CENG443
Heterogeneous Parallel Programming

Introduction to CUDA

Işıl ÖZ, IZTECH, Fall 2023

20 October 2023

3 Ways to Accelerate Applications

Libraries	Compiler Directives	Programming Languages
Easy to use Most Performance	Easy to use Portable code	Most Performance Most Flexibility

Libraries

Ease of use: Using libraries enables GPU acceleration without in-depth knowledge of GPU programming

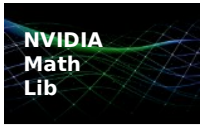
“Drop-in”: Many GPU-accelerated libraries follow standard APIs, thus enabling acceleration with minimal code changes

Quality: Libraries offer high-quality implementations of functions encountered in a broad range of applications

GPU Accelerated Libraries

Linear Algebra
FFT, BLAS,
SPARSE, Matrix

Numerical & Math
RAND, Statistics



Data Struct. & AI
Sort, Scan, Zero Sum

Visual Processing
Image & Video

NVIDIA
Video
Encode

Deep Neural Network
cuDNN

Vector Addition in Thrust

```
#include <thrust/device_vector.h>
#include <thrust/copy.h>

int main(void) {
  size_t inputLength = 500;
  thrust::host_vector<float> hostInput1(inputLength);
  thrust::host_vector<float> hostInput2(inputLength);
  thrust::device_vector<float> deviceInput1(inputLength);
  thrust::device_vector<float> deviceInput2(inputLength);
  thrust::device_vector<float> deviceOutput(inputLength);

  thrust::copy(hostInput1.begin(), hostInput1.end(), deviceInput1.begin());
  thrust::copy(hostInput2.begin(), hostInput2.end(), deviceInput2.begin());

  thrust::transform(deviceInput1.begin(), deviceInput1.end(), deviceInput2.begin(), deviceOutput.begin(),
                    thrust::plus<float>());

thrust::copy(deviceOutput.begin(), deviceOutput.end(),
       std::ostream_iterator<int>(std::cout, "\n"));
}
```

Compiler Directives

Ease of use: Compiler takes care of details of parallelism management and data movement

Portable: The code is generic, not specific to any type of hardware and can be deployed into multiple languages

Uncertain: Performance of code can vary across compiler versions

OpenACC

Compiler directives for C, C++, and FORTRAN

```
#pragma acc parallel loop copyin(input1[0:inputLength],input2[0:inputLength]),
copyout(output[0:inputLength])
    for(i = 0; i < inputLength; ++i) {
        output[i] = input1[i] + input2[i];
    }
```

Programming Languages

Performance: Programmer has best control of parallelism and data movement

Flexible: The computation does not need to fit into a limited set of library patterns or directive types

Verbose: The programmer often needs to express more details

GPU Programming Languages

Numerical analytics	MATLAB Mathematica, LabVIEW
Fortran	CUDA Fortran
C	CUDA C
C++	CUDA C++
Python	PyCUDA, Copperhead, Numba
F#	Alea.cuBase

CUDA C

Easy to use
Most Performance

Easy to use
Portable code

Most Performance
Most Flexibility

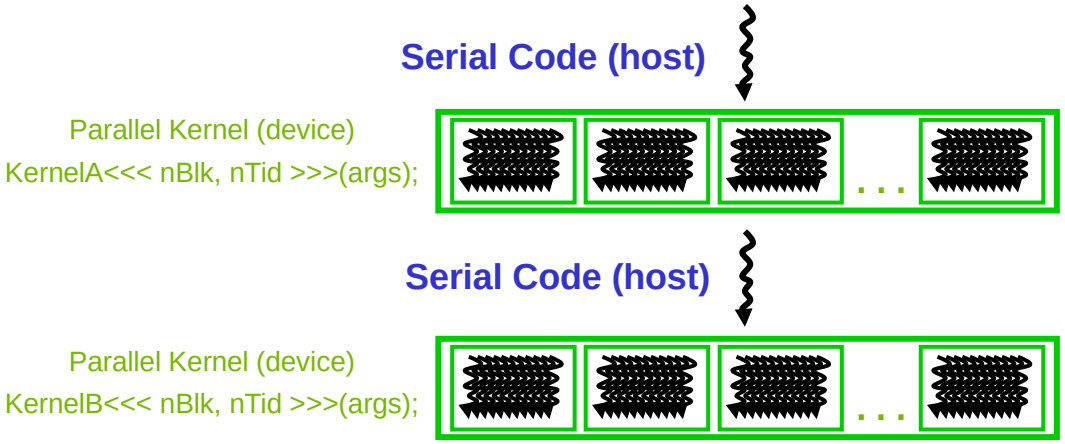
CUDA Programming

CUDA (Compute Unified Device Architecture)

