GPGPU programming
General-purpose Processing on Graphics Processing Units

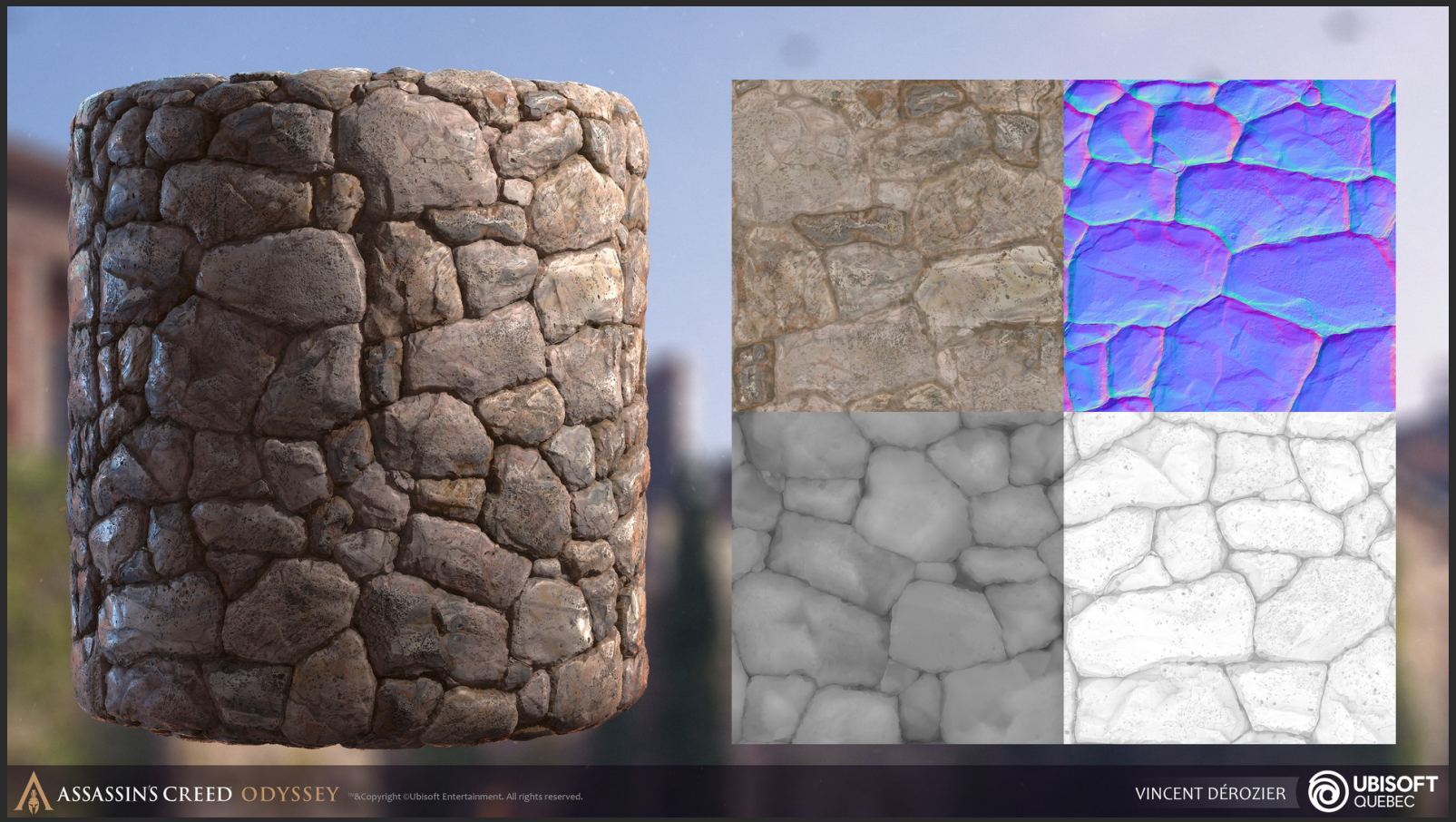
Robin Faury
robinfaurypro@gmail.com
robin.faury@allegorithmic.com

12-12-2018

Introduction

- The purpose of parallel processing
- What is a graphic card?
- The CUDA language
- GPGPU usage in the industry
- Q&A

Allegorithmic



PBR render and its maps

The purpose of parallel processing

Moore's Law

Every two years, the density of transistors in an integrated circuit doubles. That means we can compute the critical path of an algorithm faster.

