Nvidia's Compute Unified Device Architecture (CUDA)

Oregon State University

Mike Bailey

mjb@cs.oregonstate.edu

Oregon State University
Computer Graphics

The CUDA Paradigm

C/C++ Program with both host and CUDA code in it

Host code

CUDA code

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.

C/C++ Compiler and Linker

CUDA Compiler and Linker

CPU binary on the Host

CUDA binary on the Device

1. Run CPU code

2. Send data to GPU
3. Run GPU kernel

4. Get data back from GPU

5. Run CPU code

6. Send data to GPU
7. Run GPU kernel

8. Get data back from GPU

9. Run CPU code

Oregon State
University
Computer Graphics

mjb – March 18, 2023

CUDA wants you to break the problem up into Pieces

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

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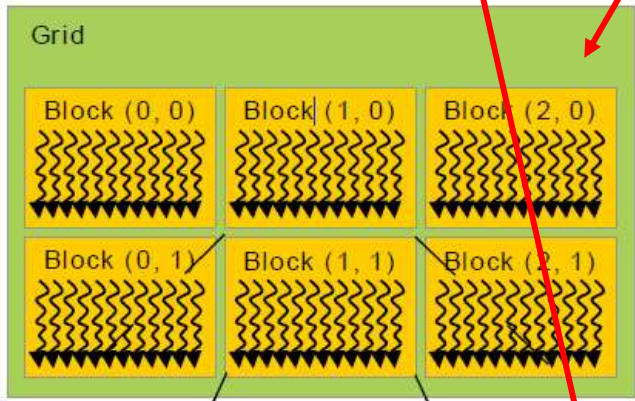
```
__global__
void
ArrayMult( float *dA, float *dB, float *dC )
{
          int gid = blockIdx.x*blockDim.x + threadIdx.x;
          dC[gid] = dA[gid] * dB[gid];

}
```

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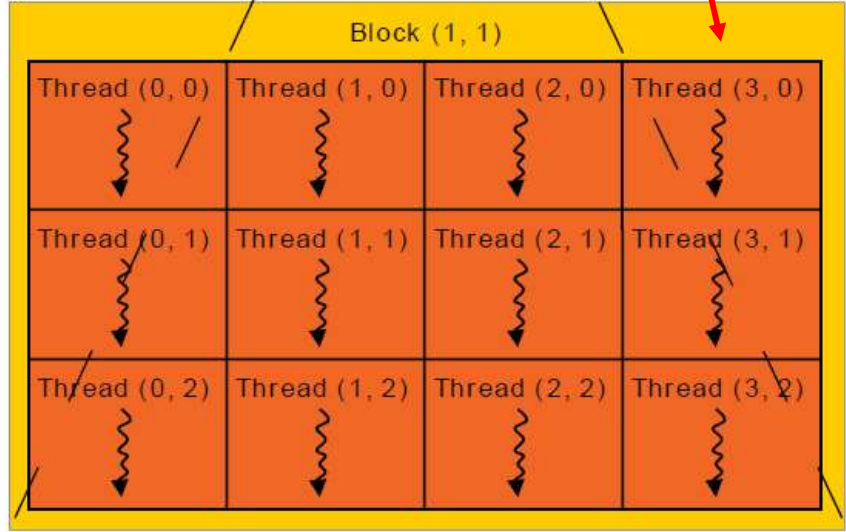

Organization: Blocks are Arranged in Grids

- The GPU's workload is divided into a **Grid of Blocks**
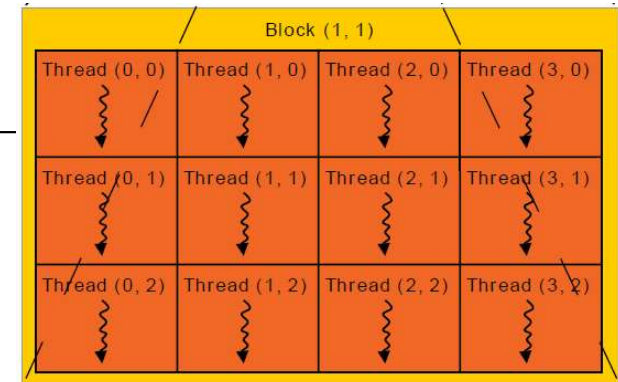- Each Block's workload is divided into a **Grid of Threads**