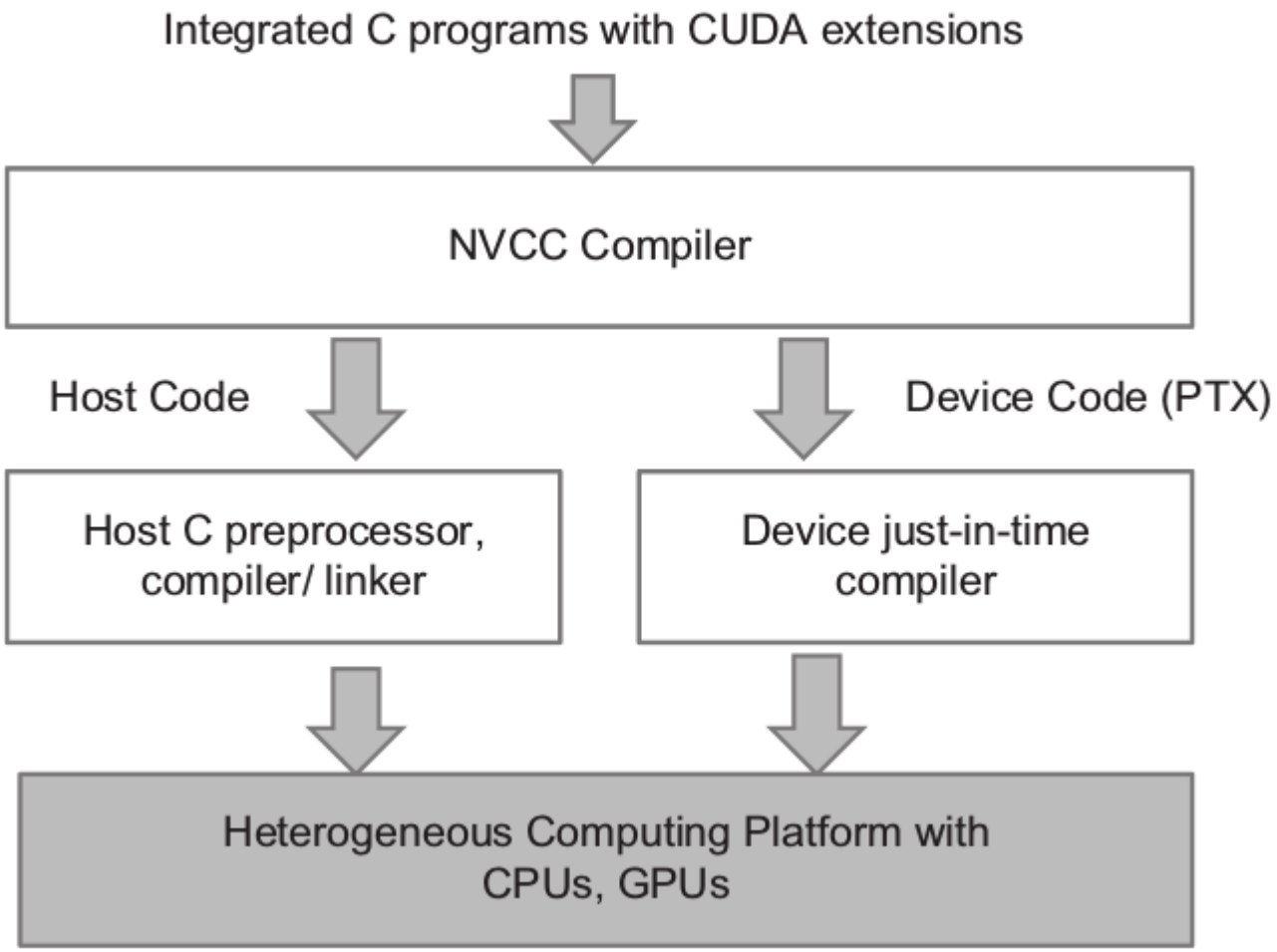
Integrated host+device app C program

Serial or modestly parallel parts