

POLYTECH PARIS-SACLAY

GP-GPU

CUDA basics

Stéphane Vialle

universit  PARIS-SACLAY
Centralesupelec

Sciences et technologies de l'information et de la communication (STIC)

LISN

Stephane.Vialle@centralesupelec.fr
http://www.metz.supelec.fr/~vialle

1

POLYTECH PARIS-SACLAY

CUDA basics

1. Principle of execution of a CUDA program
2. Variable definitions and CPU/GPU data transfers
3. Definition of a grid of blocks (of threads)
4. Definition of a 1st CUDA kernel
5. Execution of a CUDA kernel
6. Compiling a CUDA application

2

POLYTECH PARIS-SACLAY

Principle of execution of a CUDA program

Launching GPU calculations from the CPU

Steps for running a CUDA application:

- CPU: launch of a C++ pgm (*main* function)
- CPU: initialization of variables, light calculations
- CPU: data transfer from CPU memory to GPU memory
- CPU: launch of remote calculations on the GPU
→ GPU: execution of massively parallel "CUDA kernels"
- CPU: transfer of results from GPU memory to CPU memory

The GPU kernel codes are transferred from the CPU.

Warning : CPU/GPU transfers are time consuming !

3

POLYTECH PARIS-SACLAY

Principle of execution of a CUDA program

Execution of thread block grids

The CPU program requests the execution of a grid of blocks of CUDA threads:

- identical threads
- threads organized in blocks, each block running on a single Stream Multiprocessor
- blocks organized within a grid, which distributes its blocks on all Stream Multiprocessors
- it exists 1D, 2D and 3D grids and blocks:
 - 1D grid of 2D blocks
 - 2D grid of 1D blocks
 - 2D grid of 2D blocks
 - ...

4

POLYTECH PARIS-SACLAY

Principle of execution of a CUDA program

Execution of thread block grids

The CPU program requests the execution of a grid of blocks of CUDA threads:

- the block scheduler distributes the blocks on the different Streams Multiprocessors (SMs)
- different GPUs will execute the same grid of blocks at different rates

5

POLYTECH PARIS-SACLAY

Principle of execution of a CUDA program

Thread block execution, warp by warp

Block scheduler, SIMT model

one Stream Multiprocessor

6

Principle of execution of a CUDA program

Thread block execution, *warp* by *warp*

Block scheduler, **SIMT** model

Instruction decoder.

a basic Stream Multiprocessor

Thread scheduler: schedules **warps** of 32 threads consecutive on x dimension. **SIMD** model

0 31 32 63 x

0 1 2 3 y

one 2D block of 256 threads
= 8 **warps**

7

Principle of execution of a CUDA program

Constraints on block size

- an instruction decoder drives 32 "hardware threads" (32 ALUs)
- the thread scheduler enables warps of 32 thread
- if possible, warps are made of threads consecutive in x (privileged dimension)

- **Create blocks of at least 32 threads**
otherwise part of the ALU will always be unused
- **Create blocks with a thread count multiple of 32**
otherwise the last warp will be incomplete and ALUs will be unused.
- **Create blocks with an x-dimension multiple of 32**
otherwise the "coalescence" will be poor (see further)

But sometimes it still works very well with a block x-dimension multiple of 16...!

8

Principle of execution of a CUDA program

Kernel ↔ data structures ↔ grid of blocks

A grid of blocks deploys warps of threads of a CUDA kernel which access data structures on the GPU

Algorithm & code

Data structures

→ Need to **coherently develop** a GPU kernel, its grid of blocks and its data structures on the GPU

Ex: A thread of a 2D block must not make the same calculations as a thread of a 1D block to identify the array box it has to process (see further)

9

Principle of execution of a CUDA program

CUDA basics

1. Principle of execution of a CUDA program
2. Variable definitions and CPU/GPU data transfers
3. Definition of a grid of blocks (of threads)
4. Definition of a 1st CUDA kernel
5. Execution of a CUDA kernel
6. Compiling a CUDA application

10

Variable definitions and CPU/GPU data transfers

CUDA « qualifiers »