Grid of Blocks

Grid of Threads

Oregon State
University
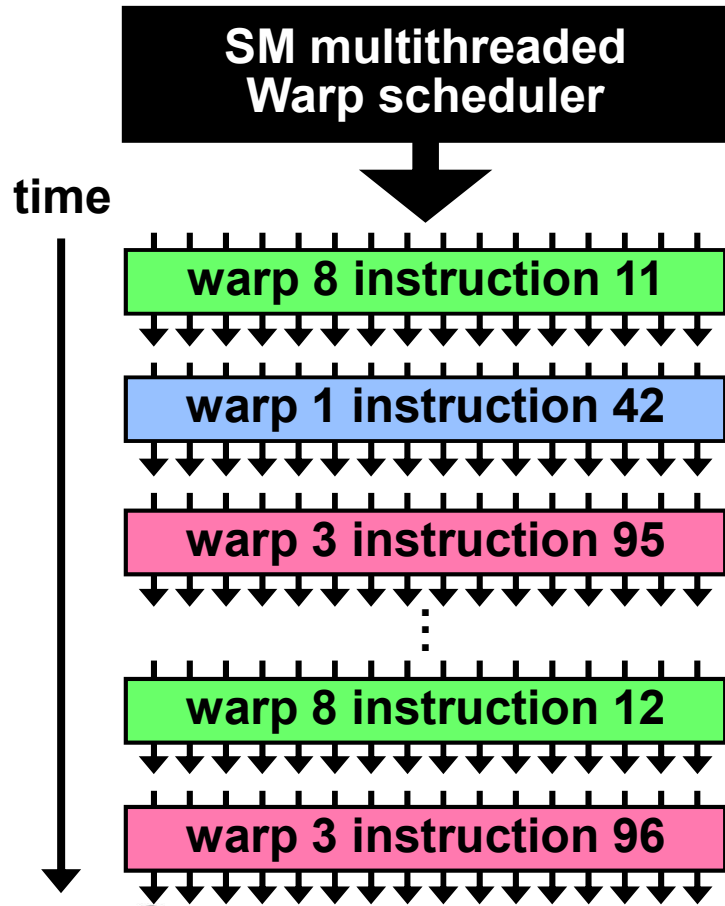Computer Graphics

mjb – March 18, 2023

A Block is made up of 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

University
Computer Graphics

Scheduling

SM multithreaded Warp scheduler

time

| warp 8 instruction 11 |
| warp 1 instruction 42 |
| warp 3 instruction 95 |
| warp 8 instruction 12 |
| warp 3 instruction 96 |

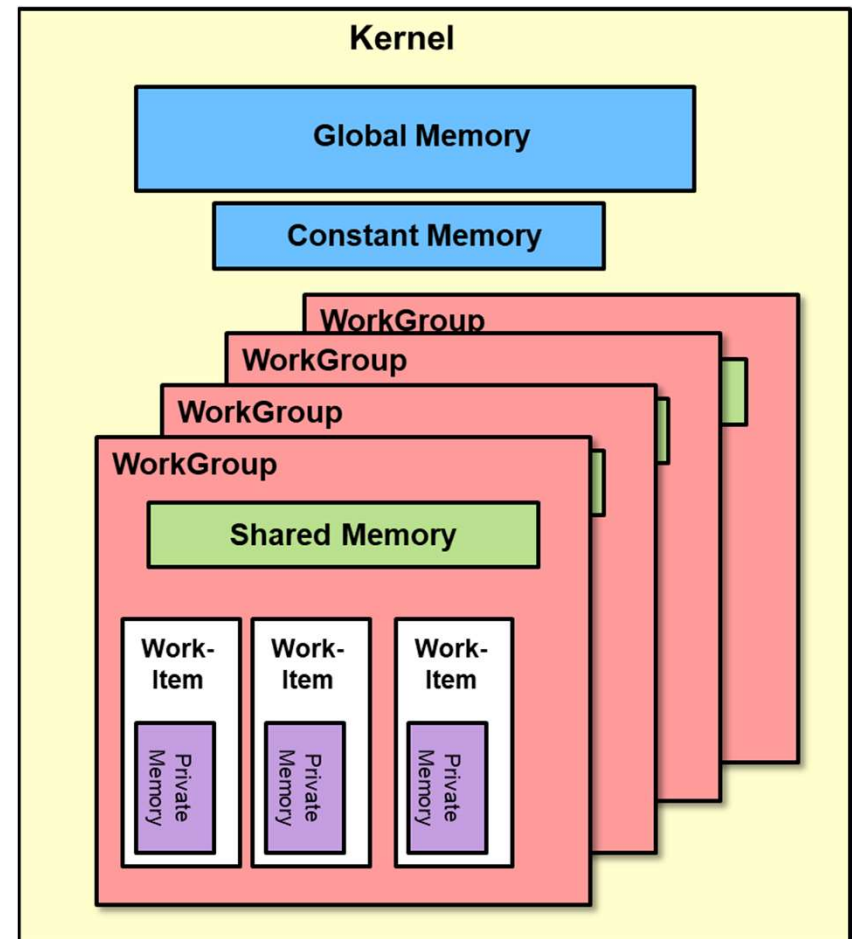- The hardware implements low-overhead Warp switching