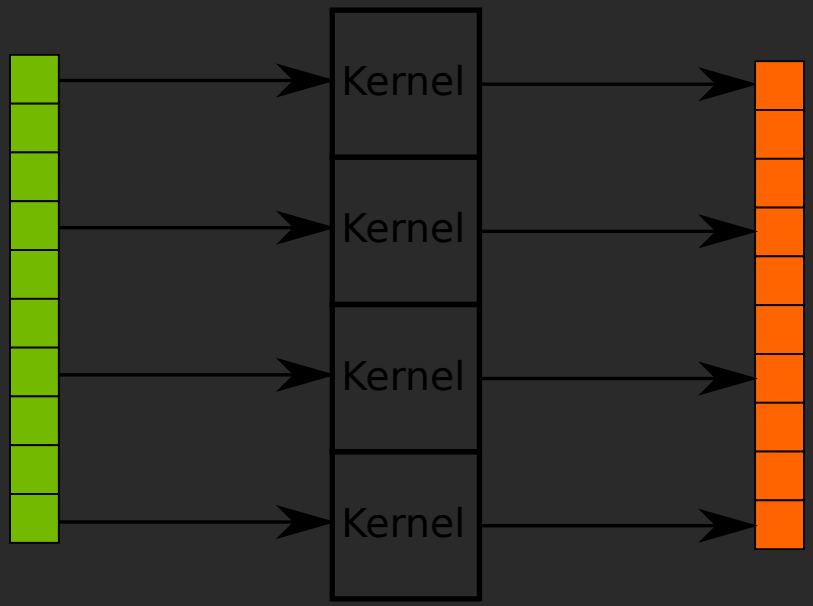
To infinity and beyond!

Critical path

Sometimes, algorithms process data one by one. When applicable, it is necessary to find the critical path and execute it in parallel. Modern CPUs offer the ability to run some threads at the same time. However, CPUs don't have a lot of cores available. For massive parallel computation we will use GPUs.

A world of buffers

The aim of parallel computing is solving heavy arithmetic computation on buffer. Single instruction on multiple data. One process is called a kernel for the GPGPU or a shader for the graphics pipeline.



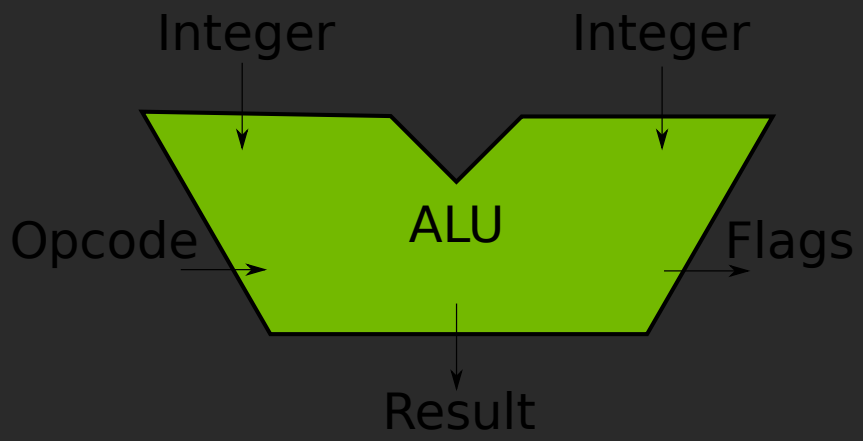
What is a graphic card?

History

The first Graphics Processing Unit (GPU) was used for drawing game sprites. It was a dedicated device for formatted data. Ten years after we had the ability to draw lines, fill areas and control the blitter. In 1990, the graphical API appears and allows us to send assembly code to the device.

