**Advanced Architectures** 



### **Master Informatics Eng.**

2019/20 *A.J.Proença* 

#### Data Parallelism 2 (NVidia GPUs)

(most slides are borrowed)

### **Beyond Vector/SIMD architectures**

#### $\sim$

- Vector/SIMD-extended architectures are hybrid approaches
  - mix (super)scalar + vector op capabilities on a single device
  - highly pipelined approach to reduce memory access penalty
  - tightly-closed access to shared memory: lower latency
- Evolution of Vector/SIMD-extended architectures
  - PU (Processing Unit) cores with wider vector units
    - <u>x86</u> many-core: ...
    - other many-core: ...

#### - coprocessors (require a host scalar processor): accelerator devices

- <u>x86</u> on disjoint physical memories ...
- ISA-free architectures: ...
- focus on SIMT/SIMD to hide memory latency: **GPU**-type approach

• ...

#### - heterogeneous PUs in a SoC: multicore PUs with GPU-cores

• .

## **Graphical Processing Units**

- Question to GPU architects:
  - Given the hardware invested to do graphics well, how can we supplement it to improve the performance of a wider range of applications?

### Key ideas:

- Heterogeneous execution model
  - CPU is the *host*, GPU is the *device*
- Develop a C-like programming language for GPU
- Unify all forms of GPU parallelism as CUDA\_threads
- Programming model follows SIMT: "Single Instruction Multiple Thread"



### *# cores/processing elements* in several devices



AJProença, Advanced Architectures, MiEI, UMinho, 2019/20

公义

# Theoretical peak performance in several computing devices (DP)



AJProença, Advanced Architectures, MiEI, UMinho, 2019/20

# Theoretical peak FP Op's per clock cycle in several computing devices (DP)



AJProença, Advanced Architectures, MiEI, UMinho, 2019/20

## **NVIDIA GPU Architecture**

- Similarities to vector machines:
  - Works well with data-level parallel problems
  - Scatter-gather transfers
  - Mask registers
  - Large register files
- Differences:
  - No scalar processor
  - Uses multithreading to hide memory latency
  - Has many functional units, as opposed to a few deeply pipelined units like a vector processor



### **Early NVidia GPU Computing Modules**

公



## **NVIDIA GPU Memory Structures**

- Each SIMD Lane has private section of off-chip DRAM
  - "Private memory" (Local Memory)
  - Contains stack frame, spilling registers, and private variables
- Each multithreaded SIMD processor (SM) also has local memory (Shared Memory)
  - Shared by SIMD lanes / threads within a block
- Memory shared by SIMD processors (SM) is GPU Memory, off-chip DRAM (Global Memory)
  - Host can read and write GPU memory



SM

### The NVidia Fermi architecture



## **Fermi Architecture Innovations**

#### Each SIMD processor has

- Two SIMD thread schedulers, two instruction dispatch units
- 16 SIMD lanes (SIMD width=32, chime=2 cycles), 16 load-store units, 4 special function units
- Thus, two threads of SIMD instructions are scheduled every two clock cycles



- Caches for GPU memory (16/64KiB\_L1/SM and global 768KiB\_L2)
- 64-bit addressing and unified address space
- Error correcting codes
- Faster context switching
- Faster atomic instructions





#### Fermi: Multithreading and Memory Hierarchy



AJProença, Advanced Architectures, MiEI, UMinho, 2019/20

公

#### TOP500 list in November 2010: 3 systems in the top4 use Fermi GPUs



公

#### HIGHLIGHTS: NOVEMBER 2010

- The Chinese Tianhe-1A system is the new No. 1 on the TOP500 and clearly in the lead with 2.57 petaflop/s performance.
- No. 3 is also a Chinese system called Nebulae, built from a Dawning TC3600 Blade system with Intel X5650 processors and NVIDIA Tesla C2050 GPUs
- There are seven petaflop/s systems in the TOP10
- The U.S. is tops in petaflop/s with three systems performing at the petaflop/s level
- The two Chinese systems and the new Japanese Tsubame 2.0 system at No. 4 are all using NVIDIA GPUs to accelerate computation and a total of 28 systems on the list are using GPU technology.