  - A Warp whose next instruction has operands ready for consumption is eligible to be executed.

  - All threads in one Warp execute the same instruction at any given time, but on different data.

  - Threads in different Warps will usually be executing different instructions at any given time

This tells you that there needs to be a bunch of Warps to work on so that something is always ready to run
If you can help it, these should be multiples of 32.

Oregon State University
Computer Graphics

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

Oregon State
University
Computer Graphics

Different Types of CUDA Memory

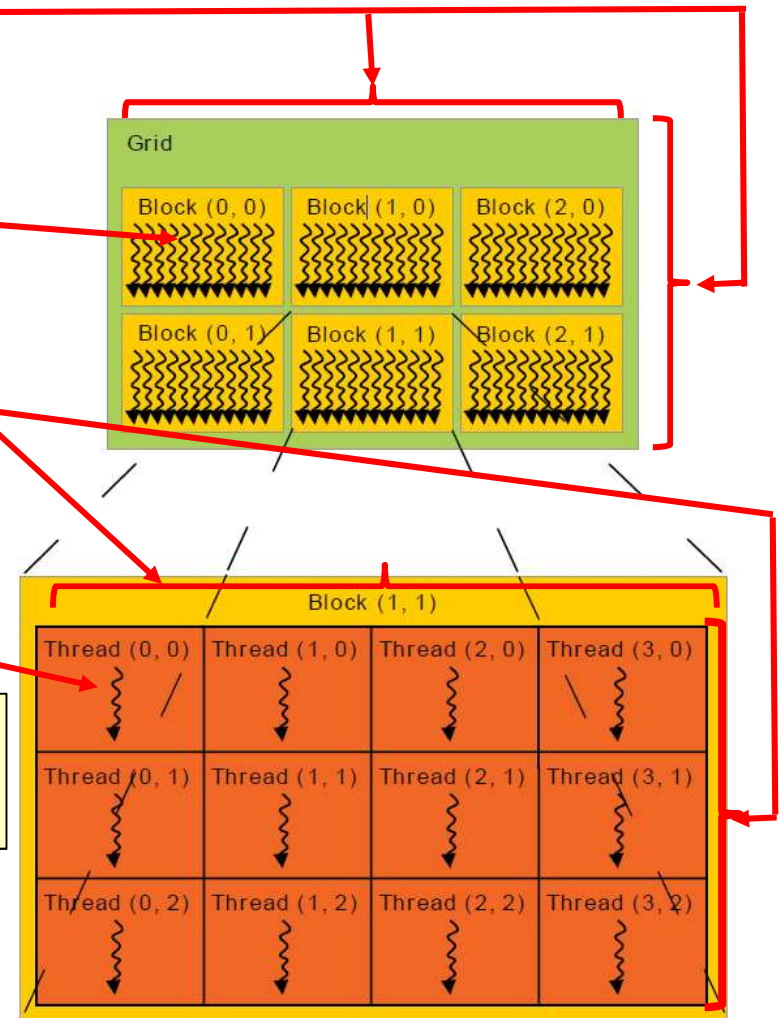| Memory	Location	Who Uses
Registers	On-chip	One thread
Private	On-chip	One thread
Shared	On-chip	All threads in that block
Global	Off-chip	All threads + Host
Constant	Off-chip	All threads + Host

