GPU Architecture

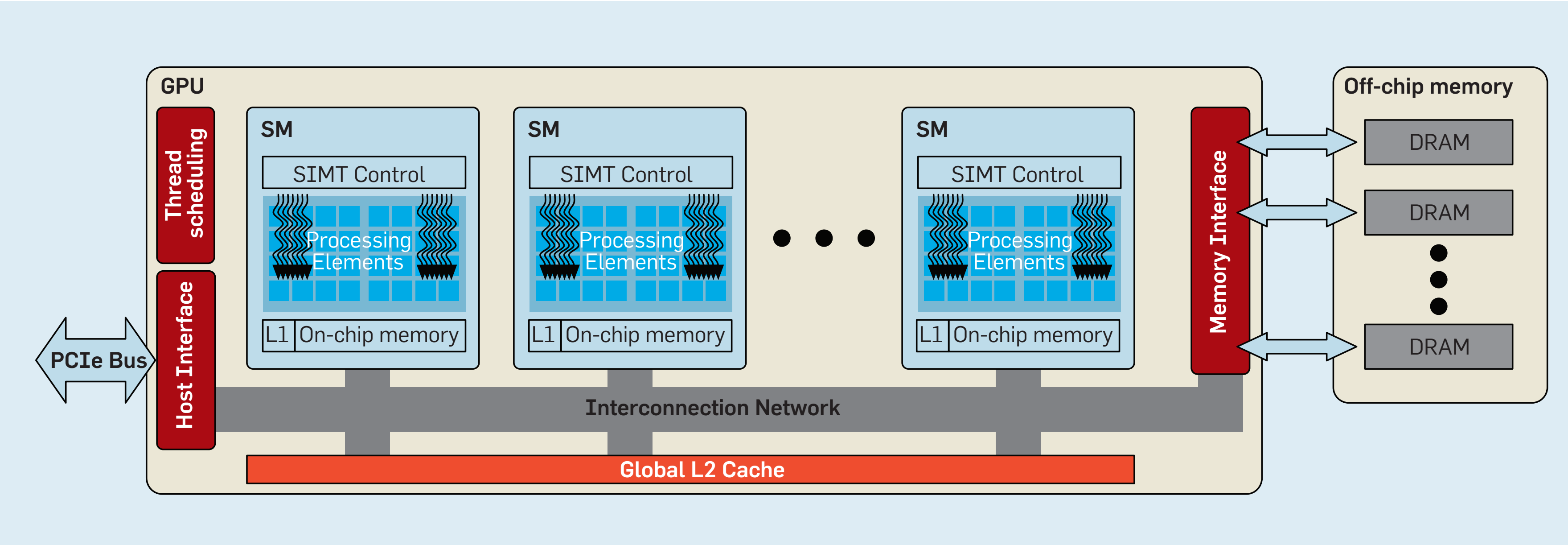
Stefano Markidis

Three Key-Points

1. GPUs are highly parallel processors combining one or more streaming multiprocessors, called SM
2. Each SM architecture typically includes hundred simple cores
3. The SM handles all the management of thread creation and management in hardware using the SIMT architecture

GPUs are **Array of Multiprocessors** …

… referred as **streaming multiprocessors**, referred as **SMs**. Each SM supports the execution of thousands of threads.



Fermi-generation GPU Architecture, i.e. the GF100 used in Tesla C2050

How Many SMs does a GPU have?

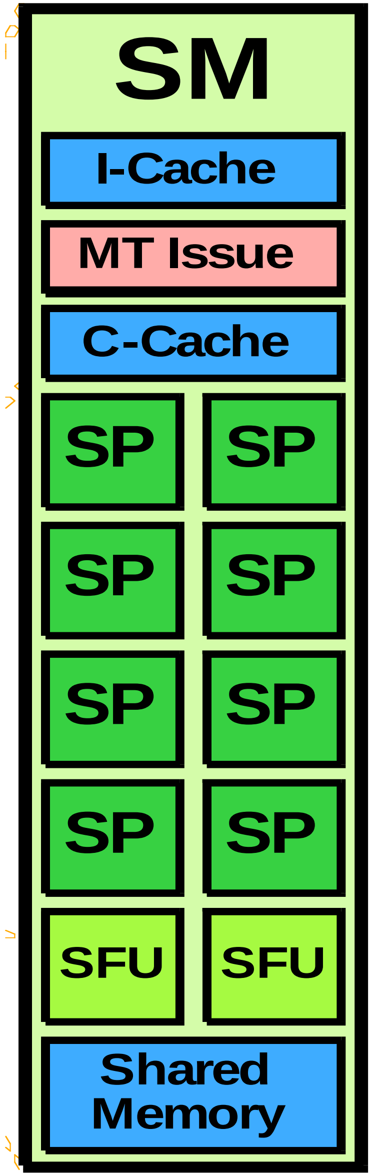
- **Quadro K420**: 1 SM.
- **Tesla K80** consists of 2 integrated GPUs with 13 SMs each
- The new Tesla **V100 GV100 GPU** has 80 SMs.

