NVIDIA FERMI GRAPHICS PROCESSOR GTX-580.

fermi processor has 32768 registers. divided into lanes, each SIMD thread is limited to 64 registers. the SIMD thread has upto (64 vector registers of 32-bit 32 elements)

32 vector register of 32 64-bit elements.

fermi has 16 physical SIMD lanes, each containing 2048 registers

each SIMD lane has private section of off chip DRAM.

private memory
contains stack frame, spilling registers, and private variables.

each multithreaded SIMD processor has also local memory.

shared by SIMD lanes / threads within a block.
memory shared memory by SIMD processor in GPU memory
host can read and write GPU memory.

each SIMD processor has
- 2 SIMD thread schedulers, two instruction dispatch units.

16 SIMD lanes, (SIMD width = 32, chime = 2 cycles)
16 load-store units, 4 SFU units.

thus, 2 threads of SIMD instructions are scheduled every 2 clock cycles.

IEEE 754 fast double prec. fast 2-level caches, 64-bit addressing and unified address space.

error correcting codes
faster context switching
faster atomic instructions.
16 SMs 32 cores/SM
512 cores.
3 billion transistors.
768 KB shared L2 cache.
6 DRAM channels.
host interfaces.
Grega thread scheduler.
2 ~~large~~ warp schedulers.
16 KB register files.
shared instruction cache.

64 KB shared L1 data cache. and shared (local) memory.
each SM processor has got.
LD/ST 16 units each. (2 cycle latency)
4 SFUs (8 cycle latency)
Separate INT, FP units in each core.

kepler gk110 7.1 billion transistors
1 TFlop double precision throughput.

power efficiency, upto 3x performance per watt of fermi.

DYNAMIC PARALLELISM

GPU to generate new work for itself.; synchronize on results, and control the scheduling of the work via dedicated accelerated hardware paths, all without involving the CPU.

HYPER-Q

glitch: enables multiple CPU cores to launch work on a single GPU simultaneously thereby dramatically increasing GPU utilization and significantly reducing CPU idle time.

GK 110 gpu by allowing 32 simultaneous hardware-managed connections.

grid-management unit, manages and prioritizes grids to be executed on the gpu.

printout.

NVIDIA GPU direct.

this is a capability that enables GPUs within a single computer, or GPUs in different servers located across the network to directly exchange the data without needing to go to CPU system memory

it will consume significantly less power and generate much less heat output.

15 SM units, each 192 CUDA cores.
and 6 64bit memory controllers. TSMC. ~ 28nm

each of the Kepler GK-110 SMX units feature 192 single precision CUDA cores, and each core has fully pipelined floating point and integer arithmetic logic units, 64 DP units, 32 SFU.

full IEEE 754 2008 compliant SP and DP compliant arithmetic - fermi.

each SMX features 4 warp schedulers and 8 instruction dispatch units. allowing 4 warps to be issued and executed concurrently.

shuffle instruction: allows threads within a warp to share data, there are also atomic ops.

NVIDIA PASCAL GP100 GPU

extreme performance, powering HPC., deep learning. and many more GPU computing areas.

NV link, NVIDIAs new high speed, high bandwidth inter-connected for max. app. stability.

it is connected to HBM2 (high bw mem)
fastest, high capacity, extremely stacked GPU memory architecture,

UNIFIED MEMORY AND COMPUTE PREEMPTION

significantly improved programming model.
16 nm FINFET, enables more features, high perf. and improved power efficiency.

extreme performance for HPC and deep learning.

5.3 TFlops of DP FP performance (FP64).
10.6 TFlops of SP FP performance (FP32)
21.2 TFlops of HP FP (FP16) performance.

NV-link extraordinary bandwidth for multi-GPU-to-CPU connectivity. GPU-to-GPU data transfers at upto 160 GB/s of bidirectional bandwidth.

classmate
Date 14.3.2020
Page 1

	GPU GK110 (Kepler)	Tesla P100 GP100 (Pascal)
SMs	15	56
TPCs	15	28
FP32 CUDA/SM	192	64
FP32 CUDA cores	2880	3584
FP64 CUDA/SM	64	32
FP64 CUDA cores/	960	1792
base clock	745 MHz	1328 MHz
boost clock	810/875 MHz	1480 MHz
peak FP32 GFLOPS	5040	10600
peak FP64 GFLOPS	1680	5300
texture units	240	224
mem. interface	384-bit GDDR5	4096-bit HBM2.
mem. size	upto 12GB	16 GB.
L2 cache size	1536 KB	4096 KB
TDP	235 W	300 W
transistors	7.1 billion	15.3 billion.
GPU die size	551 mm²	610 mm²
manufac process	28 nm	16 nm FinFET.
compute capab.	3.5 × kepler	6.0
Threads/warp	32	32
max warps/multiprocc	64	64
max threads/multiproc	2048	2048

~~Compute~~

max threadblocks.	16	32
max 32-bit regs./SM	65536	65536
max. regs. / block	65536	65536
max. regs. / thread	255	255
max. thread block size	1024	1024
Shared mem Size/SM	16K/32K/48K	64K.