in host C code

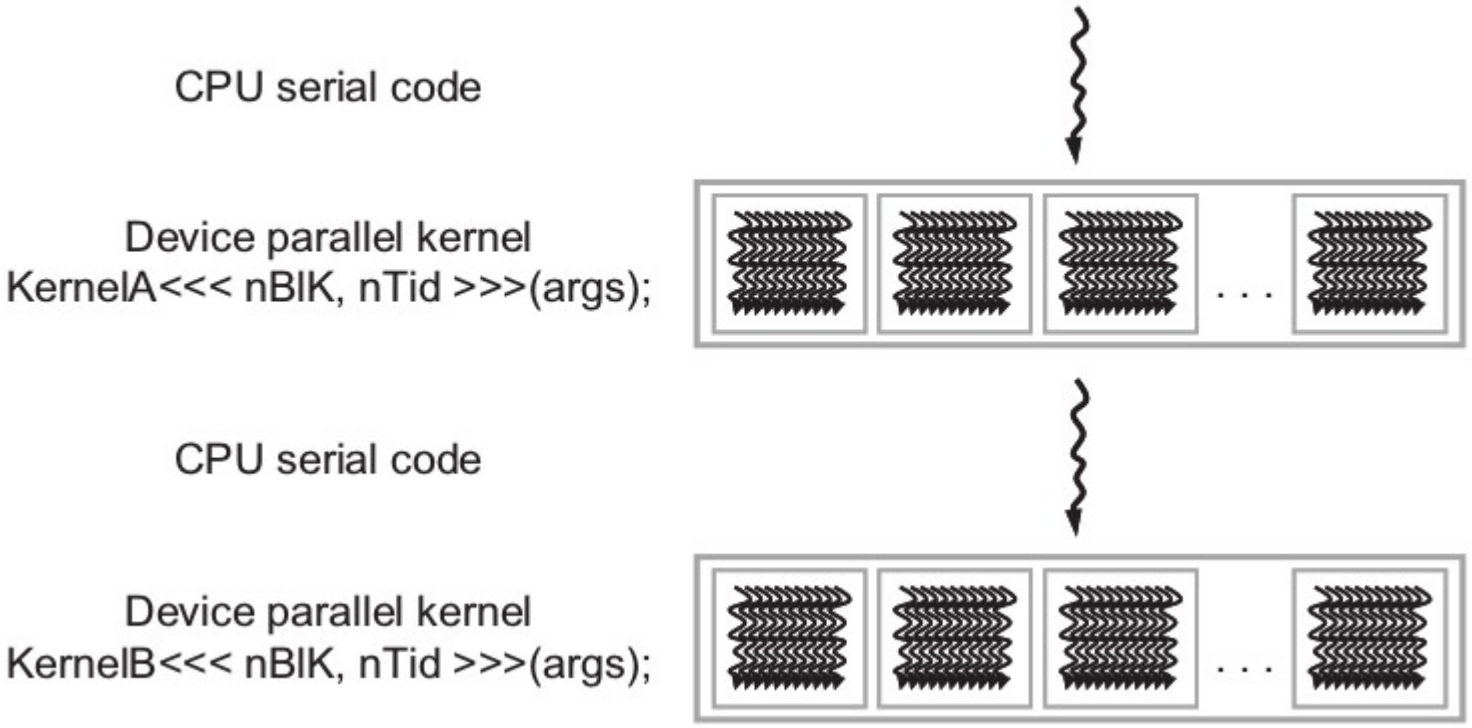
Highly parallel parts in device SPMD kernel C code



