CSEN 1013 Seminar Multi-Core & High Performance Computing

Nvidia Fermi

Ahmed Labib

February 28, 2010

1. Introduction
2. The Fermi Architecture
   - The Stream Multiprocessor
3. Fermi architecture again
4. Software support

Competition & Long Term Strategy

- Competition with Intel and AMD.
- The competitors markets.
- Trying to enter the chipset market in C2D & Atom.
- The Hybrid SLI
- Locking out & legal issues.
- GPGPU (SQL, MRI, stock options)
- G80 to GT200
- Problems with G80 / GT200's GPGPU approach
- CUDA C
- From GPGPU to GPU Computing.

Areas where changes are needed

- Double Percision Performance
- ECC Support
- Cache (from prev. shared memory)
- Shared Memory (increase its size)
- Faster Context Switching
- Faster Atomic Operations (Read - Modify - Write)

General overview of the Fermi Architecture

- 3 Billion Transistors
- 40nm TSMC
- 384 bit memory interface
- 512 Shader Cores (CUDA Cores)
- 32 CUDA cores per shader cluster
- 16 Shader clusters
- 1MB L1 Cache (64KB per shader cluster)

General overview of the Fermi Architecture contd

- 768KB Unified L2 Cache
- Upto 6GB GDDR5 Memory
- Six 64 bit Memory Controllers
- IEEE 754 - 2008 Double Percision Standard
- ECC Support
- 512 FMA in SP Mode
- 256 FMA in DP Mode

Transistor Count

- 3 Billion Transistors
- Huge die & the need for the 40nm
- Fabrication Processes
- Costs & delay

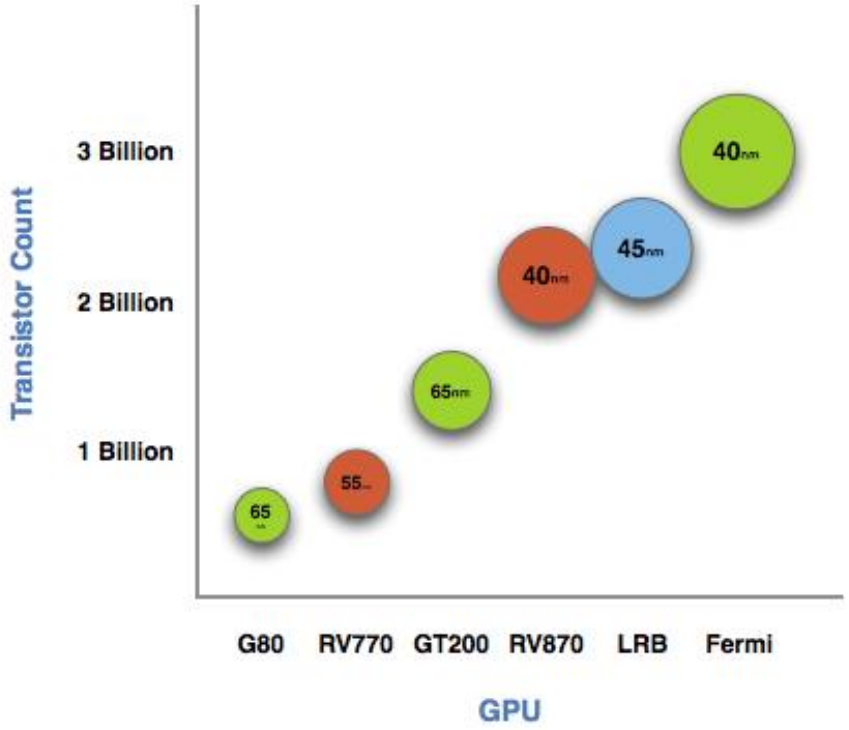

Figure: Transistor Count

Graphics Processing Cluster

- Different scalability options along GPC & SM
- 4 SMs / GPC
- 1 Raster Engine / GPC

