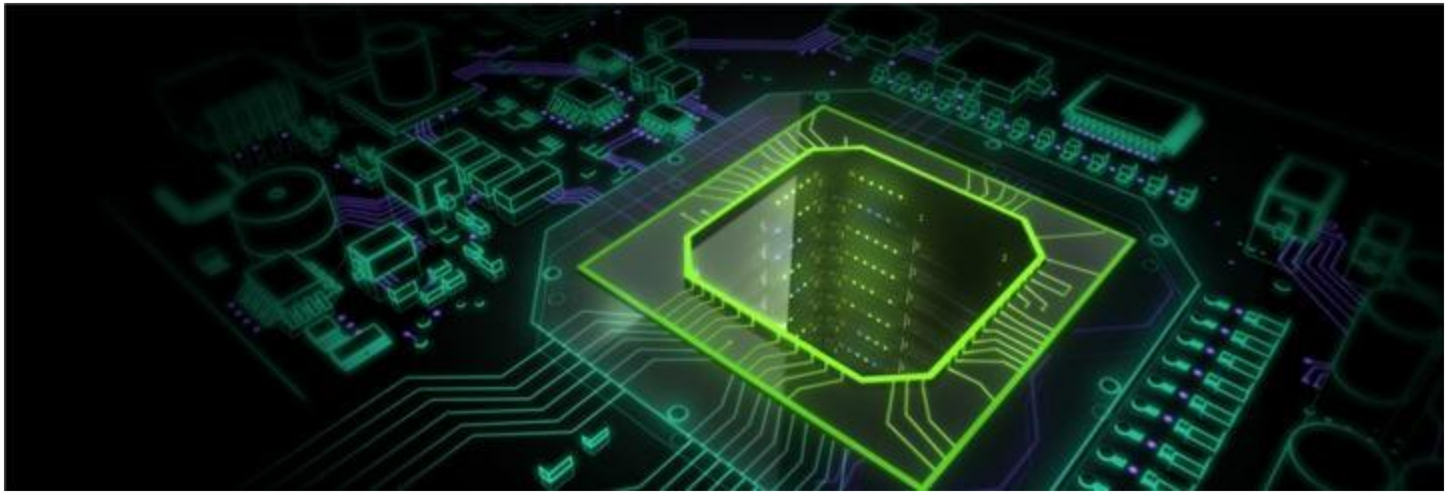
CSCI-GA.3033-004

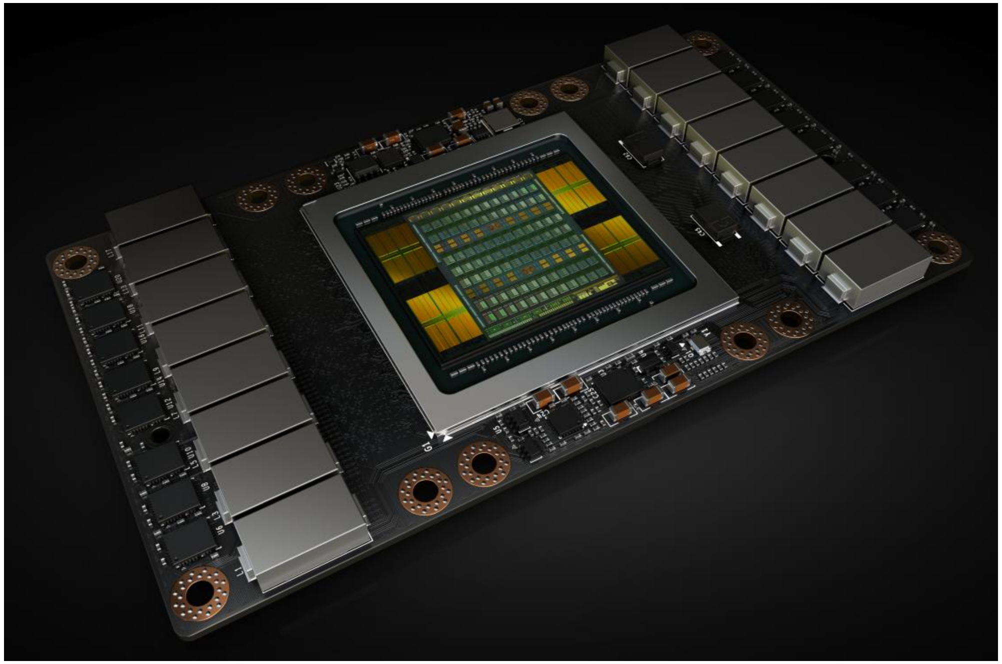
Graphics Processing Units (GPUs): Architecture and Programming