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)

Oregon State
University
Computer Graphics

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;**
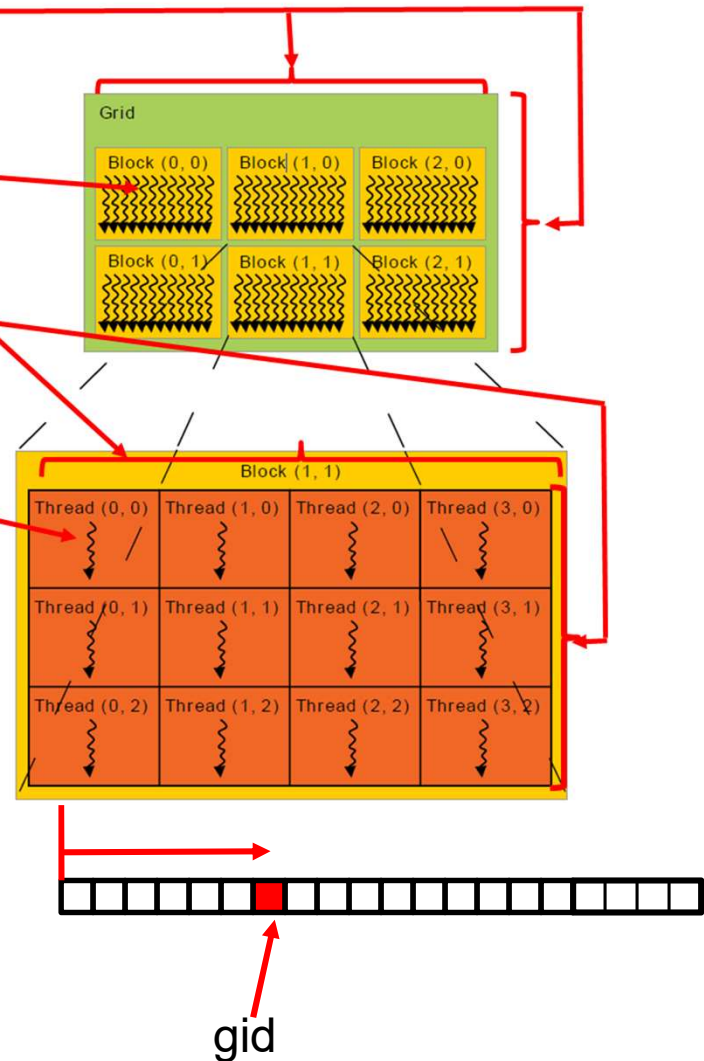
  – This thread's indexes within the block

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)

Oregon State
University
Computer Graphics

A CUDA Thread needs to know where it Lives 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

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];

gid

Computer Graphics

Types of CUDA Functions

	Executed on the:	Only callable from the:
__device__ float DeviceFunc()	GPU	GPU
__global__ void KernelFunc()	GPU	Host
__host__ float HostFunc()	Host	Host