How CUDA "qualifiers" work:

	<u>__device__</u>	<u>__constant__</u>	<u>__shared__</u>
Variables	Global memory of the GPU Lifetime of the application R/W by GPU and CPU codes	Constant memory of the GPU Lifetime of the application W by CPU code R by GPU code	Shared memory of a multiprocessor Lifetime of the block of threads cache memory of the Global GPU memory R/W by GPU code:
Functions	<u>__device__</u> Call from GPU run on GPU	<u>__host__</u> (default) Call from CPU run on CPU	<u>__global__</u> Call on CPU run on GPU

11

Variable definitions and CPU/GPU data transfers

Data transfers CPU ↔ GPU

Variables allocated on the GPU at compile time ("symbol"):

```
float TabCPU[N]; // Array on CPU
__device__ float TabGPU[N]; // Array on GPU (symbol)
```

Transfer of a variable on the CPU to a "symbol" on the GPU:

```
// Copy all TabCPU array into TabGPU array
cudaMemcpyToSymbol(TabGPU, &TabCPU[0],
    sizeof(float)*N, 0,
    cudaMemcpyHostToDevice);
// Copy 2nd half of TabCPU array into 2nd half TabGPU array
cudaMemcpyToSymbol(TabGPU, &TabCPU[N/2],
    sizeof(float)*N/2, sizeof(float)*N/2,
    cudaMemcpyHostToDevice);
```

Transfer of a "symbol" on the GPU to a variable on the CPU:

```
// Copy all TabGPU array into TabCPU array
cudaMemcpyFromSymbol(&TabCPU[0], TabGPU,
    sizeof(float)*N, 0,
    cudaMemcpyDeviceToHost);
```

12

Variable definitions and CPU/GPU data transfers

Specificities of static variables

Static variables on GPU

- Fully defined at compilation time
- Directly accessible from the GPU code
→ no need to pass their addresses as kernel parameters
- BUT: any GPU code that uses these variables must be in the same file as the variable definition!** (you quickly get to a single big file that contains all the GPU code!).
- In fact, these variables/symbols can be shared between files. (declared external in .h files), but only by activating the mode of separate compilation

```

Compilation steps:
gcc -c ... X.c
nvcc -dc ... Y.cu
nvcc -dlink ... Y.o -o __gpu_rdc.o -lculadavrt
gcc -o pgm ... Y.o __gpu_rdc.o X.o ...
    
```

13

Variable definitions and CPU/GPU data transfers

Data transfers CPU↔GPU

Variables allocated on GPU at runtime:

```

float *TabCPU;           // Dynamic array on CPU
float *TabGPU;          // Dynamic array on GPU
cudaError_t cudaStat;   // Result of op on dynamic CUDA vars.

// Allocation of the dynamic arrays from the CPU
TabCPU = (float *) malloc(N*sizeof(float));
cudaStat = cudaMalloc((void **) &TabGPU, N*sizeof(float));
    
```

Copying dynamic variables from CPU to GPU:

```

// Copy TabCPU dynamic array into TabGPU dynamic array
cudaStat = cudaMemcpy(TabGPU, TabCPU, sizeof(float)*N,
                      cudaMemcpyHostToDevice);
    
```

Copying dynamic variables from GPU to CPU:

```

// Copy TabGPU dynamic array into TabCPU dynamic array
cudaStat = cudaMemcpy(&TabCPU[N/2], &TabGPU[N/2],
                      sizeof(float)*N/2, cudaMemcpyDeviceToHost);
    
```

Release of dynamic allocations:

```

free(TabCPU);
cudaStat = cudaFree(TabGPU);
    
```

14

Variable definitions and CPU/GPU data transfers

Specificities of dynamic variables

Dynamic variables on the GPU

Most are allocated and released by the CPU:

- their pointers are stored on the CPU: the CPU maps the GPU memory!
- their pointers must be passed as parameters when calling GPU kernels

