NVIDIA®

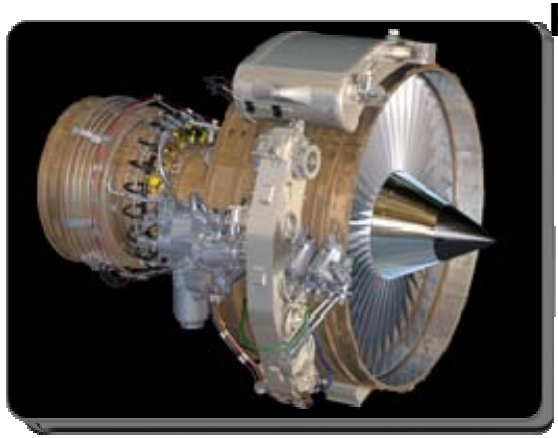
G8x Hardware Architecture

G80 Architecture

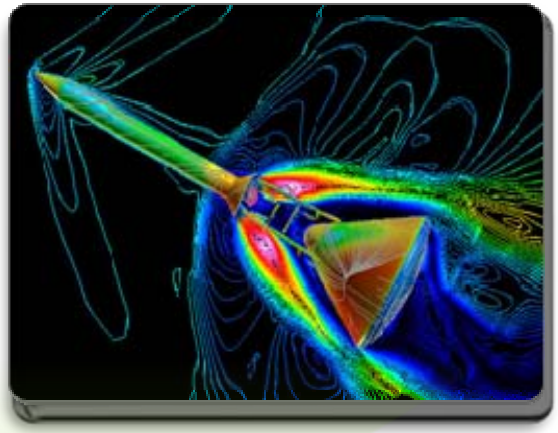
- First DirectX 10 compatible GPU
- Unified shader architecture
- Scalar processors
- Includes new hardware features designed for general purpose computation
  - shared memory (parallel data cache)
  - global load/store memory model (scattered writes)

