Pascal GPU Architecture

A.Zamani

Supervised by: Dr. Motamedi

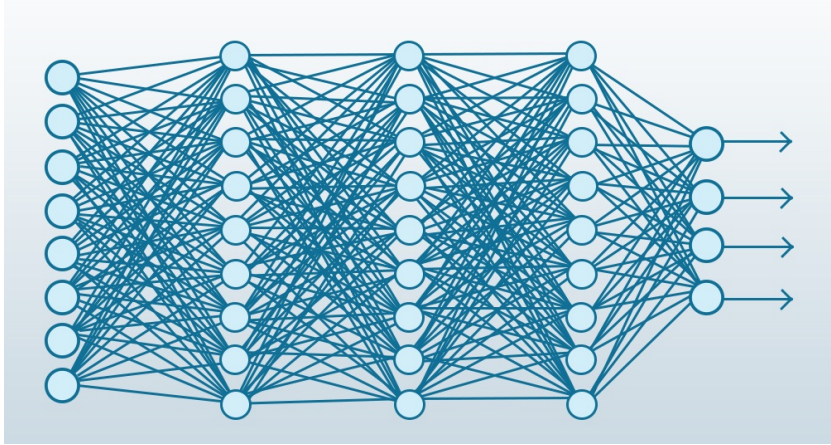
Amirkabir University of Technology

February 2018

Outline

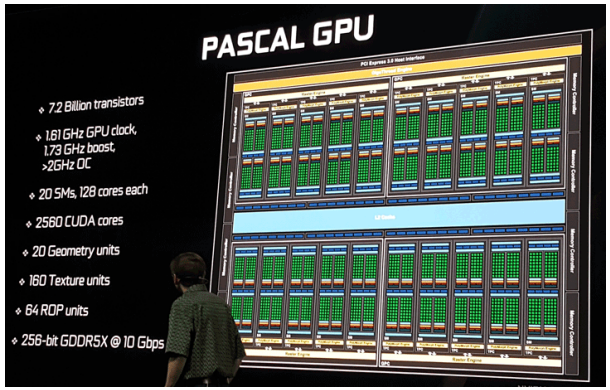
Introduction

- Graphic Processor Unit(GPU)
  - Games
  - Graphical softwares
    - Photoshop
    - corel
  - Deep learning and Artificial Intelligence



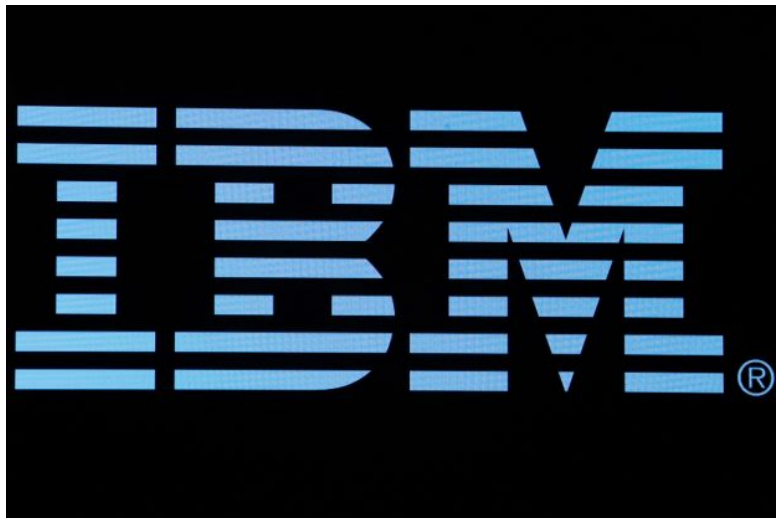
Introduction

- Nvidia
  - Pascal architecture
  - Facebook and Google
  - Audi and Benz - self drive