These variables can be easily shared between several files (variables declared "extern" in .h files)

```

CPU:
#include "x.h"
float *adrData;
cudaMalloc(&adrData, SIZE*sizeof(float));
...
kernel<<<Dg,Db>>>(adrData);

GPU:
#include "x.h"
global void kernel(float *Tab) {
    ...
    Tab[SIZE/2]++;
    ...
}
    
```

15

POLYTECH PARIS SACLAY

CUDA basics

- Principle of execution of a CUDA program
- Variable definitions and CPU/GPU data transfers
- Definition of a grid of blocks (of threads)**
- Definition of a 1st CUDA kernel
- Execution of a CUDA kernel
- Compiling a CUDA application

16

Definition of the block grid

Limits on grid and block sizes

CUDA datatype for grid and block dimensions:

- dim3** is a structured datatype of 3 int (`_x`, `_y` and `_z`)
- For grid and block descriptors (**Dg** and **Db**)

Limits of 3D block size:

- $Db.x \leq 1024, Db.y \leq 1024, Db.z \leq 64$
- Total nb of threads / bloc ≤ 1024
- $Db.z = 1 \rightarrow 2D$
- $Db.z = Db.y = 1 \rightarrow 1D$

Limits of 3D grid size:

- $Dg.x \leq 2^{31} - 1$
- $Dg.y \leq 65535, Dg.z \leq 65535$
- $Dg.z = 1 \rightarrow 2D$
- $Dg.z = Dg.y = 1 \rightarrow 1D$

```

// GPU thread management
dim3 Dg, Db;
// Block of 128x4 threads
Db.x = 128;
Db.y = 4;
Db.z = 1;
// Grid of 512 blocks
Dg.x = 512;
Dg.y = 1;
Dg.z = 1;
// -> 262144 threads!
    
```

Launching a kernel from its block grid (see further):

```

// Classical call from a CPU routine
MyKernel<<< Dg, Db >>>(parameters);
    
```

17

Definition of the block grid

Block grid without overflow

Basic example of a 1D grid, with 1D blocks:

Tab[N]

For each box « i »:
Tab[i] = Tab[i] * 2.0

Strategy:

- 1 GPU thread will process 1 box
- 1D grid of 1D blocks of threads will process the array

1 - Definition of the blocks (1D)

```

Db.x = BLOCK_SIZE_X; // = 32; 64; 128; 256; 512; 1024
Db.y = BLOCK_SIZE_Y; // = 1
Db.z = BLOCK_SIZE_Z; // = 1
    
```

18


Definition of the block grid

Block grid without overflow


Basic example of a 1D grid, with 1D blocks:

2 - Definition of the grid, when: $(N \% Db.x) = 0$

We "pave" the data with blocks.



```
Dg.x = N/Db.x; // Dg.x = N/BLOCK_SIZE_X
Dg.y = 1;
Dg.z = 1;
```

What if $(N \% Db.x) \neq 0$? 


19

Definition of the block grid

Block grid with overflow

Basic example of a 1D grid, with 1D blocks:

2 - Definition of the grid, when: $(N \% Db.x) \neq 0$:



We "pave" the data with whole blocks, even if it overflows!

```
if (N%BLOCK_SIZE_X == 0)
    Dg.x = N/BLOCK_SIZE_X;
else
    Dg.x = N/BLOCK_SIZE_X + 1;
Dg.y = 1;
Dg.z = 1;
```

Warning : we're going to create too many threads
→ we'll have to take this into account in the code...


20

Definition of the block grid

Block grid with overflow

Basic example of a 1D grid, with 1D blocks:

2 - Definition of the grid, when: $(N \% Db.x) \neq 0$:



We "pave" the data with whole blocks, even if it overflows!

```
Dg.x = N/BLOCK_SIZE_X + (N%BLOCK_SIZE_X ? 1 : 0);
Dg.y = 1;
Dg.z = 1;
```

Warning : we're going to create too many threads
→ we'll have to take this into account in the code...