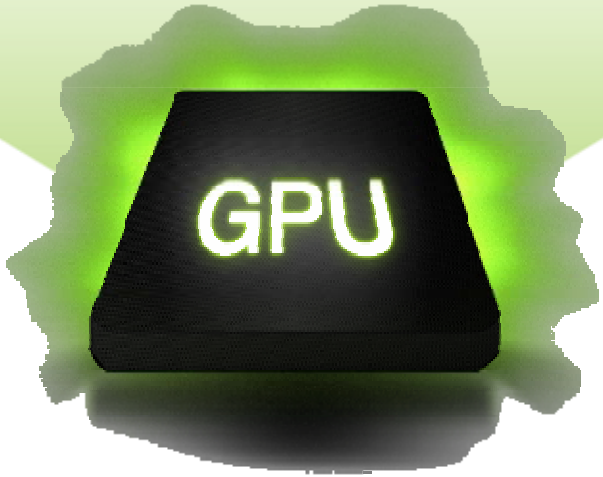
GeForce®
Entertainment

Quadro®
Design & Creation



Tesla™
High Performance Computing

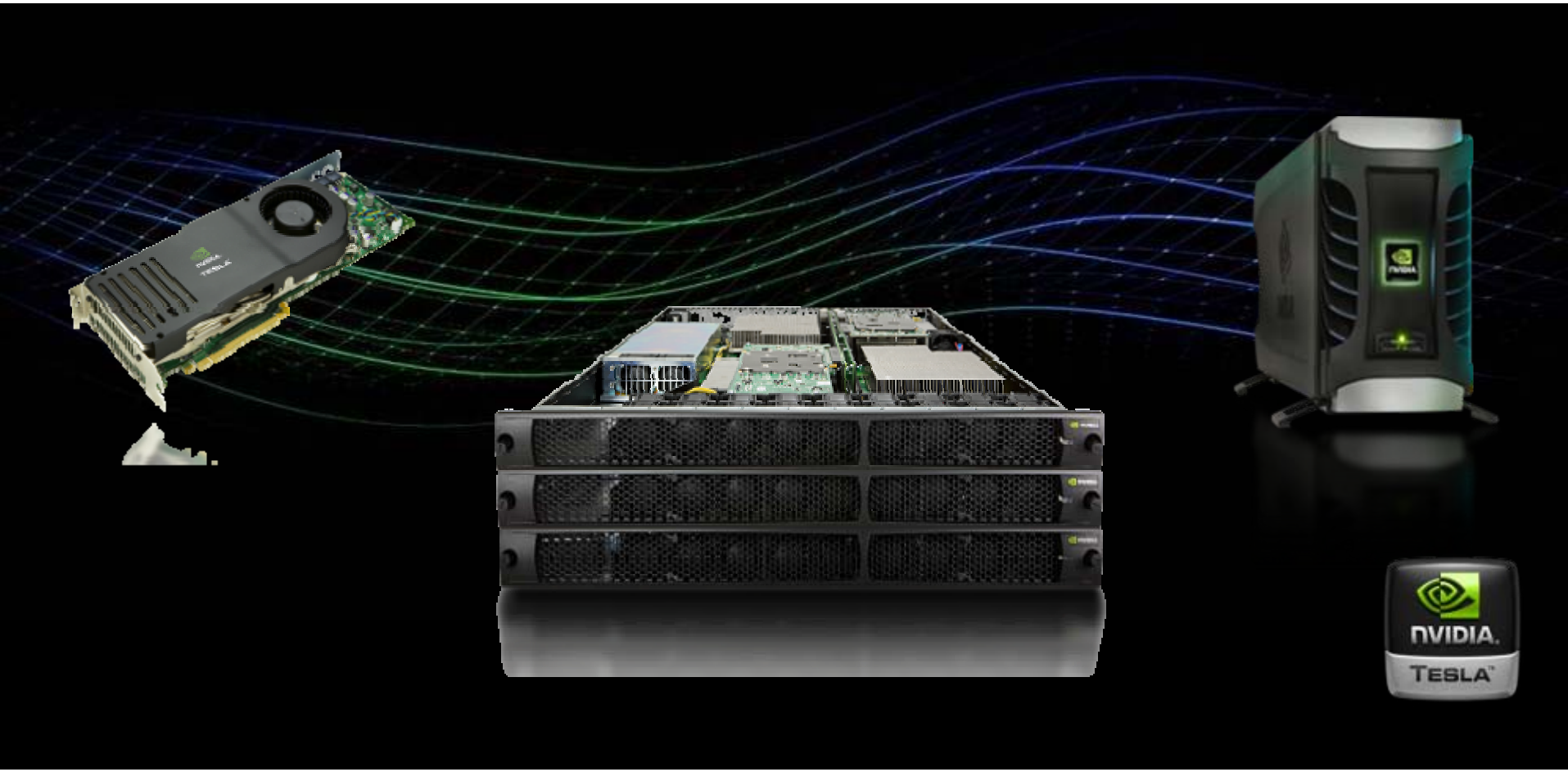




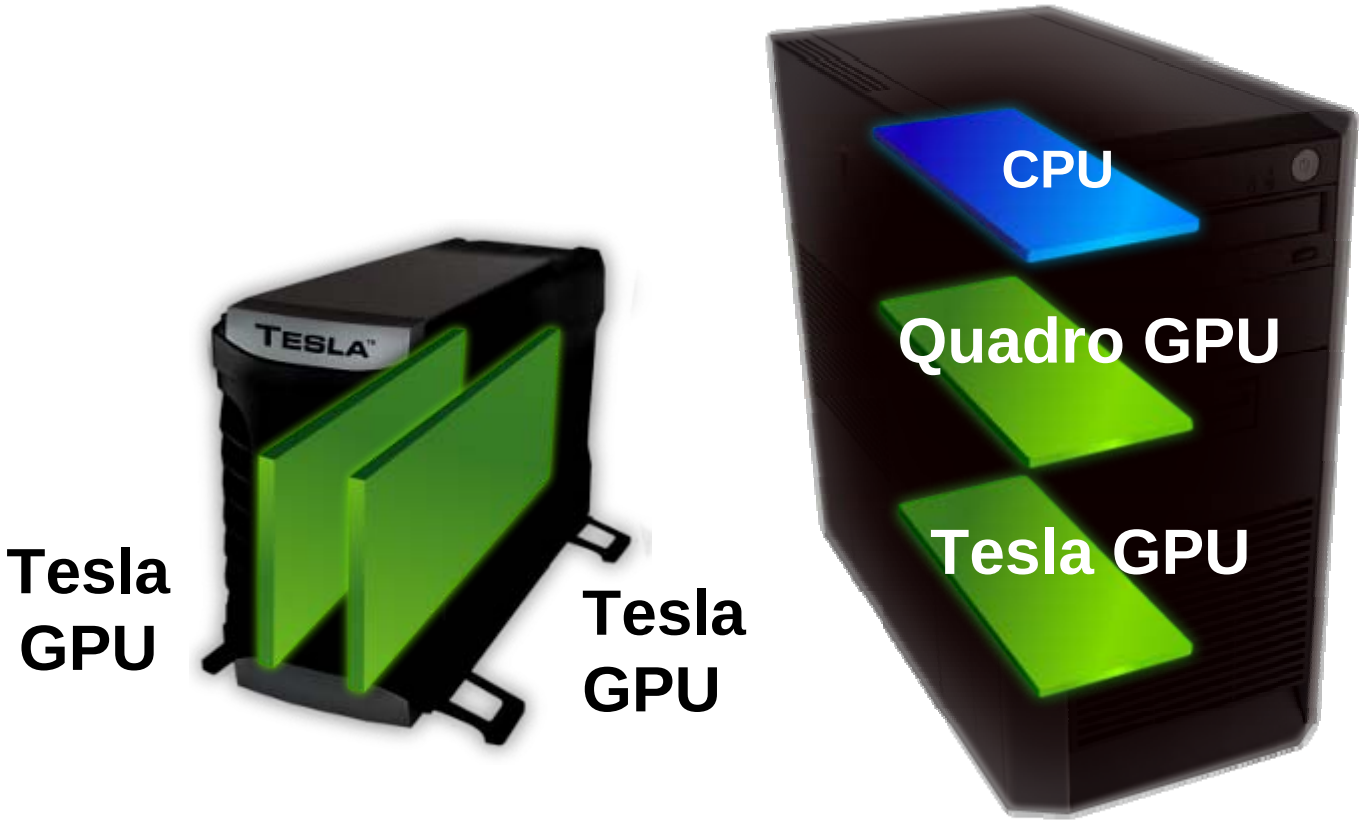
NVIDIA Tesla

Scalable High Density Computing