A Glimpse at the State-of-the-art

Mohamed Zahran (aka Z)
mzahran@cs.nyu.edu
http://www.mzahran.com



Volta Architecture

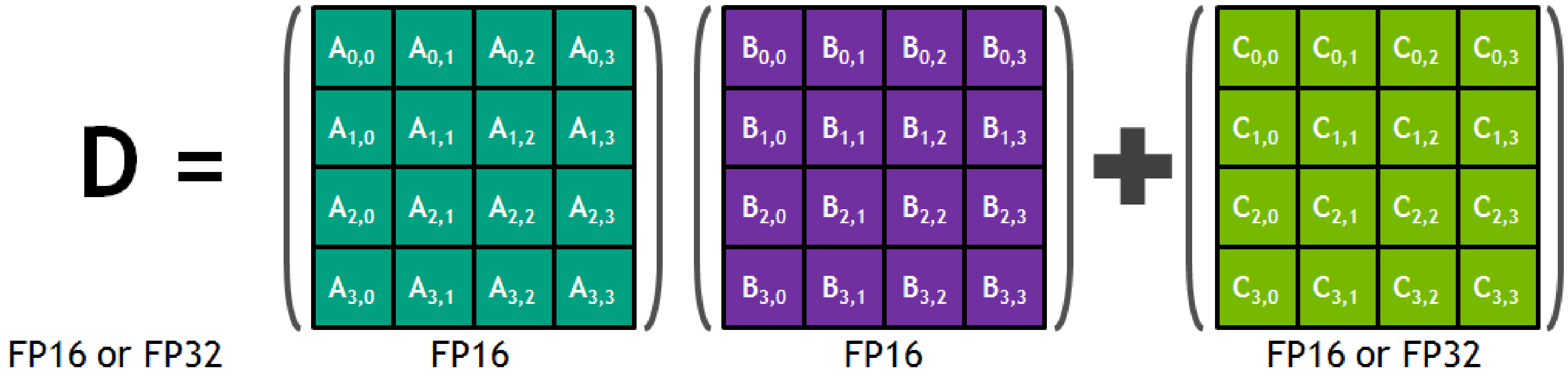


Quick Info

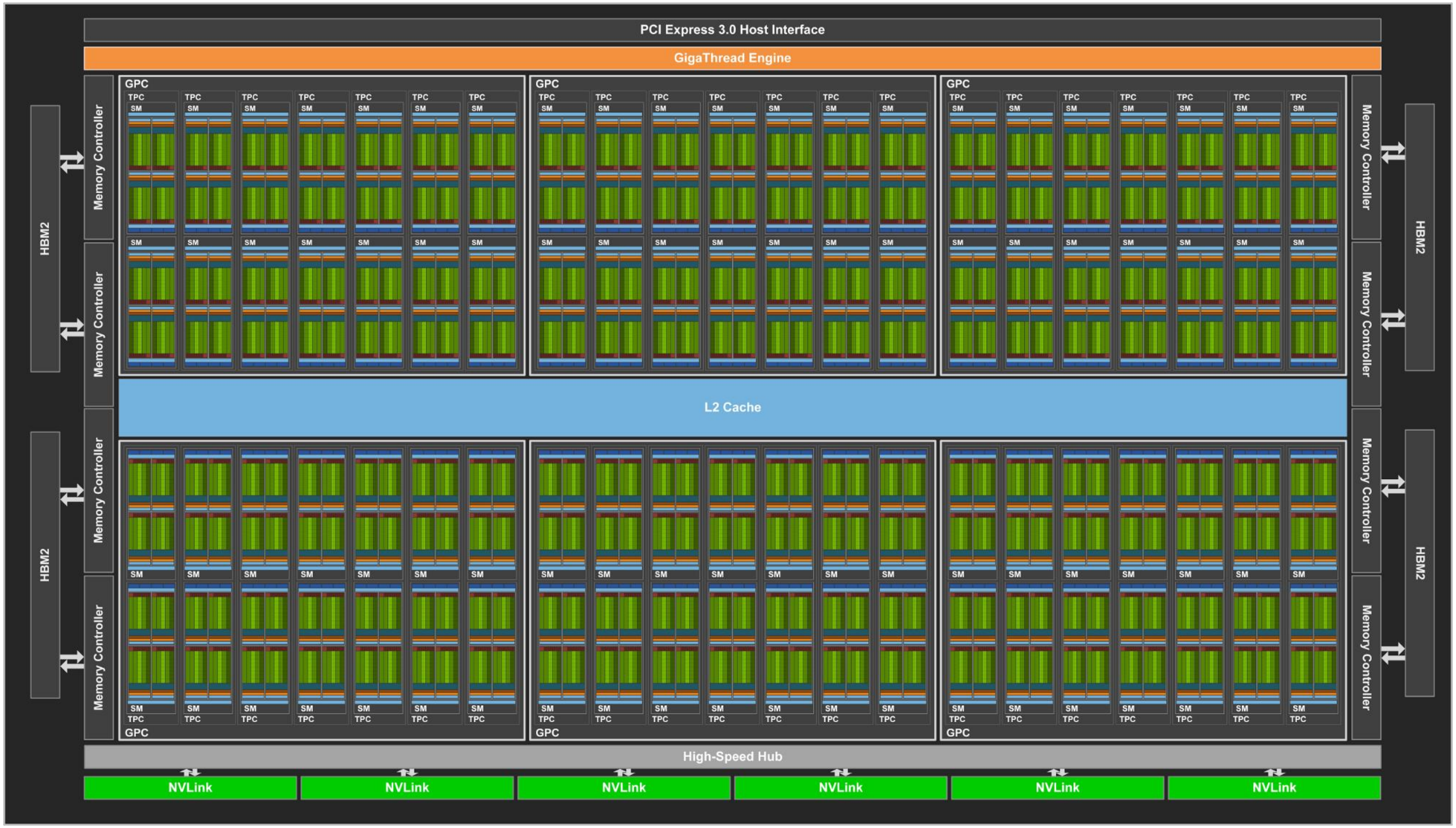
- Introduced in 2017
- Compute capability 7.0
- SM has tensor cores in addition to traditional ones
- 21.1 billion transistors
- 12 nm process technology
- NVLINK 2
  - V100 supports up to 6 NVLink links
  - 1 NVLINK provides 25 GB/s
- HBM2 global memory
  - 16GB
  - delivers 900 GB/sec peak memory bandwidth
- TDP (Thermal Design Power) level of 300W