Figure: GPC

The Stream Multiprocessor

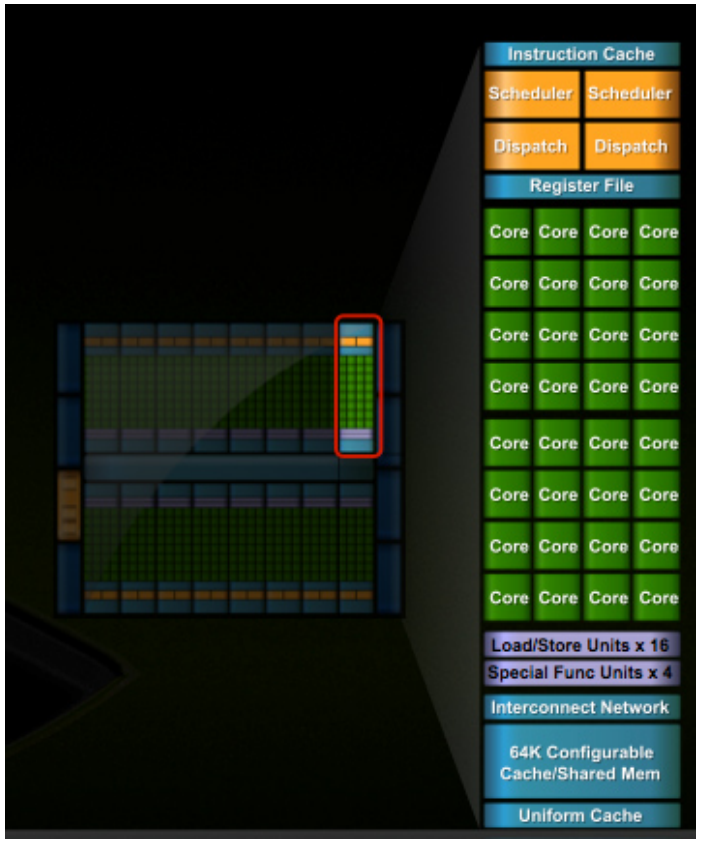

Figure: SM

The Stream Multiprocessor contd

- 32 CUDA Cores (4x The previous amount)
- 4 SFU (Special Function Units)
- 32K FP32 Registers (2x The previous amount)
- 4 Texture Units
- A PolyMorph Engine
- 64K L1 Shared Memory / L1 Cache
- 2 Warp Schedulers
- 2 Dispatch Units
- 16 load / store units

The CUDA Core

- 1 Integer ALU
- 1 FPU
- Fully pipelined ALU & FLU
- 1 Integer / FP Opr. per clock per thread in SP
- 0.5 in DP mode
- Improved compared to 1/8 in previous architectures
- Inst. can be mixed (FP + Int, FP + FP, SFU + FP)

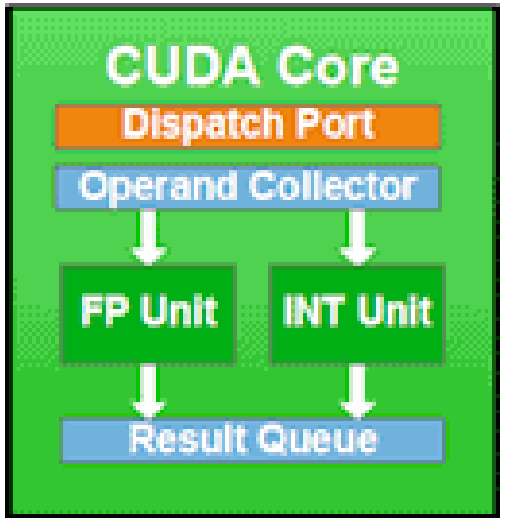

Figure: CUDA Core

The FMA and the IEEE 754 - 2008

- IEEE 754 - 1984
- MAD (truncation, rounding to nearst even)
- Inaccurate yet fast (1 clock cycle)
- IEEE 754 - 2008 (subnormal numbers, nearest, zero, +/- infinity)
- FMA (Fused Multiply Add)
- Advantages to HPC, MRI & other GPU computing apps