### Families in NVidia Tesla GPUs



#### From Fermi into Kepler: The Memory Hierarchy



AJProença, Advanced Architectures, MiEI, UMinho, 2019/20

公义



July'11

# From the GF110 to the GK110 Kepler Architecture





SMX: 192 CUDA-cores

Ratio DPunit : SPunit --> 1 : 3

AJProença, Advanced Architectures, MiEI, UMinh

#### From Fermi to Kepler core: SM and the SMX Architecture

| Instruction Cache                       |      |      |          |       |                   |      |         |                   |                |      |                   |      |         |                |      |      |         |       |    |
|---|------|------|----------|-------|-------------------|------|---------|-------------------|----------------|------|-------------------|------|---------|----------------|------|------|---------|-------|----|
| Warp Scheduler                          |      |      |          |       | Warp Scheduler    |      |         |                   | Warp Scheduler |      |                   |      |         | Warp Scheduler |      |      |         |       |    |
| Dispatch                                |      |      | Dispatch |       | Dispatch Dispatch |      |         | Dispatch Dispatch |                |      | Dispatch Dispatch |      |         |                |      |      |         |       |    |
| Register File (65.536 x 32-bit)         |      |      |          |       |                   |      |         |                   |                |      |                   |      |         |                |      |      |         |       |    |
| + |      |      |          |       |                   |      |         |                   |                |      |                   |      |         |                |      |      |         |       |    |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | s  |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | s  |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | s  |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | s  |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | s  |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | s  |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | s  |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | ٤  |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | \$ |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | ٤  |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | \$ |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | \$ |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | \$ |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | \$ |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | \$ |
| Core                                    | Core | Core | DP Unit  | Core  | Core              | Core | DP Unit | LD/ST             | SFU            | Core | Core              | Core | DP Unit | Core           | Core | Core | DP Unit | LD/ST | ٤  |
| Interconnect Network                    |      |      |          |       |                   |      |         |                   |                |      |                   |      |         |                |      |      |         |       |    |
| 48 KB Read-Only Data Cache              |      |      |          |       |                   |      |         |                   |                |      |                   |      |         |                |      |      |         |       |    |
|   | Tex  |      | Tex      |       |                   | Tex  |         | Tex               | c              |      | Tex               |      | Tex     | (              |      | Tex  |         | Tex   | 2  |
|   | Tex  |      | Tex      | c Tex |                   | Тех  | Tex     |                   | Tex            |      | Тех               |      | Tex     |                |      | Ter  |         |       |    |

# From the GK110 to the GM200 Maxwell Architecture

Kepler: 2880 CUDA-cores *October'13* 

PCI Express 3.0 Host Interface



Maxwell: 3072 CUDA-cores *November'15* 

AJProença, Advanced Architectures, MiEI, UMinho, 2019/20

18

#### The move from Kepler to Maxwell : from 15 SMXs to 48 SMMs in 6 GPCs

| e Core Core DPUnit Core Core Core DPUnit   | LDIST SFU Core Core Core DP Unit Core Core D  |                  | SWW           | 128 CLIDA-cores                   | SMM           |                       |               |                     |
|--|---|------------------|---------------|-----------------------------------|---------------|-----------------------|---------------|---------------------|
| e Core Core DP Unit Core Core Core DP Unit | LDIST SFU Core Core Core DP Unit Core Core DI | P Unit LDIST SFU | Detie Dourit  |                                   | Vertex Fe     | tch Tess              | ellator       | Viewport Transform  |
| e Core Core DP Unit Core Core DP Unit      | LDIST SFU Core Core Core DP Unit Core Core DI | P Unit LD/ST SFU | Ralio DPunit  | $: SPUNIT \longrightarrow 1 : 32$ |               | Attribute Setup       | Stream Output |                     |
| e Core Core DP Unit Core Core DP U         | GPC   |                  |               |                                   |               | Instructi             | on Cache      | alles D. Ker        |
| e Core Core DPUnit Core Core Core DPU      |   | Dester           | Facility      |                                   | Wa            | rp Scheduler          | Warg          | Scheduler           |
| e Core Core DP Unit Core Core DP U         |   |                  | Dispatch Unit | Dispatch Unit                     | Dispatch Unit | Dispatch Unit         |               |                     |
| e Core Core DP Unit Core Core Core DP U    |   |                  | CMM           | SMM                               | Register F    | ile (16,384 x 32-bit) | Register Fil  | e (16,384 x 32-bit) |
| e Core Core DP Unit Core Core Core DP U    |   | SMM              | SMIM          |                                   | Core Core C   | ore Core LDIST SFU    | Core Core Cor | e Core LDIST SFU    |
| 64 K                                       |   |                  |               |                                   | Core Core C   | ore Core LDIST SFU    | Core Core Cor | e Core LDIST SFU    |
| 48<br>Tex Tex Tex                          |   |                  |               |                                   | Core Core C   | ore Core LDIST SFU    | Core Core Cor | e Core LDIST SFU    |
| Tex Tex Tex                                |   |                  |               |                                   | Core Core C   | ore Core LDIST SFU    | Core Core Cor | e Core LDIST SFU    |
|  |   |                  |               |                                   | Core Core C   | ore Core LDIST SFU    | Core Core Cor | Core LDIST SFU      |
|  |   |                  |               |                                   | Core Core C   | ore Core LOST SFU     | Core Core Co  | e Core LDST SFU     |
|  |   |                  |               |                                   | Core Core C   | ore Core LDIST SFU    | Core Core Cor | e Core LDIST SFU    |
|  |   |                  |               |                                   |               | Texture               | L1 Carba      |                     |
|  |   |                  |               |                                   | Tex           | Tex                   | Tex           | Tex                 |
|  |   |                  |               |                                   | Inst          | ruction Buffer        | Instru        | ction Buffer        |
|  |   |                  |               |                                   | Dispatch Unit | Dispatch Unit         | Dispatch Unit | Dispatch Unit       |
|  |   |                  |               |                                   | Rogister F    | ile (16,384 x 32-bit) | Register Fi   | e (16,384 x 32-bit) |
|  |   |                  |               |                                   | Core Core C   | ore Core LD/ST SFU    | Core Core Co  | e Core LDST SFU     |
|  |   |                  |               |                                   | Core Core C   | ore Core LDIST SFU    | Core Core Co  | re Core LDIST SFU   |
|  |   |                  |               |                                   | Core Core C   | ore Core LDIST SFU    | Core Core Co  | re Core LDST SFU    |
|  |   |                  |               |                                   | Core Core C   | ore Core LDIST SFU    | Core Core Co  | re Core LDIST SFU   |
|  |   |                  |               |                                   | Core Core C   | ore Core LDIST SFU    | Core Core Co  | re Core LDIST SFU   |
|  |   |                  |               |                                   | Core Core C   | ore Core LDIST SFU    | Core Core Co  | re Core LDIST SFU   |
|  |   |                  |               |                                   | Core Core C   | ore Core LDIST SFU    | Core Core Co  | re Core LDIST SFU   |
|  |   |                  |               |                                   | Core Core C   | ore Core LDIST SFU    | Core Core Co  | re Core LDST SFU    |
|  |   |                  |               |                                   | Tex           | Texture               | L1 Cache      | Ter                 |
|  | 2   |                  |               |                                   |               | 96KB Sha              | red Memory    |                     |