Tensor Cores

- Tensor Cores are programmable matrix-multiply-and-accumulate units
- Each Tensor Core provides a matrix processing array which performs the operation D = A * B + C, where A, B, C and D are 4×4 matrices
- The Tesla V100 GPU contains 640 Tensor Cores
- 8 per SM



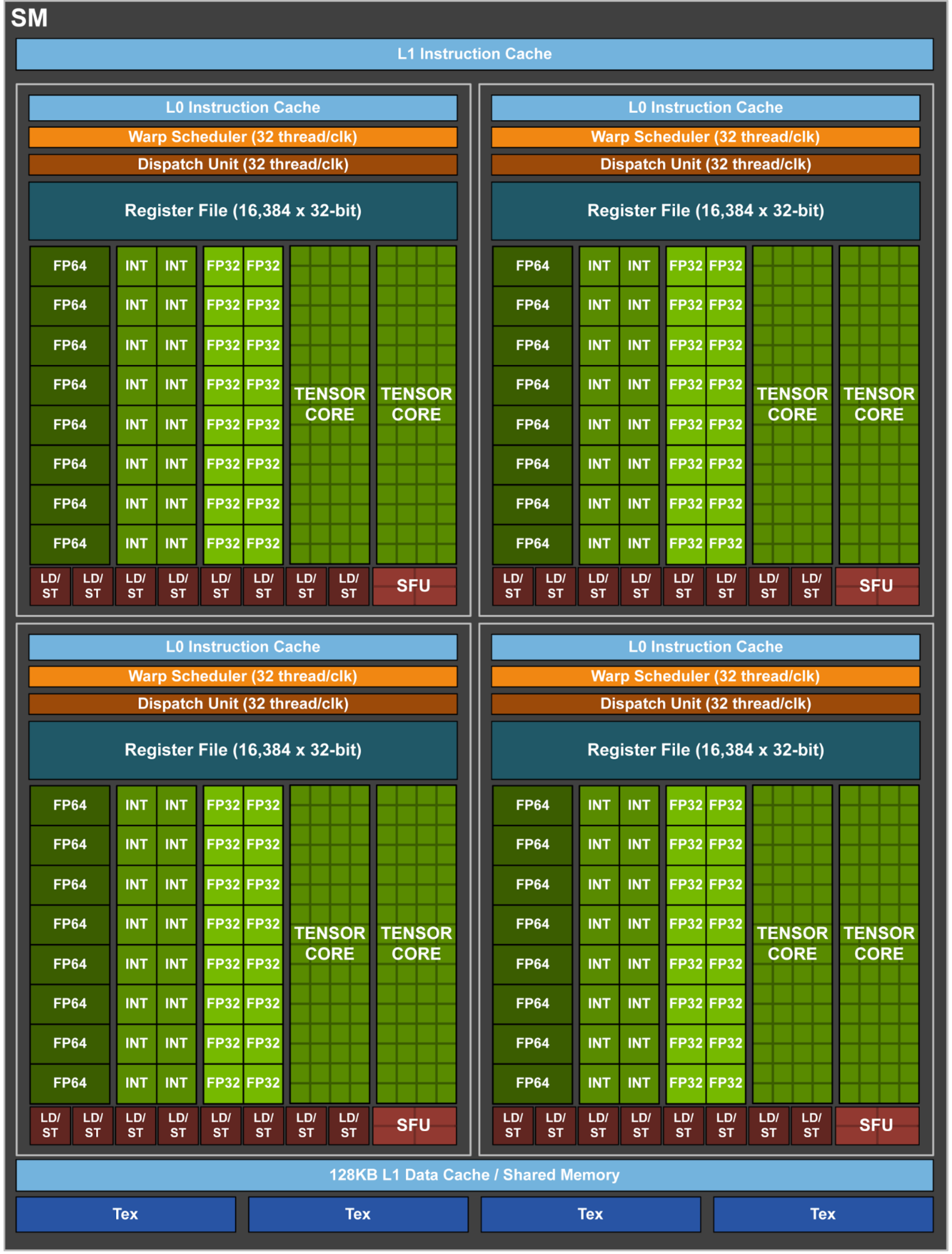
$$\mathbf{D} = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3}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