__global__ defines a kernel function – it must return **void**
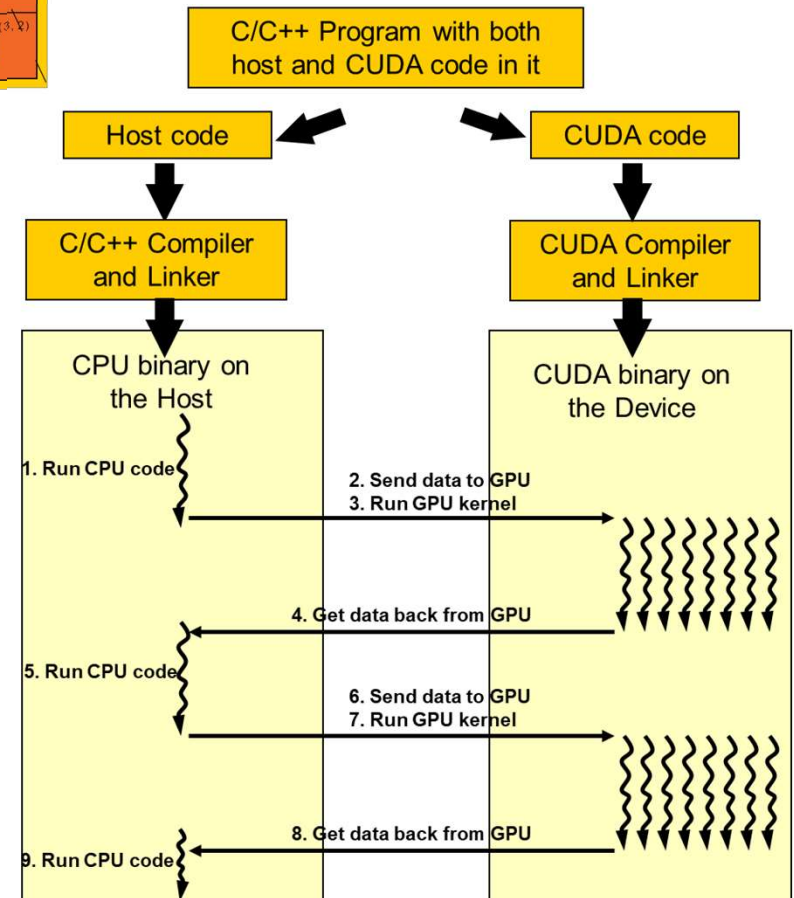
Note: "__" is *2 underscore characters*

Oregon State
University
Computer Graphics

The C/C++ Program Calls a CUDA Kernel using a Special <<<…>>> Syntax

dim3   **dim3**

These are called "chevrons"

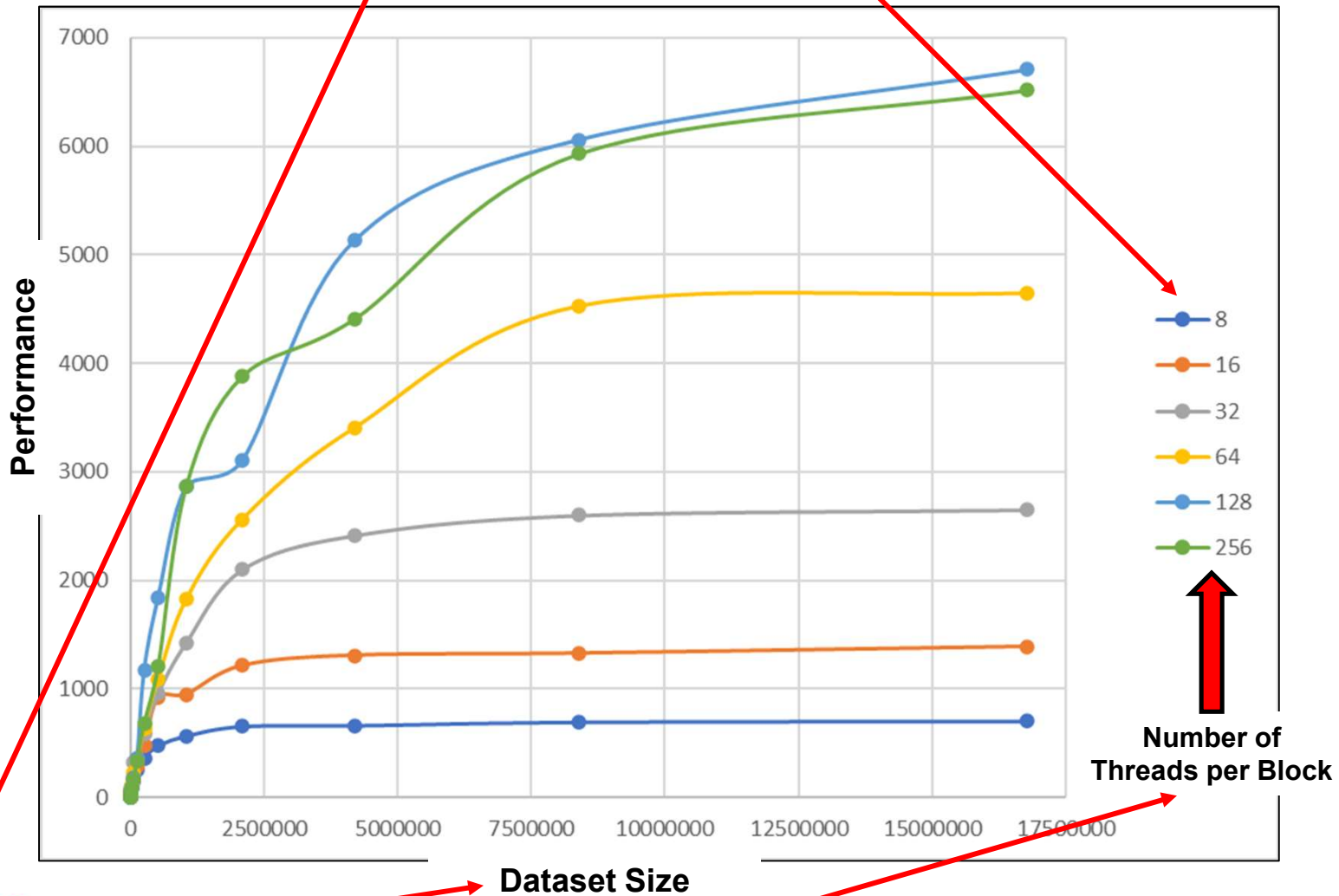**KernelFunction<<< NumBlocks, NumThreadsPerBlock >>>(arg1, arg2, …) ;**