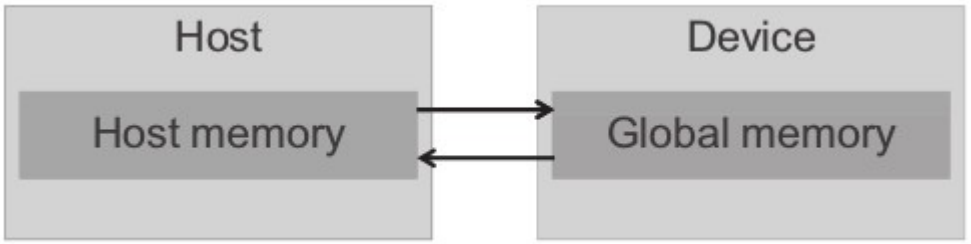
CUDA C Program Compilation

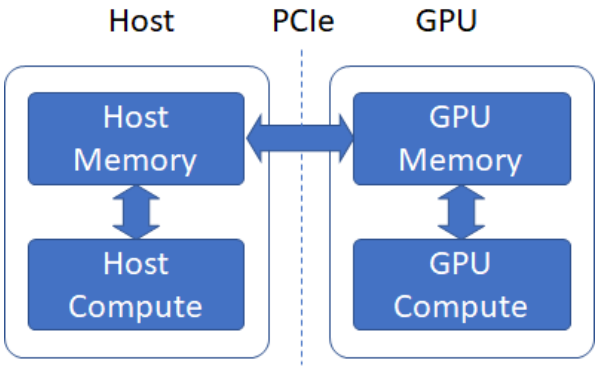


CUDA C Program Execution

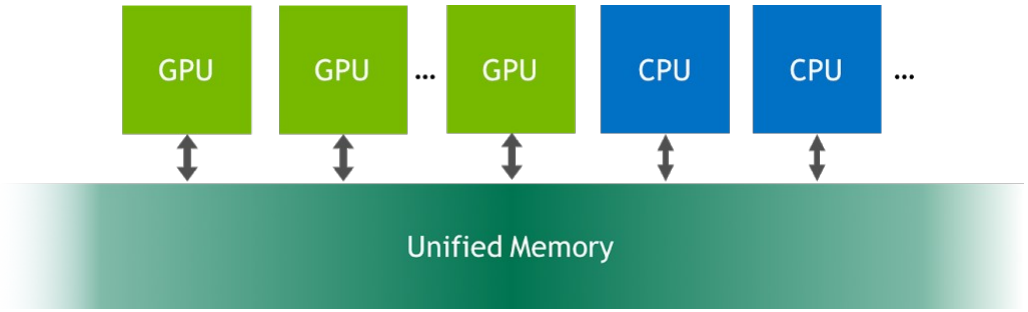


Host (CPU) Memory and Device (GPU) Memory

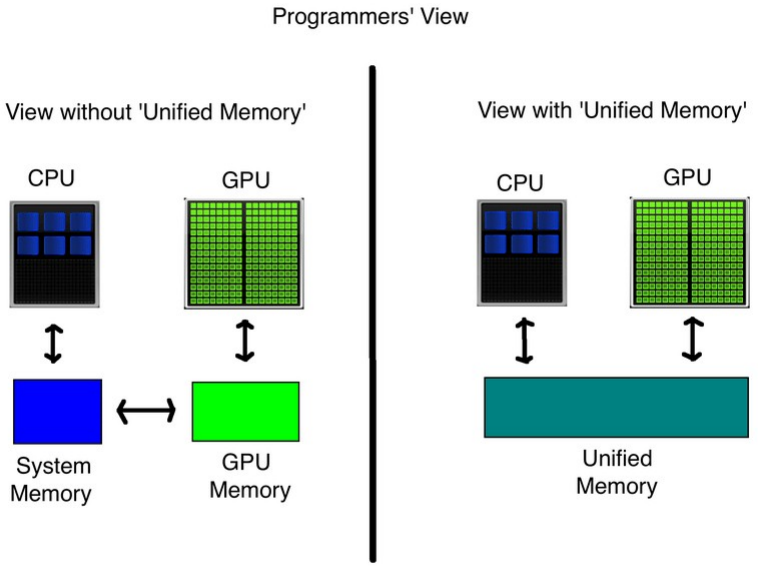




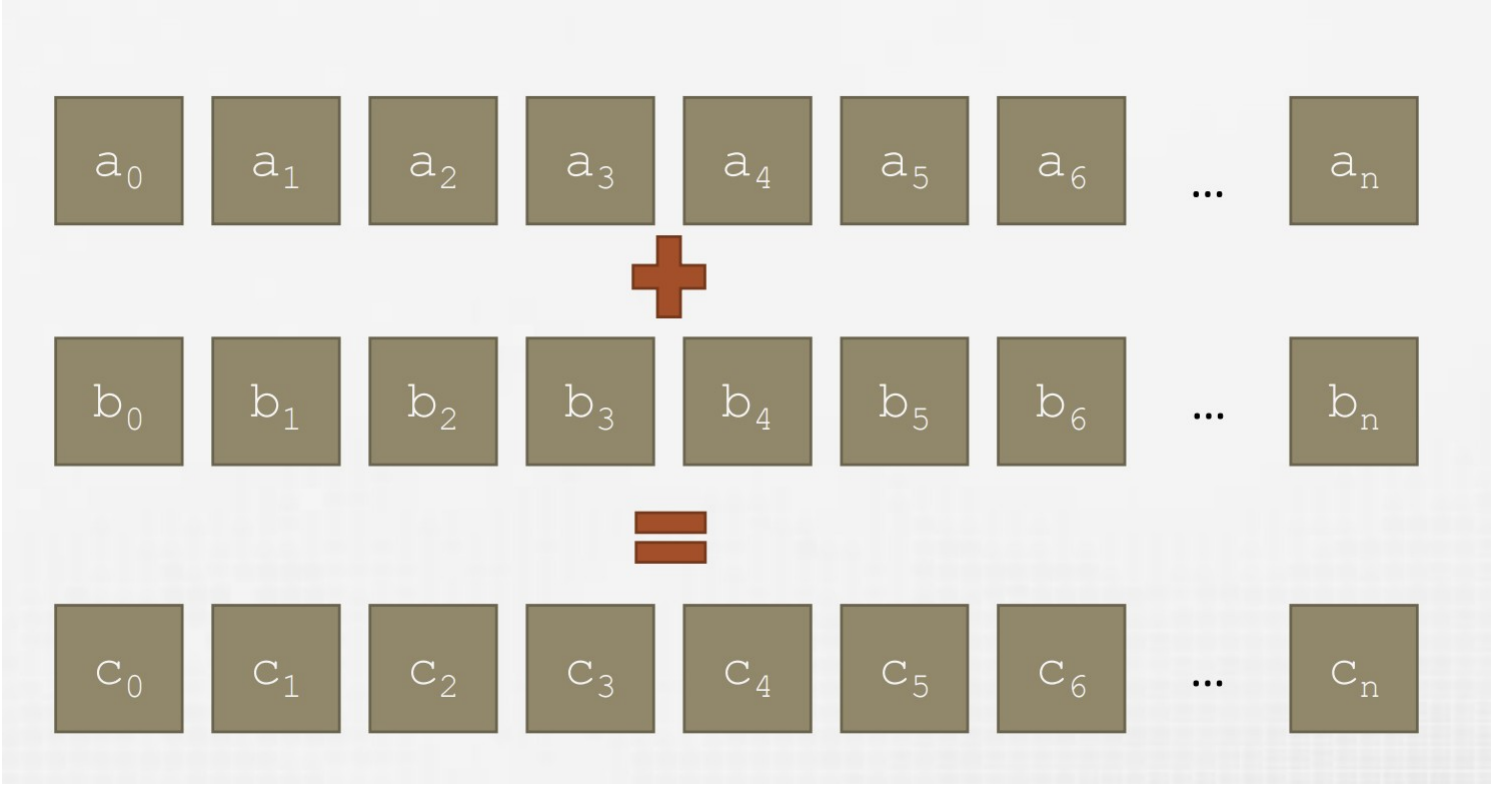
Unified Memory



Programmer’s Memory View



Vector Addition



Vector Addition – Traditional C Code

```
// Compute vector sum C = A + B
void vecAdd(float *h_A, float *h_B, float *h_C, int n)
{
    int i;