Graphics Processing Clusters (GPCs) → Texture Processing Clusters (TPCs) → SMs

V100 in numbers

- Six GPCs
- 42 TPCs (each including two SMs)
- 84 Volta SMs, each SM contains:
  - 64 FP32 Cores
  - 64 INT32 Cores
  - 32 FP64 Cores
  - 8 Tensor Cores
- Eight 512-bit memory controllers (4096 bits total).
- 6144 KB of L2 cache


SM
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
FP64
INT
FP32
TENSOR CORE
LD/ ST
SFU
128KB L1 Data Cache / Shared Memory
Tex

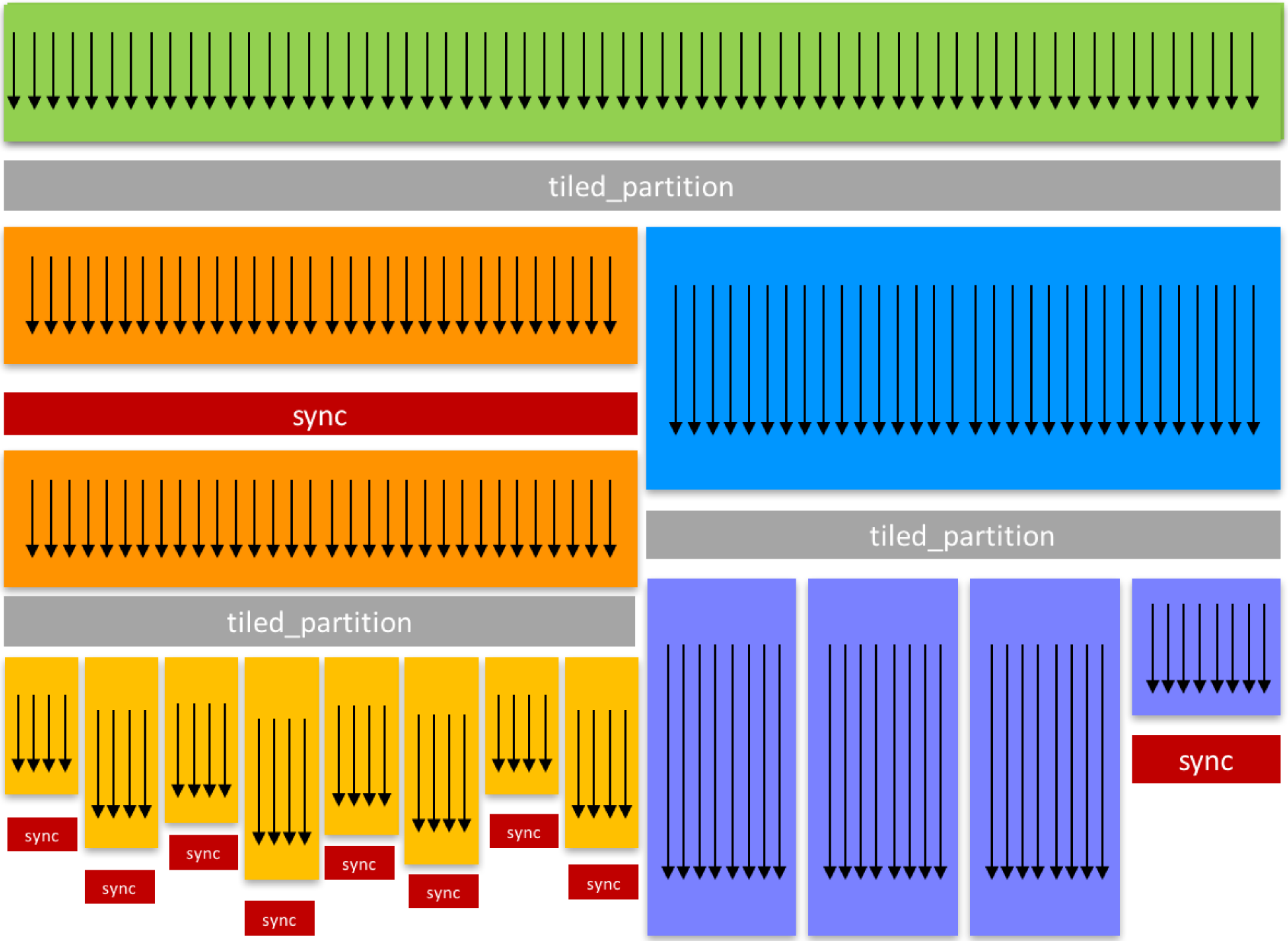
SM is partitioned into:

- L1/shared combined:
  - 128 KB/SM
  - 96 KB Shared Memory
  - All of it used as cache is no shared mem.
- Four processing blocks
- Each with:
  - 16 FP32 Cores
  - 8 FP64 Cores
  - 16 INT32 Cores
  - two Cores
  - a new L0 instruction cache
  - one warp scheduler
  - one dispatch unit
  - 64 KB Register File

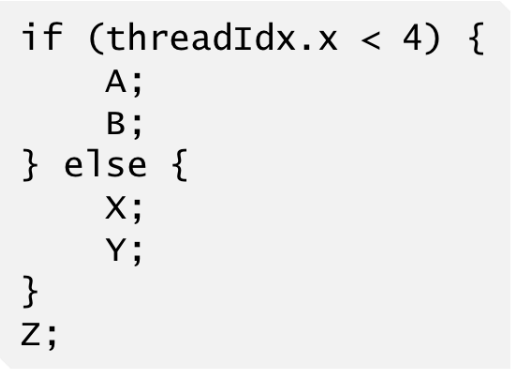
Programming Wise: Cooperative Launch APIs

- What if you want to do synchronization with a smaller number of threads than a block? or bigger?
- The Cooperative Groups programming model describes synchronization patterns both within and across CUDA thread blocks.
- Also provides host-side APIs to launch grids whose threads are all guaranteed to be executing concurrently to enable synchronization across thread blocks.

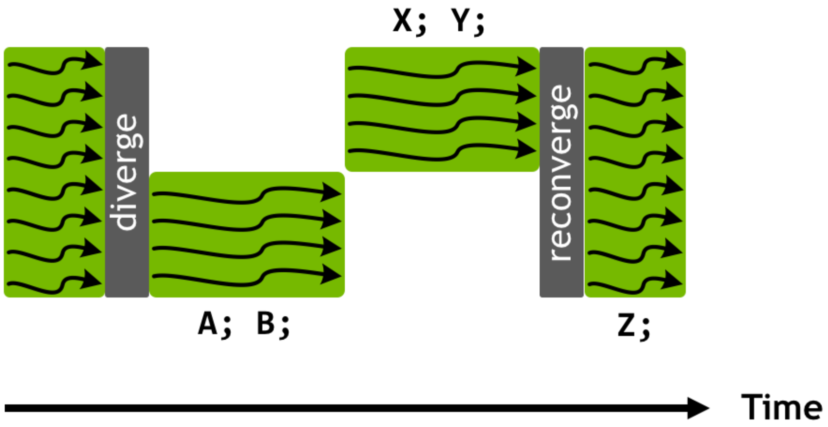
Programming Wise: Cooperative Launch APIs



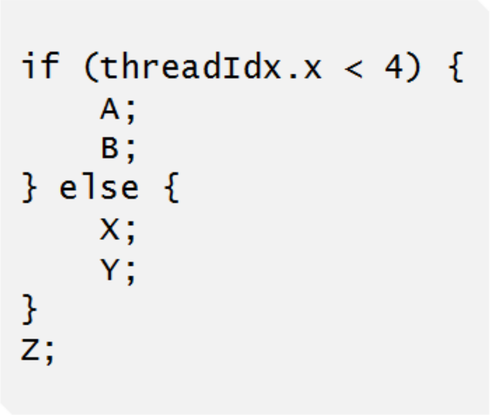
Programming Wise: Independent Thread Scheduling



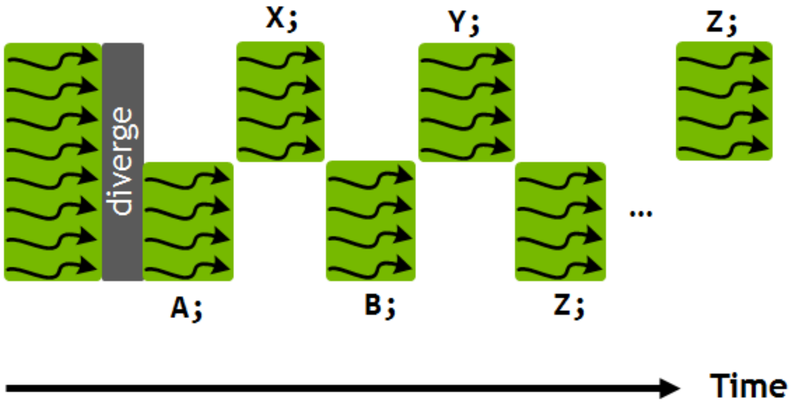
```
if (threadIdx.x < 4) {
    A;
    B;
} else {
    X;
    Y;
}
Z;
```