21

Definition of the block grid

Block grid with overflow

Basic example of a 1D grid, with 1D blocks:

2 - Definition of the grid, when: $(N \% Db.x) \neq 0$:



We "pave" the data with whole blocks, even if it overflows!

```
Dg.x = (N-1)/BLOCK_SIZE_X + 1;
Dg.y = 1;
Dg.z = 1;
```

Warning : we're going to create too many threads
→ we'll have to take this into account in the code...

22

Definition of the block grid


Create many small threads

Masking of GPU memory access times :


- a GPU switches from one thread warp to another very quickly
- a GPU masks the latency of its memory accesses by multi-threading

→ Do not hesitate to create large numbers of small GPU threads

Ex.: to process an array of N elements:

- Threads dealing with ONE element each 
- and a Grid of blocks of N threads in total

or:

- Threads dealing with n elements each 
- and a Grid of blocks of N/n threads in total

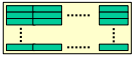
23

Definition of the block grid

Granularity of the grid and blocks


How many threads/block and blocks/grid to create ?


The **thread scheduler** of each block likes to have many warps ready to be activated, in order to hide the memory access times.

→ "big blocks of threads" 

The **block scheduler** likes to have several blocks ready to be loaded at the same time in each SM (*resident blocks*), in order to hide memory access times.

But a SM can only load a new block when it still has enough registers and other resources.

→ "lots of not too big blocks" 

 So, is it better to make a grid with few big blocks or many small blocks?

24

Definition of the block grid

Granularity of the grid and blocks

How many threads/block and blocks/grid to create ?

Several possible strategies:

- Calculate the size of the blocks leading to the maximum occupation of the GPU resources...
- Make medium size blocks (128/256 threads) to enable both schedulers to optimize execution
- Experimenting with various block sizes: 32/64/.../512/1024 threads

→ Optimal block grid granularity can be tedious to identify (depends on the CUDA code and GPU model)

But it remains an important source of performance improvement

→ Labs

25

Definition of the block grid

Granularity of the grid and blocks

Source : <https://en.wikipedia.org/wiki/CUDA>

Technical specifications	Compute capability (version)															
	1.0	1.1	1.2	1.3	2.x	3.0	3.2	3.5	3.7	5.0	5.2	5.3	6.0	6.1	6.2	7.0 (7.2)
Maximum number of resident grids per device (concurrent kernel execution)	16															
Maximum dimensionality of grid of thread blocks	3															
Maximum x- or y- or z-dimension of a grid of thread blocks	65535															
Maximum y- or z-dimension of a grid of thread blocks	65535															
Maximum dimensionality of thread block	3															
Maximum x- or y-dimension of a block	512															
Maximum z-dimension of a block	1024															
Maximum number of threads per block	512															
Warp size	32															
Maximum number of resident blocks per multiprocessor	8															
Maximum number of resident warps per multiprocessor	24															
Maximum number of resident threads per multiprocessor	768															
Number of 32-bit registers per multiprocessor	8 K															
Maximum number of 32-bit registers per thread block	124															

Some characteristics are very stable, others less so...

26

POLYTECH PARIS-SACLAY

CUDA basics

- Principle of execution of a CUDA program
- Variable definitions and CPU/GPU data transfers
- Definition of a grid of blocks (of threads)
- Definition of a 1st CUDA kernel
- Execution of a CUDA kernel
- Compiling a CUDA application

27

Definition and execution of a 1st kernel

« Qualifiers » de CUDA

Fonctionnement des « qualifiers » de CUDA :

	<u>__device__</u>	<u>__constant__</u>	<u>__shared__</u>
Variables	Mémoire globale GPU	Mémoire constante GPU	Mémoire partagée d'un multiprocesseur
	Durée de vie de l'application	Durée de vie de l'application	Durée de vie de la block de threads
	Accessible par les codes GPU et CPU	Ecrit par code CPU, lu par code GPU	Accessible par le code GPU, sert à cacher la mémoire globale GPU
	<u>__device__</u>	<u>__host__</u> (default)	<u>__global__</u>
Fonctions	Appel sur GPU Exec sur GPU	Appel sur CPU Exec sur CPU	Appel sur CPU Exec sur GPU

→ Les « qualifiers » différencient les parties de code GPU et CPU.

