Praktische Aspekte der Informatik

Moritz Mühlhausen
Prof. Marcus Magnor
GPGPU-Programming

Cuda
Further Reading

Warning!
The following slides are meant to give you a very superficial introduction.

If you want to learn more, have a look at:
https://docs.nvidia.com/cuda/
Outline

Introduction

Programming Model

GPU Memory

Synchronization

NVCC
Introduction – Why GPU?

Benefits

• Highly parallel, multithreaded, manycore processors

• very high memory bandwidth

• well-suited for data-parallel computations

• waste less ressources

Drawbacks

• memory synchronization between CPU and GPU

• synchronization between GPU threads

• extra compiler

• only for NVIDIA GPUs (Cuda)

https://graphics.tu-bs.de/teaching/ss18/padi/
Introduction – Why GPU?

Theoretical GFLOP/s at base clock

- NVIDIA GPU Single Precision
- NVIDIA GPU Double Precision
- Intel CPU Single Precision
- Intel CPU Double Precision

https://graphics.tu-bs.de/teaching/ss18/padi/
Introduction – Why GPU?

Theoretical Peak GB/s

- GeForce GPU
- Tesla GPU
- Intel CPU

https://graphics.tu-bs.de/teaching/ss18/padi/
Introduction – Why GPU?

- Element-wise sum of two vectors/matrices

```
A  B  C
2  5  
4  1  
... ... ...
1  8  ...
7  2  ...
... ...
1  8  ...
2  1  ...
5  7  ...
```

https://graphics.tu-bs.de/teaching/ss18/padi/
Introduction – Why GPU?

- Element-wise sum of two vectors/matrices
  - Naive implementation: simple for-loop over all indices

```
A = 
| 2 |
| 4 |
| ... |
| 1 |
| 7 |
| ... |
| 8 |
| 2 |

B = 
| 5 |
| 1 |
| ... |
| 8 |
| 2 |
| ... |
| 1 |
| 7 |

C = 
| ... |
| ... |
| ... |
| 1 |
| ... |
| 2 |
| ... |
```

https://graphics.tu-bs.de/teaching/ss18/padi/
Introduction – Why GPU?

• Element-wise sum of two vectors/matrices
  - Naive implementation: simple for-loop over all indices

\[
\begin{array}{c}
A \\
\begin{array}{c}
2 \\
4 \\
\vdots \\
1 \\
7 \\
\vdots \\
8 \\
2 \\
\end{array}
+ \\
B \\
\begin{array}{c}
5 \\
1 \\
\vdots \\
8 \\
2 \\
\vdots \\
1 \\
7 \\
\end{array}
= \\
C \\
\begin{array}{c}
\vdots \\
\vdots \\
\vdots \\
7 \\
\end{array}
\end{array}
\]
Introduction – Why GPU?

• Element-wise sum of two vectors/matrices
  - Naive implementation: simple for-loop over all indices

A
2
4
...
1
7
...
8
2

B
5
1
...
8
2
...
1
7

C
7
5
...

https://graphics.tu-bs.de/teaching/ss18/padi/
Introduction – Why GPU?

• Element-wise sum of two vectors/matrices
  ▪ Naive implementation: simple for-loop over all indices

\[
\begin{array}{c|c|c}
A & B & C \\
\hline
2 & 5 & 7 \\
4 & 1 & 5 \\
… & … & … \\
1 & 8 & … \\
7 & 2 & … \\
… & … & … \\
8 & 1 & … \\
2 & 7 & … \\
\end{array}
\]

https://graphics.tu-bs.de/teaching/ss18/padi/
Introduction – Why GPU?

- Element-wise sum of two vectors/matrices
  - Naive implementation: simple for-loop over all indices
  - Better: several threads iterate over all indices
Introduction – Why GPU?

- Element-wise sum of two vectors/matrices
  - Naive implementation: simple for-loop over all indices
  - Better: several threads iterate over all indices
Introduction – Why GPU?

