFUs each
- 2 Turing tensor cores each
- Each has: warp scheduler,  
dispatch unit, 16K register file





# NVIDIA Ampere Architecture

## 2020

(compute capability 8.0/8.6/8.7)

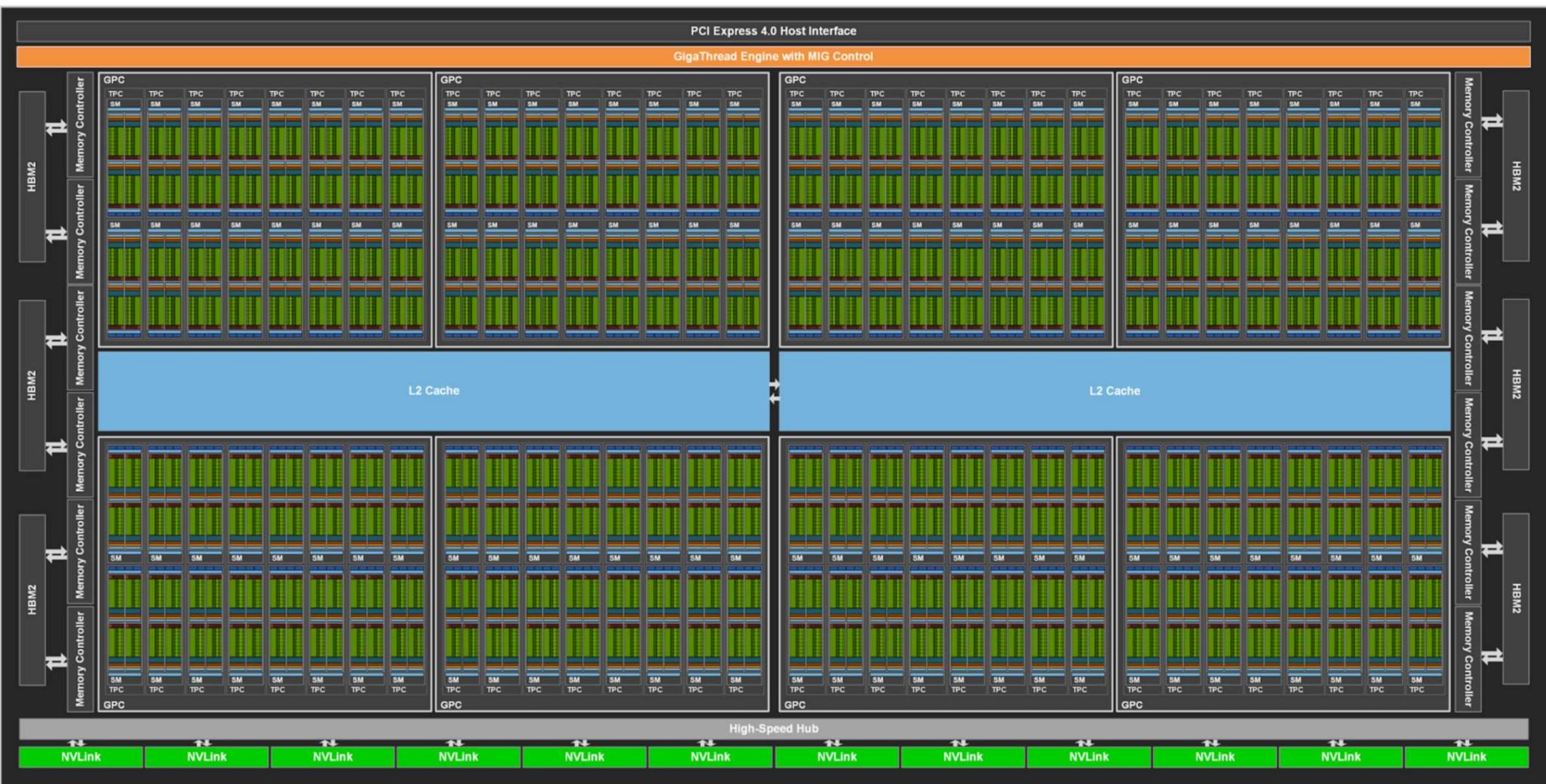
- (x=2,3,4,6,7)    GA100 (cc 8.0), ... (A100, ...)
- GA10x (cc 8.6), ... (RTX 3070, RTX 3080, RTX 3090, ...)
- GA10B (cc 8.7), ... (Jetson, DRIVE, ...)

# NVIDIA Ampere GA100 Architecture (2020)



GA 100 (A100 Tensor Core GPU)

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





# Instruction Throughput

Instruction throughput numbers in CUDA C Programming Guide (Chapter 8.4)

|   | Compute Capability |             |     |     |     |     |     |     |     |     |                       |
|---|--------------------|-------------|-----|-----|-----|-----|-----|-----|-----|-----|-----------------------|
|   | 3.5,<br>3.7        | 5.0,<br>5.2 | 5.3 | 6.0 | 6.1 | 6.2 | 7.x | 8.0 | 8.6 | 8.9 | 9.0                   |
| 16-bit floating-point add, multiply, multiply-add | N/A                |             | 256 | 128 | 2   | 256 | 128 | 256 | 128 | 256 | 128 for __nv_bfloat16 |
| 32-bit floating-point add, multiply, multiply-add | 192                | 128         |     | 64  | 128 |     | 64  |     | 128 |     | 128 for __nv_bfloat16 |
| 64-bit floating-point add, multiply, multiply-add | 64                 |             | 4   | 32  | 4   |     | 32  | 32  | 2   | 2   | 64                    |

8 for GeForce GPUs, except for Titan GPUs

2 for compute capability 7.5 GPUs

# ALU Instruction Latencies and Instructs. / SM



| CC                                 | 2.0<br>(Fermi)       | 2.1<br>(Fermi)       | 3.x<br>(Kepler) | 5.x<br>(Maxwell) | 6.0<br>(Pascal) | 6.1/6.2<br>(Pascal) | 7.x<br>(Volta, Turing) | 8.0/8.6<br>(Ampere) | 8.9/9.0<br>(Ada/Hopper) |
|------------------------------------|----------------------|----------------------|-----------------|------------------|-----------------|---------------------|------------------------|---------------------|-------------------------|
| # warp sched.<br>/ SM              | 2                    | 2                    | 4               | 4                | 2               | 4                   | 4                      | 4                   | 4                       |
| # ALU<br>dispatch<br>/ warp sched. | 1 (over<br>2 clocks) | 2 (over<br>2 clocks) | 2               | 1                | 1               | 1                   | 1                      | 1                   | 1                       |
| SM busy with<br># warps + inst     | L                    | 2L                   | 8L              | 4L               | 2L              | 4L                  | 4L                     | 4L                  | 4L                      |
| inst. pipe<br>latency (L)          | 22                   | 22                   | 11              | 9                | 6               | 6                   | 4                      | 4                   | 4                       |
| SM busy with<br># warps            | 22                   | 22<br>+ ILP          | 44<br>+ ILP     | 36               | 12              | 24                  | 16                     | 16                  | 16                      |

see NVIDIA CUDA C Programming Guides (different versions)  
 performance guidelines/multiprocessor level; compute capabilities

# NVIDIA GA100 SM

## Multiprocessor: SM (CC 8.0)

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

## 4 partitions inside SM

- 16 FP32 + 16 INT32 cores
- 8 FP64 cores
- 8 LD/ST units; 4 SFUs each
- 1 3<sup>rd</sup> 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 (CC 8.6)

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

## 4 partitions inside SM

- 32<sub>(16+16)</sub> FP32 + 16 INT32 cores
- 4 LD/ST units; 4 SFUs each
- 1 3<sup>rd</sup> gen tensor core each
- Each has: warp scheduler, dispatch unit, 16K register file



