Nvidia's Compute Unified Device Architecture (CUDA)

Mike Bailey
mjb@cs.oregonstate.edu

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

CUDA wants you to break the problem up into Pieces

If you were writing in C/C++, you would say:

```c
void ArrayMult( int n, float *a, float *b, float *c) {
    for ( int i = 0; i < n; i++ )
        c[i] = a[i] * b[i];
}
```

If you were writing in CUDA, you would say:

```c
__global__
void ArrayMult( float *dA, float *dB, float *dC )
{
    int gid = blockIdx.x*blockDim.x + threadIdx.x;
}
```

Think of this as having an implied for-loop around it, looping through all possible values of gid.

The CUDA Paradigm

CUDA is an NVIDIA-only product. It is very popular, and got the whole GPU-as-CPU ball rolling, which has resulted in other packages like OpenCL.

CUDA also comes with several libraries that are highly optimized for applications such as linear algebra and deep learning.

Organization: Blocks are Arranged in Grids

- The GPU's workload is divided into a Grid of Blocks.
- Each Block's workload is divided into a Grid of Threads.
• The threads in a block each have Thread ID numbers within the Block.
• Your CUDA program will use these Thread IDs to select work to do and pull the right data from memory.
• Threads share data and synchronize while doing their share of the work.
• Every 32 threads constitute a “Warp.” Each thread in a Warp simultaneously executes the same instruction on different pieces of data.
• But, it is likely that a Warp’s execution will need to stop at some point, waiting for a memory access. This would make the execution go idle – bad! So, it is worthwhile to have multiple Warps worth of threads available so that when one Warp blocks, another Warp can be swapped in.
• The threads in a Thread Block can cooperate with each other by:
  – Synchronizing their execution
  – Efficiently sharing data through a low latency shared memory.
• Threads from different blocks cannot cooperate.

The threads in a block each have Thread ID numbers within the Block. Your CUDA program will use these Thread IDs to select work to do and pull the right data from memory. Threads share data and synchronize while doing their share of the work. Every 32 threads constitute a “Warp.” Each thread in a Warp simultaneously executes the same instruction on different pieces of data. But, it is likely that a Warp’s execution will need to stop at some point, waiting for a memory access. This would make the execution go idle – bad! So, it is worthwhile to have multiple Warps worth of threads available so that when one Warp blocks, another Warp can be swapped in. The threads in a Thread Block can cooperate with each other by: Synchronizing their execution and Efficiently sharing data through a low latency shared memory. Threads from different blocks cannot cooperate.

Threads can access various types of storage:

• Each thread has access to:
  – Its own R/W per-thread registers
  – Its own R/W per-thread private memory
• Each thread has access to:
  – Its block’s R/W per-block shared memory
• Each thread has access to:
  – The entire R/W per-grid global memory
  – The entire read-only per-grid constant memory
  – The entire read-only per-grid texture memory
• The CPU can read and write global and constant memories.

Different types of CUDA memory:

<table>
<thead>
<tr>
<th>Memory</th>
<th>Location</th>
<th>Who Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registers</td>
<td>On-chip</td>
<td>One thread</td>
</tr>
<tr>
<td>Private</td>
<td>On-chip</td>
<td>One thread</td>
</tr>
<tr>
<td>Shared</td>
<td>On-chip</td>
<td>All threads in that block</td>
</tr>
<tr>
<td>Global</td>
<td>Off-chip</td>
<td>All threads + Host</td>
</tr>
<tr>
<td>Constant</td>
<td>Off-chip</td>
<td>All threads + Host</td>
</tr>
</tbody>
</table>
Thread Rules

- Each Thread has its own registers and private memory
- Each Block can use at most some maximum number of registers, divided equally among all Threads
- Threads can share local memory with the other Threads in the same Block
- Threads can synchronize with other Threads in the same Block
- Global and Constant memory is accessible by all Threads in all Blocks
- 192 or 256 are good numbers of Threads per Block (multiples of the Warp size)

A CUDA Thread can Query where it Fits in its “Community” of Threads and Blocks

- dim3 gridDim;
  - Dimensions of the blocks in this grid
- dim3 blockIdx;
  - This block’s indexes within this grid
- dim3 blockDim;
  - Dimensions of the threads in this block
- dim3 threadIdx;
  - This thread’s indexes within the block

Note: It is as if dim3 is defined as:
typedef int[3] dim3;
(it’s not really – it is actually defined within the CUDA compiler)

A CUDA Thread needs to know where it Lives in its “Community” of Threads and Blocks

For a 1D problem:
int blockThreads = blockIdx.x*blockDim.x;
int gid = blockThreads + threadIdx.x;
C[gid] = A[gid]*B[gid];

