Pascal GPU Architecture

A.Zamani

Supervised by: Dr. Motamedi

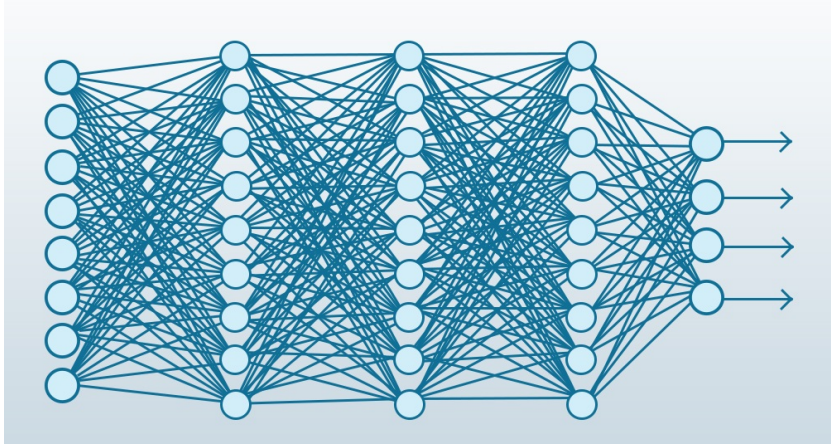
Amirkabir University of Technology

February 2018

Outline

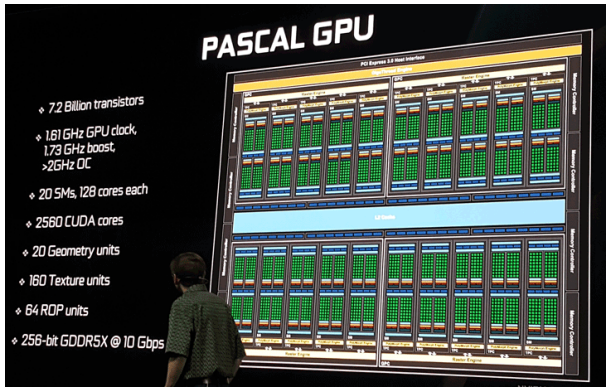
Introduction

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



