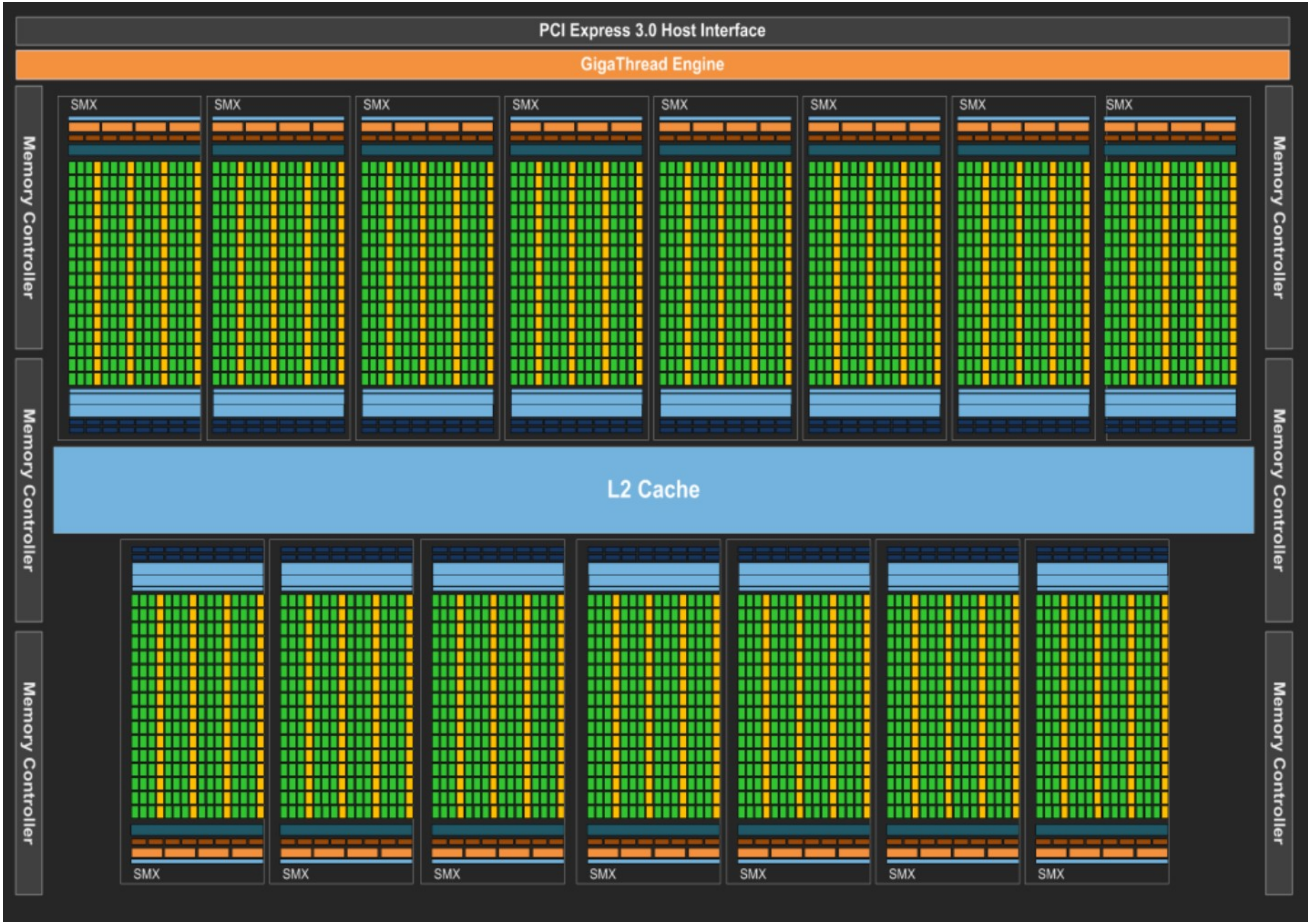
# Nvidia GPU microarchitecures

# Nvidia microarchitectures

- Tesla
- **Fermi**
- **Kepler**
- Maxwell
- **Pascal**
- **Volta**
- Turing
- **Ampere**