# Specs CC 3.5 – 9.0



Ampere: 8.0, 8.6, 8.7  
(8.9 / Ada missing in table)

(CUDA C Programming Guide  
11.8, Table 15)

| Technical Specifications   | Compute Capability |       |     |     |      |      |      |     |     |     |     |            |      |      |      |
|--|--------------------|-------|-----|-----|------|------|------|-----|-----|-----|-----|------------|------|------|------|
|  | 3.5                | 3.7   | 5.0 | 5.2 | 5.3  | 6.0  | 6.1  | 6.2 | 7.0 | 7.2 | 7.5 | 8.0        | 8.6  | 8.7  | 9.0  |
| Maximum number of resident grids per device <small>[Concurrent Kernel Execution]</small> | 32                 |       | 16  | 128 | 32   | 16   | 128  | 16  |     |     |     | 128        |      |      |      |
| Maximum dimensionality of grid of thread blocks  |                    |       |     |     |      |      |      |     |     |     |     | 3          |      |      |      |
| Maximum x-dimension of a grid of thread blocks   |                    |       |     |     |      |      |      |     |     |     |     | $2^{31}-1$ |      |      |      |
| Maximum y- or z-dimension of a grid of thread blocks                                     |                    |       |     |     |      |      |      |     |     |     |     | 65535      |      |      |      |
| Maximum dimensionality of a 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 SM   | 16                 |       |     |     |      | 32   |      |     |     |     |     | 16         | 32   | 16   | 32   |
| Maximum number of resident warps per SM  |                    |       |     |     |      | 64   |      |     |     |     |     | 32         | 64   | 48   | 64   |
| Maximum number of resident threads per SM  |                    |       |     |     |      | 2048 |      |     |     |     |     | 1024       | 2048 | 1536 | 2048 |
| Number of 32-bit registers per SM  | 64 K               | 128 K |     |     |      |      |      |     |     |     |     | 64 K       |      |      |      |
| Maximum number of 32-bit registers per thread block                                      |                    | 64 K  |     |     | 32 K | 64 K | 32 K |     |     |     |     | 64 K       |      |      |      |
| Maximum number of 32-bit registers per thread  |                    |       |     |     |      |      |      |     |     |     |     | 255        |      |      |      |

# NVIDIA Ampere GA102 Architecture (2020)



GA 102 (RTX 3070, 3080, 3090, A40)    Full GPU: 84 SMs (in 7 GPCs/42 TPCs)

- 64K 32-bit registers / SM = 256 KB register storage per SM
- 128 KB shared memory / L1 per SM

For 84 SMs on full GPU [RTX 3090: 82 SMs]

- 21 MB register storage, 10.5 MB shared mem / L1 storage = **31.5 MB context+”shared context” storage !**
- L2 cache size on A40, RTX 3090: 6 MB
- 10,752 FP32 cores (128 FP32 cores per SM) [RTX 3090: 10,496]
- 129,024 max threads in flight (max warps / SM = 48) [RTX 3090: 125,952]

# NVIDIA Ampere GA100 Architecture (2020)



GA 100 (A100)

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

- 64K 32-bit registers / SM = 256 KB register storage per SM
- 192 KB shared memory / L1 per SM

For 128 SMs on full GPU [A100: 108 SMs]

- 32 MB register storage, 24 MB shared mem / L1 storage = **56 MB context+”shared context” storage !**
- L2 cache size on A100: 40 MB
- 8,912 FP32 cores (64 FP32 cores per SM) [A100: 6,912]
- 262,144 max threads in flight (max warps / SM = 64) [A100: 221,184]



# 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) <sup>1</sup>  | 10.6                              | 11.2                                    | 29.8                                    |
| Peak FP16 TFLOPS (non-Tensor) <sup>1</sup>  | 21.2                              | 22.3                                    | 29.8                                    |
| Peak BF16 TFLOPS (non-Tensor) <sup>1</sup>  | NA                                | NA                                      | 29.8                                    |
| Peak INT32 TOPS (non-Tensor) <sup>1,3</sup> | 10.6                              | 11.2                                    | 14.9                                    |



# Turing vs. Ampere GA102

|   |                  |                  |                        |
|---|------------------|------------------|------------------------|
| <b>Peak FP16 Tensor TFLOPS<br/>with FP16 Accumulate<sup>1</sup></b> | 84.8             | 89.2             | 119/238 <sup>2</sup>   |
| <b>Peak FP16 Tensor TFLOPS<br/>with FP32 Accumulate<sup>1</sup></b> | 42.4             | 44.6             | 59.5/119 <sup>2</sup>  |
| <b>Peak BF16 Tensor TFLOPS<br/>with FP32 Accumulate<sup>1</sup></b> | NA               | NA               | 59.5/119 <sup>2</sup>  |
| <b>Peak TF32 Tensor TFLOPS<sup>1</sup></b>                          | NA               | NA               | 29.8/59.5 <sup>2</sup> |
| <b>Peak INT8 Tensor TOPS<sup>1</sup></b>                            | 169.6            | 178.4            | 238/476 <sup>2</sup>   |
| <b>Peak INT4 Tensor TOPS<sup>1</sup></b>                            | 339.1            | 356.8            | 476/952 <sup>2</sup>   |
| <b>Frame Buffer Memory Size and<br/>Type</b>                        | 8192 MB<br>GDDR6 | 8192 MB<br>GDDR6 | 10240 MB<br>GDDR6X     |
| <b>Memory Interface</b>   | 256-bit          | 256-bit          | 320-bit                |
| <b>Memory Clock (Data Rate)</b>                                     | 14 Gbps          | 15.5 Gbps        | 19 Gbps                |
| <b>Memory Bandwidth</b>   | 448 GB/sec       | 496 GB/sec       | 760 GB/sec             |
| <b>ROPs</b>   | 64               | 64               | 96                     |
| <b>Pixel Fill-rate (Gigapixels/sec)</b>                             | 115.2            | 116.2            | 164.2                  |
| <b>Texture Units</b>  | 184              | 192              | 272                    |
| <b>Texel Fill-rate (Gigatexels/sec)</b>                             | 331.2            | 348.5            | 465                    |
| <b>L1 Data Cache/Shared Memory</b>                                  | 4416 KB          | 4608 KB          | 8704 KB                |