For a 2D problem:
int blockNum = blockIdx.y*gridDim.x + blockIdx.x;
int blockThreads = blockNum*blockDim.x*blockDim.y;
int gid = blockThreads + threadIdx.y*blockDim.x + threadIdx.x;
C[gid] = A[gid]*B[gid];

Types of CUDA Functions

<table>
<thead>
<tr>
<th>Function Type</th>
<th>Executed on the:</th>
<th>Only callable from the:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>device</strong></td>
<td>GPU</td>
<td>GPU</td>
</tr>
<tr>
<td><strong>global</strong></td>
<td>GPU</td>
<td>Host</td>
</tr>
<tr>
<td><strong>host</strong></td>
<td>Host</td>
<td>Host</td>
</tr>
</tbody>
</table>

__global__ defines a kernel function – it must return void

Note: “__” is 2 underscore characters
The C/C++ Program Calls a CUDA Kernel using a Special `<<<>>>` Syntax

These are called "chevrons"

```
KernelFunction<<< NumBlocks, NumThreadsPerBlock >>>( arg1, arg2, ... ) ;
```

These are called "chevrons"

```
KernelFunction<<< NumBlocks, NumThreadsPerBlock >>>( arg1, arg2, ... ) ;
```

Note that this is just like calling the C/C++ function:
```
KernelFunction( arg1, arg2, ... ) ;
```
except that we have designated it to run on the GPU
with a particular block/thread configuration.

One of my own Experiments with Number of Threads Per Block

```
KernelFunction<<< NumBlocks, NumThreadsPerBlock >>>( arg1, arg2, ... ) ;
```

Performance

Dataset Size

Number of Threads per Block

```
NumBlocks = DataSetSize / NumThreadsPerBlock
```

Getting CUDA Programs to Run under Linux

This is the path where the CUDA tools are loaded on our Oregon State University systems.
```
CUDA_PATH          =       /usr/local/apps/cuda/cuda-10.1
CUDA_BIN_PATH      =       $(CUDA_PATH)/bin
CUDA_NVCC          =       $(CUDA_BIN_PATH)/nvcc
arrayMul:           =       arrayMul.cu
$(CUDA_NVCC) -o arrayMul arrayMul.cu
```

Or, without the Makefile syntax:
```
/usr/local/apps/cuda/cuda-10.1/bin/nvcc -o arrayMul arrayMul.cu
```

We also have the CUDA-11 and CUDA-12 tools loaded for your use. You can use them if you want. But, given the wide breadth of different Nvidia cards around campus, CUDA-10 seems to be the one that will run everywhere! I recommend you use it.

CUDA-10

CUDA-11

CUDA-12

arrayMul:           =       arrayMul.cu
$(CUDA_NVCC) -o arrayMul arrayMul.cu
```

We also have the CUDA-11 and CUDA-12 tools loaded for your use. You can use them if you want. But, given the wide breadth of different Nvidia cards around campus, CUDA-10 seems to be the one that will run everywhere! I recommend you use it.
1. Install Visual Studio if you haven’t already. If you are an OSU student, go to: 
https://azureforeducation.microsoft.com/devtools
Click the blue Sign In button on the right. Login using your ONID@oregonstate.edu username and password.

2. Install the CUDA toolkit. It is available here:

From the main screen, click File → New → Project…

Then, in this templates box, type: CUDA

After a few seconds, you will then see this. Click Next.
1. Navigate to the folder you want to contain this project folder.

2. Give the name you want for the folder and project.

3. Leave this box checked.

4. Click Create.

Getting CUDA Programs to Run under Visual Studio

Using CUDA and OpenMP Together

This is the Makefile we use on Linux:

```bash
CUDA_PATH = /usr/local/apps/cuda/cuda-10.1
CUDA_BIN_PATH = $(CUDA_PATH)/bin
CUDA_NVCC = $(CUDA_BIN_PATH)/nvcc
arrayMul: arrayMul.out

CUDA_NVCC -DarrayMul -c arrayMul.cu
```

Or, on Linux, but without the Makefile syntax:

```bash
/usr/local/apps/cuda/cuda-10.1/bin/nvcc -DarrayMul -c arrayMul.cu
```

Or, in Visual Studio:

1. Go to the Project menu ➔ Project Properties
2. Change the setting Configuration Properties ➔ C/C++ ➔ Language ➔ OpenMP Support to "Yes (/openmp)"

We also have the CUDA-11 and CUDA-12 tools loaded for your use. You can use them if you want. But, given the wide breadth of different Nvidia cards around campus, CUDA-10 seems to be the one that will run everywhere! I recommend you use it.
int deviceCount;
cudaGetDeviceCount(&deviceCount);

... 

int device; // 0 ≤ device ≤ deviceCount - 1
cudaSetDevice(device);