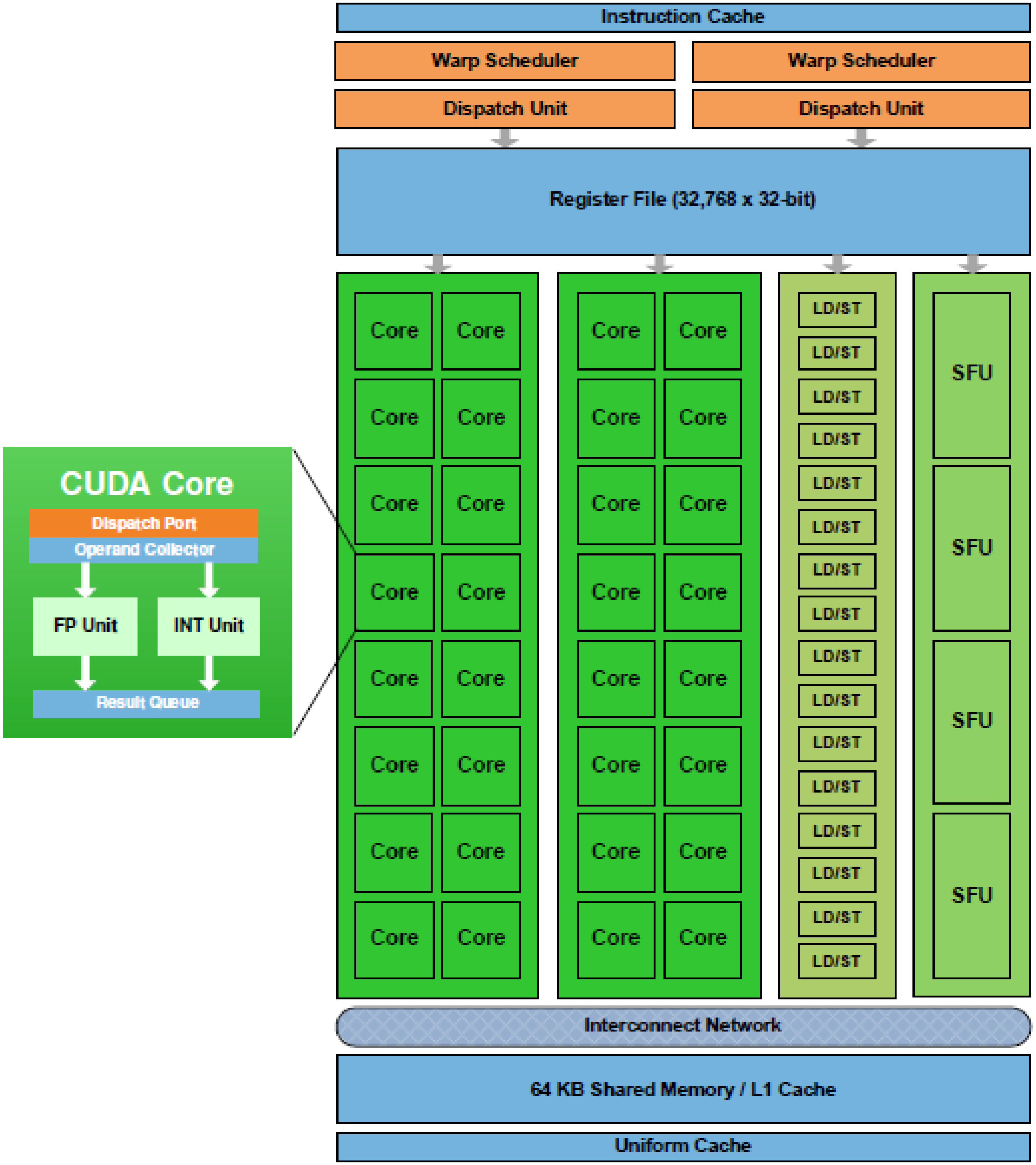
# Kepler GPU


PCI Express 3.0 Host Interface
GigaThread Engine
Memory Controller
Memory Controller
Memory Controller
Memory Controller
Memory Controller
Memory Controller
SMX
SMX
SMX
SMX
SMX
SMX
SMX
SMX
L2 Cache
SMX
SMX
SMX
SMX
SMX
SMX
SMX