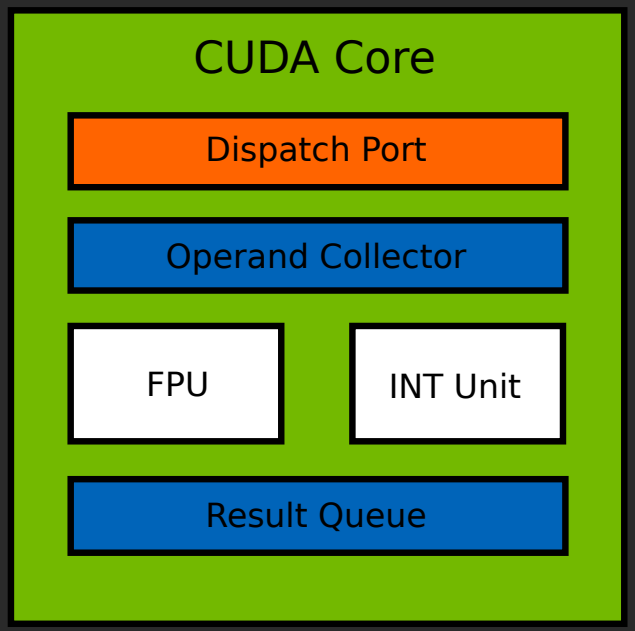
Arithmetic Logic Unit

The Arithmetic Logic Unit (ALU) is the component that performs arithmetic operations. The GPU is more focused on floating point operations, multiple ALUs are combined to create a Floating Point Unit (FPU).



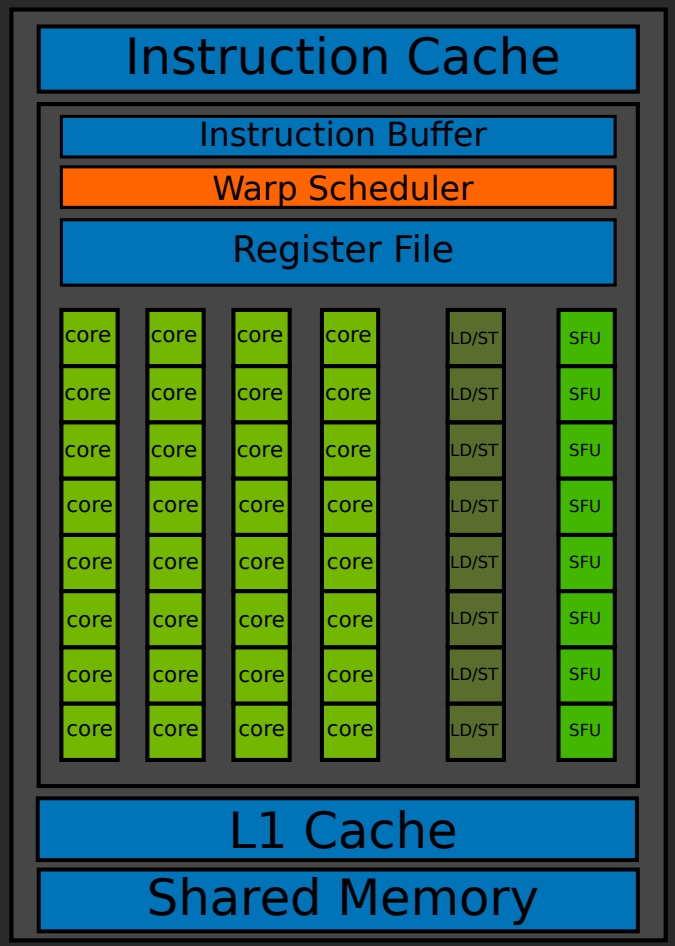
CUDA Core

CUDA cores are used to execute opcodes from compiled kernels. There are composed of an FPU, logic unit, branch unit and compare unit.