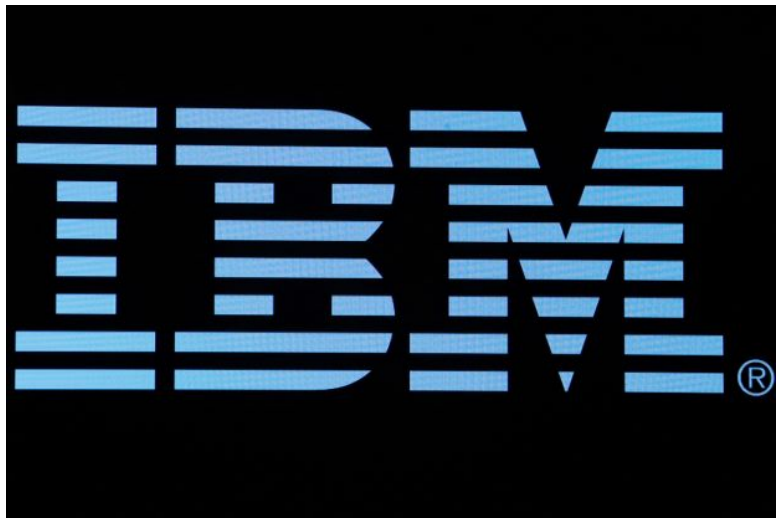
Introduction

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



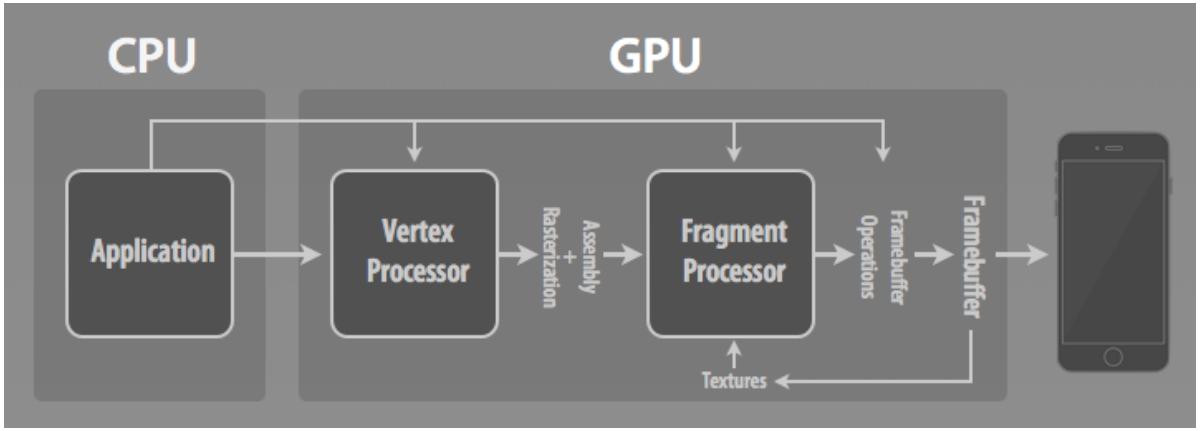
Graphics card

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



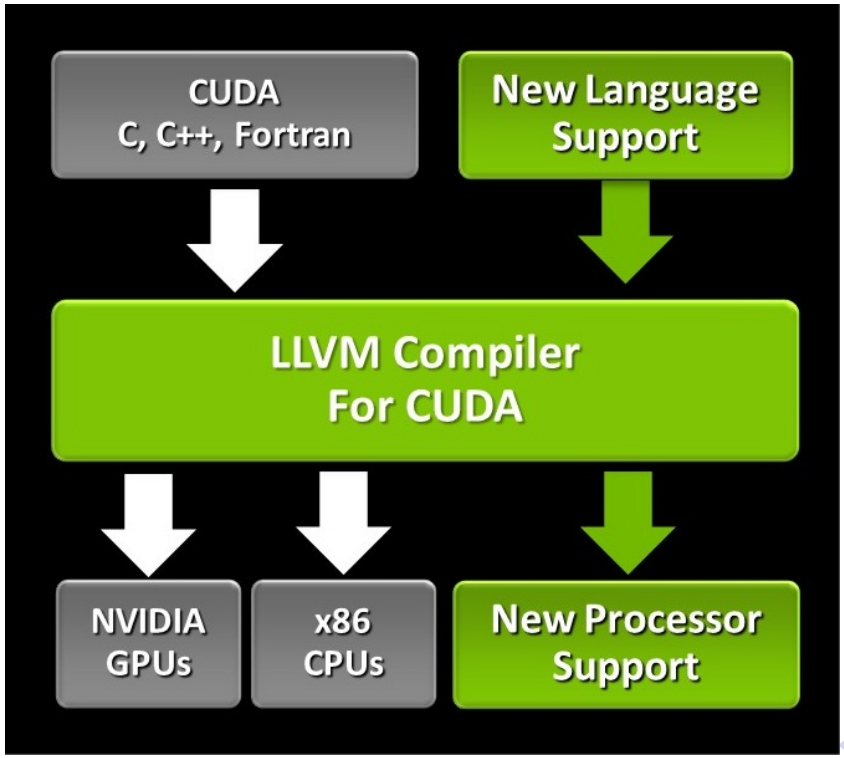
- Vertex processor
- Pixel Fragment Processor