- Element-wise sum of two vectors/matrices
  - Naive implementation: simple for-loop over all indices
  - Better: several threads iterate over all indices

$$\begin{array}{cccc}
  & A & B & C \\
 2 & 5 & 7 \\
4 & 1 & 5 \\
\ldots & \ldots & \ldots \\
1 & 8 & 9 \\
7 & 2 & 9 \\
\ldots & \ldots & \ldots \\
8 & 1 & 9 \\
2 & 7 & 9 \\
\end{array}$$

https://graphics.tu-bs.de/teaching/ss18/padi/
Introduction – Why GPU?

- Element-wise sum of two vectors/matrices
  - Naive implementation: simple for-loop over all indices
  - Better: several threads iterate over all indices
  - GPU can handle more threads concurrently!

\[
\begin{array}{c|c|c}
A & B & C \\
2 & 5 & \cdots \\
4 & 1 & \cdots \\
\ldots & \ldots & \ldots \\
1 & 8 & \cdots \\
7 & 2 & \cdots \\
\ldots & \ldots & \ldots \\
8 & 1 & \cdots \\
2 & 7 & \\
\end{array}
\]

https://graphics.tu-bs.de/teaching/ss18/padi/
### Introduction – Why GPU?

- **Element-wise sum of two vectors/matrices**
  - Naive implementation: simple for-loop over all indices
  - Better: several threads iterate over all indices
  - GPU can handle more threads concurrently!

### Example

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

\[
A + B = C
\]

1. allocate GPU memory

<table>
<thead>
<tr>
<th>A'</th>
<th>B'</th>
<th>C'</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

https://graphics.tu-bs.de/teaching/ss18/padi/
Introduction – Why GPU?

- Element-wise sum of two vectors/matrices
  - Naive implementation: simple for-loop over all indices
  - Better: several threads iterate over all indices
  - GPU can handle more threads concurrently!

1. allocate GPU memory
2. copy A and B values
### Why GPU?

- Element-wise sum of two vectors/matrices
  - Naive implementation: simple for-loop over all indices
  - Better: several threads iterate over all indices
  - GPU can handle more threads concurrently!

### Example

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

+  
=  

<table>
<thead>
<tr>
<th>A'</th>
<th>B'</th>
<th>C'</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

1. allocate GPU memory
2. copy A and B values
3. Execute adding kernel

[Link to Teaching Material](https://graphics.tu-bs.de/teaching/ss18/padi/)
Introduction – Why GPU?

- **Element-wise sum of two vectors/matrices**
  - Naive implementation: simple for-loop over all indices
  - Better: several threads iterate over all indices
  - GPU can handle more threads concurrently!

```plaintext
A: [2, 4, 1, 7, 8, 2, 2]
B: [5, 1, 8, 2, 1, 7, 7]
C: [7, 5, 9, 9, ...]

A': [2, 4, 1, 7, 8, 2, 2]
B': [5, 1, 8, 2, 1, 7, 7]
C': [7, 5, 9, 9, ...]
```

1. allocate GPU memory
2. copy A and B values
3. Execute adding kernel
4. copy C from GPU

[https://graphics.tu-bs.de/teaching/ss18/padi/](https://graphics.tu-bs.de/teaching/ss18/padi/)
Introduction – Overview

GPGPU Overview

1. allocate GPU memory
2. copy A and B values
3. Execute adding kernel
4. copy C from GPU

\[
\begin{array}{ccc}
A & + & B \\
\hline
2 & 5 & 7 \\
4 & 1 & 5 \\
... & ... & ...
\end{array}
\]  
\[
\begin{array}{ccc}
A' & + & B' \\
\hline
2 & 5 & 7 \\
4 & 1 & 5 \\
... & ... & ...
\end{array}
\]
Outline

Introduction

Programming Model

GPU Memory

Synchronization

NVCC
Programming Model – Kernels

GPGPU Overview

<table>
<thead>
<tr>
<th>A</th>
<th></th>
<th>B</th>
<th></th>
<th>C</th>
<th></th>
<th>A’</th>
<th></th>
<th>B’</th>
<th></th>
<th>C’</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td>5</td>
<td></td>
<td>7</td>
<td></td>
<td>2</td>
<td></td>
<td>5</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>1</td>
<td></td>
<td>5</td>
<td></td>
<td>4</td>
<td></td>
<td>1</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>8</td>
<td></td>
<td>9</td>
<td></td>
<td>1</td>
<td></td>
<td>8</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>2</td>
<td></td>
<td>9</td>
<td></td>
<td>2</td>
<td></td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>1</td>
<td></td>
<td>9</td>
<td></td>
<td>8</td>
<td></td>
<td>1</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>7</td>
<td></td>
<td>9</td>
<td></td>
<td>2</td>
<td></td>
<td>7</td>
<td></td>
<td>9</td>
</tr>
</tbody>
</table>