    for (i = 0; i<n; i++) h_C[i] = h_A[i] + h_B[i];
}

int main()
{
    // Memory allocation for h_A, h_B, and h_C
  // I/O to read h_A and h_B, N elements
  …
    vecAdd(h_A, h_B, h_C, N);
}
```

Heterogeneous Computing

Setup inputs on the host (CPU-accessible memory)

Allocate memory for outputs on the host

Allocate memory for inputs on the GPU

Allocate memory for outputs on the GPU

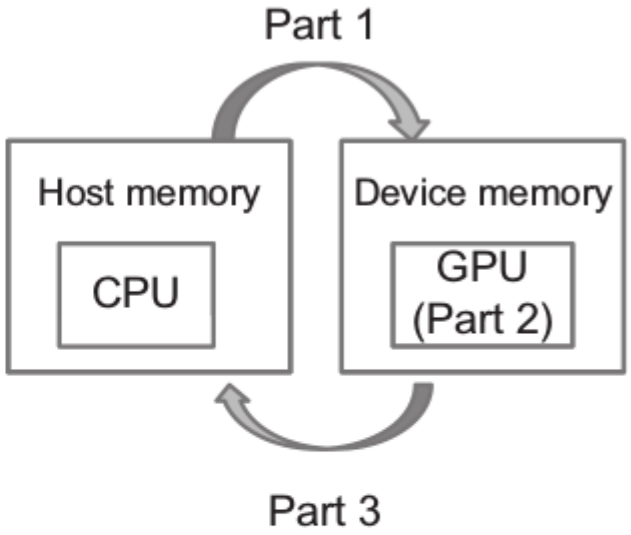
Copy inputs from host to GPU

Start GPU kernel (function that executed on gpu)