28

Definition and execution of a 1st kernel

1^{er} Kernel (traitant 1 donnée par thread)

Kernel utilisant la mémoire globale et des registres

Une barre de threads par bloc, et une barre de blocs par grille (un choix).
Un thread traite une seule donnée.
Hyp : $Nd = k \cdot BLOCK_SIZE_X$

```

__global__ void k1(void)
{
    int idx; // Registers:
    float data; // 16 Kreg per
    float res; // multipro.

    // Compute data idx of the thread
    idx = threadIdx.x +
        blockIdx.x * BLOCK_SIZE_X;
    // Read data from the global mem
    data = InGPU[idx];
    // Compute result
    res = (data + 1.0f) * data ...;
    // Write result in the global mem
    OutGPU[idx] = res;
}

```

29

Definition and execution of a 1st kernel

1^{er} Kernel (traitant 1 donnée par thread)

Calcul de l'indice de la donnée traitée par chaque thread

```

// Compute data idx of the thread
idx = blockIdx.x * BLOCK_SIZE_X
    + threadIdx.x;
// Read data from the global mem
data = InGPU[idx];

```

2 variables implicites et propres à chaque thread :

```

dim3 threadIdx
dim3 blockIdx

```

$idx = blockIdx.x * BLOCK_SIZE_X + threadIdx.x$

Indexage permettant des accès coalescents

30

POLYTECH PARIS-SACLAY

Définition et exécution d'un 1^{er} kernel

1^{er} Kernel (traitant 1 donnée par thread)

Calcul de l'indice de la donnée traitée par chaque thread

```
// Compute data idx of the thread
idx = (N-1) - (threadIdx.x +
             blockDim.x*BLOCK_SIZE_X);
// Read data from the global mem
data = InGPU[idx];
```

2 variables implicites et propres à chaque thread :

```
dim3 threadIdx
dim3 blockDim
```

Grille de blocs de threads

Tableau de données

$idx = (N-1) - (blockIdx.x*BLOCK_SIZE_X + threadIdx.x)$

Mais ce serait un indexage *moins* coalescent... !!

31

POLYTECH PARIS-SACLAY

Définition et exécution d'un 1^{er} kernel

1^{er} Kernel (traitant 1 donnée par thread)

Calcul de l'indice de la donnée traitée par chaque thread

```
// Compute data idx of the thread
idx = threadIdx.x +
      blockDim.x*BLOCK_SIZE_X;
if (idx % 2 == 1)
  idx = (N-1) - idx;
// Read data from the global mem
data = InGPU[idx];
```

2 variables implicites et propres à chaque thread :

```
dim3 threadIdx
dim3 blockDim
```

Grille de blocs de threads

Tableau de données

$idx = (N-1) - (blockIdx.x*BLOCK_SIZE_X + threadIdx.x)$

Mais ce serait un indexage *NON* coalescent... !!

32

POLYTECH PARIS-SACLAY

Définition et exécution d'un 1^{er} kernel

1^{er} Kernel (traitant 1 donnée par thread)

Kernel utilisant la mémoire globale et des registres

Une barre de threads par bloc, et une barre de blocs par grille (un choix).
Un thread traite *une seule* donnée.

Hyp : $Nd \gg k \cdot BLOCK_SIZE_X$

```
global__ void k1(void)
{
  int idx; // Registers:
  float data; // 16kreg per
  float res; // multipro.
  // Compute data idx of the thread
  idx = threadIdx.x +
        blockDim.x*BLOCK_SIZE_X;
  // If the elt indexed exists:
  if (idx < Nd) {
    // Read data from the global mem
    data = InGPU[idx];
    // Compute result
    res = (data + 1.0f)*data ...;
    // Write result in the global mem
    OutGPU[idx] = res;
  }
}
```

```
Db = {BLOCK_SIZE_X,1,1}
if (Nd%BLOCK_SIZE_X == 0)
  Dg = {Nd/BLOCK_SIZE_X,1,1}
else
  Dg = {Nd/BLOCK_SIZE_X + 1,1,1}
```

Pavage classique

Protection classique :
Les threads « en trop » ne font rien...