1. allocate GPU memory
2. copy A and B values
3. Execute adding kernel
4. copy C from GPU
Programming Model – Kernels

• CPU callable function to start computations on the GPU

• executed N times in parallel by N different CUDA threads

• Defined using __global__ declaration specifier

• Number of CUDA threads are specified using a new <<<...>>> execution configuration

• Asynchronous from CPU code
Parallel element-wise sum of two matrices

```c
// Kernel definition
__global__ void MatAdd(float* A, float* B, float* C)
{
    int i = threadIdx.x;
    C[i] = A[i] + B[i];
}

int main()
{
    ...

    // Kernel invocation with N threads
    MatAdd<<<1, N>>>(A, B, C);
    ...
}
```

`__global__` specifier defines a kernel function, only `void` return.

`MatAdd<<<1, N>>>(A, B, C)` calls the kernel function with `N` threads.

`threadIdx.x` specifies the id of the thread. This way each thread is handling a different matrix element.

*Kernels* can only return `void`, but they can alter global GPU memory.
Parallel element-wise sum of two matrices

```c
// Kernel definition
__global__ void MatAdd(float* A, float* B, float* C, int N)
{
    int i = threadIdx.x;
    int j = threadIdx.y;
}

int main()
{
    ...
    // Kernel invocation with one block of N * N * 1 threads
    int numBlocks = 1;
    dim3 threadsPerBlock(N, N);
    MatAdd<<<numBlocks, threadsPerBlock>>>(A, B, C);
    ...
}
```

_**_global__ specifier defines a kernel function, only **void** return.

MatAdd<<<numBlocks, threadsPerBlock>>>(A, B, C) calls the kernel function with N * N threads.

threadIdx.x & .y specifies the id of the thread. This way each thread is handling a different matrix element.

**Kernels** can only return **void**, but they can alter global GPU memory.

**Threads** can be identified by a one-, two- or three-dimensional threadIdx.

- Convenient for computations for a vector, matrix or volume

https://graphics.tu-bs.de/teaching/ss18/padi/
1. allocate GPU memory
2. copy A and B values
3. Execute adding kernel
4. copy C from GPU
Parallel element-wise sum of two matrices

Number of **threads** is limited.
- reside on the same processor core
- share limited memory resources
- depending on the GPU up to 1024 threads

Kernel can be executed by multiple equally-shaped **thread blocks**.
- organized in one-, two- or three-dimensional grid (**dim3** or **int**)
- required to execute independently

```c
// Kernel invocation with one block of N * N * 1 threads
int numBlocks = 1;
dim3 threadsPerBlock(N, N);
MatAdd<<<numBlocks, threadsPerBlock>>>(A, B, C);
```
1. allocate GPU memory
2. copy A and B values
3. Execute adding kernel
4. copy C from GPU

Programming Model – Blocks

GPGPU Overview

A

B

C

A'

B'

C'

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

https://graphics.tu-bs.de/teaching/ss18/padi/
Programming Model – Blocks

Parallel element-wise sum of two matrices

```c
// Kernel definition
__global__ void MatAdd(float* A, float* B, float* C, int N)
{
    int i = blockIdx.x * blockDim.x + threadIdx.x;
    int j = blockIdx.y * blockDim.y + threadIdx.y;

}

int main()
{
    ...

    // Kernel invocation with one block of N * N * 1 threads
    dim3 numBlocks(N, N);
    dim3 threadsPerBlock(N, N);
    MatAdd<<<numBlocks, threadsPerBlock>>>(A, B, C);
    ...
}
```

MatAdd<<<numBlocks, threadsPerBlock>>>(A, B, C) calls the kernel function in N * N blocks with N * N threads.

`blockIdx.x` & `.y` specifies the id of the block.

`blockDim.x` & `.y` holds the number of blocks in the respective dimension.