Massively Multi-threaded Parallel Computing

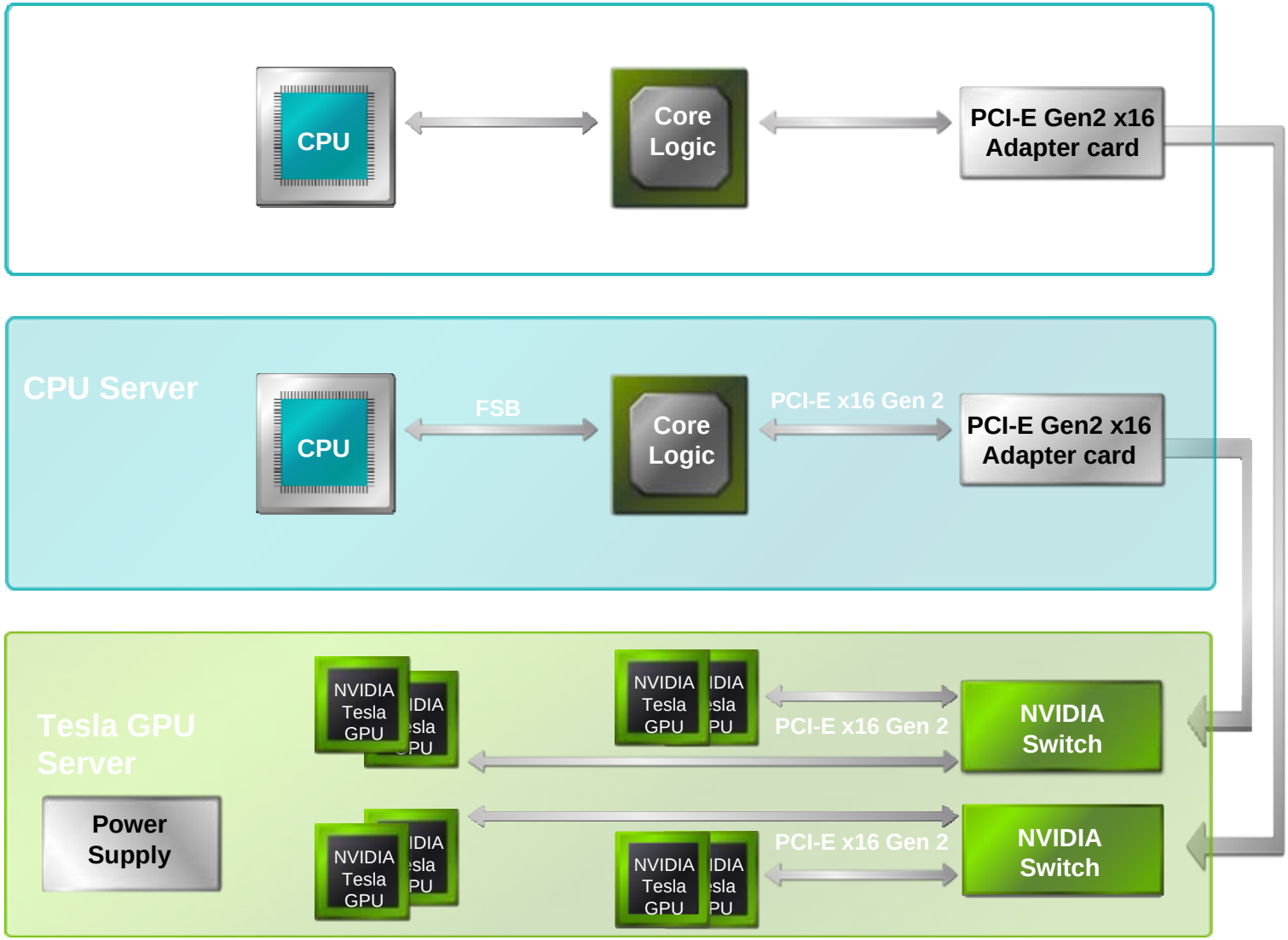


Personal Supercomputer



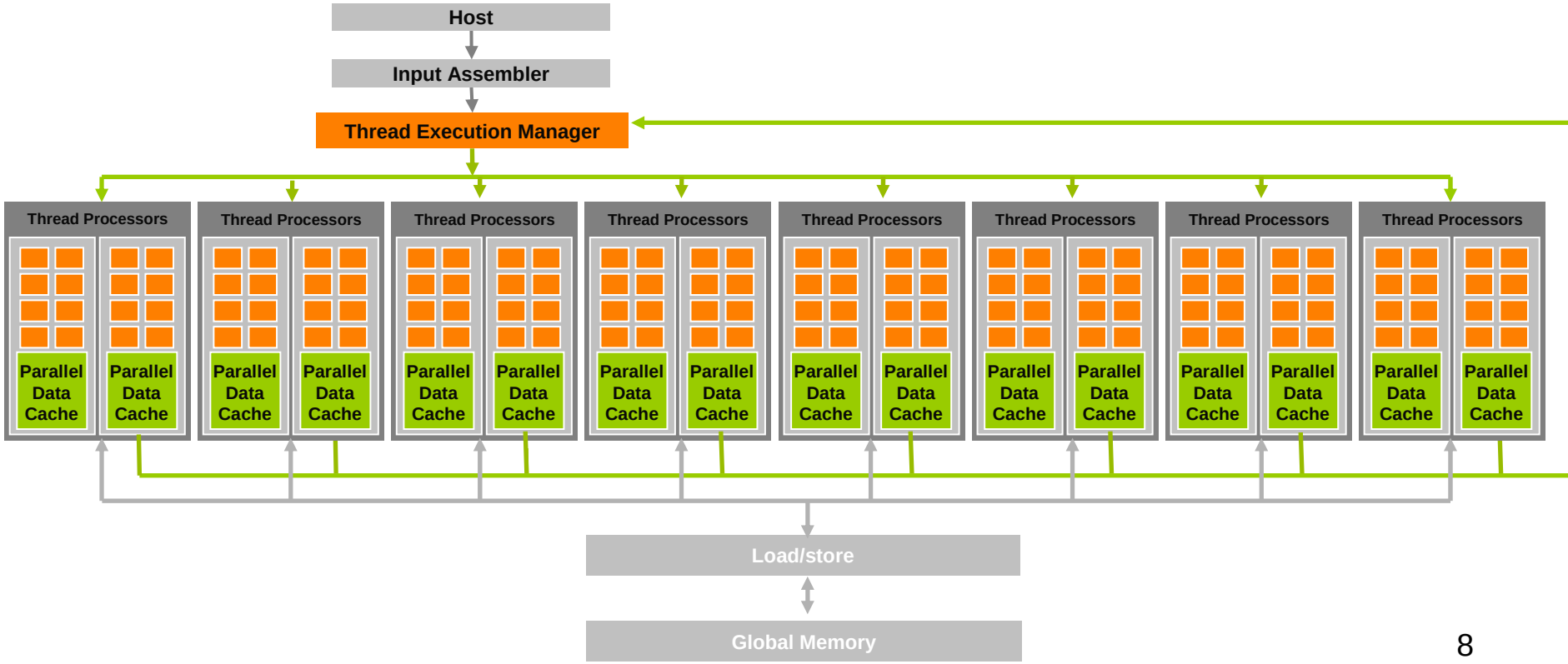