- Programming language
- NVidia
- CUDA

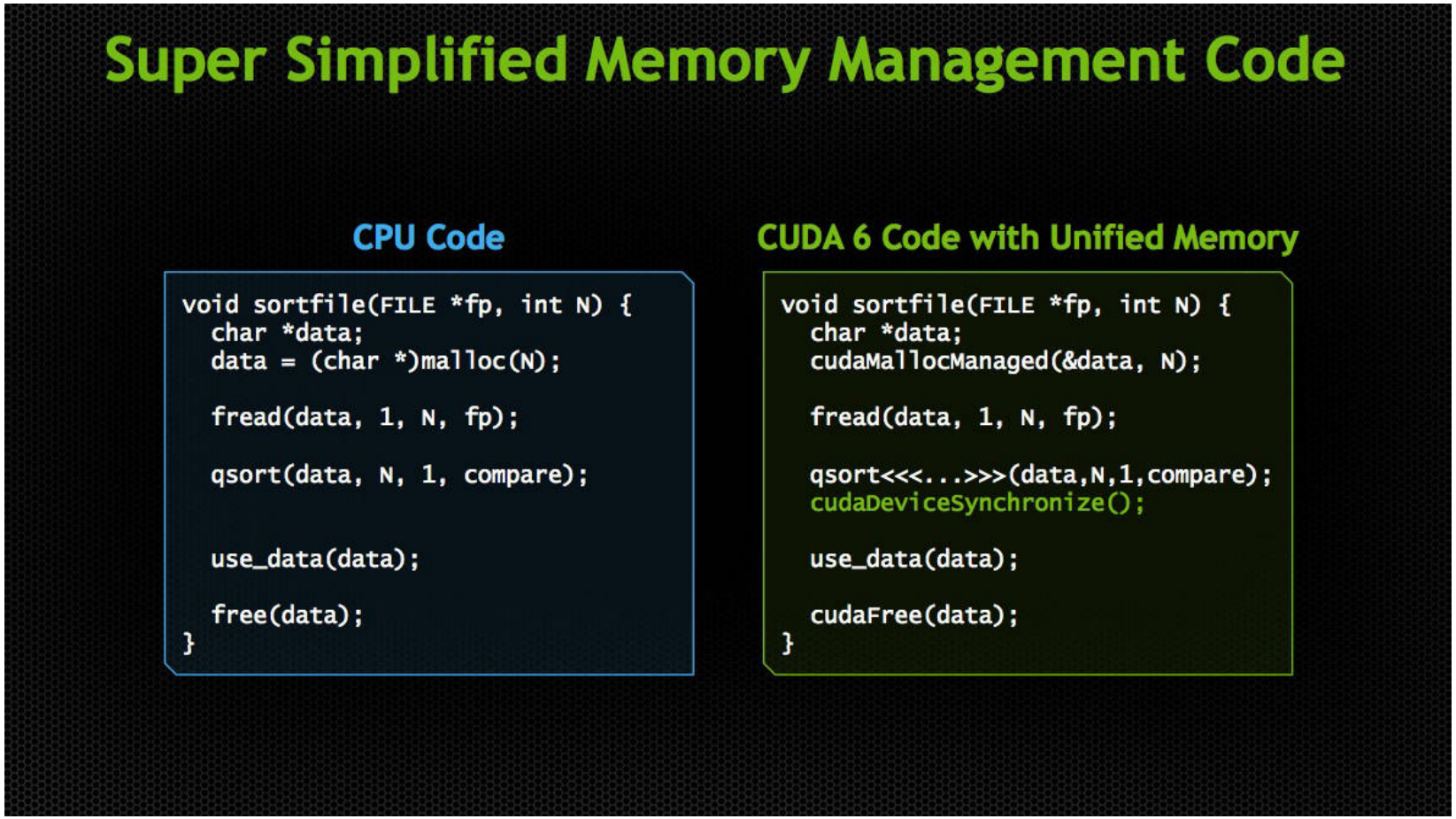


CUDA

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



CUDA

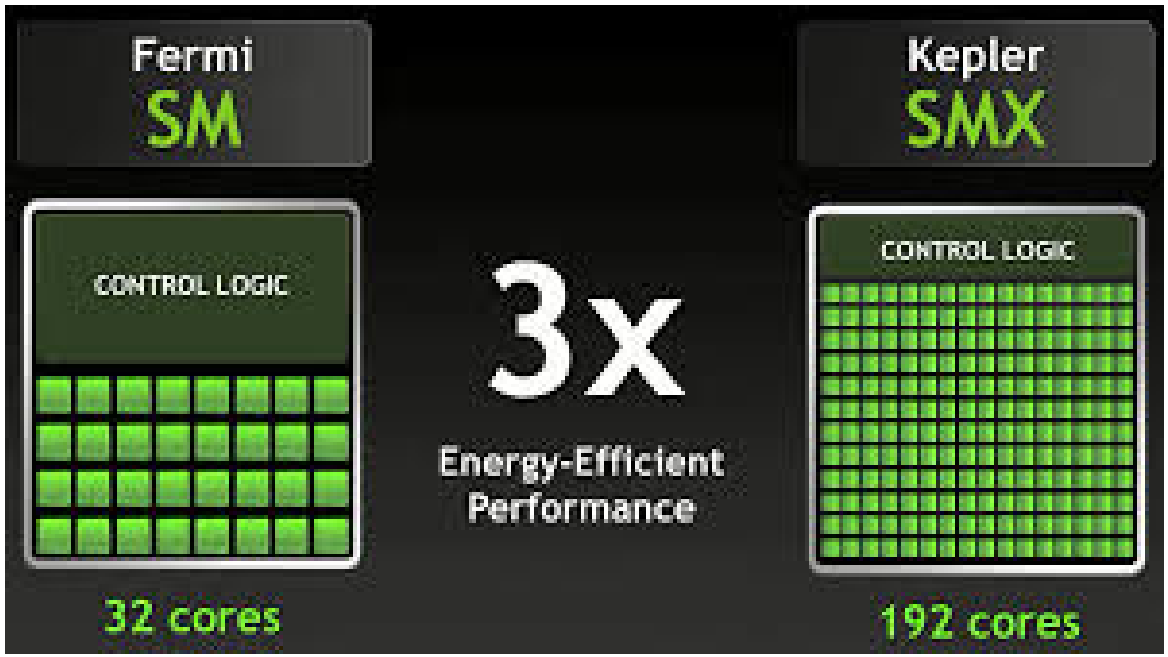

Super Simplified Memory Management Code
CPU Code
void sortfile(FILE *fp, int N) {
char *data;
data = (char *)malloc(N);
fread(data, 1, N, fp);
qsort(data, N, 1, compare);
use_data(data);