33

POLYTECH PARIS-SACLAY

CUDA basics

1. Principle of execution of a CUDA program
2. Variable definitions and CPU/GPU data transfers
3. Definition of a grid of blocks (of threads)
4. Definition of a 1st CUDA kernel
5. Execution of a CUDA kernel
6. Compiling a CUDA application

34

POLYTECH PARIS-SACLAY

Execution of a CUDA kernel

Exécution depuis le CPU d'une grille de threads sur le GPU :

```
Kernel<<< Dg, Db >>>(parameter, ..., ...);
```

→ Descripteur d'un bloc 3D de threads
→ Descripteur d'une grille 3D de blocs 3D de threads

Versión complète

```
Kernel<<< Dg, Db, Ns, S >>>(parameter, ..., ...);
```

→ Numéro du *stream* d'interaction CPU/GPU (le 0 par défaut)... utile pour du recouvrement transferts/calculs en utilisant plusieurs *streams*
→ Allocation dynamique de *shared memory* au lancement de chaque bloc sur un SM... (0 octets par défaut)

Voir plus loin

35

POLYTECH PARIS-SACLAY

CUDA basics

1. Principle of execution of a CUDA program
2. Variable definitions and CPU/GPU data transfers
3. Definition of a grid of blocks (of threads)
4. Definition of a 1st CUDA kernel
5. Execution of a CUDA kernel
6. Compiling a CUDA application

36

POLYTECH PARIS-SACLAY

Compilation d'une application CUDA

Compilation 100% CUDA

Compilation d'applications CUDA – entièrement développées en CUDA :

Définitions de variables et fonctions avec « qualificateurs » CUDA

Code C, ou C++ avec des appels à la lib CUDA

Code C, ou C++ « standard »

Fichiers **xxx.cu** et **xxx.h**, et **xxx.cc** incluant si besoin **cuda.h** et **cuda_runtime.h**

nvcc → **binaire**
Codes CPU et GPU intégrés

Pour les codes C/C++ simples :

- Il est possible de simplement tout recompiler en **nvcc** dans des fichiers **xxx.cu** (et **xxx.cc** incluant **cuda.h** et **cuda_runtime.h**)
- Mais les optimisations sérielles peuvent en souffrir...

37

POLYTECH PARIS-SACLAY

Compilation d'une application CUDA

Compilation mixte CUDA & C++

Compilation d'applications mixte CUDA & C++ avec 1 compilateur :

Legacy code
Fichiers **xxx.cc** et **xxx.h** → **nvcc -c** → Fichiers objets

Code CUDA
Fichiers **yyy.cu**, **yyy.cc** et **yyy.h** → **nvcc -c** → Fichiers objets

Fichiers objets → **Edition de liens** → **binaire**
Codes CPU et GPU intégrés

- Code C++ métier : fichiers **.cc** compilés en **nvcc**
- Code CUDA : fichiers **.cu** compilés en **nvcc**
- Edition de liens : des pbs peuvent encore apparaître avec des templates ...

38

POLYTECH PARIS-SACLAY

Compilation d'une application CUDA

Compilation mixte CUDA & C++

Compilation d'applications mixte CUDA & C++ avec 2 compilateurs :

Legacy code
Fichiers **xxx.cc** et **xxx.h** → **g++ -c** / **icc -c** → Fichiers objets

Code CUDA
Fichiers **yyy.cu**, **yyy.cc** et **yyy.h** → **nvcc -c** → Fichiers objets

Fichiers objets → **Edition de liens...** → **binaire**
Codes CPU et GPU intégrés

... pbs possibles (g++ ok) !!

- Code C++ métier : fichiers **.cc** compilés en **gcc/g++** ou en **icc** ou ...
- Code CUDA : fichiers **.cu** compilés en **nvcc**
- Edition de liens : des pbs de nommage peuvent apparaître (mais pas en **gcc**)

39

POLYTECH PARIS-SACLAY

CUDA basics

End

40