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.

Oregon State
University
Computer Graphics

mjb – March 18, 2023

One of my own Experiments with Number of Threads Per Block

KernelFunction<<< NumBlocks , NumThreadsPerBlock >>>(arg1, arg2, …) ;

Performance

Dataset Size

Number of Threads per Block

Legend: 8, 16, 32, 64, 128, 256

NumBlocks = DataSetSize / NumThreadsPerBlock

Computer Graphics

One of my own Experiments with Number of Threads Per Block

KernelFunction<<< NumBlocks , NumThreadsPerBlock >>>(arg1, arg2, …) ;

Performance (y-axis)

Number of Threads per Block (x-axis)

Legend (Dataset Size):
- 1024
- 2048
- 4096
- 8192
- 16384
- 32768
- 65536
- 131072
- 262144
- 524288
- 1048576
- 2097152
- 4194304
- 8388608
- 16777216

Dataset Size

NumBlocks = DataSetSize / NumThreadsPerBlock

Oregon State University

mjb – March 18, 2023

Getting CUDA Programs to Run under Linux

This is the Makefile we use:

```
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
```

This is the path where the CUDA tools are loaded on our Oregon State University systems.

Or, without the Makefile syntax:

```
/usr/local/apps/cuda/cuda-10.1/bin/nvcc   -o   arrayMul  arrayMul.cu
```