# Turing vs. Ampere GA102

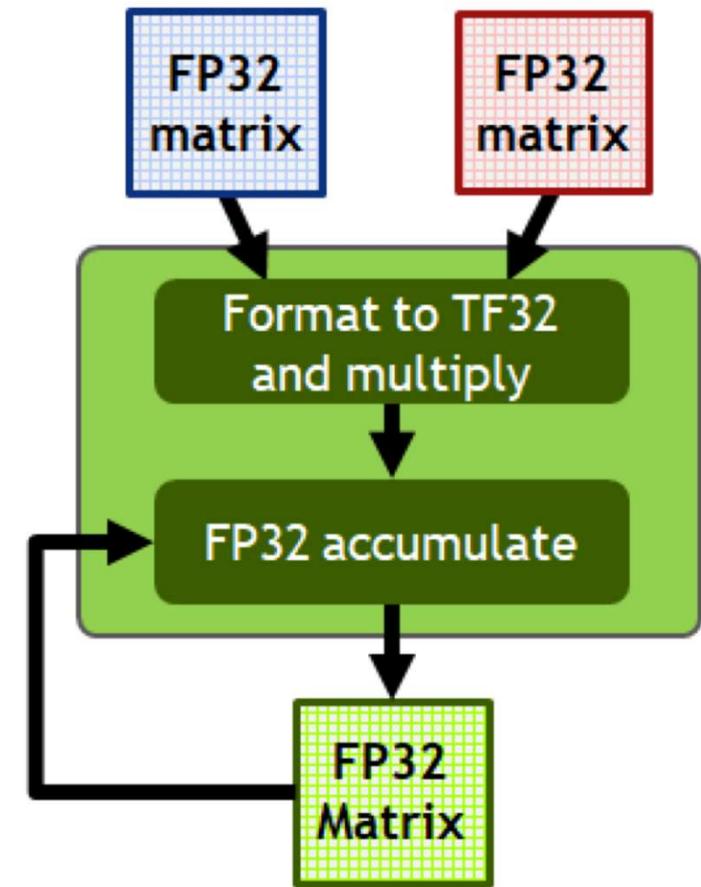
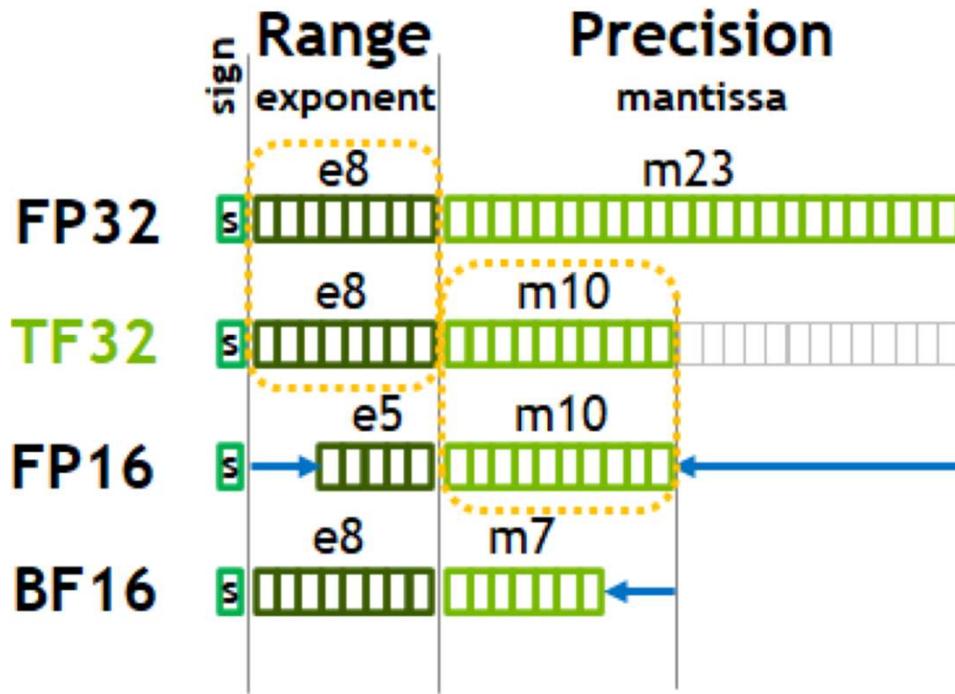
|                                   |                                   |                                   |  |
|-----------------------------------|-----------------------------------|-----------------------------------|--|
| <b>L2 Cache Size</b>              | 4096 KB                           | 4096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

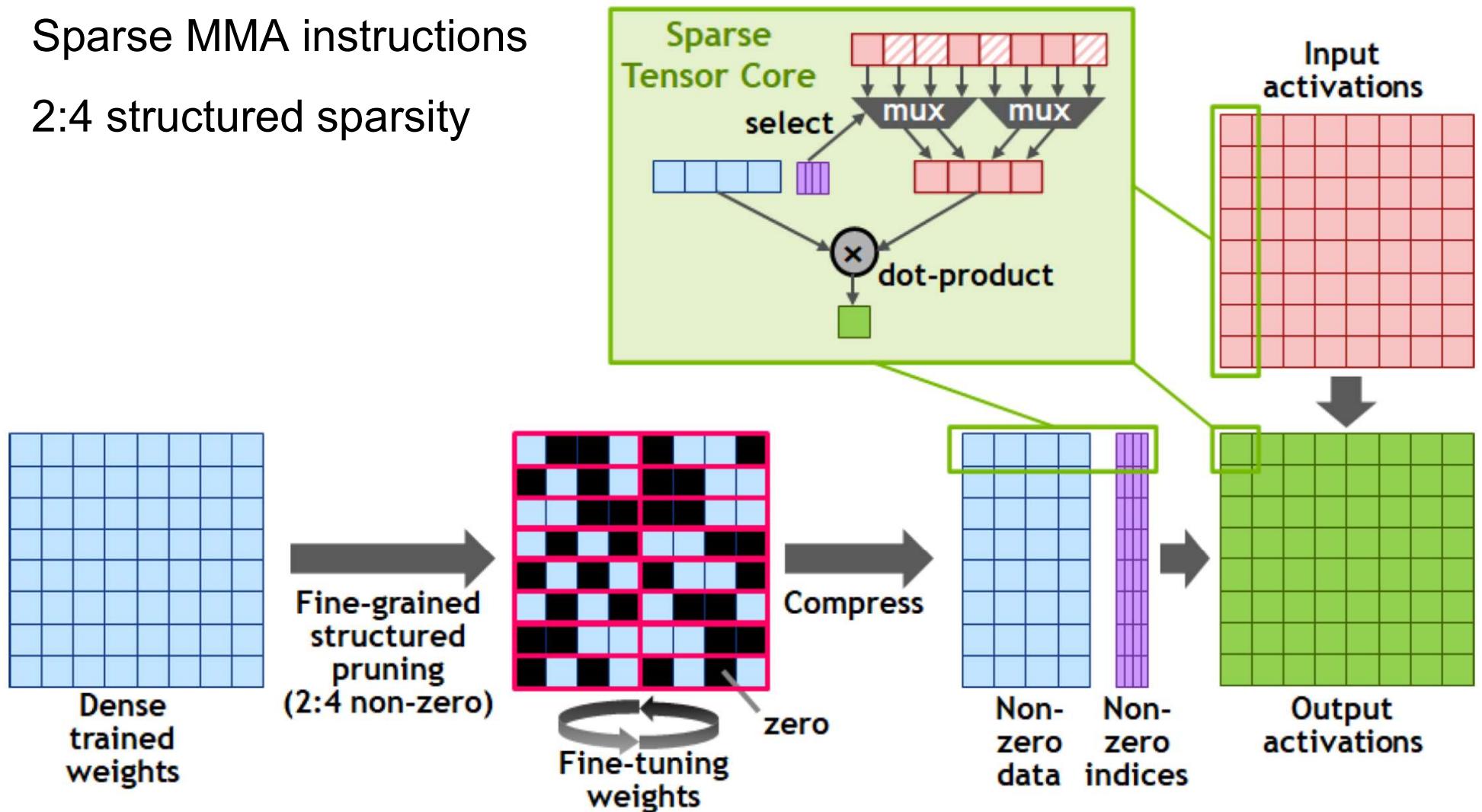
plus INT4/INT8/binary data types (already introduced in Turing)



# Tensor Cores: Sparsity Support

Sparse MMA instructions

2:4 structured sparsity





# NVIDIA Hopper Architecture

## 2022

(compute capability 9.0)