**Block** can be identified by a one-, two- or three-dimensional `blockIdx`. 

[Link to Graphics TU-Bs.de Teaching SS18 PADI](https://graphics.tu-bs.de/teaching/ss18/padi/)
Programming Model – Blocks

Parallel element-wise sum of two matrices

```c
// Kernel definition
__global__ void MatAdd(float* A, float* B, float* C, int N) {
    int i = blockIdx.x * blockDim.x + threadIdx.x;
    int j = blockIdx.y * blockDim.y + threadIdx.y;
}

int main() {
    ...
    // Kernel invocation with one block of N * N * 1 threads
    dim3 numBlocks(N, N);
    dim3 threadsPerBlock(N, N);
    MatAdd<<<numBlocks, threadsPerBlock>>>(A, B, C, N);
    ...
}
```

MatAdd<<<numBlocks, threadsPerBlock>>>(A, B, C) calls the kernel function in N * N blocks with N * N threads.

blockIdx.x & .y specifies the id of the block.

blockDim.x & .y holds the number of blocks in the respective dimension.

**Block** can be identified by a one-, two- or three-dimensional blockIdx.

What happens if it is still not enough?

https://graphics.tu-bs.de/teaching/ss18/padi/
1. allocate GPU memory
2. copy A and B values
3. Execute adding kernel
4. copy C from GPU
Programming Model – Stride

Parallel element-wise sum of two matrices

```c
// Kernel definition
__global__ void MatAdd(float* A, float* B, float* C, int N) {
    int i = blockIdx.x * blockDim.x + threadIdx.x;
    int j = blockIdx.y * blockDim.y + threadIdx.y;
    int iStride = blockDim.x * gridDim.x;
    int jStride = blockDim.y * gridDim.y;

    for(int y = j; y < N; y += jStride) {
        for(int x = i; x < N; x += iStride) {
            C[y * N + x] = A[y * N + x] + B[y * N + x];
        }
    }
}
```

**Loop** over all indices.

**Memory distance** of threads within a block should be **minimized**.

Allows the GPU to utilize cached memory and therefore **minimize** global memory lookups.

**Loop** prevents the kernel of accessing indices outside the pointer.

**Block-** and **Gridsize** can be independent from matrix/vector size.
1. allocate GPU memory
2. copy A and B values
3. Execute adding kernel
4. copy C from GPU
Programming Model – Device Function

Parallel element-wise sum of two matrices

```c
// Device function definition
__device__ float stuff(float const& a, float const& b)
{
    ...;
    return ...;
}

// Kernel definition
__global__ void MatAdd(float* A, float* B, float* C, int N)
{
    int i = blockIdx.x * blockDim.x + threadIdx.x;
    int j = blockIdx.y * blockDim.y + threadIdx.y;
    int iStride = blockDim.x * gridDim.x;
    int jStride = blockDim.y * gridDim.y;

    for(int y = j; y < N; y += jStride)
    {
        for(int x = i; x < N; x += iStride)
        {
            C[y * N + x] = stuff(A[y*N+x], B[y*N+x]);
        }
    }
}
```

**Device functions** can only be called from **global** or other **device functions**.

Any **return type** allowed.

**Split your device code** into small functions for **readability**!
• **__host__** specifies a host (CPU) callable function.
  - Default specifier, any return type possible

• **__device__** specifies a device (GPU thread) callable function.
  - Threads may call these functions
  - Not only void as return type allowed

• **__host__ __device__** specifies a function callable by the host (CPU) and the device (GPU thread)
  - No need to write functions for CPU and GPU separately

• **__global__** can not be combined with any of these

https://graphics.tu-bs.de/teaching/ss18/padi/
Outline

Introduction
Programming Model
GPU Memory
Synchronization
NVCC

https://graphics.tu-bs.de/teaching/ss18/padi/
GPU Memory – Memory Hierarchy

- Pre-thread local memory
  - No other thread or block have access to read or write
// Kernel definition
__global__ void MatAdd(float* A, float* B, float* C, int N)
{
    int i = blockIdx.x * blockDim.x + threadIdx.x;
    int j = blockIdx.y * blockDim.y + threadIdx.y;
    int iStride = blockDim.x * gridSize.x;
    int jStride = blockDim.y * gridSize.y;

    for(int y = j; y < N; y += jStride)
    {
        for(int x = i; x < N; x += iStride)
        {
            C[y * N + x] = A[y * N + x] + B[y * N + x];
        }
    }
}

All parameters are in local memory.