Tesla GPU Server

Tesla GPU Server



G80 Hardware

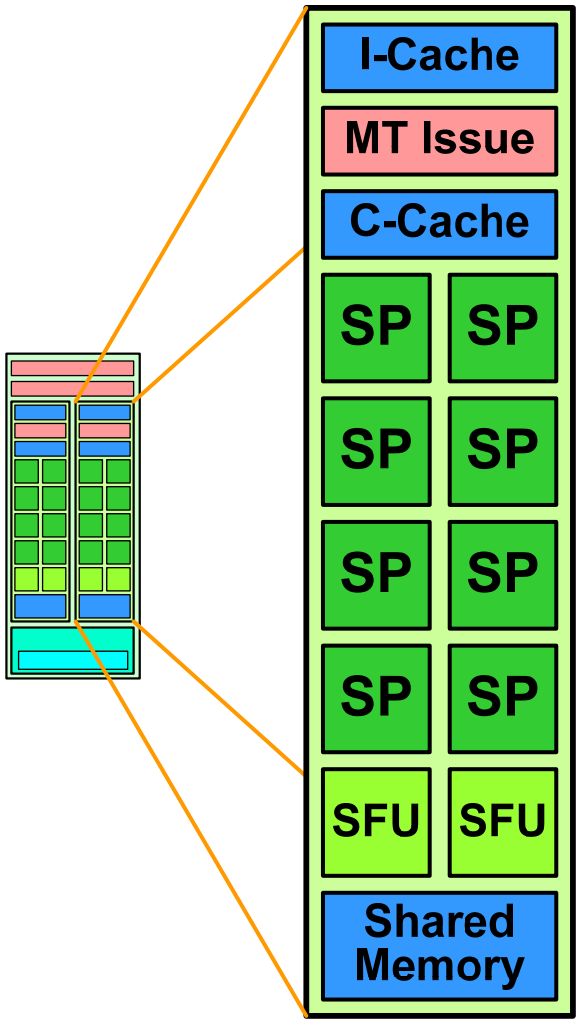
- 8 Texture Processor Clusters (TPC)
  - 1 TPC = 2 Streaming Multiprocessors (SMs) + texture
  - 1 SM = 8 streaming processors (SPs)
- 128 Thread Processors total