Copy output from GPU to host

vecAdd CUDA Host Code

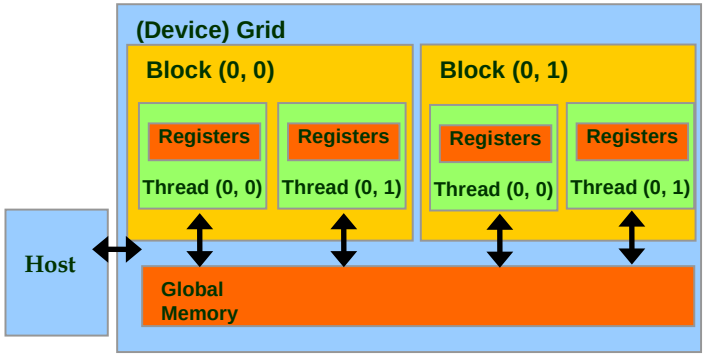
```
#include <cuda.h>
…
void vecAdd(float* A, float* B, float* C, int n)
{
   int  size = n* sizeof(float);
   float  *d_A *d_B, *d_C;
   …
1. // Allocate device memory for A, B, and C
    // copy A and B to device memory

2. // Kernel launch code – to have the device
    // to perform the actual vector addition

3. // copy C from the device memory
    // Free device vectors
}
```



Partial Overview of CUDA Memories

Device code can:

- R/W per-thread registers
- R/W shared/global memory



Host code can:

- Transfer data to/from global memory

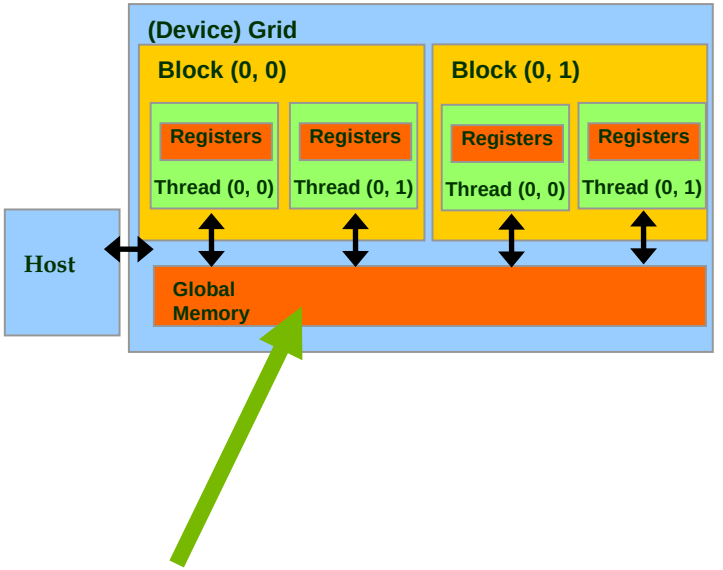
https://docs.nvidia.com/cuda/cuda-runtime-api/group__CUDART__MEMORY.html

CUDA Device Memory API functions

cudaMalloc()

- Allocates an object in the device global memory
- Two parameters
  - Address of a pointer to the allocated object
  - Size of allocated object in terms of bytes

```
cudaError_t cudaMalloc (void ** devPtr,
                        size_t    size)
```