free(data);
}
CUDA 6 Code with Unified Memory
void sortfile(FILE *fp, int N) {
char *data;
cudaMallocManaged(&data,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

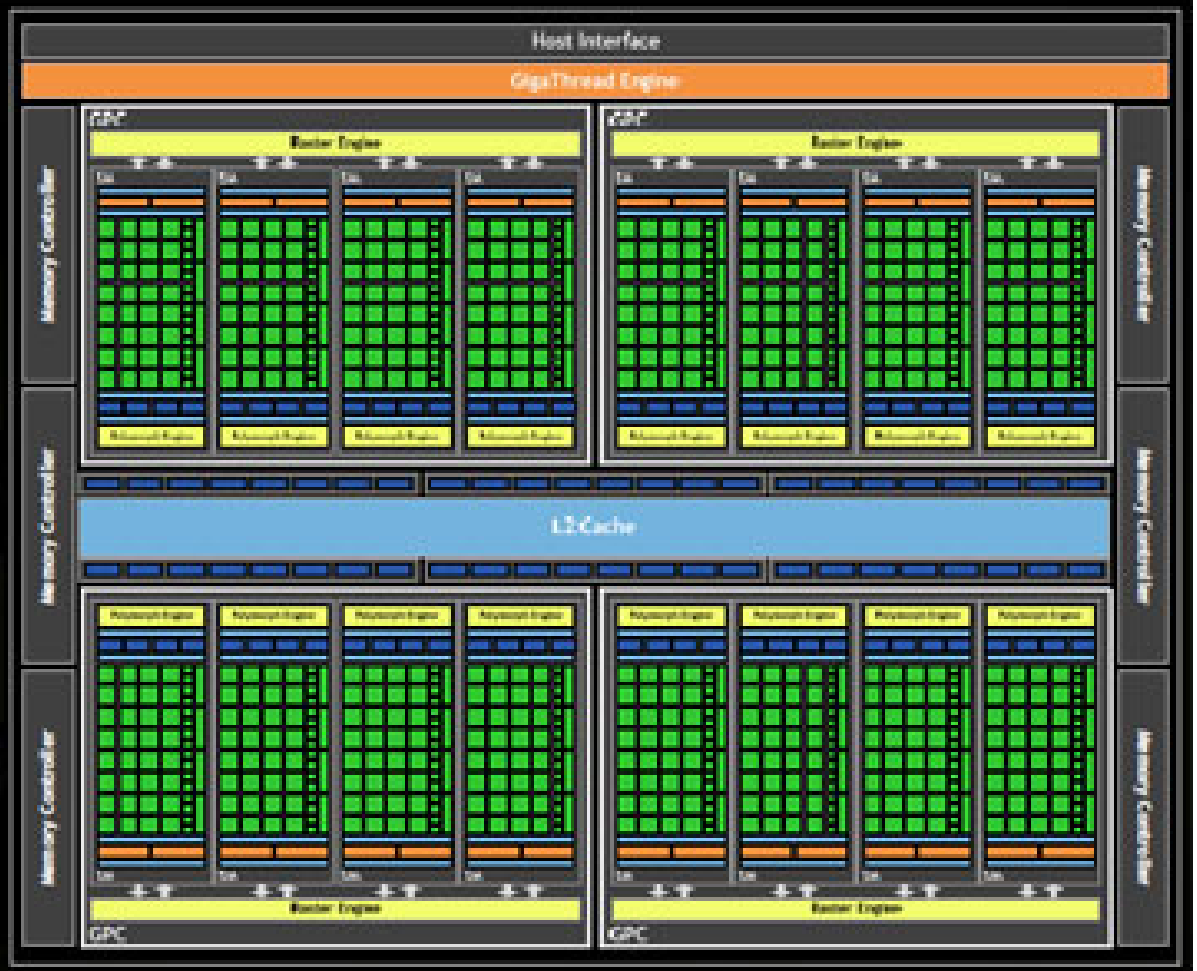
- Release date: April 2010
- Transistors: 40 nm and 28 nm
- Predecessor: Tesla 2.0
- Successor: Kepler
- used in the GeForce 400 series and GeForce 500 series