Pointer are in local memory, but point to global memory!
Comparable to CPU: pointer are on STACK but point to HEAP

Standard Variables defined in global or device functions are in local memory.

Local memory is not accessible by any other thread.

Local variables will allocate memory for each thread!
GPU Memory – Memory Hierarchy

• Pre-thread local memory
  - No other thread or block have access to read or write

• Per-block shared memory
  - All threads within a block have access to the same shared memory
  - Faster memory access than global memory
__shared__ specifier defines shared memory.

Static shared memory array size can be explicitly declared.

Dynamic shared memory array size must be given as an optional third execution configuration parameter.

Shared memory is accessible to all threads within the same block.

Shared variables will allocate memory for each block!
GPU Memory – Memory Hierarchy

- **Pre-thread local memory**
  - No other thread or block have access to read or write

- **Per-block shared memory**
  - All threads within a block have access to the same shared memory
  - Faster memory access than global memory

- **Global memory**
  - All threads within all blocks have access to this memory
  - Slowest memory access, try to copy to shared memory if possible
Global memory

// Kernel definition
__global__ void MatAdd(float* A, float* B, float* C, int N)
{
    ...
}
...

float *a_gpu, *b_gpu, *c_gpu;
cudaMalloc(&a_gpu, n * sizeof(float));
cudaMalloc(&b_gpu, n * sizeof(float));
cudaMalloc(&c_gpu, n * sizeof(float));
...

cudaFree(a_gpu);
cudaFree(b_gpu);
cudaFree(c_gpu);

Global memory is not directly accessible.

Pointer can point to global memory! Comparable to CPU: pointer are on STACK but point to HEAP

Global memory allocation needs to be done before the kernel call!

Global memory is accessible by any thread within any block if they have a pointer pointing to it.

Global memory allocation is done once for each cudaMalloc(...) call!
GPGPU Overview

1. allocate GPU memory
2. copy A and B values
3. Execute adding kernel
4. copy C from GPU

A  B  C
2  5  7
4  1  9
...  ...  ...
1  8  9
7  2  9
...  ...  ...
8  1  9
2  7  9

A'  B'  C'
2  5  7
4  1  9
...  ...  ...
1  8  9
7  2  9
...  ...  ...
8  1  9
2  7  9

https://graphics.tu-bs.de/teaching/ss18/padi/
GPU Memory – Global Memory Allocation

• Global GPU memory has to be allocated and freed with special commands
  ▪ `cudaMalloc(void** ptr, size_t sizeInBytes)` allocate memory on the GPU
  ▪ `cudaFree(void*)` will free the allocated memory

• GPU memory allocation does **not initialize** the memory!
  ▪ Always initialize it yourself before you read from it
1. allocate GPU memory
2. copy A and B values
3. Execute adding kernel
4. copy C from GPU
Global memory

... float *a_gpu, *b_gpu, *c_gpu;
cudaMalloc(&a_gpu, n * sizeof(float));
cudaMalloc(&b_gpu, n * sizeof(float));
cudaMalloc(&c_gpu, n * sizeof(float));
cudaMemcpy(a_gpu, a, n * sizeof(float), cudaMemcpyHostToDevice);
cudaMemcpy(b_gpu, b, n * sizeof(float), cudaMemcpyHostToDevice);
cudaMemset(c_gpu, 0, n * sizeof(float));

... cudaMemcpy(c, c_gpu, n * sizeof(float), cudaMemcpyDeviceToHost);
cudaFree(a_gpu);
cudaFree(b_gpu);
cudaFree(c_gpu);

Global memory is not initialized during allocation!

cudaMemcpy(...) copies values from/to global memory. Last parameter specifies direction.

cudaMemset(...) sets bytes to the specified int.

cudaMemcpy(void* dst, void* src, size_t sizeInBytes, cudaMemcpyKind kind)
cudaMemset(void* dst, int value, size_t sizeInBytes)

Both methods lock the CPU thread!