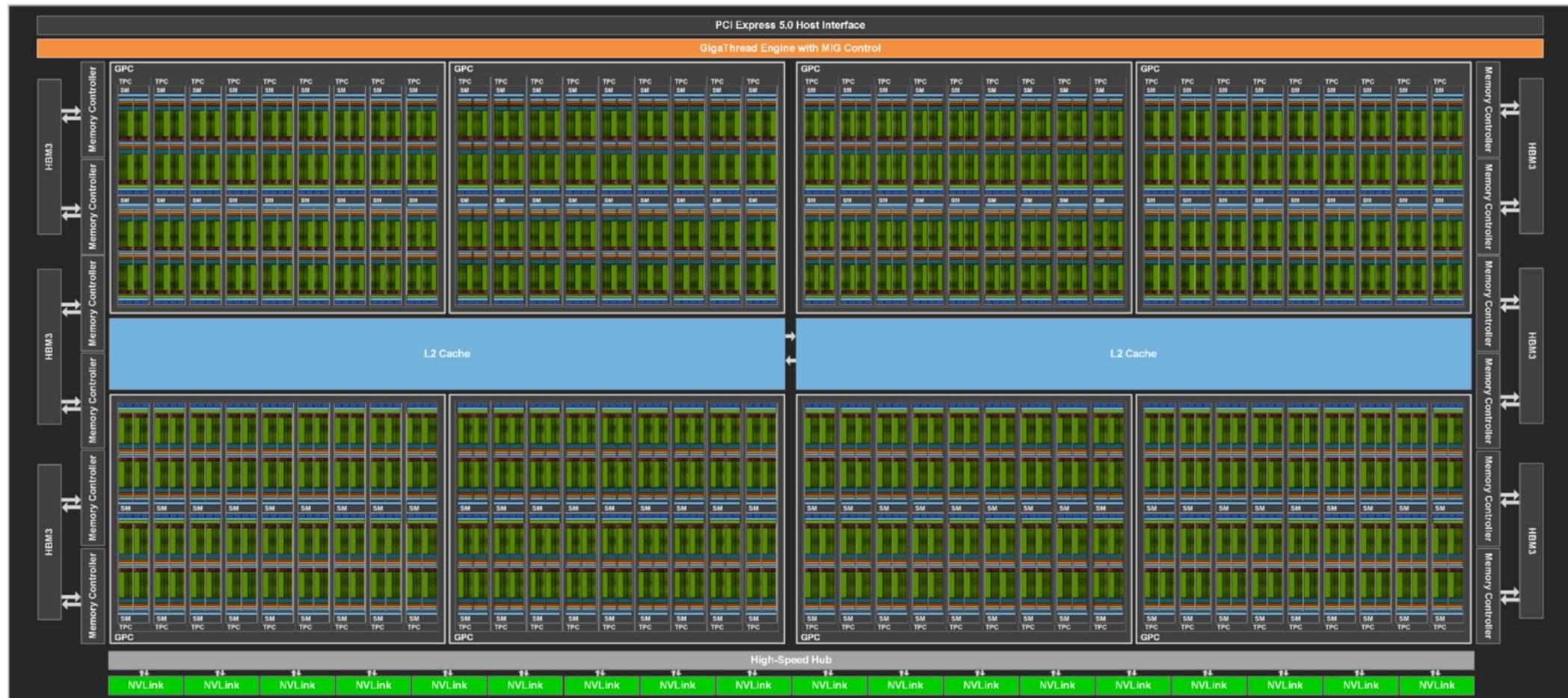
GH100 (cc 9.0), ... (H100, ...)

# NVIDIA Hopper GH100 Architecture (2022)



GH 100 (H100 Tensor Core GPU)

Full GPU: 144 SMs (in 8 GPCs/72 TPCs)





# Instruction Throughput

Instruction throughput numbers in CUDA C Programming Guide (Chapter 8.4)

|   | Compute Capability |             |     |     |     |     |     |     |     |     |                       |
|---|--------------------|-------------|-----|-----|-----|-----|-----|-----|-----|-----|-----------------------|
|   | 3.5,<br>3.7        | 5.0,<br>5.2 | 5.3 | 6.0 | 6.1 | 6.2 | 7.x | 8.0 | 8.6 | 8.9 | 9.0                   |
| 16-bit floating-point add, multiply, multiply-add | N/A                |             | 256 | 128 | 2   | 256 | 128 | 256 | 128 | 256 | 128 for __nv_bfloat16 |
| 32-bit floating-point add, multiply, multiply-add | 192                |             | 128 | 64  | 128 |     | 64  |     | 128 |     | 128 for __nv_bfloat16 |
| 64-bit floating-point add, multiply, multiply-add | 64                 |             | 4   | 32  | 4   |     | 32  | 32  | 2   | 2   | 64                    |

8 for GeForce GPUs, except for Titan GPUs

2 for compute capability 7.5 GPUs

# ALU Instruction Latencies and Instructs. / SM



| CC                                 | 2.0<br>(Fermi)       | 2.1<br>(Fermi)       | 3.x<br>(Kepler) | 5.x<br>(Maxwell) | 6.0<br>(Pascal) | 6.1/6.2<br>(Pascal) | 7.x<br>(Volta, Turing) | 8.0/8.6<br>(Ampere) | 8.9/9.0<br>(Ada/Hopper) |
|------------------------------------|----------------------|----------------------|-----------------|------------------|-----------------|---------------------|------------------------|---------------------|-------------------------|
| # warp sched.<br>/ SM              | 2                    | 2                    | 4               | 4                | 2               | 4                   | 4                      | 4                   | 4                       |
| # ALU<br>dispatch<br>/ warp sched. | 1 (over<br>2 clocks) | 2 (over<br>2 clocks) | 2               | 1                | 1               | 1                   | 1                      | 1                   | 1                       |
| SM busy with<br># warps + inst     | L                    | 2L                   | 8L              | 4L               | 2L              | 4L                  | 4L                     | 4L                  | 4L                      |
| inst. pipe<br>latency (L)          | 22                   | 22                   | 11              | 9                | 6               | 6                   | 4                      | 4                   | 4                       |
| SM busy with<br># warps            | 22                   | 22<br>+ ILP          | 44<br>+ ILP     | 36               | 12              | 24                  | 16                     | 16                  | 16                      |

see NVIDIA CUDA C Programming Guides (different versions)  
 performance guidelines/multiprocessor level; compute capabilities