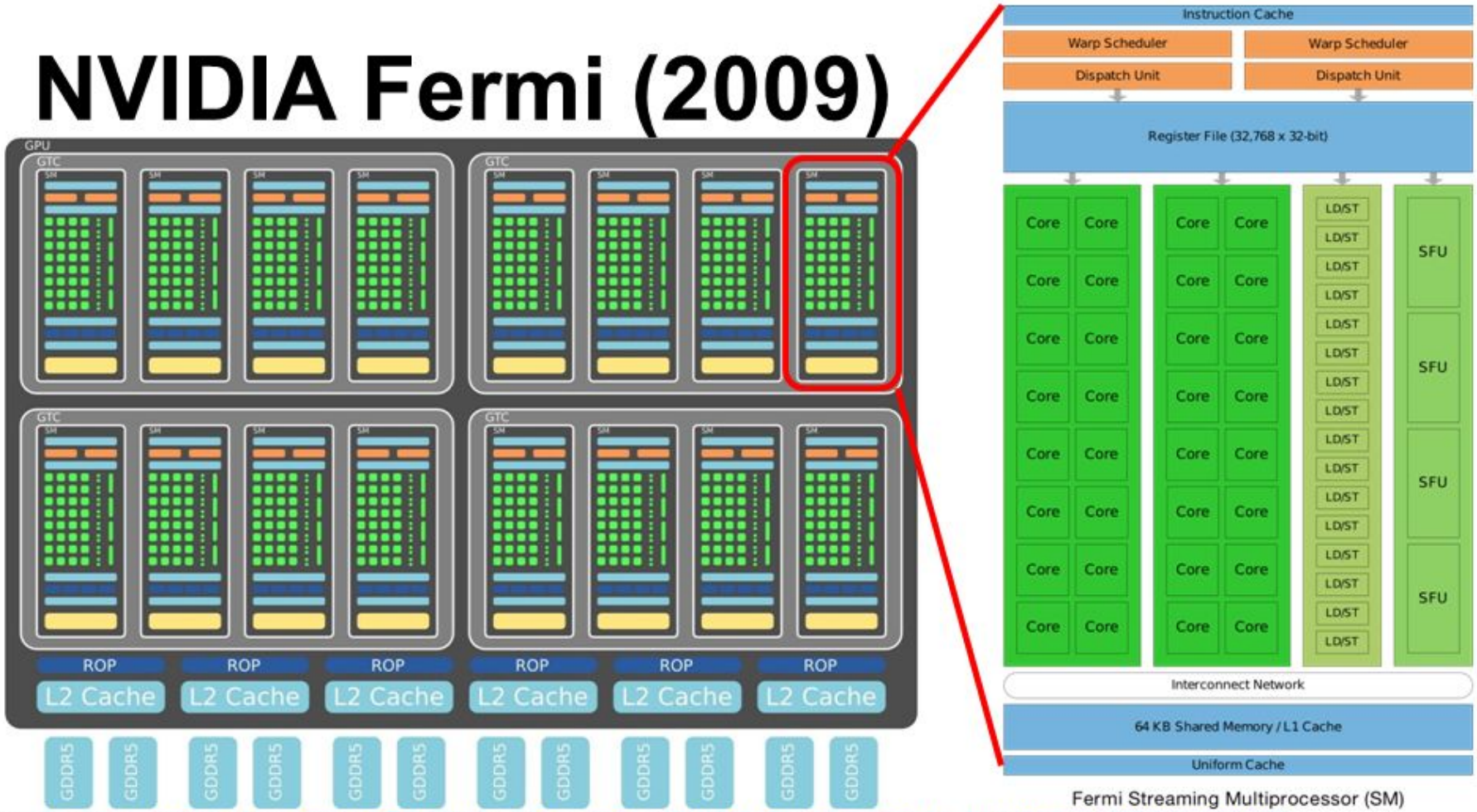
Fermi architecture

GF100 Block Diagram

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

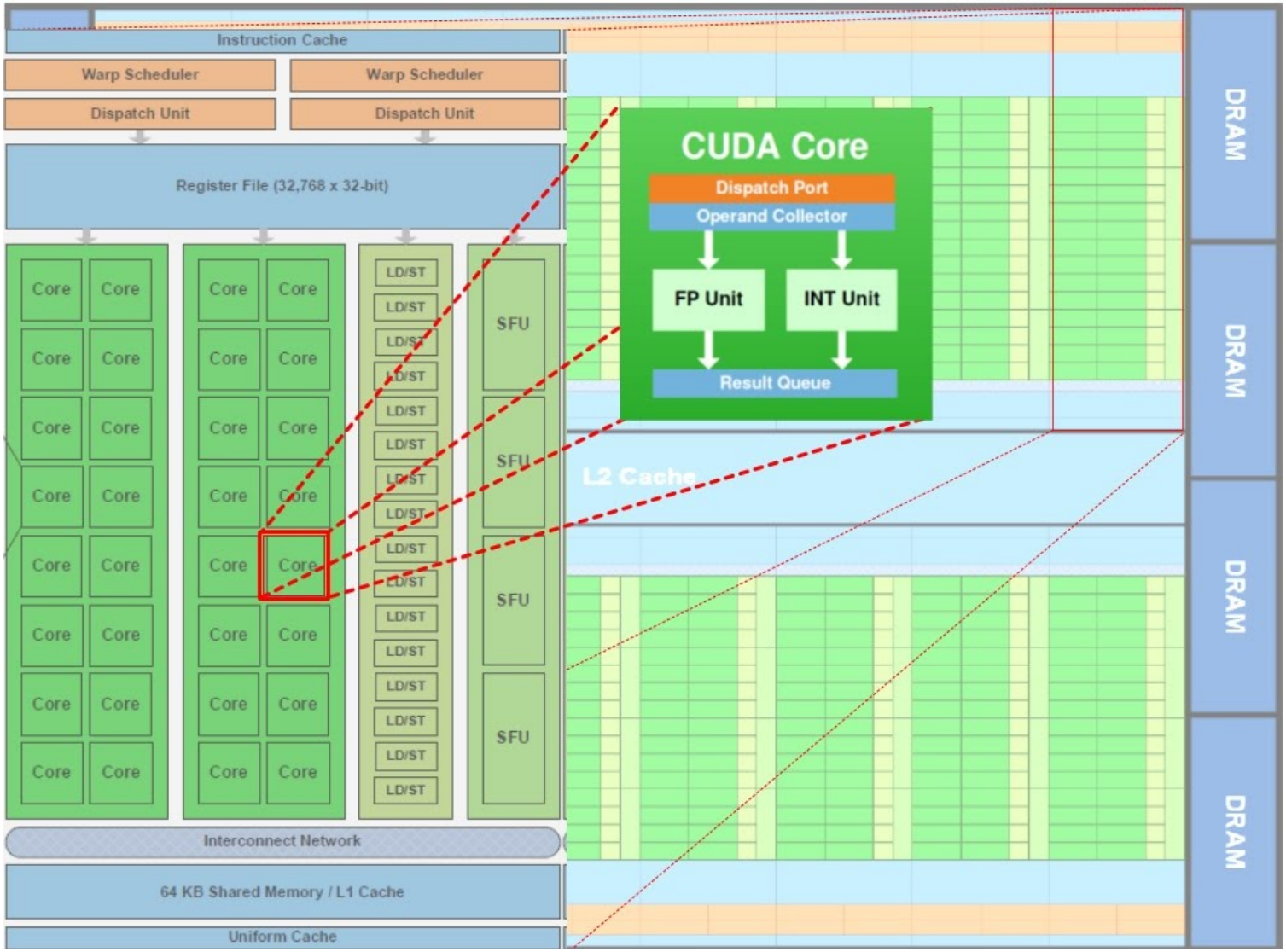


Fermi architecture



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

Fermi architecture

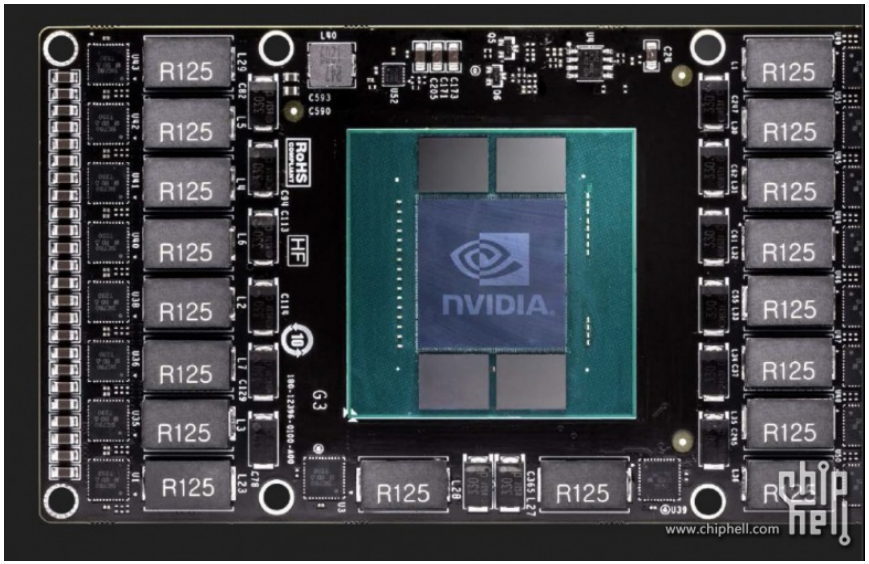


Pascal architecture

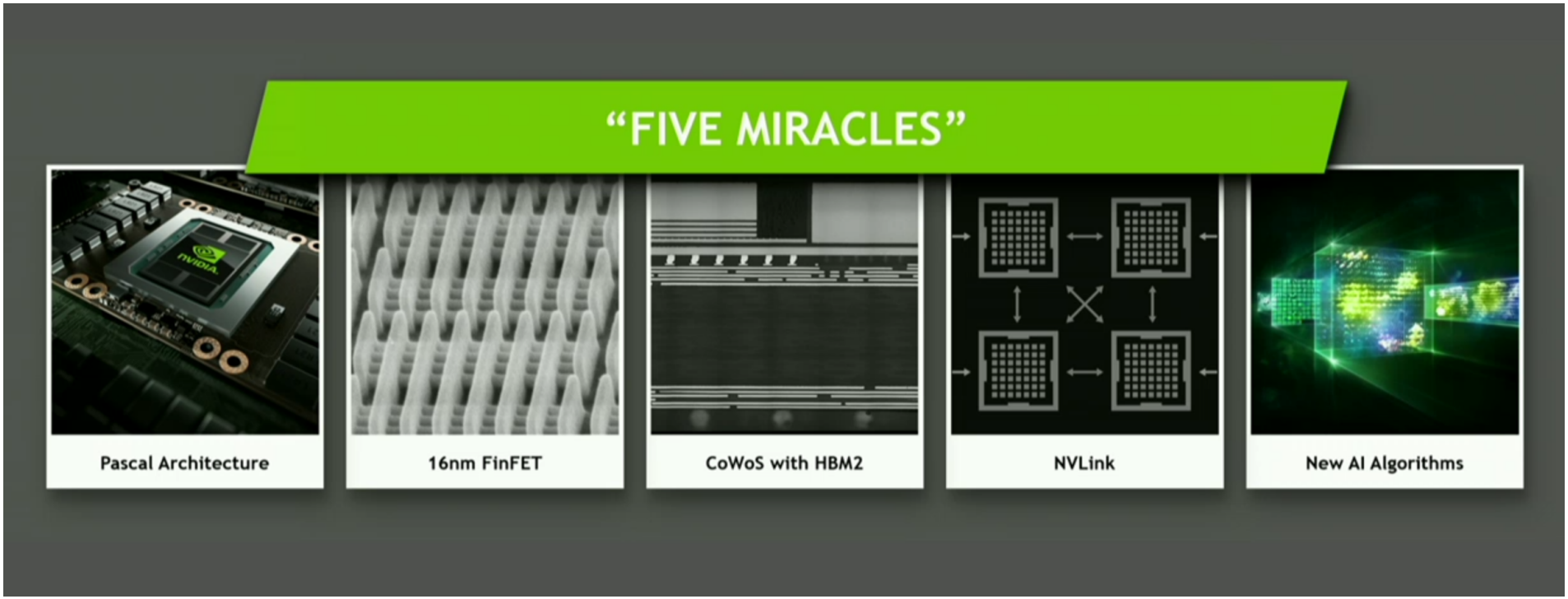
- successor to the Maxwell architecture
- April 2016 with the release of the Tesla P100 (GP100) on April 5
- primarily used in the GeForce 10 series
- Several usages in Deep learning