-> cudaMemcpyAsync(...), cudaMemcpyAsync(...)
For standard 2D or 3D global memory it can be beneficial to add padding for faster access
- \texttt{cudaMallocPitched(...), cudaMalloc3D(...)} does add padding automatically
- \texttt{cudaMemcpyPitched(...), cudaMemcpy3D(...)} are then used to copy values

Constant memory also resides in device memory
- Cached in constant cache
- Read-only memory, but lookup can be faster

Texture and Surface memory also resides in device memory
- Cached in texture cache
- Optimized for 2D spatial locality, higher bandwidth can be achieved
- Read-only memory
Outline

Introduction

Programming Model

GPU Memory

Synchronization

NVCC

https://graphics.tu-bs.de/teaching/ss18/padi/
Synchronization – CPU and GPU

- Independent tasks that can operate concurrently
  - Computation on the host
  - Computation on the device
  - Memory transfer from the host to the device
  - Memory transfer from the device to the host
  - Memory transfers within the memory of a given device
  - Memory transfers among devices

- Kernel calls are asynchronous by default
  - Host code will continue

- `CudaMemcpy()`, `cudaMemset()` etc. also have asynchronous calls
  - `cudaMemcpyAsync()`, `cudaMemsetAsync()`, ...

- `cudaDeviceSynchronize()` can be used to synchronize GPU and CPU
  - Will lock host-code until all called device code finished

https://graphics.tu-bs.de/teaching/ss18/padi/
Synchronization between kernel calls and copies?

- Kernel, which alters global memory and a memcpy from this global memory are started
- Copy should wait for the kernel to finish...
Synchronization – CPU and GPU – Streams

• Synchronization between kernel calls and copies?
  ▪ Kernel, which alters global memory and a memcpy from this global memory are started
  ▪ Copy should wait for the kernel to finish…

• Streams are used for device call synchronization
  ▪ Same stream calls run sequentially in call order

• Without specification default stream (0) is used
  ▪ All kernel and copy calls are synchronized
  ▪ No need to care about this

• `cudaStreamSynchronize(streamid)` synchronizes CPU and GPU with respect to a specific stream

• Use streams when you have calls that are independent from others

Vector reduction

A

\[2\]
\[4\]
\[\ldots\]
\[1\]
\[7\]
\[\ldots\]
\[8\]
\[2\]

Sum of all elements in a vector

\[+\rightarrow 85\]
Vector reduction

// Kernel definition
__global__ void reduce(int const* A, int* sum, int N) {
    int tid = blockIdx.x * blockDim.x + threadIdx.x;
    int stride = gridDim.x * blockDim.x;
    sum[0] = sum[0] + A[tid];
}

int *a_gpu, *sum;
cudaMalloc(&a_gpu, n * sizeof(int));
cudaMalloc(&sum, sizeof(int));
cudaMemcpy(a_gpu, b, n * sizeof(int), cudaMemcpyHostToDevice);
cudaMemset(sum, 0, sizeof(int));
Reduce<<<1, N>>>(a_gpu, sum, N);

Sum of all elements in a vector

Threads may not write to the same memory address in parallel
• Undefined behaviour

Threads may not read a memory address while another thread writes to it
• Undefined which value is read!
Vector reduction

// Kernel definition
__global__ void reduce(int const* A, int* sum, int N)
{
    int tid = blockIdx.x * blockDim.x + threadIdx.x;
    int stride = blockDim.x * blockDim.x;

    atomicAdd(sum, A[tid]);
}

int *a_gpu, *sum;
cudaMalloc(&a_gpu, n * sizeof(int));
cudaMalloc(&sum, sizeof(int));
cudaMemcpy(a_gpu, b, n * sizeof(int), cudaMemcpyHostToDevice);
cudaMemset(sum, 0, sizeof(int));
Reduce<<<1, N>>>(a_gpu, sum, N);

Atomics are guaranteed to be performed without interference from other threads!

Sum of all elements in a vector

Threads may not write to the same memory address in parallel
• Undefined behaviour

Threads may not read a memory address while another thread writes to it
• Undefined which value is read!
Synchronization – Threads – Atomics

• Atomics guarantee execution **without interference** from other threads
  - Even **across blocks**!