Before

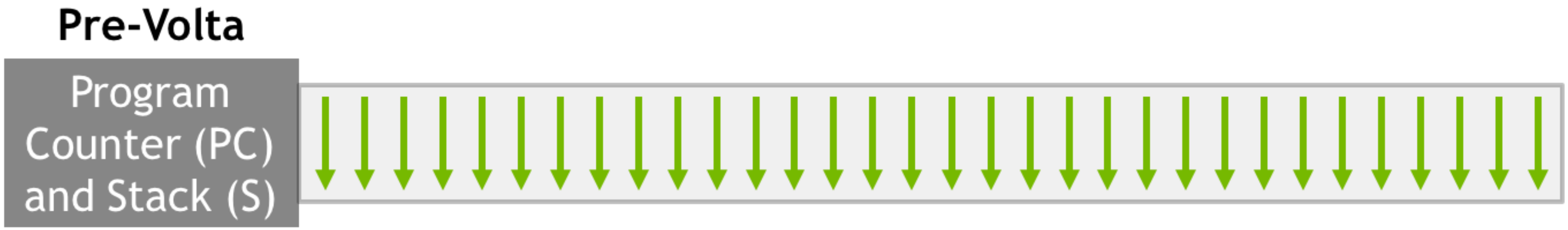


```
if (threadIdx.x < 4) {
    A;
    B;
} else {
    X;
    Y;
}
Z;
```