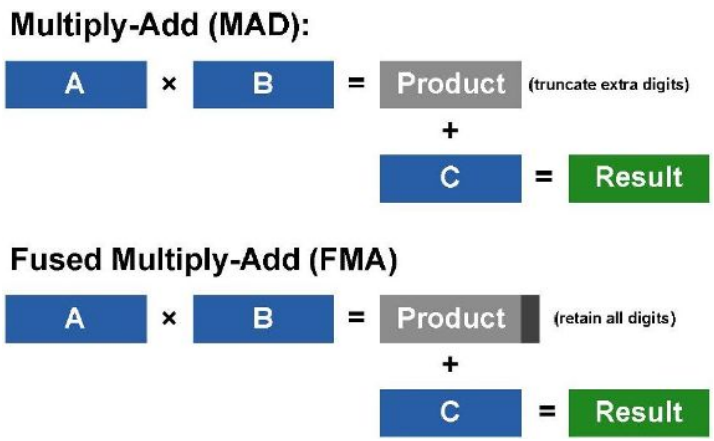

Figure: FMA vs MAD

The Fermi thread hierarchy

- Threads
- Warps
- Grid
- GPU kernel grids
- SM thread blocks
- CUDA cores threads

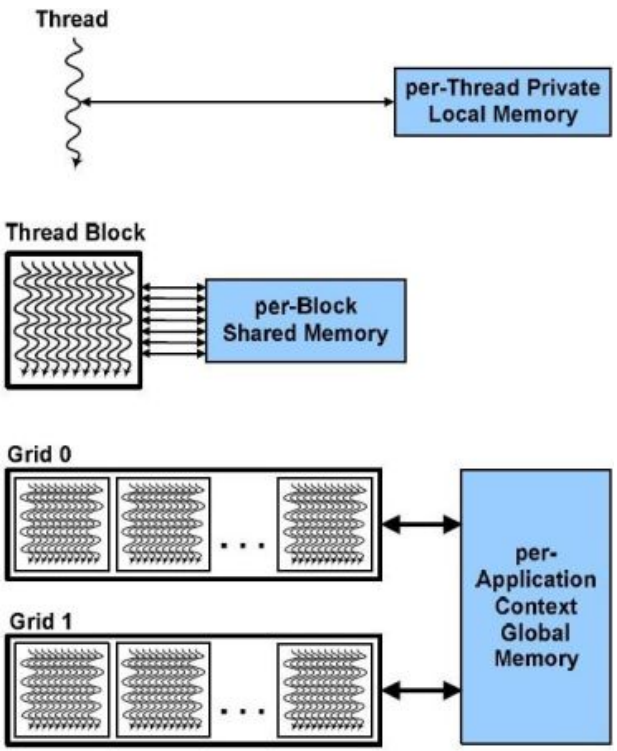

Figure: Thread Hierarchy

The Warp Scheduler

- 2 warp scheduler
- 2 warps executed at the same time on each SM
- Decoupled SFU

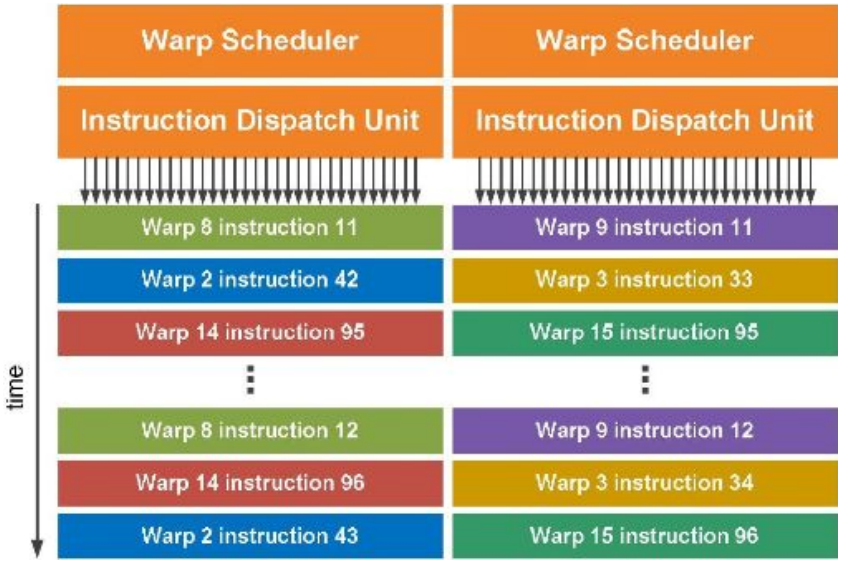

Figure: The Warp Scheduler

The 64KB Shared Memory / L1 Cache

Figure: Shared Memory / L1 Cache

- Older configuration and its limitations with the new strategy
- Nvidia Fermi's solution