Oregon State University
Computer Graphics

We also have the CUDA-11 and CUDA-12 tools loaded for your use. You can use them if you want. Bur, 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:
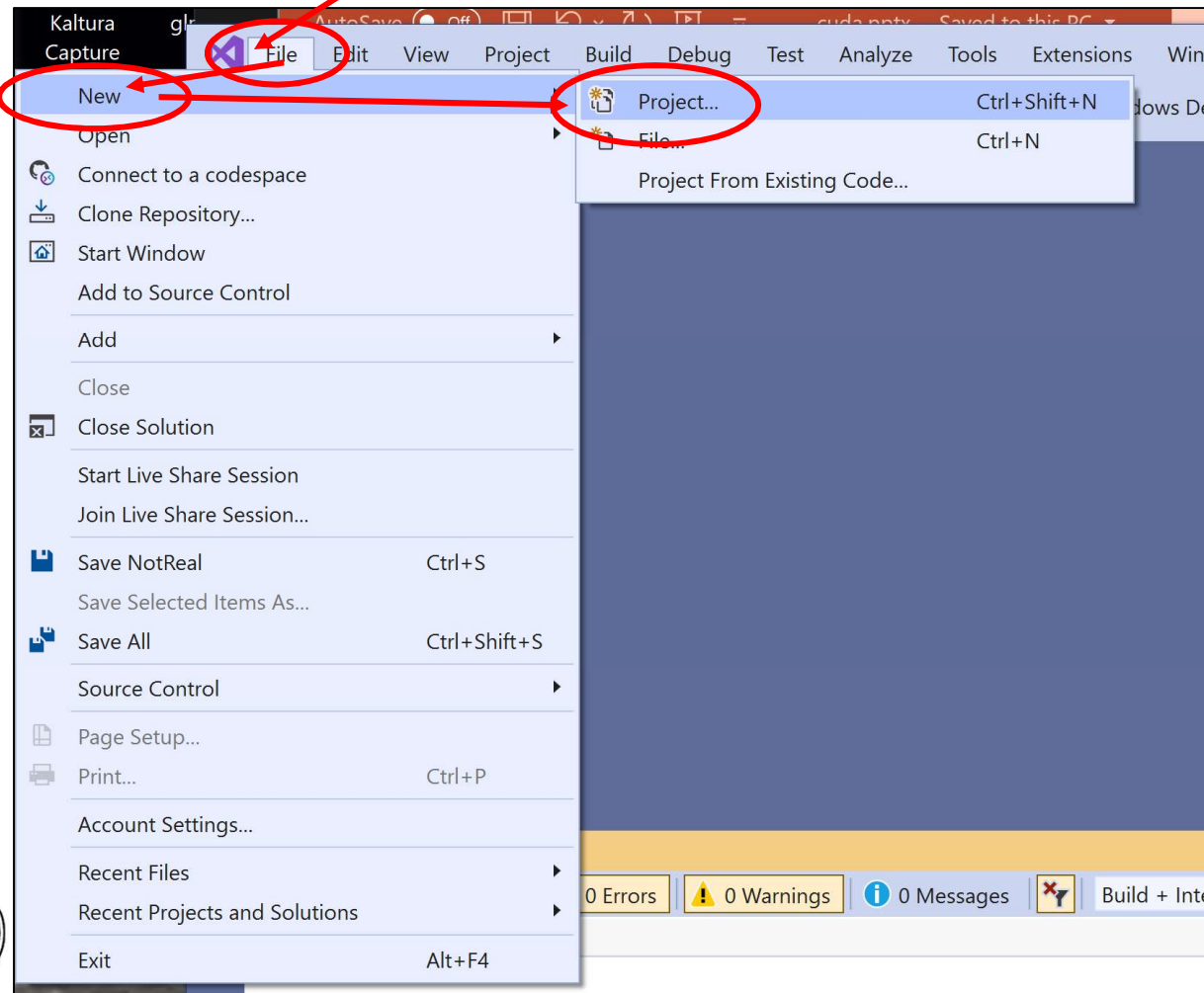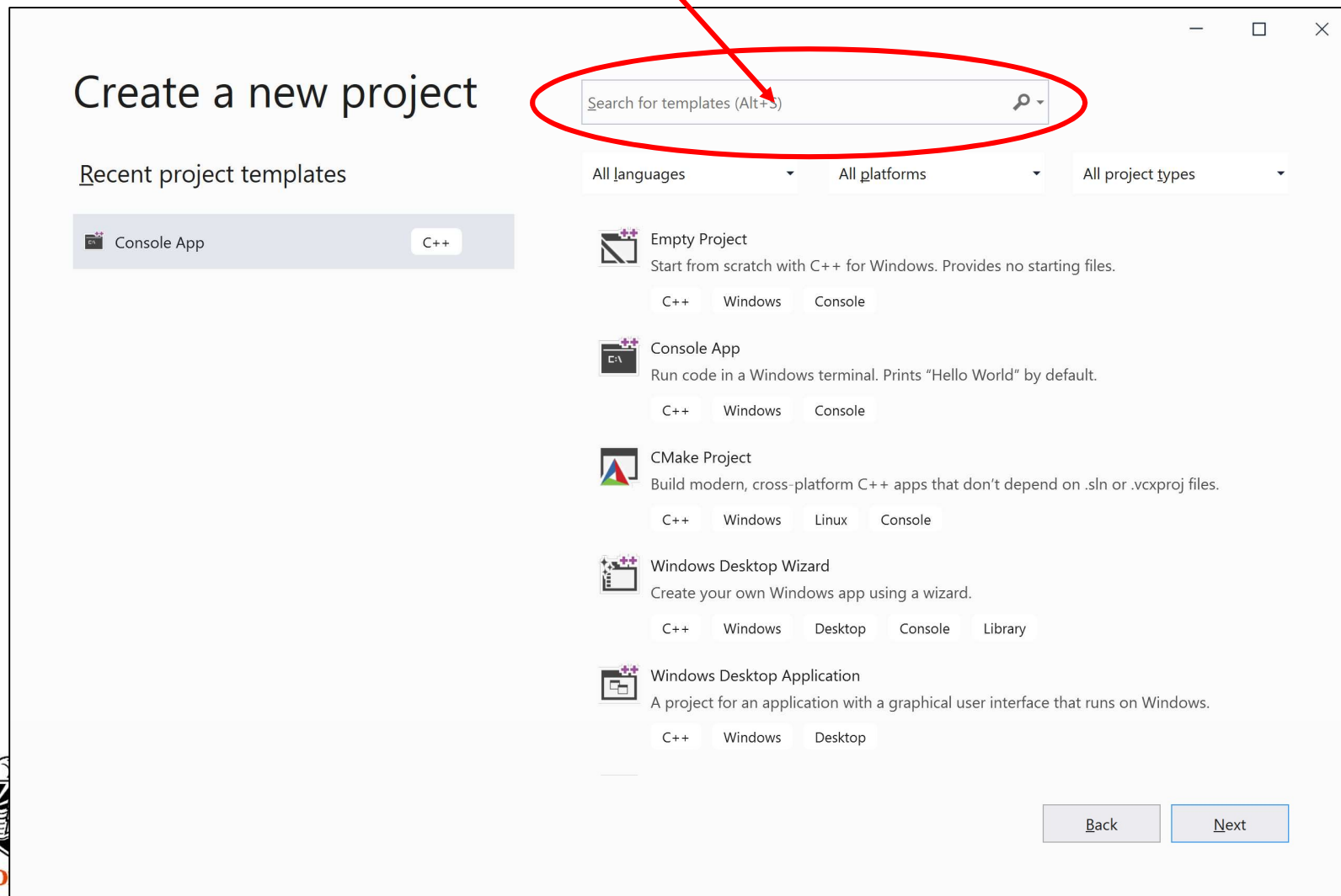
https://developer.nvidia.com/cuda-downloads