G80 Hardware Specs

- 128 thread processors
  - Up to 1.5GHz shader clock (8800 Ultra)
  - 384 GFLOPS peak (128 * 2 * 1.5)
- 384-bit memory interface
  - Up to 1080 Mhz memory clock = 103.7 GB / sec
  - Up to 1.5Gb on board memory
  - Up to 4GB/sec host to GPU memory transfers over PCI-E with NVIDIA motherboards

SM Multithreaded Multiprocessor



- SM has 8 SP Thread Processors
  - IEEE 754 32-bit floating point
  - 32-bit and 64-bit integer
  - 8K 32-bit registers
- SM has 2 SFU Special Function Units
- Scalar ISA
  - Memory load/store, texture fetch
  - Branch, call, return
  - Barrier synchronization instruction
- Multithreaded Instruction Unit
  - 768 Threads, hardware multithreaded
  - 24 SIMT warps of 32 threads
  - Independent thread execution
  - Hardware thread scheduling
- 16KB Shared Memory
  - Concurrent threads share data
  - Low latency load/store

SM SIMT Multithreaded Execution

Single-Instruction Multi-Thread instruction scheduler

warp 8 instruction 11

warp 1 instruction 42

warp 3 instruction 95

warp 8 instruction 12

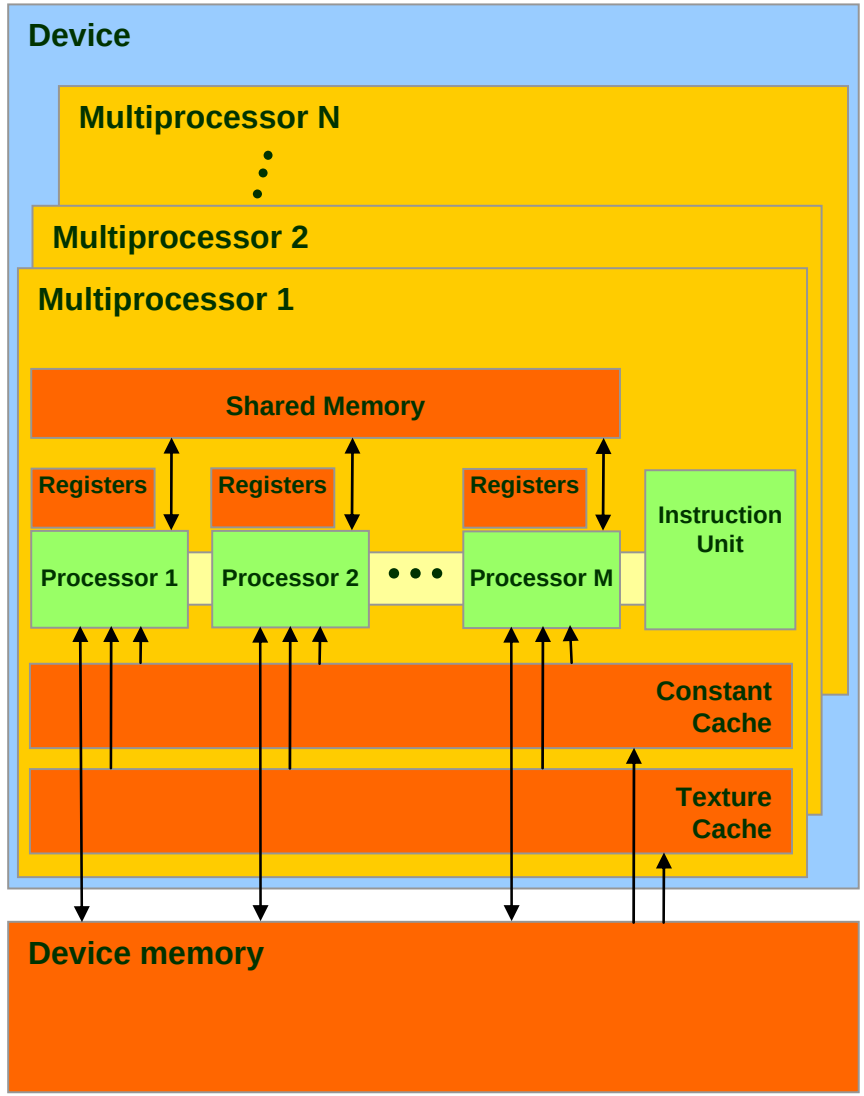
warp 3 instruction 96

- Weaving: first parallel thread technology
- Warp: the set of 32 parallel threads that execute a SIMT instruction

- SIMT: Single-Instruction Multi-Thread
- SM hardware implements zero-overhead warp and thread scheduling
- Each SM executes up to 768 concurrent threads, as 24 SIMT warps of 32 threads
- Threads can execute independently
- SIMT warp diverges and converges when threads branch independently
- Best efficiency and performance when threads of a warp execute together
- SIMT across threads (not just SIMD data) provides easy single-thread scalar programming with SIMD efficiency

Memory Architecture

- The local, global, constant, and texture spaces are regions of device memory
- Each multiprocessor has:
  - A set of 32-bit registers per processor
  - On-chip shared memory
    - Where the shared memory space resides
  - A read-only constant cache
    - To speed up access to the constant memory space
  - A read-only texture cache
    - To speed up access to the texture memory space



G8x Deviations from IEEE-754

- Addition and Multiplication are IEEE compliant
  - Maximum 0.5 ulp error
- However, often combined into multiply-add (FMAD)
  - Intermediate result is truncated
- Division is non-compliant (2 ulp)
- Not all rounding modes are supported
- Denormalized numbers are not supported
- No mechanism to detect floating-point exceptions

Floating Point Characteristics

	G8x	SSE	IBM Altivec	Cell SPE
Format	IEEE 754	IEEE 754	IEEE 754	IEEE 754
Rounding modes for FADD and FMUL	Round to nearest and round to zero	All 4 IEEE, round to nearest, zero, inf, -inf	Round to nearest only	Round to zero/truncate only
Denormal handling	Flush to zero	Supported, 1000's of cycles	Supported, 1000's of cycles	Flush to zero
NaN support	Yes	Yes	Yes	No
Overflow and Infinity support	Yes, only clamps to max norm	Yes	Yes	No, infinity
Flags	No	Yes	Yes	Some
Square root	Software only	Hardware	Software only	Software only
Division	Software only	Hardware	Software only	Software only
Reciprocal estimate accuracy	24 bit	12 bit	12 bit	12 bit
Reciprocal sqrt estimate accuracy	23 bit	12 bit	12 bit	12 bit
log2(x) and 2^x estimates accuracy	23 bit	No	12 bit	No

Questions?