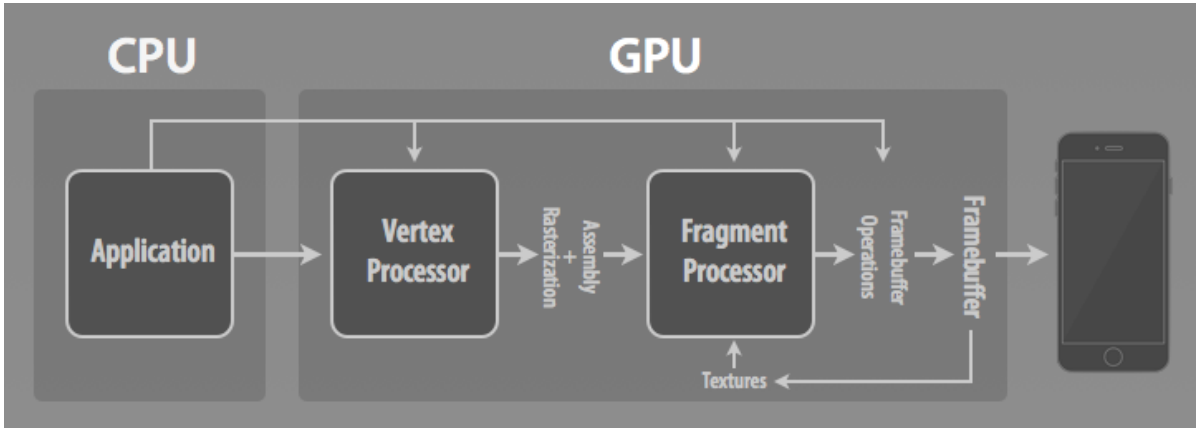
Graphics card

- First graphci card: IBM 1960 / 4 kb RAM / green
- Graphic Card Components
  - Graphic processor: Main componet
  - Memory
  - Peripherals



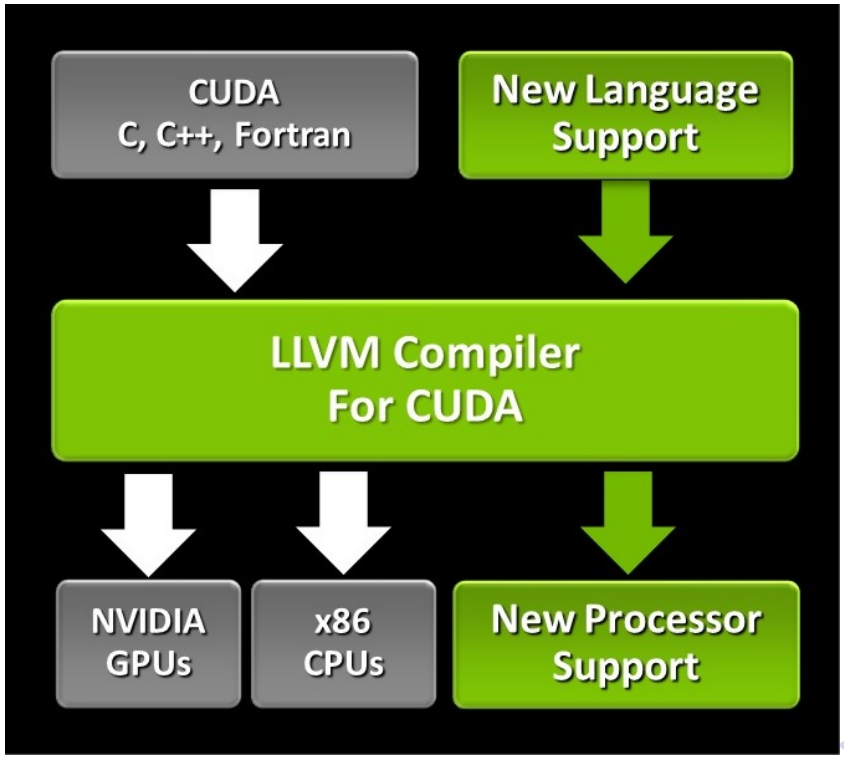
- Vertex processor
- Pixel Fragment Processor