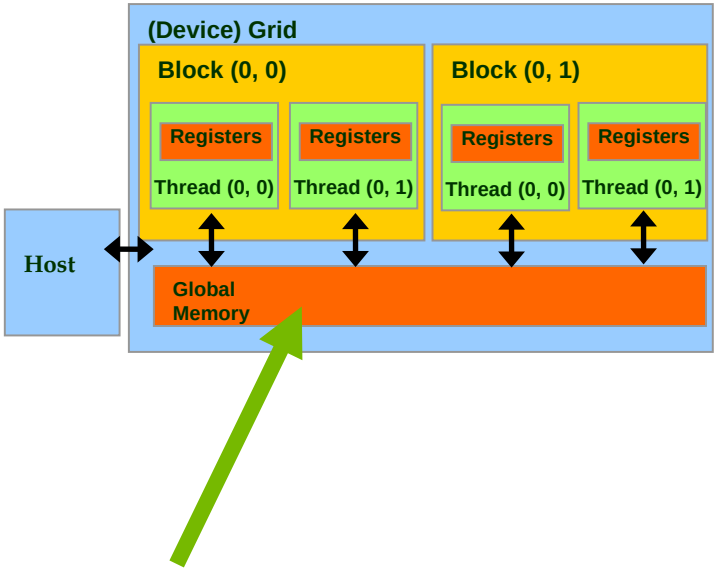


CUDA Device Memory API functions

cudaFree()

- Frees object from device global memory
- One parameter
  - Pointer to freed object

```
cudaError_t cudaFree(void* devPtr)
```



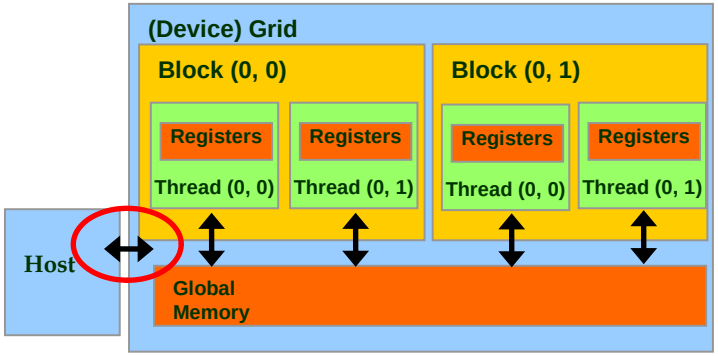
Vector Addition Host Code

```
void vecAdd(float *h_A, float *h_B, float *h_C, int n)
{
    int size = n * sizeof(float); float *d_A, *d_B, *d_C;

    cudaMalloc((void **) &d_A, size);
    //transfer data to device
     cudaMalloc((void **) &d_B, size);
    //transfer data to device
     cudaMalloc((void **) &d_C, size);

     // Kernel invocation code – to be shown later

    //get data from device
     cudaFree(d_A); cudaFree(d_B); cudaFree (d_C);
}
```

Host-Device Data Transfer API functions

cudaMemcpy

- Memory data transfer
- Requires four parameters
  - Pointer to destination
  - Pointer to source
  - Number of bytes copied
  - Type/Direction of transfer




```
cudaError_t cudaMemcpy (void *   dst,
                        const void * src,
                        size_t    count,
                  enum cudaMemcpyKind  kind    )
```

Host-Device Data Transfer API functions

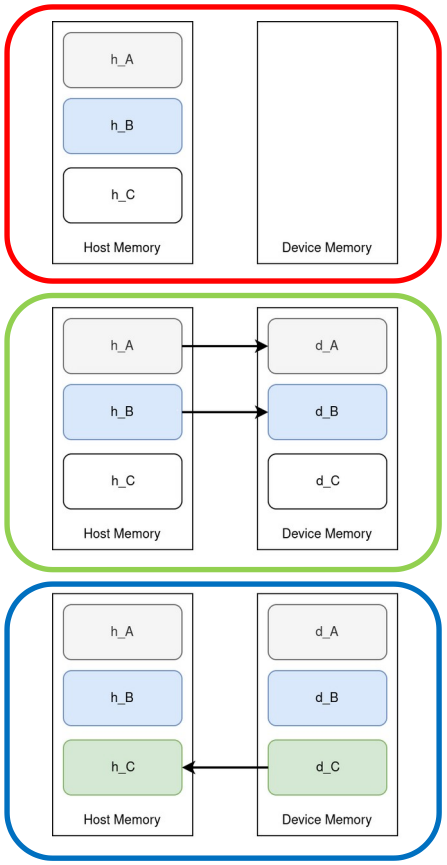
cudaMemcpy

kind is one of *cudaMemcpyHostToHost*, *cudaMemcpyHostToDevice*, *cudaMemcpyDeviceToHost*, or *cudaMemcpyDeviceToDevice*

Calling cudaMemcpy with dst and src pointers that do not match the direction of the copy results in an undefined behavior