# Streaming Multiprocessor(SM)

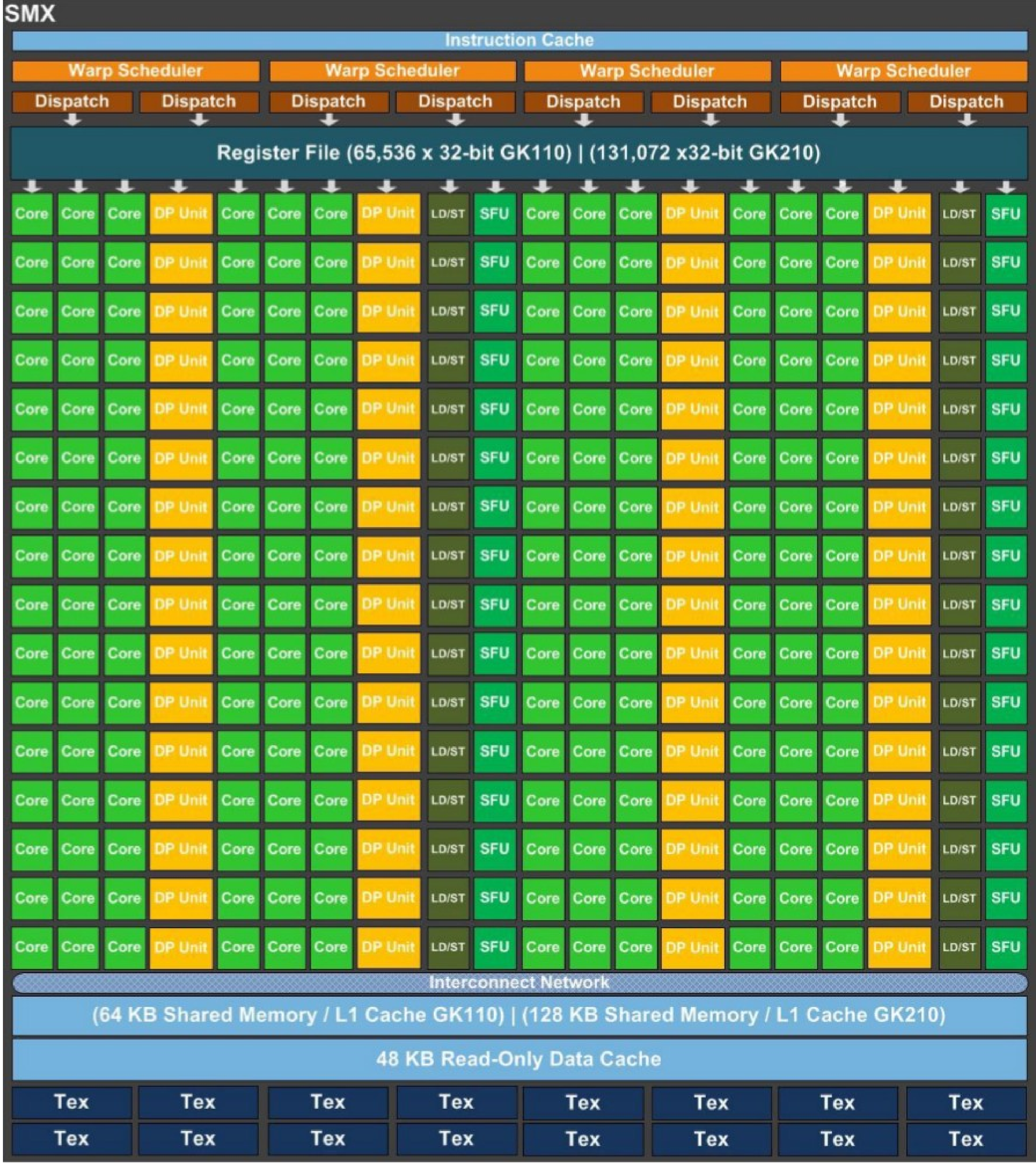
- Basic building block
- Contains
  - Multiple CUDA cores
  - Special Function Units (SFUs)
  - Register file
  - L1 cache / Shared memory
  - ...