• Several atomic-functions are implemented directly

• In principle it is possible to implement a **mutex** for threads and make any functionality 'atomic'

• When atomic functions are used and another thread is accessing the memory it will have to **wait** for the atomic to end

• **Try to resolve synchronization elsewise if possible!**

https://graphics.tu-bs.de/teaching/ss18/padi/
### Vector reduction

<table>
<thead>
<tr>
<th>A</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sum of all elements in a vector.

Each thread only **sum 2 elements** in each iteration -> **less active waiting** for a write/read operation!

**Synchronization** between **iterations** must be done!
Vector reduction

Sum of all elements in a vector.

Each thread only \textbf{sum 2 elements} in each iteration -> \textbf{less active waiting} for a write/read operation!

\textbf{Synchronization} between \textbf{iterations} must be done!
Vector reduction

Sum of all elements in a vector.

Each thread only sum 2 elements in each iteration -> less active waiting for a write/read operation!

Synchronization between iterations must be done!
Vector reduction

Sum of all elements in a vector.

Each thread only **sum 2 elements** in each iteration -> **less active waiting** for a write/read operation!

**Synchronization** between **iterations** must be done!
Vector reduction

Sum of all elements in a vector.

Each thread only **sum 2 elements** in each iteration -> **less active waiting** for a write/read operation!

**Synchronization** between **iterations** must be done!
Vector reduction

Sum of all elements in a vector.

Each thread only **sum 2 elements** in each iteration -> **less active waiting** for a write/read operation!

**Synchronization** between **iterations** must be done!
Synchronization – Threads

Vector reduction

A

2 → 6
4 → 20
… → 24
1 → 8
7 → 26
… → 15
8 → 10
2 → 44

Sum of all elements in a vector.

Each thread only **sum 2 elements** in each iteration -> **less active waiting** for a write/read operation!

**Synchronization** between **iterations** must be done!
Synchronization – Threads – Blocks

Vector reduction

// Kernel definition
__global__ void reduce(int const* A, int* sum, int N) {
    extern __shared__ int sdata[];
    int tid = blockIdx.x * blockDim.x + threadIdx.x;
    int stride = gridDim.x * blockDim.x;

    sdata[tid] = A[tid];
    __syncthreads();

    for(int s=1; s < blockDim.x; s*=2) {
        if(tid % (2*s) == 0)
            sdata[tid] += sdata[tid + s];
        __syncthreads();
    }

    if(tid==0) sum[0] = sdata[0];
}

Reduce<<<1, N, N*sizeof(int)>>>(a_gpu, sum, N);

Sum of all elements in a vector

__syncthreads() will synchronize all threads within the same block

No synchronization across blocks
• may not run in parallel

For even more speedups (up to 30x):
Outline

Introduction
Programming Model
GPU Memory
Synchronization

NVCC

https://graphics.tu-bs.de/teaching/ss18/padi/
• CUDA files should end with .cu and compiled with NVCC

• NVCC – NVIDIA Cuda Compiler
  ▪ Proprietary compiler by Nvidia for CUDA

• Separates the code into two parts
  ▪ Host code, which is forwarded to a C compiler
  ▪ Device code, which is further compiled by NVCC

• CMake will automatically use NVCC for .cu files
  ▪ Needs CMake 3.8 or newer
  ▪ For older versions use find_package() and cuda_add_executable()
• CUDA API calls return a cudaError that can be checked
  - cudaMemcpy(), cudaMemcpy(), ...

• NVCC – NVidia Cuda Compiler
  - Proprietary compiler by Nvidia for CUDA

• Separates the code into two parts
  - Host code, which is forwarded to a C compiler
  - Device code, which is further compiles by NVCC

• CMake will automatically use NVCC for .cu files
  - Needs CMake 3.8 or newer
  - For older versions use find_package() and cuda_add_executable()
CUDA API calls return a cudaError that can be checked
• cudaMalloc(), cudaMemcpy(), ...

CUDA kernel calls do not return any value
• asynchronous calls cannot know whether an error will occur
• After execution it can be checked if an error occurred with cudaGetLastError()

Error message can be checked with cudaGetErrorString()