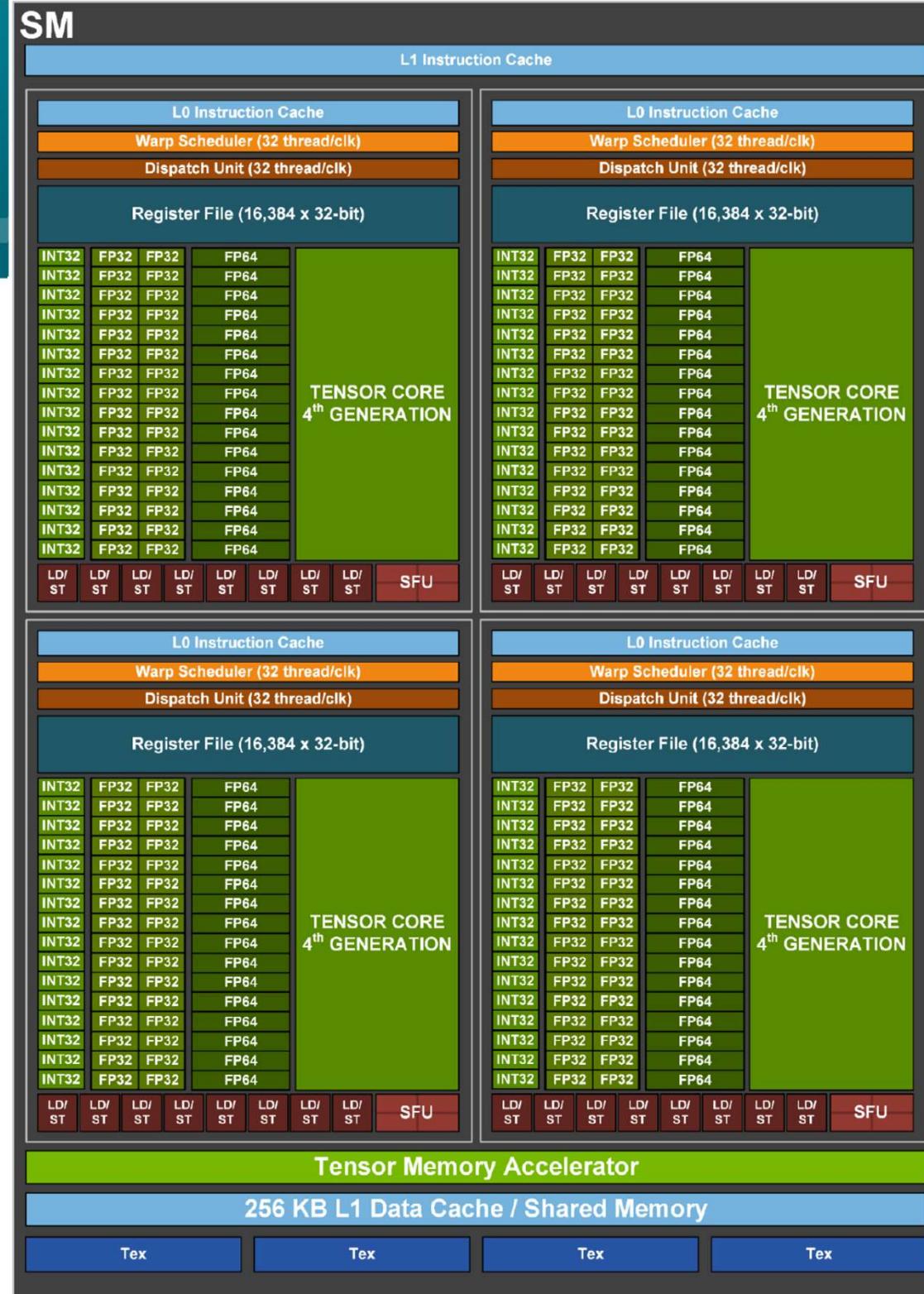
# NVIDIA GH100 SM

## Multiprocessor: SM (CC 9.0)

- 128 FP32 + 64 INT32 cores
- 64 FP64 cores
- 4x 4<sup>th</sup> gen tensor cores
- ++ thread block clusters, DPX insts., FP8, TMA

## 4 partitions inside SM

- 32 FP32 + 16 INT32 cores
- 16 FP64 cores
- 8x LD/ST units; 4 SFUs each
- 1x 4<sup>th</sup> gen tensor core each
- Each has: warp scheduler, dispatch unit, 16K register file



# NVIDIA Hopper GH100 Architecture (2022)



GH 100 (H100)

Full GPU: 144 SMs (in 8 GPCs/72 TPCs)

- 64K 32-bit registers / SM = 256 KB register storage per SM
- 256 KB shared memory / L1 per SM

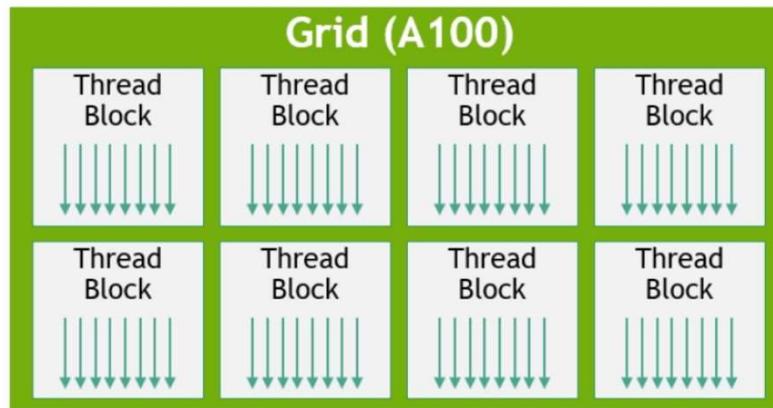
For 144 SMs on full GPU [SXM5: 132; PCIe: 114]

- 36 MB register storage, 36 MB shared mem / L1 storage = **72 MB context+”shared context” storage !**
- L2 cache size on H100: 50 MB
- 18,432 FP32 cores (128 FP32 cores per SM) [SXM5: 16,896]
- 294,912 max threads in flight (max warps / SM = 64) [SXM5: 270,336]

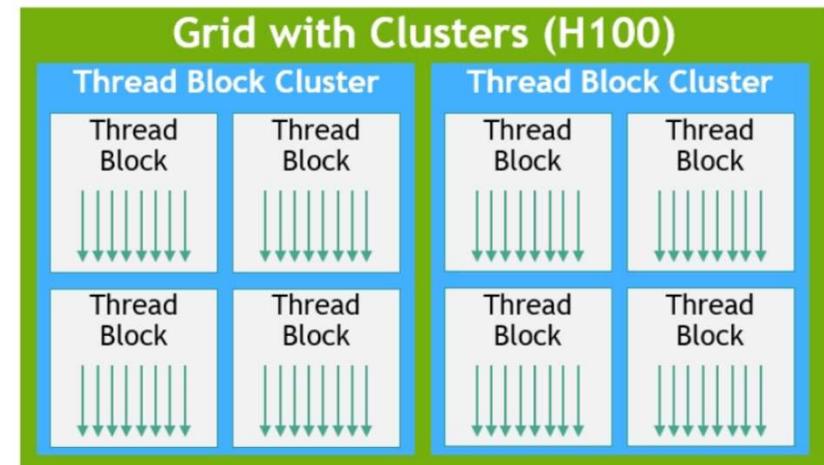
# New in CC 9.0: Thread Block Clusters



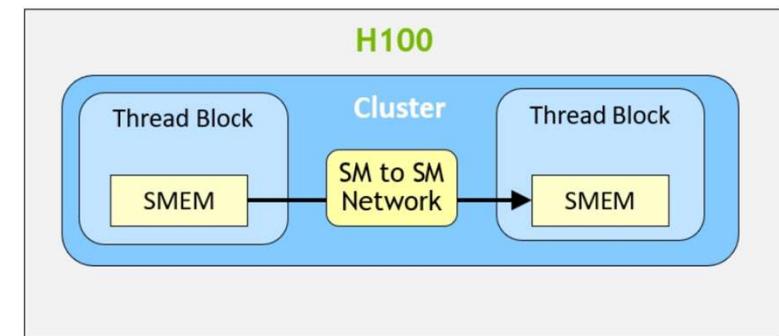
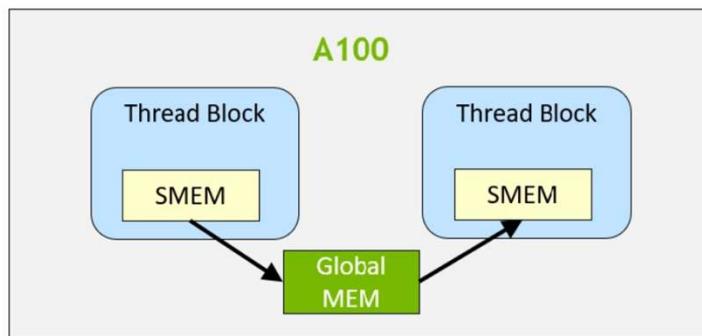
New thread hierarchy level!