# Fermi

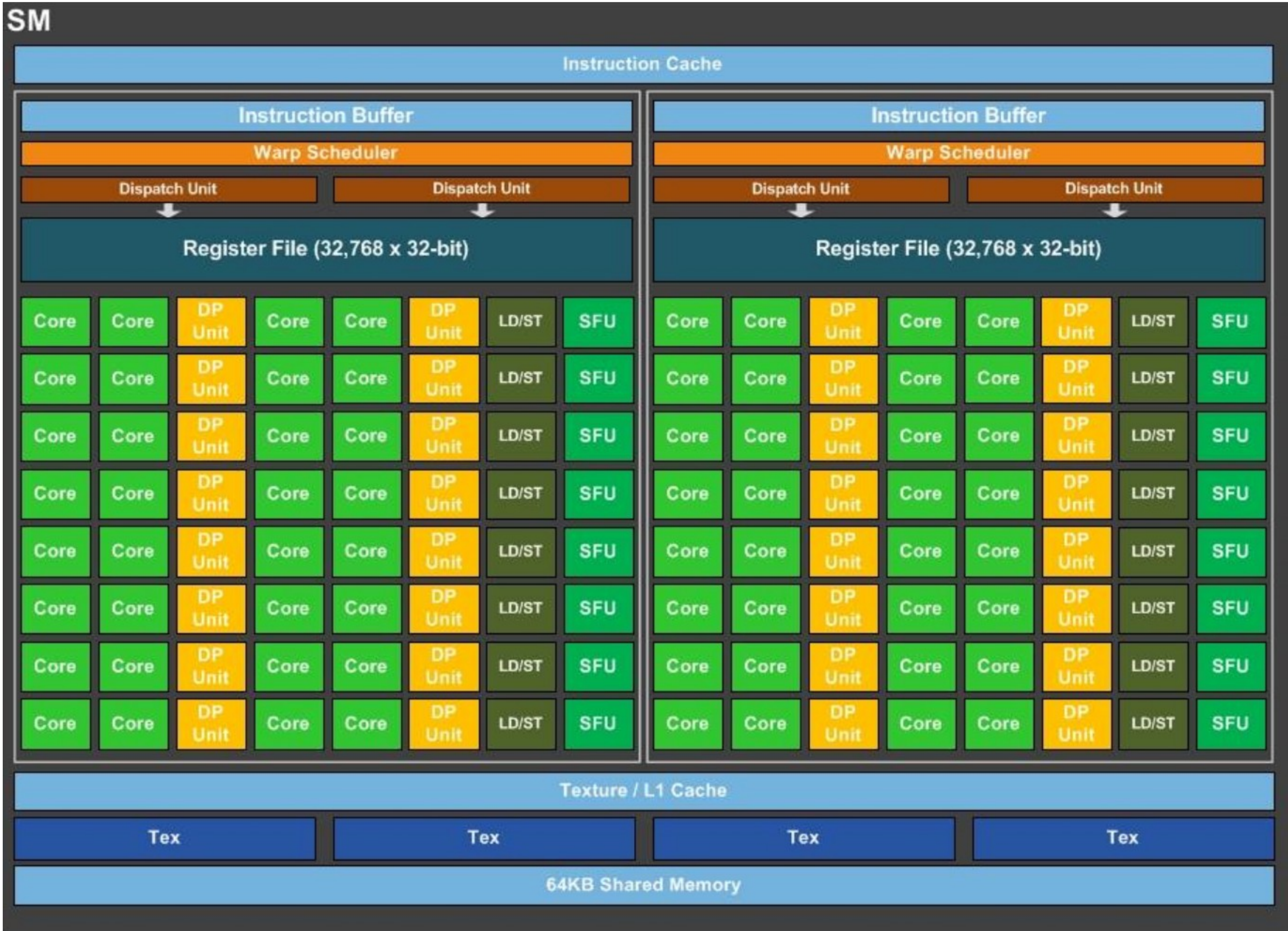

Instruction Cache
Warp Scheduler
Warp Scheduler
Dispatch Unit
Dispatch Unit
Register File (32,768 x 32-bit)
CUDA Core
Dispatch Port
Operand Collector
FP Unit
INT Unit
Result Queue
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
Core
LD/ST
LD/ST
LD/ST
LD/ST
LD/ST
LD/ST
LD/ST
LD/ST
LD/ST
LD/ST
LD/ST
LD/ST
LD/ST
LD/ST
LD/ST
LD/ST
SFU
SFU
SFU
SFU
Interconnect Network
64 KB Shared Memory / L1 Cache
Uniform Cache

# Kepler

- 192 single-precision CUDA cores
- 64 double-precision units
- 32 special function units
- 32 load/store units
- 4 warp schedulers
- 8 instruction dispatch units
- 48 KB Read-Only Data Cache
- 64/128 KB L1 cache / shared memory


SMX
Instruction Cache
Warp Scheduler Warp Scheduler Warp Scheduler Warp Scheduler
Dispatch Dispatch Dispatch Dispatch Dispatch Dispatch Dispatch Dispatch
Register File (65,536 x 32-bit GK110) | (131,072 x32-bit GK210)
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU Core Core Core DP Unit Core Core Core DP Unit LD/ST SFU
Interconnect Network
(64 KB Shared Memory / L1 Cache GK110) | (128 KB Shared Memory / L1 Cache GK210)
48 KB Read-Only Data Cache
Tex Tex Tex Tex Tex Tex Tex Tex
Tex Tex Tex Tex Tex Tex Tex Tex

