GPU ARCHITECTURE

For sleep deprived college students…

And also dummies

Jeffrey Grange

A GP WHAT? WHAZZAT?

A GPU (Graphics Processing Unit) is used to accelerate graphics processes in a computer

The earliest predecessors appeared in the 1970s, they helped, but still relied on the CPU for calculations

In the 1990s, the first actual GPUs appeared, which had their own hardware for processing

HOW GPU BETTER?

GPUs achieve higher throughput by sacrificing control and single instruction latency

There's a lot of similar calculations in graphics, so you can group them together and perform one calculation on the whole group

This is called SIMD (Single Instruction Multiple Data)

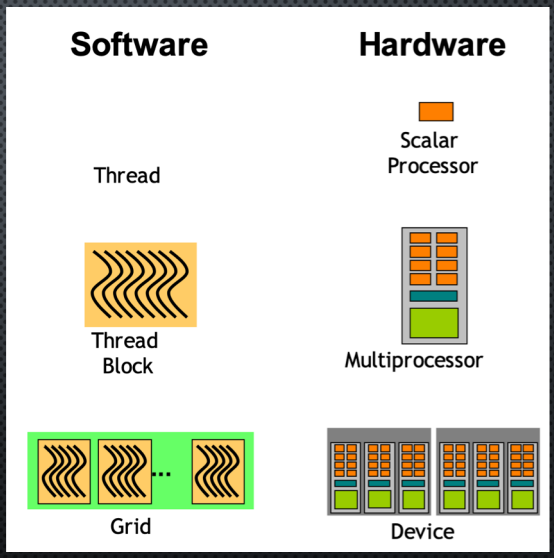
DATA (CUDA 2.0)

Threads: Data and a series of operations done on it

Warps: Sets of 32 similar threads run on a streaming multiprocessor

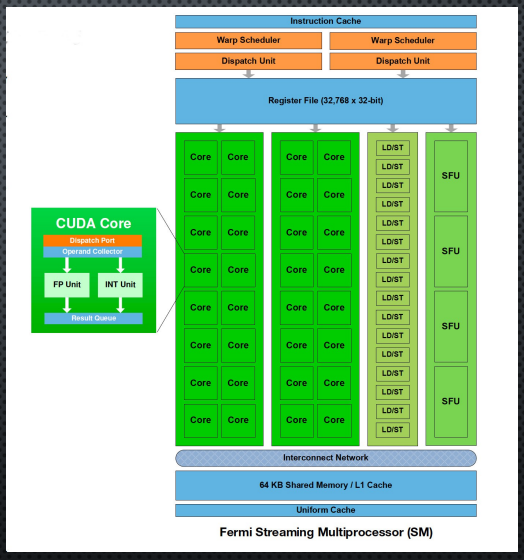
Blocks: A whole bunch of threads, multiple of 32

Grid: Whole bunch of blocks



STREAMING MULTIPROCESSOR

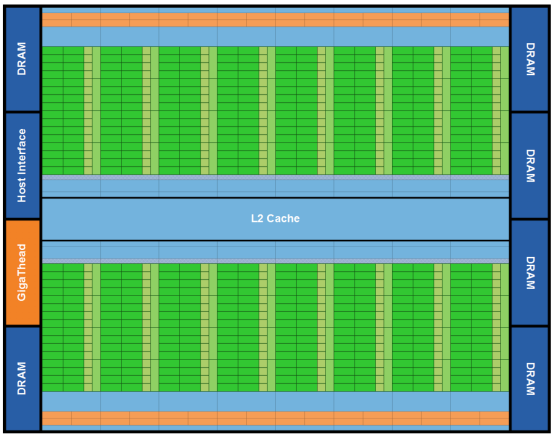
- Has 32 CUDA cores, which can do floating point or integer calculations
- A whole bunch of registers to hold Blocks and warps
- 2 Warp Schedulers, each for half the cores
- 4 Special Function Units, do sine, cosine…etc
- L1 Cache
- Uniform Cache (holds duplicate of required ROM contents)



Fermi Streaming Multiprocessor (SM)

GPU - ALMOST

- Bunch of streaming multiprocessors
- L2 Cache
- Access to VRAM
- Access to other stuff



Fermi's 16 SM are positioned around a common L2 cache. Each SM is a vertical rectangular strip that contain an orange portion (scheduler and dispatch), a green portion (execution units), and light blue portions (register file and L1 cache).

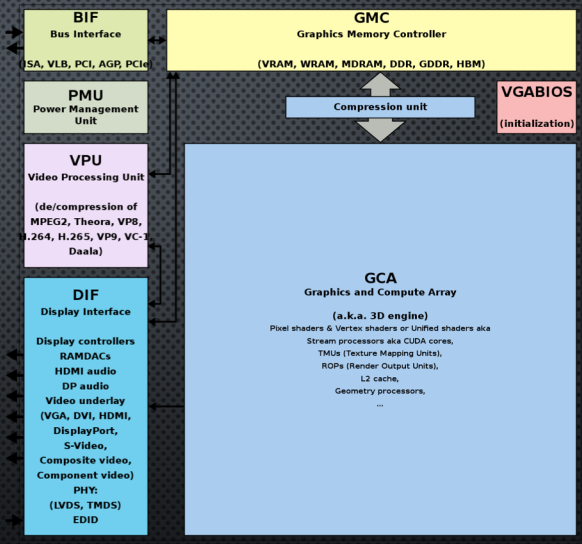
THE OTHER STUFF

GCA: Everything we just covered

VPU: Used for compressing and decompressing different video storage formats

Display Controller: Converts VRAM data to something interpretable by screen and outputs

Memory Controller: Controls memory…



REFERENCES:

- http://www.icl.utk.edu/~luszczek/teaching/courses/fall2016/cosc462/pdf/GPU_Fundamentals.pdf
- http://developer.download.nvidia.com/compute/cuda/3_1/toolkit/docs/NVIDIA_CUDA_C_ProgrammingGuide_3.1.pdf
- http://meseec.ce.rit.edu/551-projects/spring2015/3-2.pdf
- https://en.wikipedia.org/wiki/Graphics_processing_unit
- https://en.wikipedia.org/wiki/Video_display_controller
- https://medium.com/@smallfishbigsea/basic-concepts-in-gpu-computing-3388710e9239
- https://en.wikipedia.org/wiki/CUDA

MORE REFERENCES:

- https://www.quora.com/What-is-a-warp-and-how-is-it-different-from-a-thread-block-or-wave-in-CUDA
- https://electronics.stackexchange.com/questions/102695/how-does-a-gpu-cpu-communicate-with-a-standard-display-output-hdmi-dvi-etc
- https://en.wikipedia.org/wiki/Video_codec
- https://itstillworks.com/video-controller-1731.html
- https://www.nvidia.com/en-us/geforce/20-series/