Oregon State
University
Computer Graphics

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**.

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

Configure your new project

CUDA 11.3 Runtime   C++   CUDA   Windows   Linux   Cloud   Console   DataScience   Desktop   Machine Learning

Project name

CudaRuntime1

Location

C:\Users\Mike Bailey\source\repos

Solution name

CudaRuntime1

☑ Place solution and project in the same directory

4. Click **Create**

3. Leave this box checked.

Back   Create

1. Navigate to the folder you want to contain this project folder.

University
Computer Graphics

1. Visual Studio then "writes" a program for you. It has both CUDA and C++ code in it. Its structure looks just like our notes' examples.
2. You can click **Build** → **Build** to compile it, both the C++ and the CUDA code.
3. You can click **Debug** → **Start Without Debugging** to run it.
4. You can then either modify this file, or clear it and paste your own code in.

Getting CUDA Programs to Run under Visual Studio

Computer Graphics

Using CUDA and OpenMP Together

This is the Makefile we use on Linux:

```
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 -Xcompiler -fopenmp
```

Or, on Linux, but without the Makefile syntax:

```
/usr/local/apps/cuda/cuda-10.1/bin/nvcc  -o  arrayMul  arrayMul.cu  -Xcompiler  -fopenmp
```

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)"**

Oregon State University
Computer Graphics

We also have the CUDA-11 and CUDA-12 tools loaded for your use. You can use them if you want. Bur, 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.

Using Multiple GPU Cards with CUDA

```
int  deviceCount;
cudaGetDeviceCount(  &deviceCount  );


. . .


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

Oregon State
University
Computer Graphics