# From the M200 to the GP100 Pascal Architecture

Maxwell: 3072 CUDA-cores *November'15* 

Pascal: 3584 CUDA-cores HBM on-package September'16







# From the GP100 to the GV100 Volta Architecture

Pascal: 3584 CUDA-cores *November'15* 



Volta: 5120 CUDA-cores HBM on-package *June'17* 

### Volta Architecture: 6x GPCs, 80 SMs

Tex

|  | SM   | ing Oppha  |
|--|--|--|
|  |  |  |
|  | L0 Instruction Cache   | L0 Instruction Cache   |
|  | Warp Scheduler (32 thread/clk)   | Warp Scheduler (32 thread/clk)                                   |
|  | Dispatch Unit (32 thread/clk)  | Dispatch Unit (32 thread/clk)                                    |
|  | Register File (16,384 x 32-bit)  | Register File (16,384 x 32-bit)                                  |
|  | FP64 INT INT FP32 FP32   | FP64 INT INT FP32 FP32   |
| lanc lanc lanc lanc lanc                     | FP64 INT INT FP32 FP32   | FP64 INT INT FP32 FP32   |
| NVLink NVLink NVLink                         | FP64 INT INT FP32 FP32   | FP64 INT INT FP32 FP32   |
|  | FP64 INT INT FP32 FP32 TENSOR TENSOR                                   | FP64 INT INT FP32 FP32 TENSOR TENSOR                             |
|  | FP64 INT INT FP32 FP32 CORE CORE                                       | FP64 INT INT FP32 FP32 CORE CORE                                 |
| Volta SM:                                    | FP64 INT INT FP32 FP32   | FP64 INT INT FP32 FP32   |
| 64 CLIDA coros                               | FP64 INT INT FP32 FP32   | FP64 INT INT FP32 FP32   |
| 04 CODA-COIES                                | FP64 INT INT FP32 FP32   | FP64 INT INT FP32 FP32   |
| New: 8 Tensor-cores                          | LD/ LD/ LD/ LD/ LD/ LD/ LD/ LD/ ST | LD/ LD/ LD/ LD/ LD/ LD/ LD/ LD/ ST |
| $\sim 10^{-1}$                               | L 0 Instruction Cache  | L 0 Instruction Cache  |
| Ratio <b>DP</b> unit : <b>SP</b> unit> 1 : 2 | Warp Scheduler (32 thread/clk)   | Warp Scheduler (32 thread/clk)                                   |
|  | Dispatch Unit (32 thread/clk)  | Dispatch Unit (32 thread/clk)                                    |
|  | Register File (16,384 x 32-bit)  | Register File (16,384 x 32-bit)                                  |
|  | FP64 INT INT FP32 FP32   | FP64 INT INT FP32 FP32   |
| Volta V100 w/ 16GiB HBM2                     | FP64 INT INT FP32 FP32   | FP64 INT INT FP32 FP32   |
|  | FP64 INT INT FP32 FP32   | FP84 INT INT FP32 FP32   |
|  | FP64 INT INT FP32 FP32 TENSOR TENSOR                                   | FP64 INT INT FP32 FP32 TENSOR TENSOR                             |
| 7251<br>P244                                 | FP64 INT INT FP32 FP32 CORE CORE                                       | FP64 INT INT FP32 FP32 CORE CORE                                 |
|  | FP64 INT INT FP32 FP32   | FP64 INT INT FP32 FP32   |
|  | FP64 INT INT FP32 FP32   | FP64 INT INT FP32 FP32   |
| 1961<br>1961                                 | FP64 INT INT FP32 FP32   | FP64 INT INT FP32 FP32   |
|  |  |  |
|  | LD/ LD/ LD/ LD/ LD/ LD/ LD/ LD/ ST    | LD/ LD/ LD/ LD/ LD/ LD/ LD/ LD/ ST ST ST ST ST ST ST ST ST       |