*all threads of a block are on the same SM !*



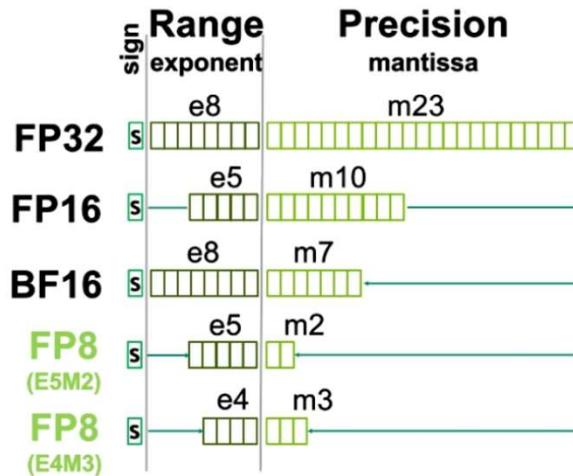
*all blocks of a cluster are on the same GPC !*



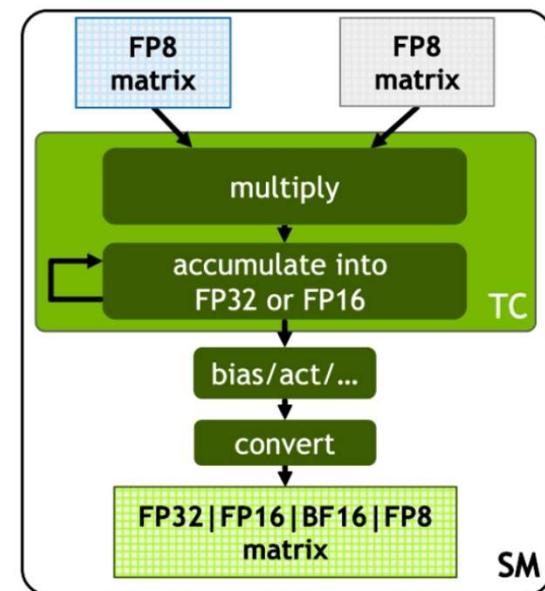
# Tensor Cores: More Mixed Precision Options



New in Hopper: FP8



Allocate 1 bit to either range or precision



Support for multiple accumulator and output types

plus other data types from before (INT4/INT8/binary, ...)



# Tensor Cores: Hopper vs. Ampere

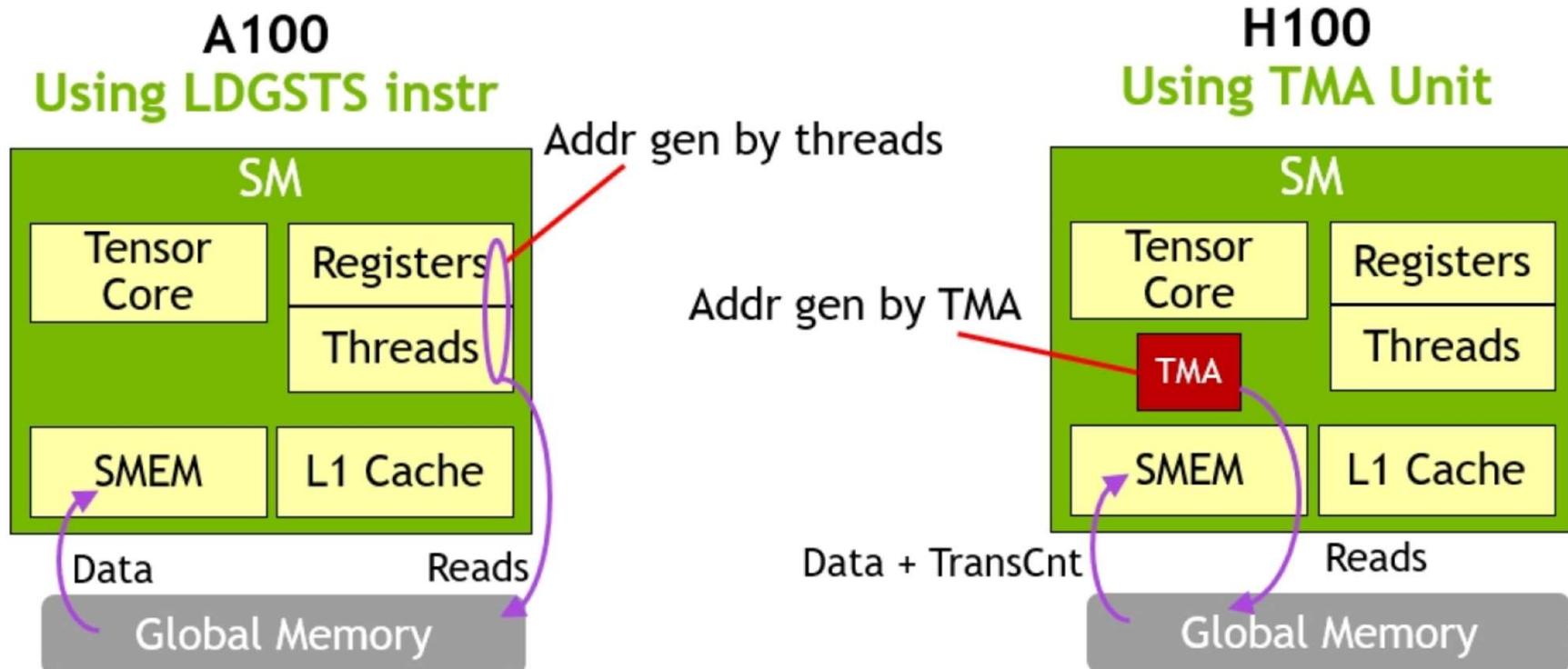
(preliminary)

|                  | A100        | A100 Sparse | H100 SXM5 <sup>1</sup> | H100 SXM5 <sup>1</sup> Sparse | H100 SXM5 <sup>1</sup> Speedup vs A100 |
|------------------|-------------|-------------|------------------------|-------------------------------|--|
| FP8 Tensor Core  | NA          | NA          | 2000 TFLOPS            | 4000 TFLOPS                   | 6.4x vs A100 FP16                      |
| FP16             | 78 TFLOPS   | NA          | 120 TFLOPS             | NA                            | 1.5x                                   |
| FP16 Tensor Core | 312 TFLOPS  | 624 TFLOPS  | 1000 TFLOPS            | 2000 TFLOPS                   | 3.2x                                   |
| BF16 Tensor Core | 312 TFLOPS  | 624 TFLOPS  | 1000 TFLOPS            | 2000 TFLOPS                   | 3.2x                                   |
| FP32             | 19.5 TFLOPS | NA          | 60 TFLOPS              | NA                            | 3.1x                                   |
| TF32 Tensor Core | 156 TFLOPS  | 312 TFLOPS  | 500 TFLOPS             | 1000 TFLOPS                   | 3.2x                                   |
| FP64             | 9.7 TFLOPS  | NA          | 30 TFLOPS              | NA                            | 3.1x                                   |
| FP64 Tensor Core | 19.5 TFLOPS | NA          | 60 TFLOPS              | NA                            | 3.1x                                   |
| INT8 Tensor Core | 624 TOPS    | 1248 TOPS   | 2000 TFLOPS            | 4000 TFLOPS                   | 3.2x                                   |



# Tensor Memory Accelerator (TMA)

## Asynchronous transfers





# Hopper vs. Ampere (1)

(preliminary)

| GPU Features   | NVIDIA A100   | NVIDIA H100 SXM5 <sup>1</sup> | NVIDIA H100 PCIe <sup>1</sup> |
|--|---------------|-------------------------------|-------------------------------|
| GPU Architecture   | NVIDIA Ampere | NVIDIA Hopper                 | NVIDIA Hopper                 |
| GPU Board Form Factor                                    | SXM4          | SXM5                          | PCIe Gen 5                    |
| SMs  | 108           | 132                           | 114                           |
| TPCs   | 54            | 66                            | 57                            |
| FP32 Cores / SM  | 64            | 128                           | 128                           |
| FP32 Cores / GPU   | 6912          | 16896                         | 14592                         |
| FP64 Cores / SM (excl. Tensor)                           | 32            | 64                            | 64                            |
| FP64 Cores / GPU (excl. Tensor)                          | 3456          | 8448                          | 7296                          |
| INT32 Cores / SM   | 64            | 64                            | 64                            |
| INT32 Cores / GPU  | 6912          | 8448                          | 7296                          |
| Tensor Cores / SM  | 4             | 4                             | 4                             |
| Tensor Cores / GPU                                       | 432           | 528                           | 456                           |
| GPU Boost Clock<br>(Not Finalized for H100) <sup>3</sup> | 1410 MHz      | Not Finalized                 | Not Finalized                 |



