

Технології графічного процесінгу & розподілених обчислень

Лекція 5: Пам'ять та місце зберігання даних

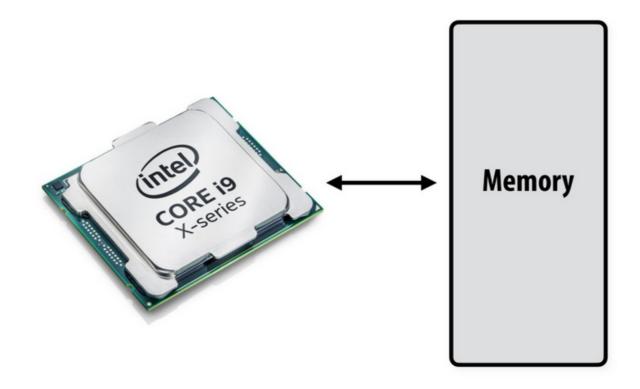
Кочура Юрій Петрович iuriy.kochura@gmail.com @y_kochura

Сьогодні

Мета— навчитися ефективно використовувати різні типи пам'яті CUDA у паралельному програмуванні

- Важливість ефективного доступу до пам'яті
- Регістри, спільна (shared) пам'ять, глобальна (global) пам'ять
- Область видимості та час життя

Що таке пам'ять?



Адресний простір програми

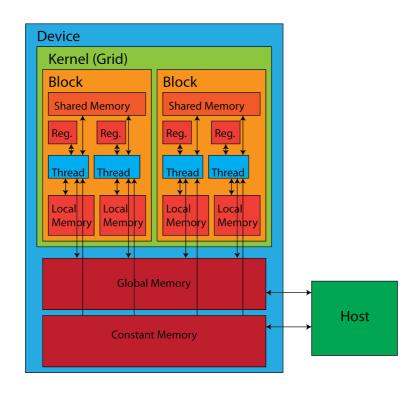
- Пам'ять комп'ютера організована як масив байтів
- Кожен байт ідентифікується за адресою в пам'яті

За адресою 0х8 зберігається значення 32

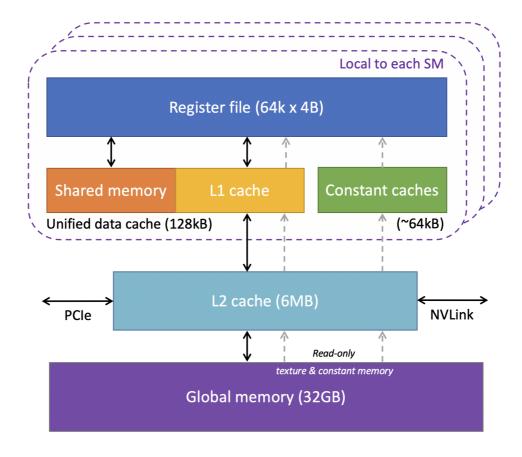
За адресою 0х10 зберігається значення 128

Address	Value		
0x0	16		
0x1	255		
0x2	14		
0x3	0		
0x4	0		
0x5	0		
0x6	6		
0x7	0		
0x8	32		
0x9	48		
0xA	255		
0xB	255		
0xC	255		
0xD	0		
0xE	0		
0xF	0		
0x10	128		
:	:		
0x1F	0		

Доступ до пам'яті



Рівні пам'яті



Ієрархія пам'яті в Tesla V100

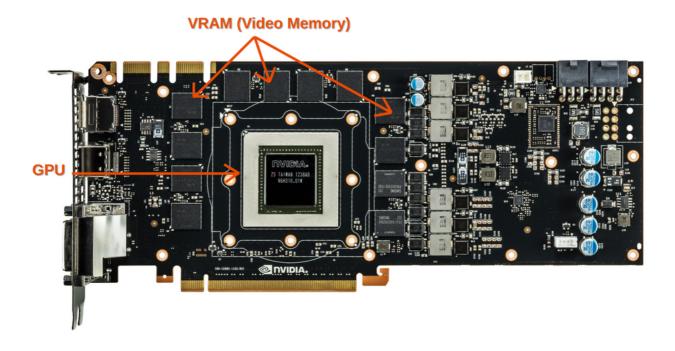
Вбудована пам'ять

- Register File
- L1 Cache
- Shared Memory
- Constant Caches

Віддалена пам'ять

- L2 Cache
- Global Memory
- Local Memory
- Texture and Constant Memory

Розмір і пропускна здатність



Кваліфікатори

Qualifiers	Location	Scope	Lifespan
Automatic variable	Register	Thread	Kernel
Automatic array	Local	Thread	Kernel
device,shared, int SharedVar	Shared	Block	Kernel
device, int GlobalVar	Global	Grid	Application
device,constant, int ConstVar	Constant	Grid	Application

Automatic Variables

int autovar;

Automatic Arrays

```
int autoarr[];
```

__shared__

```
__shared__ int shvar;

//is the same as

__device__ _shared__ int shvar2;
```

__device__

__device__ int dvvar;

__constant__

```
__constant__ int cnvar;

//is the same as

__device__ __constant__ int cnvar2;
```

Оголошення за адресою

```
float *ptr= &GlobalVar;
```

Література

PMPP Ch. 4, pp. 71-101