Tex

Tex

Tex

AJProença, Advanced Architectures, MiEl, UM

nee 2.0 Moet

| Tesla Product               | Tesla K40            | Tesla M40          | Tesla P100          | Tesla V100                  |  |
|-----------------------------|----------------------|--------------------|---------------------|-----------------------------|--|
| GPU                         | GK180 (Kepler)       | GM200<br>(Maxwell) | GP100<br>(Pascal)   | GV100 (Volta)               |  |
| SMs                         | 15                   | 24                 | 56                  | 80                          |  |
| TPCs                        | 15                   | 24                 | 28                  | 40                          |  |
| FP32 Cores / SM             | 192                  | 128                | 64                  | 64                          |  |
| FP32 Cores / GPU            | 2880                 | 3072               | 3584                | 5120                        |  |
| FP64 Cores / SM             | 64                   | 4                  | 32                  | 32                          |  |
| FP64 Cores / GPU            | 960                  | 96                 | 1792                | 2560                        |  |
| Tensor Cores / SM           | NA                   | NA                 | NA                  | 8                           |  |
| Tensor Cores / GPU          | NA                   | NA                 | NA                  | 640                         |  |
| GPU Boost Clock             | 810/875 MHz          | 1114 MHz           | 1480 MHz            | 1530 MHz                    |  |
| Peak FP32 TFLOP/s           | 5.04                 | 6.8                | 10.6                | 15.7                        |  |
| Peak FP64 TFLOP/s           | 1.68                 | .21                | 5.3                 | 7.8                         |  |
| Peak Tensor Core<br>TFLOP/s | NA                   | NA                 | NA                  | 125                         |  |
| Texture Units               | 240                  | 1 <b>92</b>        | 224                 | 320                         |  |
| Memory Interface            | 384-bit GDDR5        | 384-bit GDDR5      | 4096-bit<br>HBM2    | 4096-bit HBM2               |  |
| Memory Size                 | Up to 12 GB          | Up to 24 GB        | 16 GB               | 16 GB                       |  |
| L2 Cache Size               | 1536 KB              | 3072 KB            | 4096 KB             | 6144 KB                     |  |
| Shared Memory Size /<br>SM  | 16 KB/32 KB/48<br>KB | 96 KB              | 64 KB               | Configurable up to 96<br>KB |  |
| Register File Size / SM     | 256 KB               | 256 KB             | 256 KB              | 256KB                       |  |
| Register File Size / GPU    | 3840 KB              | 6144 KB            | 14336 KB            | 20480 KB                    |  |
| TDP                         | 235 Watts            | 250 Watts          | 300 Watts           | 300 Watts                   |  |
| Transistors                 | 7.1 billion          | 8 billion          | 15.3 billion        | 21.1 billion                |  |
| GPU Die Size                | 551 mm²              | 601 mm²            | 610 mm <sup>2</sup> | 815 mm²                     |  |
| Manufacturing Process       | 28 nm                | 28 nm              | 16 nm<br>FinFET+    | 12 nm FFN                   |  |

#### Tesla accelerators: recent evolution

#### ANNOUNCING TESLA V100 GIANT LEAP FOR AI & HPC VOLTA WITH NEW TENSOR CORE

21B xtors | TSMC 12nm FFN | 815mm<sup>2</sup> 5,120 CUDA cores 7.5 FP64 TFLOPS | 15 FP32 TFLOPS NEW 120 Tensor TFLOPS 20MB SM RF | 16MB Cache | 16GBHB 300 GB/s NVLink

https://devblogs.nvidia.com/parallelforall/inside-volta/