- Programming language
- NVidia
- CUDA

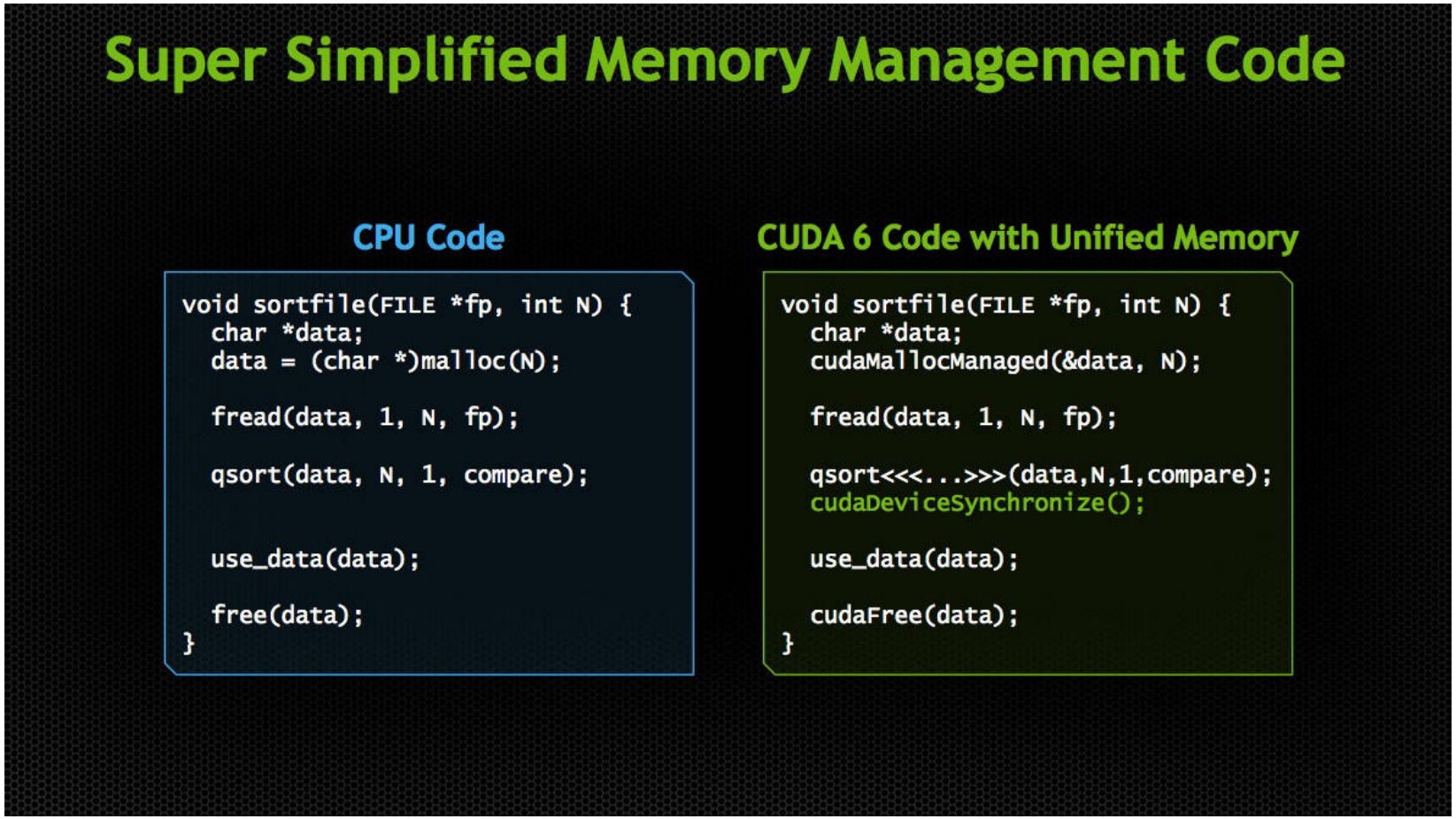


CUDA

- 2006 / GeForce 8800
- parallel programming in NVidia processors
- programming like CPU (GPGPU)
- Fortran / C++ / C
- OpenCL / MATLAB / LabVIEW