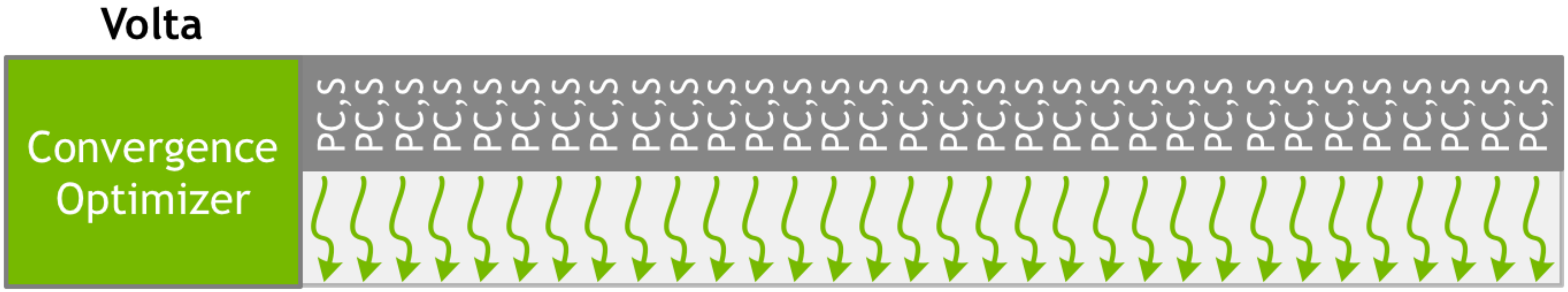


After

Programming Wise: Independent Thread Scheduling

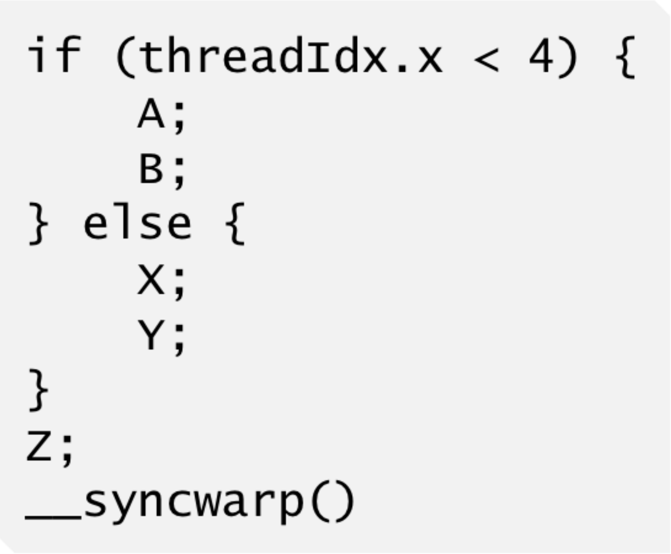


32 thread warp

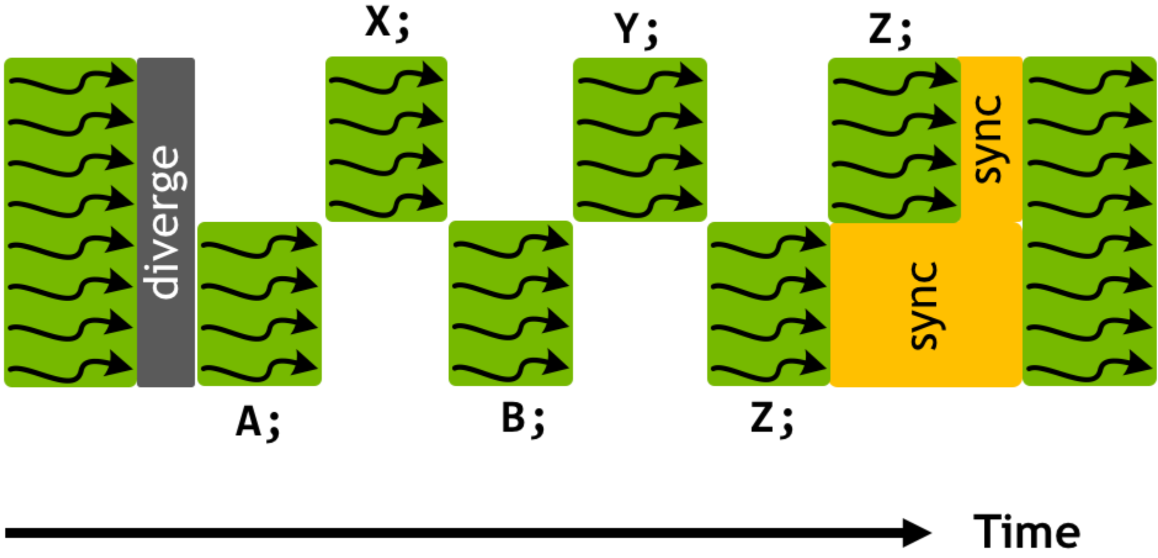

Volta
Convergence Optimizer
PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S PC,S

32 thread warp with independent scheduling

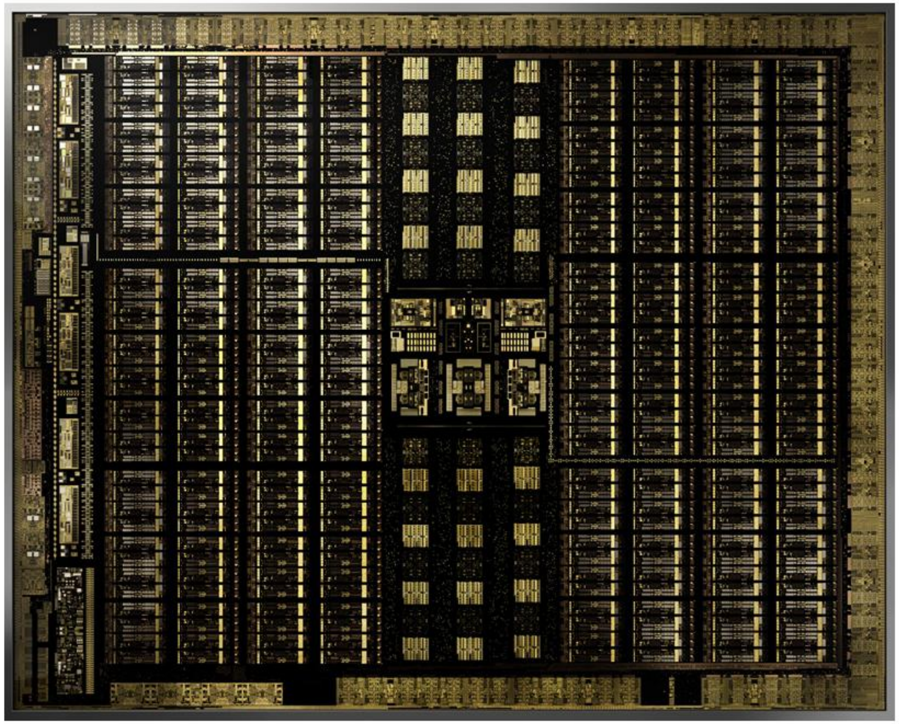
Programming Wise:
Independent Thread Scheduling



```
if (threadIdx.x < 4) {
    A;
    B;
} else {
    X;
    Y;
}
Z;
__syncwarp()
```