# Pascal

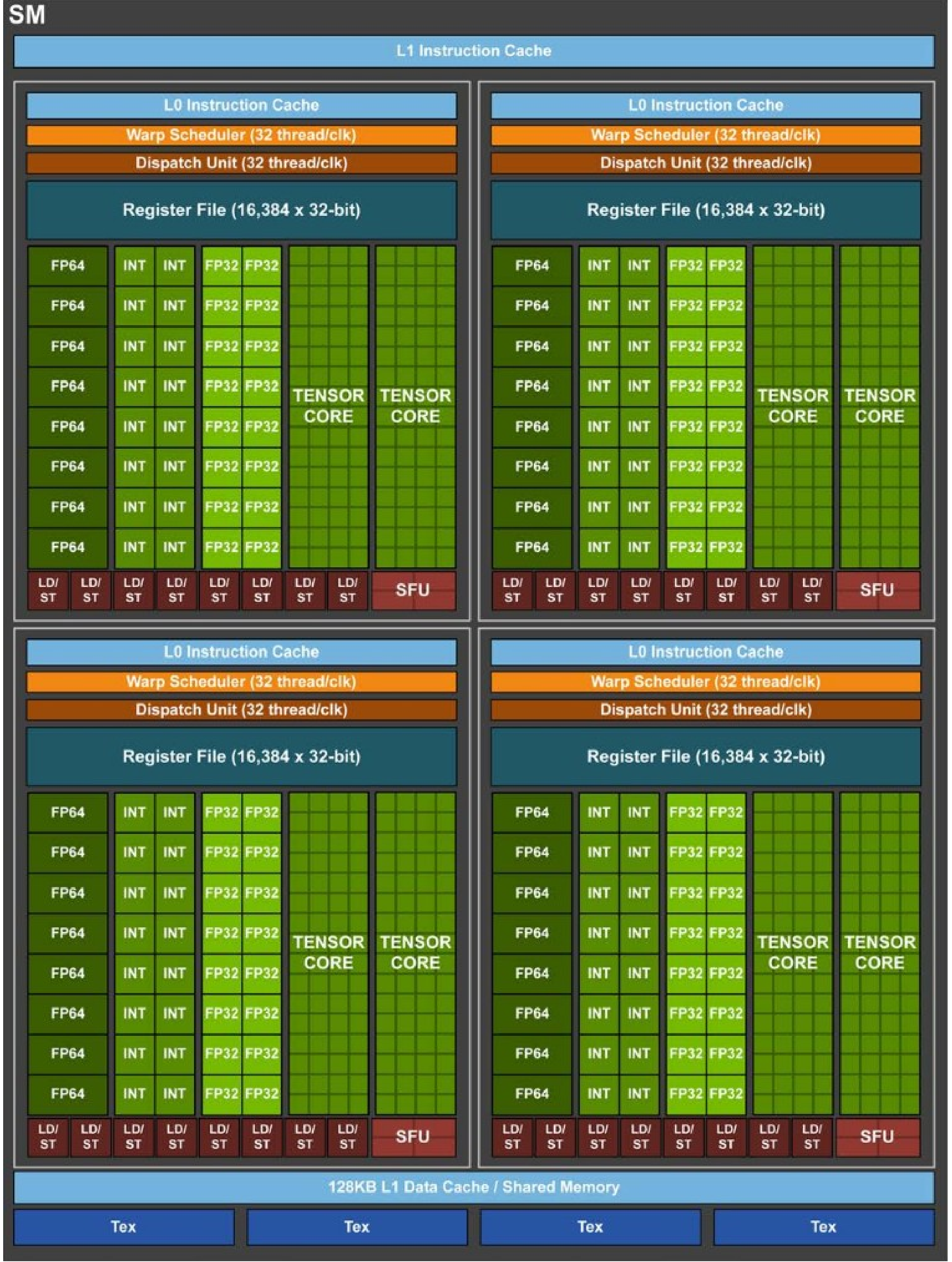

SM
Instruction Cache
Instruction Buffer
Warp Scheduler
Dispatch Unit
Dispatch Unit
Register File (32,768 x 32-bit)
Core
Core
DP Unit
Core
Core
DP Unit
LD/ST
SFU
Instruction Buffer
Warp Scheduler
Dispatch Unit
Dispatch Unit
Register File (32,768 x 32-bit)
Core
Core
DP Unit
Core
Core
DP Unit
LD/ST
SFU
Texture / L1 Cache
Tex
Tex
Tex
Tex
64KB Shared Memory

# Pascal (contd)

- two processing blocks, each with
  - 32 single-precision CUDA Cores
  - 16 double precision (FP64) CUDA Cores
  - one instruction buffer
  - one warp scheduler
  - two dispatch units