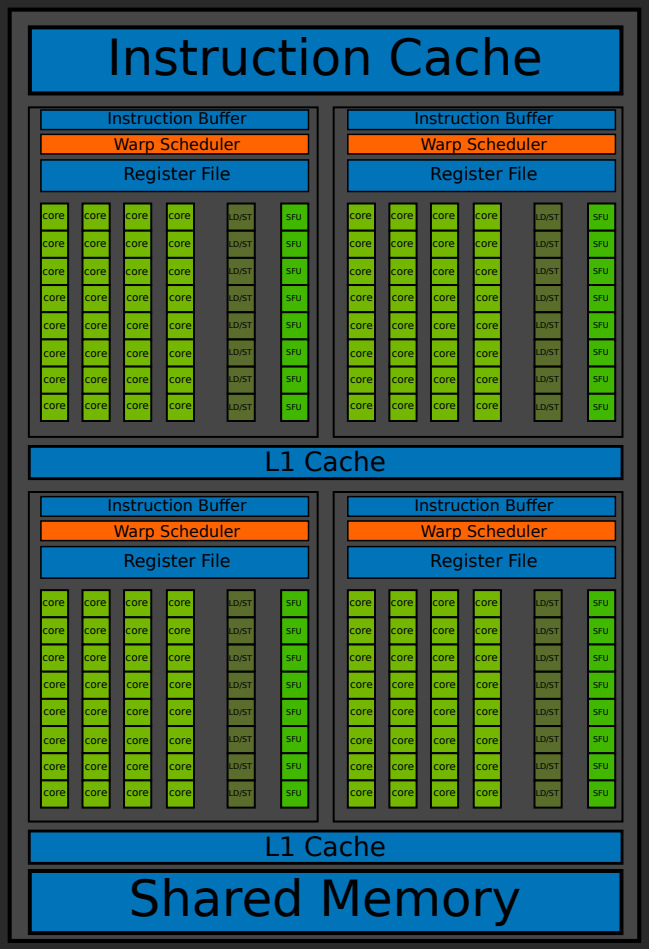
Streaming Multiprocessor

The Streaming Multiprocessor (SM) organizes threads in groups of 32 called warp. This architecture is called SPMD (Single Program, Multiple Data).



Streaming Multiprocessor

On the GP104 (The GPU of GTX 1080) each SM has four warps.

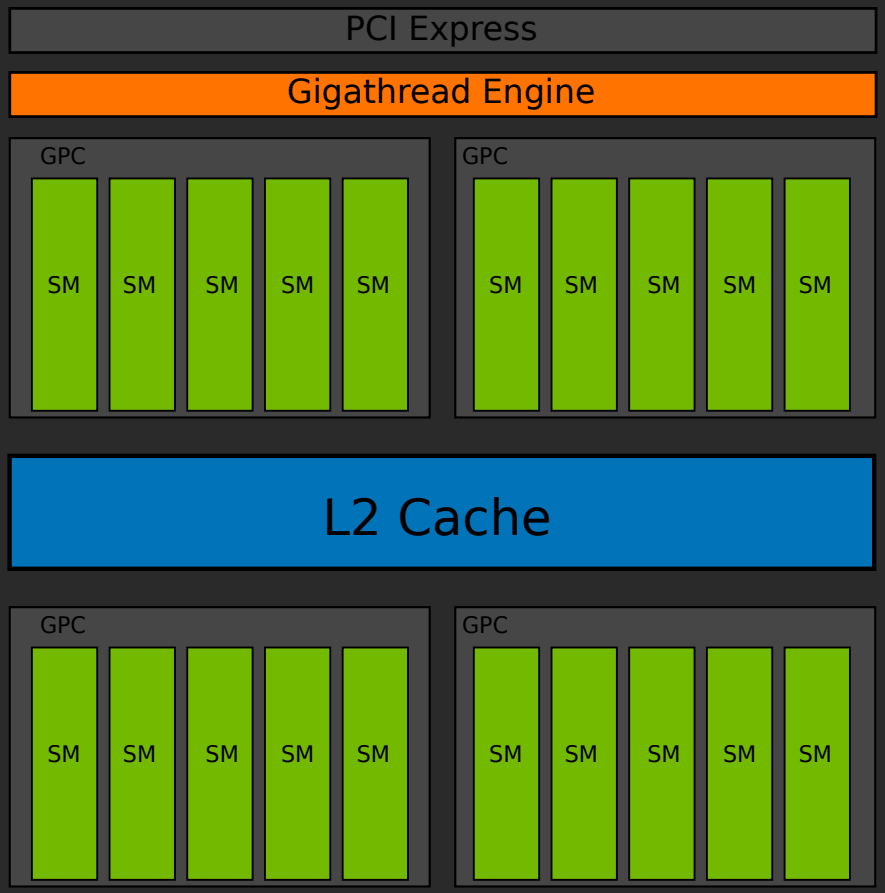


Graphics Processing Clusters

A Graphics Processing Clusters (GPC) is a collection of streaming multiprocessors. In the case of the GP104, there are four clusters.

GP104

All the GPC are connected to the L2 cache memory. The Gigathread engine distributes block threads to streaming multiprocessor. This device has 32 cores * 4 warps * 5 SMs * 4 GPCs = 2560 CUDA cores.



The CUDA language

Host and Devices

- Host: The CPU and its memory. The host can manage the memory on both the host and the device. The executed code can launch kernels.
- Devices: The GPU and its memory. Kernels are executed on many GPU threads in parallel.