The PolyMorph Engine

- Performance gap
- Reason for this gap
- Nvidia's Solution
- PolyMorph Engine advantages

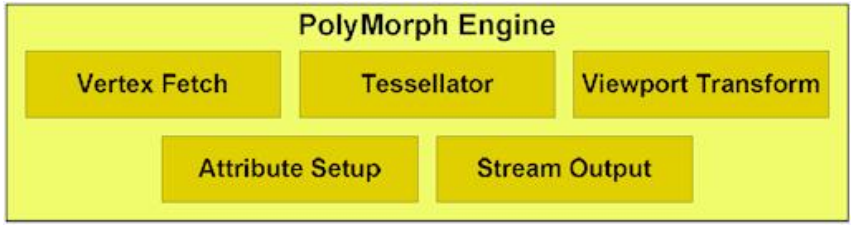

Figure: PolyMorph Engine

Texture Units

- 4 Texturing Units per SM
- Uses of the Texturing units

Figure: Texture Units

Memory Hierarchy

- Shared Memory / L1 Cache
- L2 Cache
- Memory Controllers & DRAM
- ECC Protection

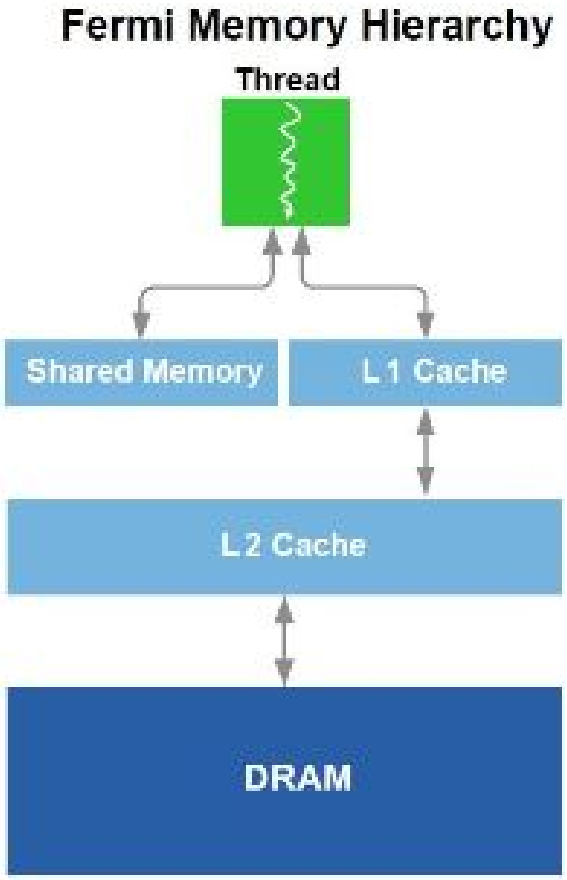

Figure: Memory Hierarchy

The unified address space

- Old configuration
- Unification of thread private, block shared and global
- Advantages
- 40bit addressing
- Supports 64bit addressing for future growth