# Hopper vs. Ampere (2)

(preliminary)

| GPU Features  | NVIDIA A100               | NVIDIA H100 SXM <sup>1</sup> | NVIDIA H100 PCIe <sup>1</sup> |
|---|---------------------------|------------------------------|-------------------------------|
| Texture Units   | 432                       | 528                          | 456                           |
| Memory Interface  | 5120-bit HBM2             | 5120-bit HBM3                | 5120-bit HBM2e                |
| Memory Size   | 40 GB                     | 80 GB                        | 80 GB                         |
| Memory Data Rate <sup>1</sup>                             | 1215 MHz DDR              | Not Finalized                | Not Finalized                 |
| Memory Bandwidth<br>(Not Finalized for H100) <sup>1</sup> | 1555 GB/sec               | 3000 GB/sec                  | 2000 GB/sec                   |
| L2 Cache Size   | 40 MB                     | 50 MB                        | 50 MB                         |
| Shared Memory Size / SM                                   | Configurable up to 164 KB | Configurable up to 228 KB    | Configurable up to 228 KB     |
| Register File Size / SM                                   | 256 KB                    | 256 KB                       | 256 KB                        |
| Register File Size / GPU                                  | 27648 KB                  | 33792 KB                     | 29184 KB                      |
| TDP <sup>1</sup>  | 400 Watts                 | 700 Watts                    | 350 Watts                     |
| Transistors   | 54.2 billion              | 80 billion                   | 80 billion                    |
| GPU Die Size  | 826 mm <sup>2</sup>       | 814 mm <sup>2</sup>          | 814 mm <sup>2</sup>           |
| TSMC Manufacturing Process                                | 7 nm N7                   | 4N customized for NVIDIA     | 4N customized for NVIDIA      |



# Compute Capabilities

| Data Center GPU                           | NVIDIA Tesla V100        | NVIDIA A100               | NVIDIA H100               |
|---|--------------------------|---------------------------|---------------------------|
| GPU Architecture                          | NVIDIA Volta             | NVIDIA Ampere             | NVIDIA Hopper             |
| Compute Capability                        | 7.0                      | 8.0                       | 9.0                       |
| Threads / Warp                            | 32                       | 32                        | 32                        |
| Max Warps / SM                            | 64                       | 64                        | 64                        |
| Max Threads / SM                          | 2048                     | 2048                      | 2048                      |
| Max Thread Blocks (CTAs) / SM             | 32                       | 32                        | 32                        |
| Max Thread Blocks / Thread Block Clusters | NA                       | NA                        | 16                        |
| Max 32-bit Registers / SM                 | 65536                    | 65536                     | 65536                     |
| Max Registers / Thread Block (CTA)        | 65536                    | 65536                     | 65536                     |
| Max Registers / Thread                    | 255                      | 255                       | 255                       |
| Max Thread Block Size (# of threads)      | 1024                     | 1024                      | 1024                      |
| FP32 Cores / SM                           | 64                       | 64                        | 128                       |
| Ratio of SM Registers to FP32 Cores       | 1024                     | 1024                      | 512                       |
| Shared Memory Size / SM                   | Configurable up to 96 KB | Configurable up to 164 KB | Configurable up to 228 KB |



# NVIDIA Ada Lovelace Architecture

## 2022/2023

(compute capability 8.9)

GA10x (cc 8.9), ... (RTX 4080 12 GB, RTX 4080 16GB,  
(x=2,3,4,6,7)                    RTX 4090, RTX 6000, L40, ...)



# NVIDIA Ada Lovelace AD10x Architecture (2022)

Full AD 10x

Full GPU: 144 SMs (in 12 GPCs/72 TPCs)





# NVIDIA Ada Lovelace AD102 Architecture (2022)

AD 102 (RTX 4090, ...)

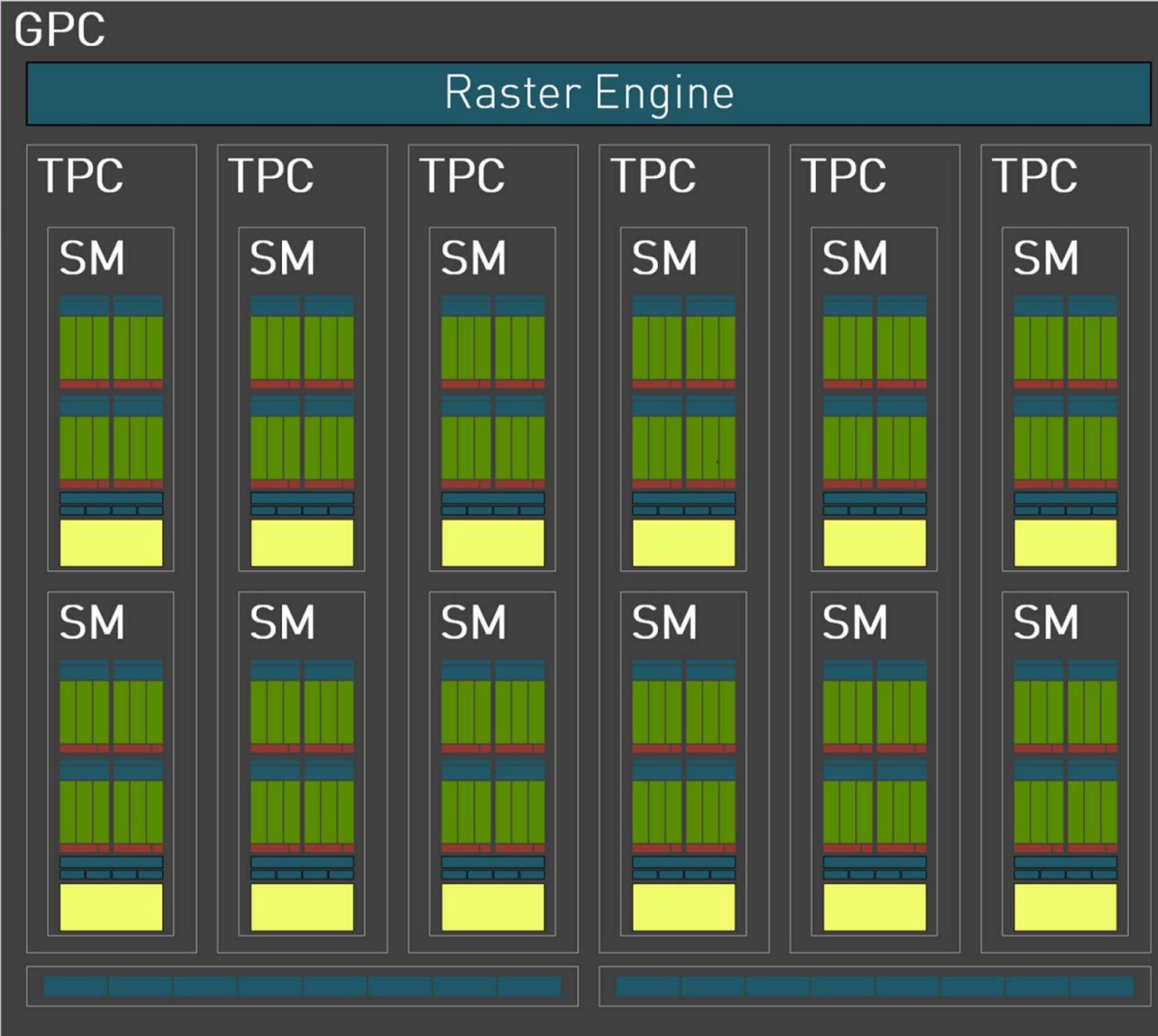
Full RTX 4090: 128 SMs (in 11 GPCs/64 TPCs)