Kernel

With CUDA, the kernel declaration is easy. the keyword __global__ has to be added before the kernel function.
The number of threads that execute the kernel is specified by this syntax:

```
int blocks = 16; // blocks per grid
int threads = 128; // threads per block
kernelName<<<blocks, threads>>>();
```

Indexing

In the kernel, the threadIdx, the blockIdx and the blockDim allow the user to compute the unique thread id.

```
int index = blockIdx * blockDim + threadIdx;
```

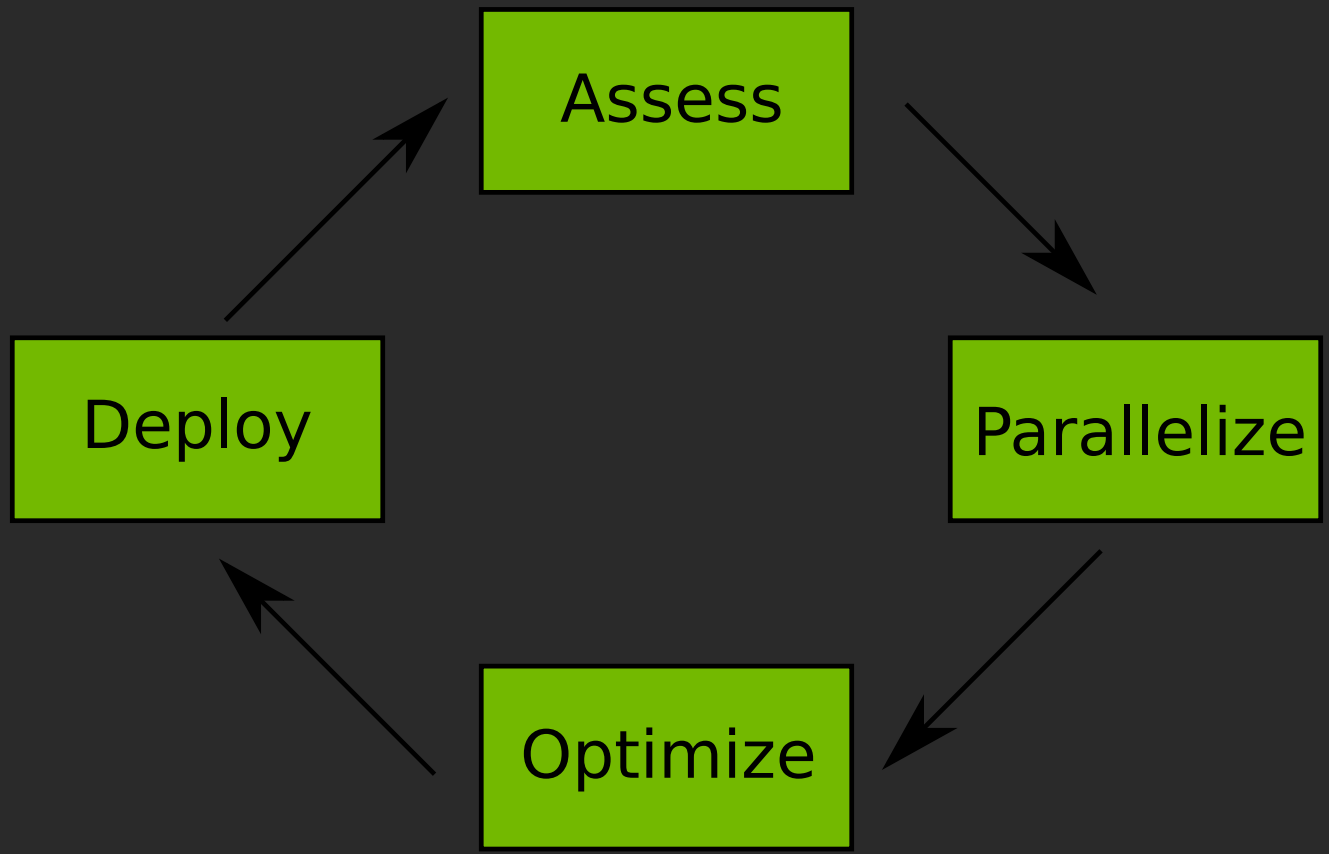
If data is stored into 2D or 3D array, it is possible to launch the kernel using a dim3 instead of an integer and the index becomes:

```
int x = blockIdx.x * blockDim.x + threadIdx.x;
int y = blockIdx.y * blockDim.y + threadIdx.y;
int z = blockIdx.z * blockDim.z + threadIdx.z;
```

GPGPU usage in the industry

APOD

The Assess, Parallelize, Optimize, Deploy (APOD) design cycle's goal is to identify and correct bottlenecks into the application.



Domain Specific

- Deep Learning
- Linear Algebra and Math: Solver, Random function, Finite element method, etc...
- Signal
- Image and video
- Data structure

Q&A