CUDA

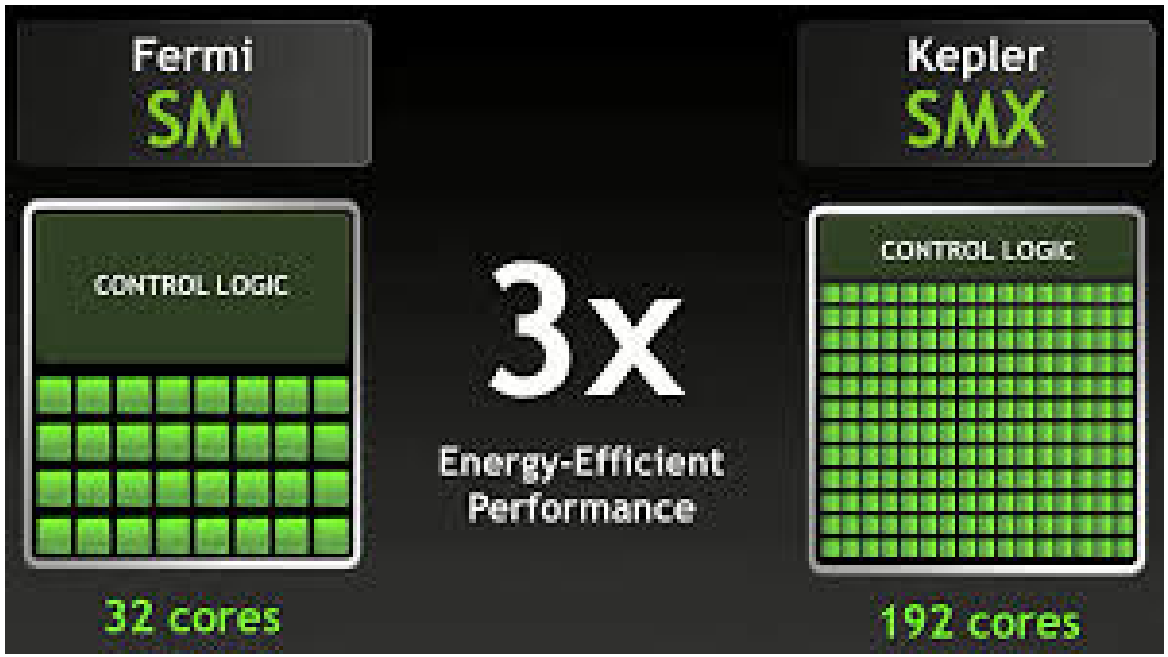

Super Simplified Memory Management Code
CPU Code
void sortfile(FILE *fp, int N) {
char *data;
data = (char *)malloc(N);
fread(data, 1, N, fp);
qsort(data, N, 1, compare);
use_data(data);
free(data);
}
CUDA 6 Code with Unified Memory
void sortfile(FILE *fp, int N) {
char *data;
cudaMallocManaged(&data, N);
fread(data, 1, N, fp);
qsort<<<...>>>(data,N,1,compare);
cudaDeviceSynchronize();
use_data(data);
cudaFree(data);
}

Architectures

- On Tesla 1 SM combines 8 single-precision (FP32) shader processors
- On Fermi 1 SM combines 32 single-precision (FP32) shader processors
- On Kepler 1 SM combines 192 single-precision (FP32) shader processors and also 64 double-precision units (at least the GK110 GPUs)
- On Maxwell 1 SM combines 128 single-precision (FP32) shader processors
- On Pascal it depends

Fermi architecture

- Release date: April 2010
- Transistors: 40 nm and 28 nm
- Predecessor: Tesla 2.0
- Successor: Kepler
- used in the GeForce 400 series and GeForce 500 series