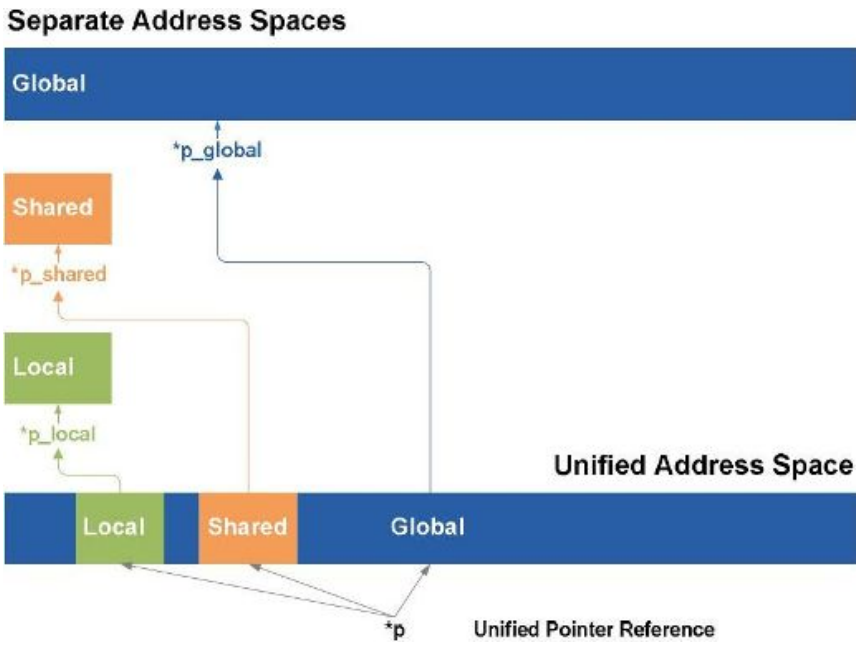

Figure: Unified address space

The GigaThread scheduler

- Two thread schedulers
- Scope of each thread scheduler
- Advantages of the GigaThread Scheduler

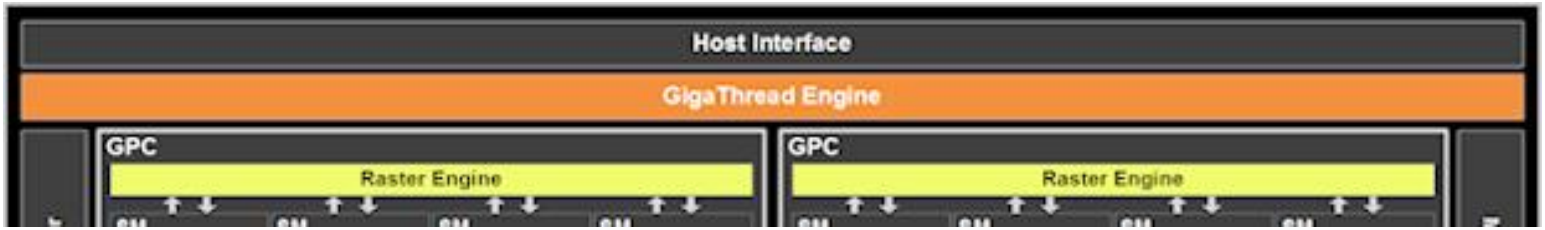

Figure: The GigaThread Scheduler

ROPs - Raster Operator

- 48 ROPs
- ROP function inside the GPU

Figure: ROPs

Nvidia Nexus

- Purpose
- Microsoft Visual Studio
- Code and Debug
- Co-processing applications between CPU and GPU

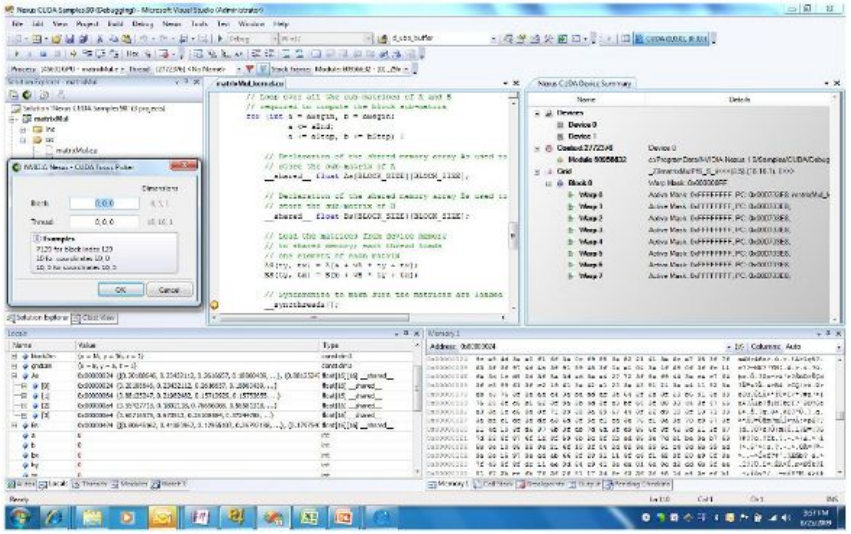

Figure: Nvidia Nexus

Blog Entry

- http://nvidiafermi.wordpress.com/

References

www.brightsideofnews.com

www.xbitlabs.com

beyond3d.com

www.techreport.com

www.semiaccurate.com

www.hardocp.com

www.gpureview.com

www.behardware.com

www.nvidia.com