- 16nm FinFET

Pascal architecture

- Architectural improvements
  - CUDA Compute Capability 6.1
  - new memory standard supporting 10Gbit/s data rates, updated memory controller
  - DisplayPort 1.4, HDMI 2.0b
  - GPU Boost 3.0



Pascal architecture



Pascal architecture

Pascal architecture

	TITAN X PASCAL	TITAN X MAXWELL
CUDA cores	3584	3072
Boost Clock	1.53 GHZ	1.08GHZ
Memory	12GB G5X	12GB G5
Memory Speed (Gb/s)	10	7
Memory Bandwidth (GB/s)	480	336
Texture Rate (GT/s)	343	206
TFLOPS (INT8)	44	-
TFLOPS (FP32)	11	7

Pascal architecture

Nvidia Tesla Workstation GPU Specifications Comparison					
	Full P2xx	**P100**	**M40**	**K80**	**K40**
Family	Pascal (2nd-gen)	Pascal	Maxwell (2nd-gen)	Kepler (2nd-gen)	Kepler
Architecture	N/A	GP100	GM200	GK210	GK110
Cores	3840	3584	3072	2 x 2496	2880
ROPs	N/A	N/A	96	96	48
Texture Units	240	224	192	416	240
Core Clock	N/A	1328MHz	948MHz	562MHz	745MHz
Memory Clock	N/A	1400MHz	1500MHz	2500MHz	3004MHz
Memory Bandwidth	N/A	N/A	288GB/s	480GB/s	288GB/s
Memory Bus Width	4096-bit	4096-bit	384-bit	2x 384-bit	384-bit
Memory Type	HBM2	HBM2	GDDR5	GDDR5	GDDR5
Memory Size	16GB or higher	16GB	24GB	2 x 12GB	12GB
Die Size	N/A	$610mm^2$	$601mm^2$	$561mm^2$	$551mm^2$
Transistors	N/A	15.3 billion	8 billion	2 x 7.08 billion	7.08 billion
Register File Size / SM	256KB	256KB	256KB	512KB	256KB
L2 Cache	N/A	4MB	3MB	1.5MB	1.5MB
TDP	N/A	300W	250W	300W	235W
Manufacturing Process	TSMC 16nm	TSMC 16nm	TSMC 28nm	TSMC 28nm	TSMC 28nm
Release Date	N/A	Jul-16	Nov-15	Nov-14	Nov-13

References

[1] GP100 Datasheet
[2] jetron Datasheet
[3] professional CUDA programming
[4] CUDA for Engineers

Thanks for your attention.