# GPC

## Full GPC

- 6 TPCs
- 12 SMs
- 16 ROPs





# Instruction Throughput

Instruction throughput numbers in CUDA C Programming Guide (Chapter 8.4)

|   | Compute Capability |             |     |     |     |     |     |     |     |     |                       |
|---|--------------------|-------------|-----|-----|-----|-----|-----|-----|-----|-----|-----------------------|
|   | 3.5,<br>3.7        | 5.0,<br>5.2 | 5.3 | 6.0 | 6.1 | 6.2 | 7.x | 8.0 | 8.6 | 8.9 | 9.0                   |
| 16-bit floating-point add, multiply, multiply-add | N/A                |             | 256 | 128 | 2   | 256 | 128 | 256 | 128 | 256 | 128 for __nv_bfloat16 |
| 32-bit floating-point add, multiply, multiply-add | 192                |             | 128 | 64  | 128 |     | 64  |     | 128 |     | 128 for __nv_bfloat16 |
| 64-bit floating-point add, multiply, multiply-add | 64                 |             | 4   | 32  | 4   |     | 32  | 32  | 2   | 2   | 64                    |

8 for GeForce GPUs, except for Titan GPUs

2 for compute capability 7.5 GPUs

# ALU Instruction Latencies and Instructs. / SM



| CC                                 | 2.0<br>(Fermi)       | 2.1<br>(Fermi)       | 3.x<br>(Kepler) | 5.x<br>(Maxwell) | 6.0<br>(Pascal) | 6.1/6.2<br>(Pascal) | 7.x<br>(Volta, Turing) | 8.0/8.6<br>(Ampere) | 8.9/9.0<br>(Ada/Hopper) |
|------------------------------------|----------------------|----------------------|-----------------|------------------|-----------------|---------------------|------------------------|---------------------|-------------------------|
| # warp sched.<br>/ SM              | 2                    | 2                    | 4               | 4                | 2               | 4                   | 4                      | 4                   | 4                       |
| # ALU<br>dispatch<br>/ warp sched. | 1 (over<br>2 clocks) | 2 (over<br>2 clocks) | 2               | 1                | 1               | 1                   | 1                      | 1                   | 1                       |
| SM busy with<br># warps + inst     | L                    | 2L                   | 8L              | 4L               | 2L              | 4L                  | 4L                     | 4L                  | 4L                      |
| inst. pipe<br>latency (L)          | 22                   | 22                   | 11              | 9                | 6               | 6                   | 4                      | 4                   | 4                       |
| SM busy with<br># warps            | 22                   | 22<br>+ ILP          | 44<br>+ ILP     | 36               | 12              | 24                  | 16                     | 16                  | 16                      |

see NVIDIA CUDA C Programming Guides (different versions)  
 performance guidelines/multiprocessor level; compute capabilities

# NVIDIA AD102 SM

## Multiprocessor: SM (CC 8.9)

- 128 (64+64) FP32 + 64 INT32 cores
- 2 (!) FP64 cores (not in diagram)
- 4x 4<sup>th</sup> gen tensor cores
- 1x 3<sup>rd</sup> gen RT (ray tracing) core
- ++ thread block clusters, FP8, ... (?)

## 4 partitions inside SM

- 32 (16+16) FP32 + 16 INT32 cores
- 4x LD/ST units; 4 SFUs each
- 1x 4<sup>th</sup> gen tensor core each
- Each has: warp scheduler, dispatch unit, 16K register file



# NVIDIA Ada Lovelace AD10x Architecture (2022)



AD 10x / AD 102 (RTX 4090)

Full GPU: 144 SMs (in 12 GPCs/72 TPCs)

- 64K 32-bit registers / SM = 256 KB register storage per SM
- 128 KB shared memory / L1 per SM

For 144 SMs on full GPU [*RTX 4090: 128; RTX 4080 16GB: 76; RTX 4080 12GB: 60*]

- 36 MB register storage, 18 MB shared mem / L1 storage =  
**54 MB context+”shared context” storage !**
- L2 cache size on RTX 4090: 72 MB
- 18,432 FP32 cores (128 FP32 cores per SM) [*RTX 4090: 16,384*]
- 294,912 max threads in flight (max warps / SM = 64) [*RTX 4090: 262,144*]

# Comparisons

## RTX GPUs



| Graphics Card            | GeForce RTX 2080 Ti | GeForce RTX 3090 Ti | GeForce RTX 4090          |
|--------------------------|---------------------|---------------------|---------------------------|
| CUDA Cores               | 4352                | 10752               | 16384                     |
| GPCs                     | 6                   | 7                   | 11                        |
| TPCs                     | 34                  | 42                  | 64                        |
| SMs                      | 68                  | 84                  | 128                       |
| GPU Boost Clock (MHz)    | 1635                | 1860                | 2520                      |
| FP32 TFLOPS              | 14.2                | 40                  | 82.6                      |
| Tensor Cores             | 544 (2nd Gen)       | 336 (3rd Gen)       | 512 (4th Gen)             |
| Tensor TFLOPS (FP8)      | N/A                 | N/A                 | 660.6/1321.2 <sup>1</sup> |
| RT Cores                 | 68 (1st Gen)        | 84 (2nd Gen)        | 128 (3rd Gen)             |
| RT TFLOPS                | 42.9                | 78.1                | 191                       |
| Texture Units            | 272                 | 336                 | 512                       |
| Texture Fill Rate        | 444.7               | 625                 | 1290.2                    |
| ROPs                     | 88                  | 112                 | 176                       |
| Pixel Fill Rate          | 143.9               | 208.3               | 443.5                     |
| Memory Size and Type     | 11 GB GDDR6         | 24 GB GDDR6X        | 24 GB GDDR6X              |
| Memory Clock (Data Rate) | 14 Gbps             | 21 Gbps             | 21 Gbps                   |
| Memory Bandwidth         | 616 GB/sec          | 1008 GB/sec         | 1008 GB/sec               |



# Comparisons

## RTX GPUs

| Graphics Card                 | GeForce RTX 2080 Ti               | GeForce RTX 3090 Ti                         | GeForce RTX 4090                 |
|-------------------------------|-----------------------------------|---|----------------------------------|
| <b>L1 Cache/Shared Memory</b> | 6528 KB                           | 10752 KB                                    | 16384 KB                         |
| <b>L2 Cache</b>               | 5632 KB                           | 6144 KB                                     | 73728 KB                         |
| <b>TGP</b>                    | 260 W                             | 450 W                                       | 450 W                            |
| <b>Transistor Count</b>       | 18.6 Billion                      | 28.3 Billion                                | 76.3 Billion                     |
| <b>Die Size</b>               | 754 mm <sup>2</sup>               | 628.4 mm <sup>2</sup>                       | 608.5 mm <sup>2</sup>            |
| <b>Manufacturing Process</b>  | TSMC 12 nm FFN<br>(FinFET NVIDIA) | Samsung 8 nm 8N<br>NVIDIA Custom<br>Process | TSMC 4N NVIDIA<br>Custom Process |

1- Using Sparsity feature

Thank you.