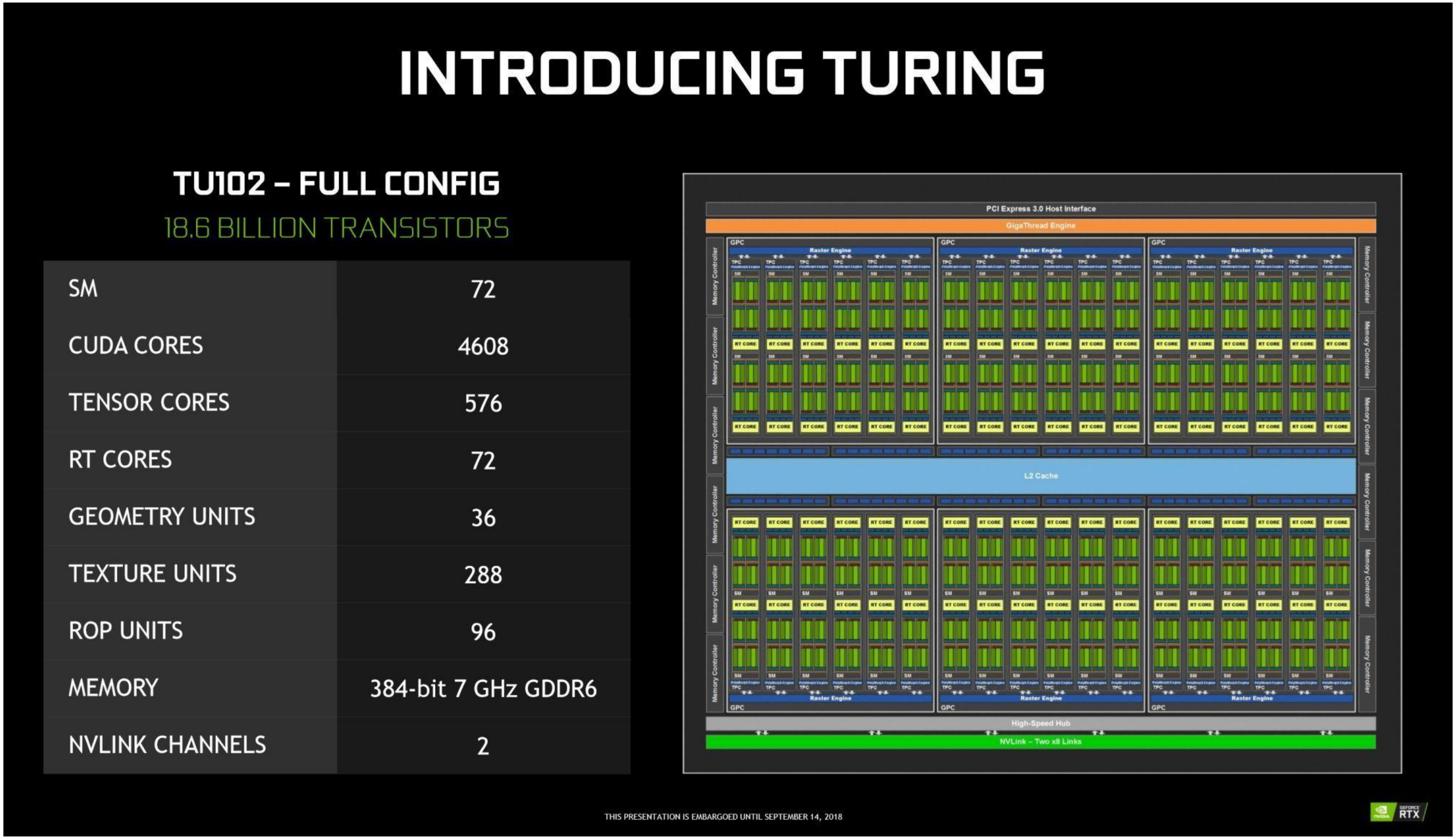


Turing Architecture



Quick Info

- Introduced in 2018
- Introducing Ray-Tracing cores in addition to other core types
- Designed to handle real-time ray tracing.



INTRODUCING TURING

TU102 – FULL CONFIG

18.6 BILLION TRANSISTORS

SM	72
CUDA CORES	4608
TENSOR CORES	576
RT CORES	72
GEOMETRY UNITS	36
TEXTURE UNITS	288
ROP UNITS	96
MEMORY	384-bit 7 GHz GDDR6
NVLINK CHANNELS	2

THIS PRESENTATION IS EMBARGOED UNTIL SEPTEMBER 14, 2018

6 GPC → 36 TPC → 72 SMs

Each SM

- 1 Ray-Tracing core
- 64 CUDA Cores (i.e. SPs)
- 8 Tensor Cores
- 256 KB register file
- 96 KB of L1/shared memory

Conclusions

- More specialized cores are being added in new architectures.
- More fine grain synchronization are giving more control to the programmer.
- Still computation is the cheapest compared to memory access and communication.