- How many SMs does your GPU have?

SMs have Simple Cores, called SP

A Streaming Multiprocessor (SM):

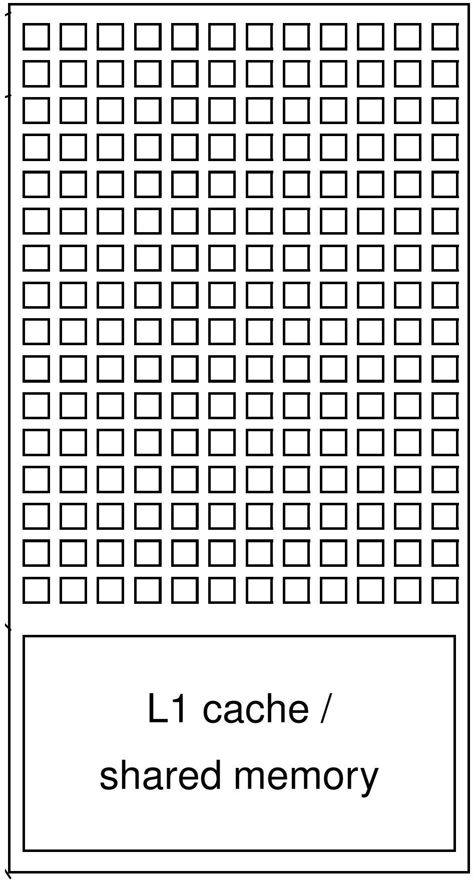
- A collection of 8/32/192 SP (depends on SM architecture) simple cores, called SP.
- All the SP cores in SM run the same instructions
- Has some fast cache shared memory
- Can synchronize



SMX = Next Generation SM (2012)

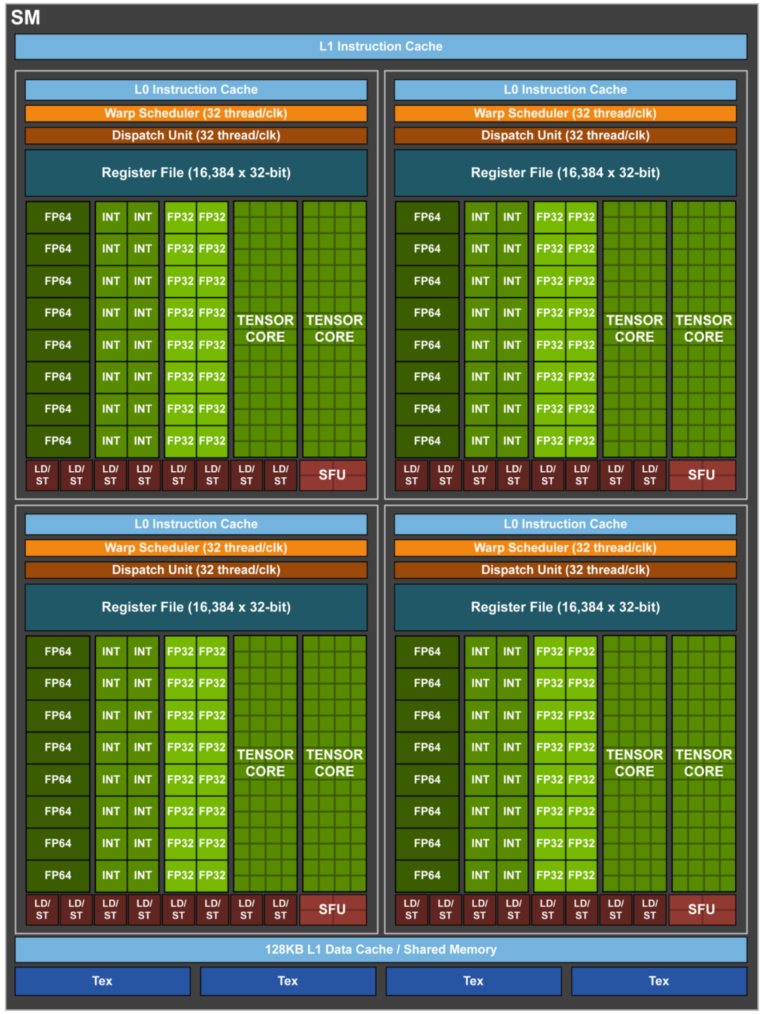
SM Architecture introduced in Kepler GPU:

- **192 cores per SMX**



The New Volta Streaming Processor (2017)

- Major new features:
    - New **mixed-precision tensor cores**
- GV100 SM has 64 CUDA FP32 cores


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

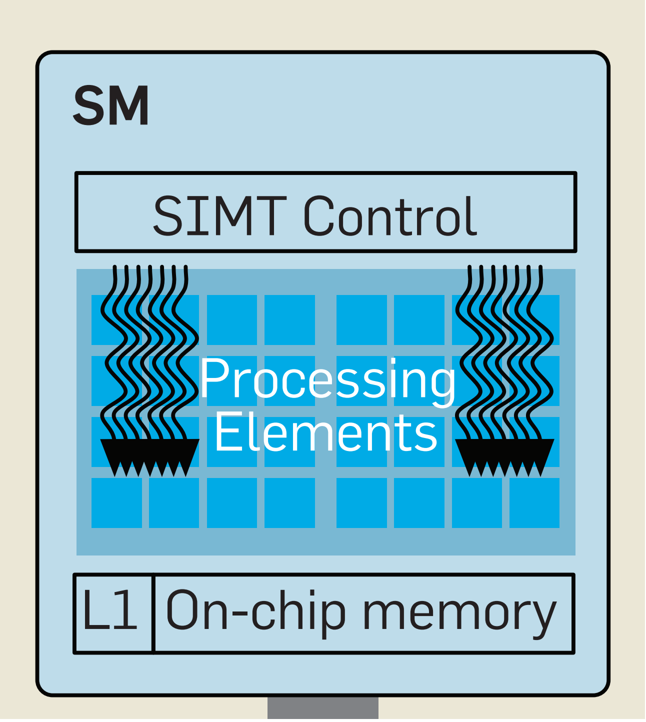
A GPU has lots of computing cores!

- NVIDIA Quadro K420 has **192** cores
- NVIDIA Tesla K80 has **4992** cores
- New Volta GV100 has **5120** cores
- DGX-1 server with 8 Tesla V100 has **40,960** cores!

These are small supercomputers! Our supercomputers at KTH, Beskow, has 53k cores

Management of Hardware Threads

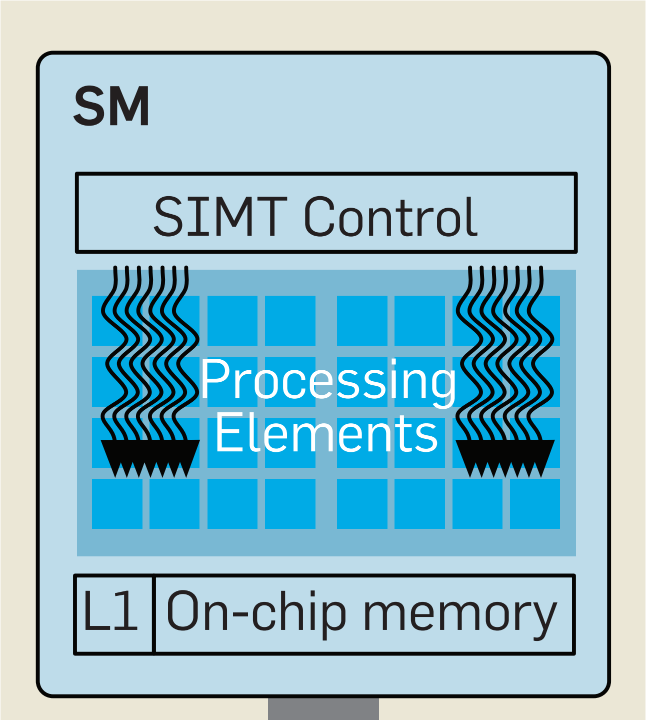
Each SM supports the execution of thousands of threads.



Each SM handles all thread creation, resource allocation and scheduling in hardware

SIMT Architecture

- To manage such large number of threads, the GPUs uses a Single-Instruction, Multiple-Thread (SIMT) architecture:
  - each thread executes a single instruction at the time across all the threads.
- Warps are the basic unit of thread scheduling.
  - Each thread resident on a single SM are executed in groups (**warps**) of 32



To summarize

1. GPUs include one or more several multiprocessors, called SMs
2. The SM architecture evolves in time and includes hundreds of cores each
3. The SM handles all the thread creation and management in hardware using the SIMT architecture