# Volta

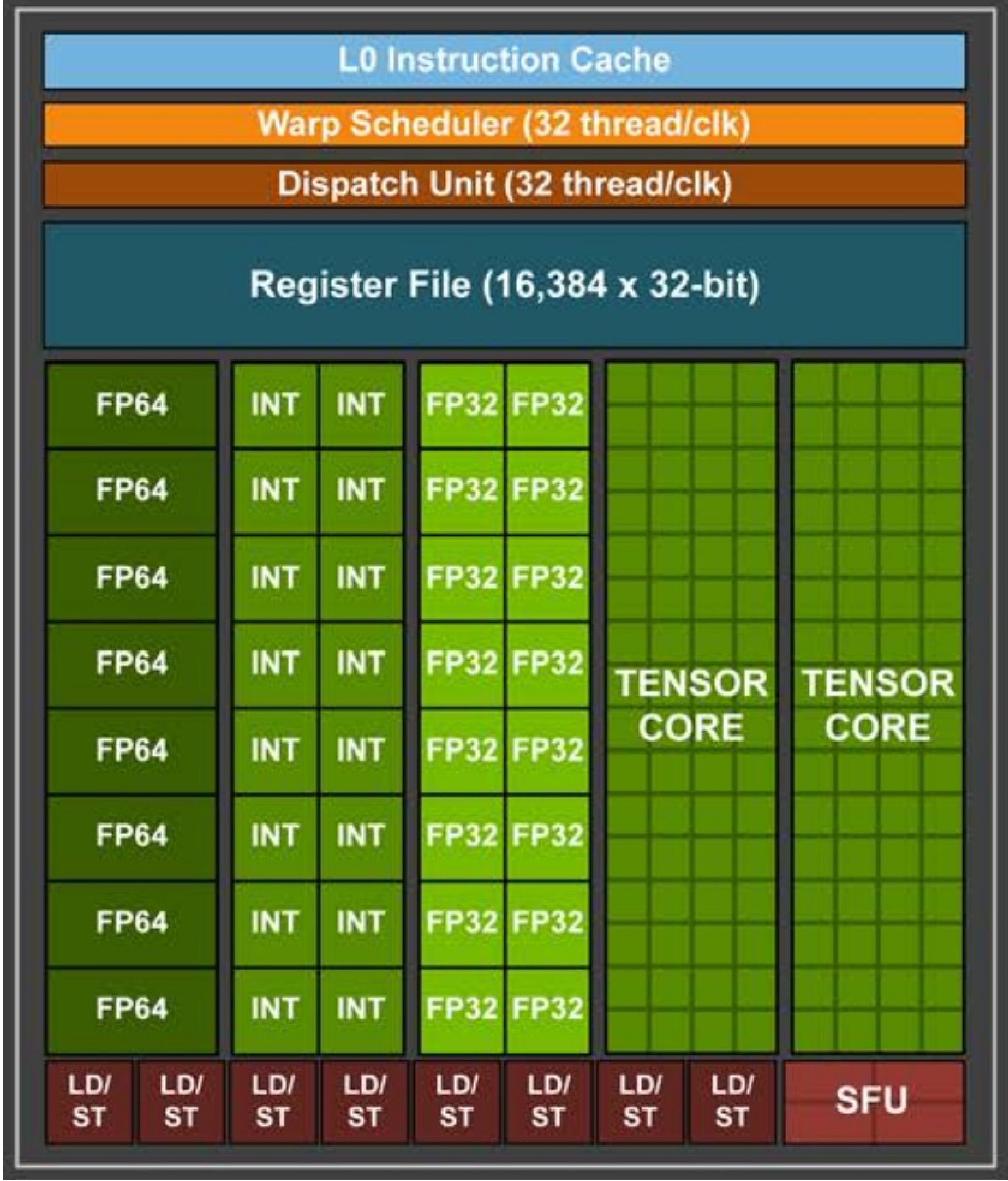
Four processing blocks


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

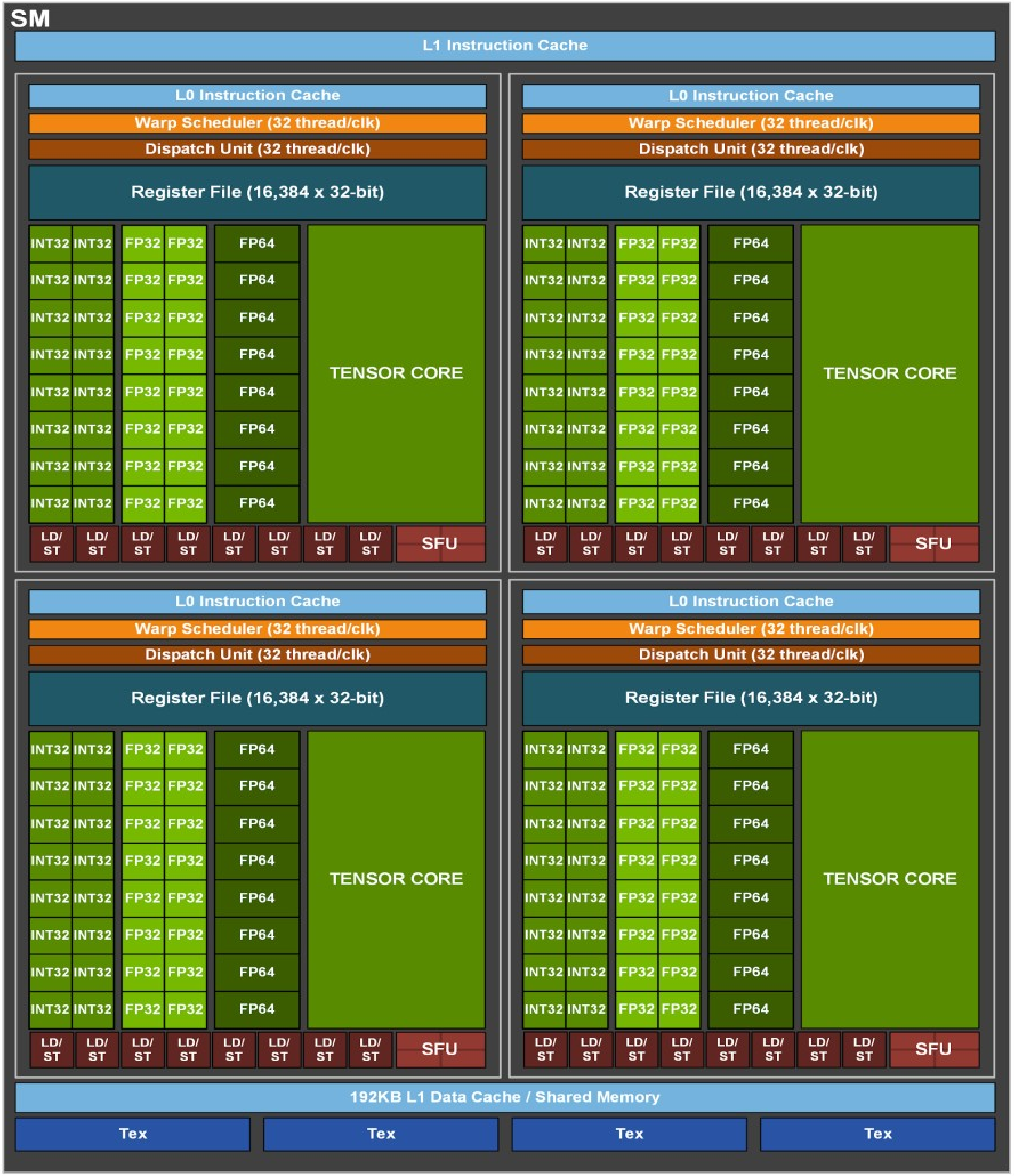
# Volta(contd)

each processing block has

- 16 FP32 Cores
- 8 FP64 Cores
- 16 INT32 Cores
- 2 new mixed-precision Tensor Cores
- one L0 instruction cache
- one warp scheduler
- one dispatch unit
- 64 KB Register File


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
TENSOR CORE
LD/ ST
SFU

# Ampere



SM

L1 Instruction Cache

L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)

INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64

TENSOR CORE

LD/ST LD/ST LD/ST LD/ST LD/ST LD/ST LD/ST LD/ST SFU

L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)

INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64

TENSOR CORE

LD/ST LD/ST LD/ST LD/ST LD/ST LD/ST LD/ST LD/ST SFU

L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)

INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64

TENSOR CORE

LD/ST LD/ST LD/ST LD/ST LD/ST LD/ST LD/ST LD/ST SFU

L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)

INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64
INT32 INT32 FP32 FP32 FP64

TENSOR CORE

LD/ST LD/ST LD/ST LD/ST LD/ST LD/ST LD/ST LD/ST SFU

192KB L1 Data Cache / Shared Memory

Tex Tex Tex Tex