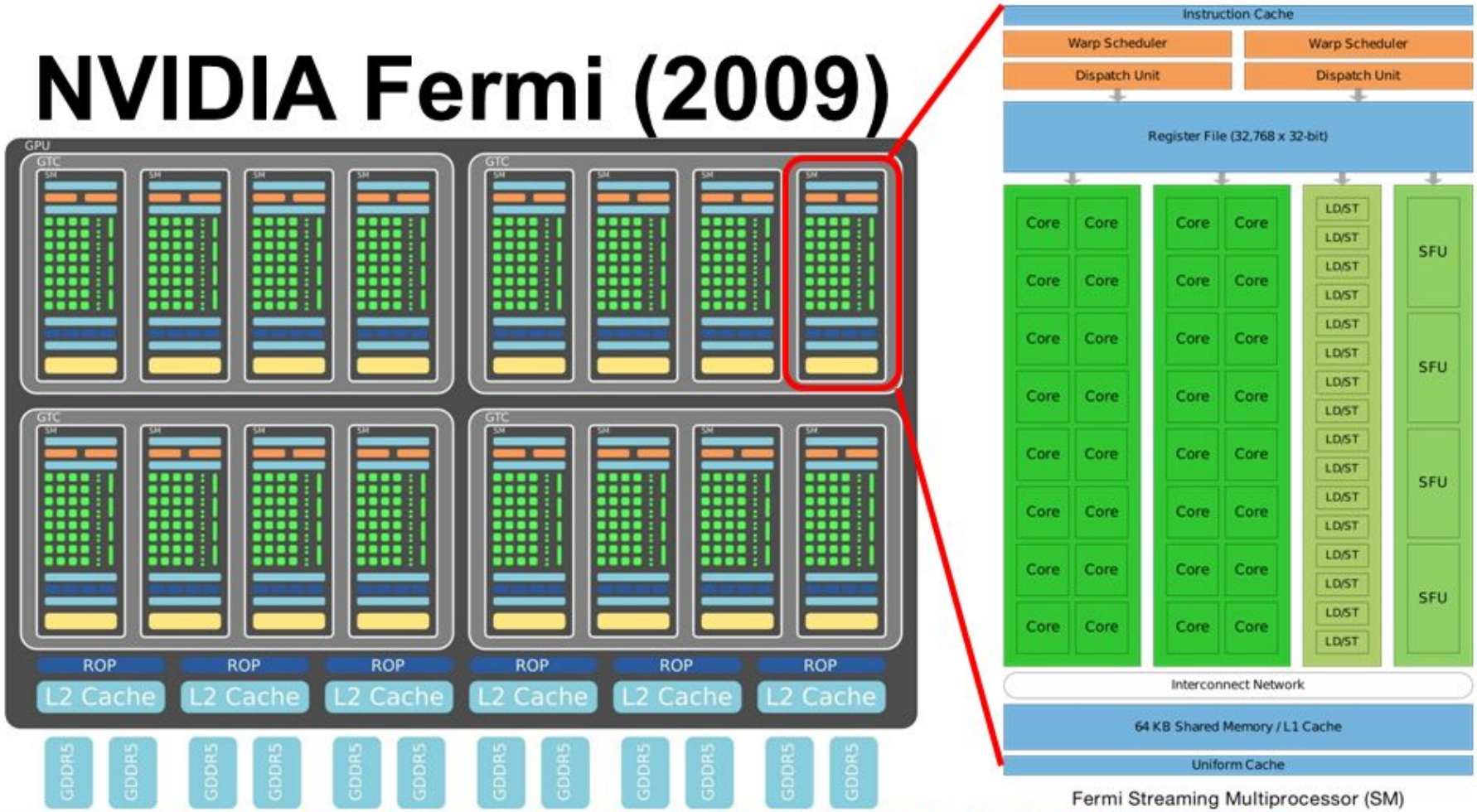


Fermi architecture

GF100 Block Diagram

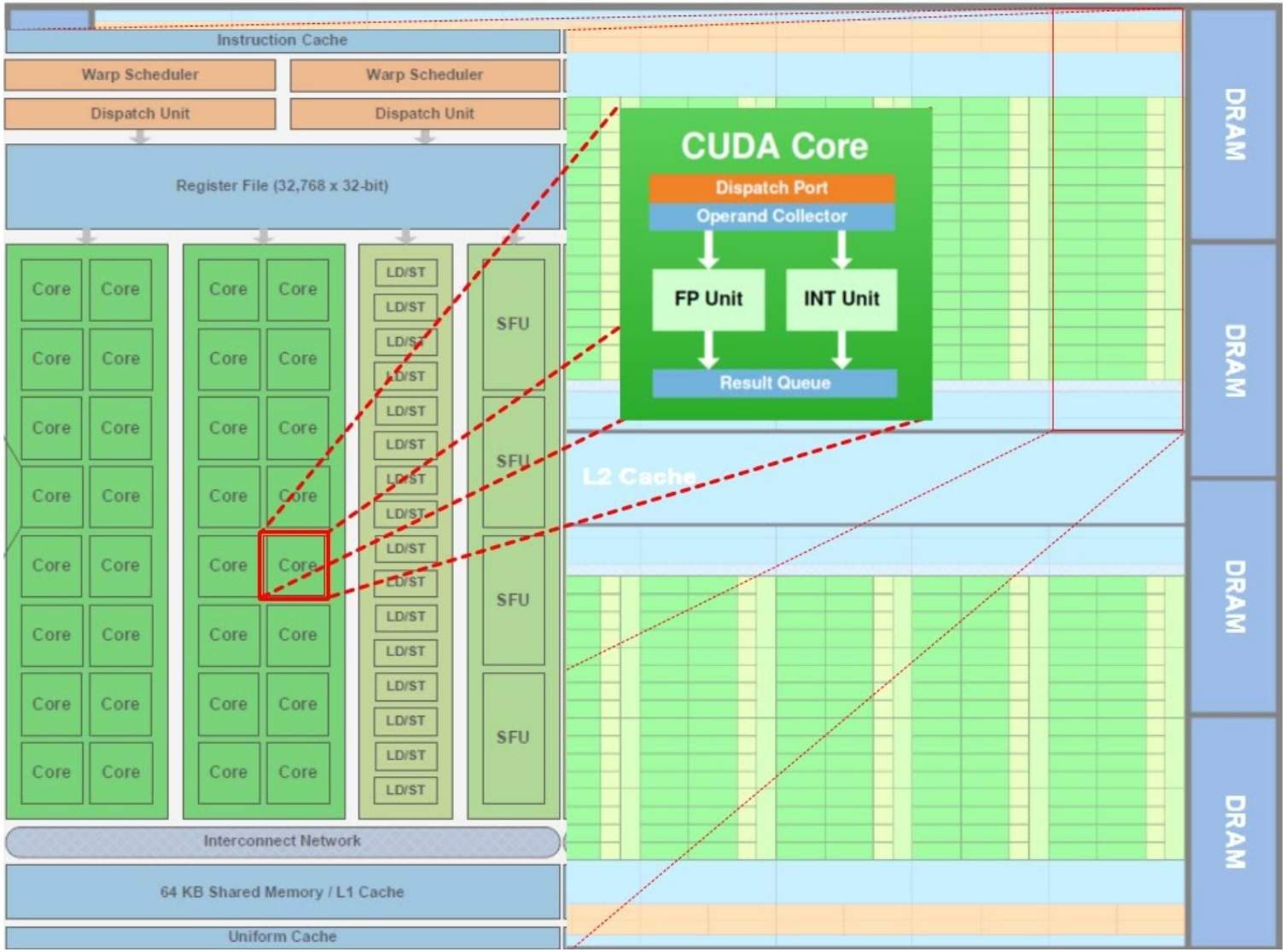
- 512 CUDA cores
- 16 geometry units
- 4 raster units
- 64 texture units
- 48 ROP units
- 384-bit GDDR5

Fermi architecture



ref: http://www.nvidia.com/content/PDF/fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf

Fermi architecture



Pascal architecture

- successor to the Maxwell architecture
- April 2016 with the release of the Tesla P100 (GP100) on April 5
- primarily used in the GeForce 10 series

References

[7] Drãxler, S., H. Karl, and Z.Á. Mann. Joint Optimization of Scaling and Placement of Virtual Network Services. in 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID). 2017.
[8] Huin, N., B. Jaumard, and F. Giroire, Optimal Network Service Chain Provisioning. IEEE/ACM Transactions on Networking, 2018. 26(3): p. 1320-1333.
[9] Masri, W., et al. Minimizing delay in IoT systems through collaborative fog-to-fog (F2F) communication. in 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN). 2017
[10] Fan, J., et al. Deadline-Aware Task Scheduling in a Tiered IoT.Infrastructure. in GLOBECOM 2017 - 2017 IEEE Global Communications Conference. 2017.
[11] Gupta, A., et al. Service Chain (SC) Mapping with Multiple SC Instances in a Wide Area Network. in GLOBECOM 2017 - 2017 IEEE Global Communications Conference. 2017.

Thanks for your attention.