Vector Addition Host Code

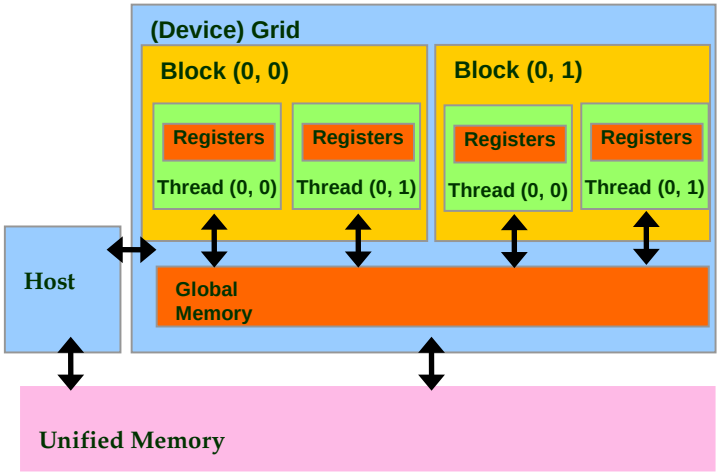
```
void vecAdd(float *h_A, float *h_B, float *h_C, int n)
{
    int size = n * sizeof(float); float *d_A, *d_B, *d_C;

    cudaMalloc((void **) &d_A, size);
    cudaMemcpy(d_A, h_A, size, cudaMemcpyHostToDevice);
     cudaMalloc((void **) &d_B, size);
     cudaMemcpy(d_B, h_B, size, cudaMemcpyHostToDevice);
     cudaMalloc((void **) &d_C, size);

     // Kernel invocation code – to be shown later

     cudaMemcpy(h_C, d_C, size, cudaMemcpyDeviceToHost);
     cudaFree(d_A); cudaFree(d_B); cudaFree (d_C);
}
```



Unified Memory

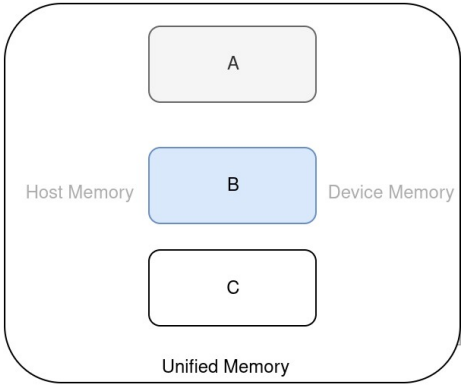
cudaMallocManaged(void ptr, size_t size)**

- Single memory space for all CPUs/GPUs
  - Maintain single copy of data
- CUDA-managed data
- On-demand page migration
- Compatible with cudaMalloc(), cudaFree()



Vector Addition, Unified Memory

```
float *A, *B, *C
cudaMallocManaged(&A, n * sizeof(float));
cudaMallocManaged(&B, n * sizeof(float));
cudaMallocManaged(&C, n * sizeof(float));

// Initialize A, B

void vecAdd(float *A, float *B, float *C, int n)
{
 // Kernel invocation code – to be shown later
}

cudaFree(A);
cudaFree(B);
cudaFree(C);
```



Error Check

```
cudaError_t  err = cudaMalloc((void **) &d_A, size);

if (err != cudaSuccess)  {
   printf("%s in %s at line %d\n",   cudaGetErrorString(err), __FILE__, __LINE__);
   exit(EXIT_FAILURE);
}
```

-For brevity, we will omit this!

Kernel Function

Our “parallel” function

Given to each thread

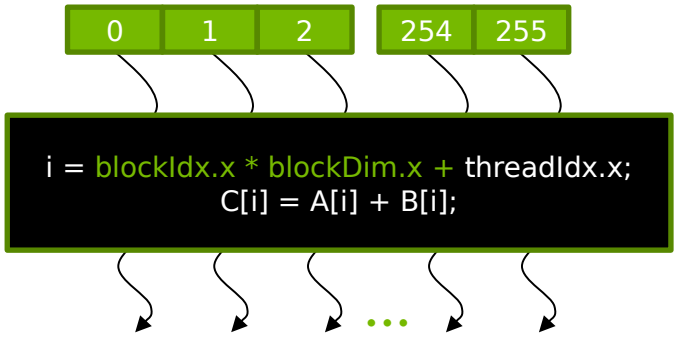
Simple implementation:

```
__global__ void vecAddKernel(float* A, float* B, float* C)
{
        //decide index somehow
        C[index] = A[index] + B[index];
}
```

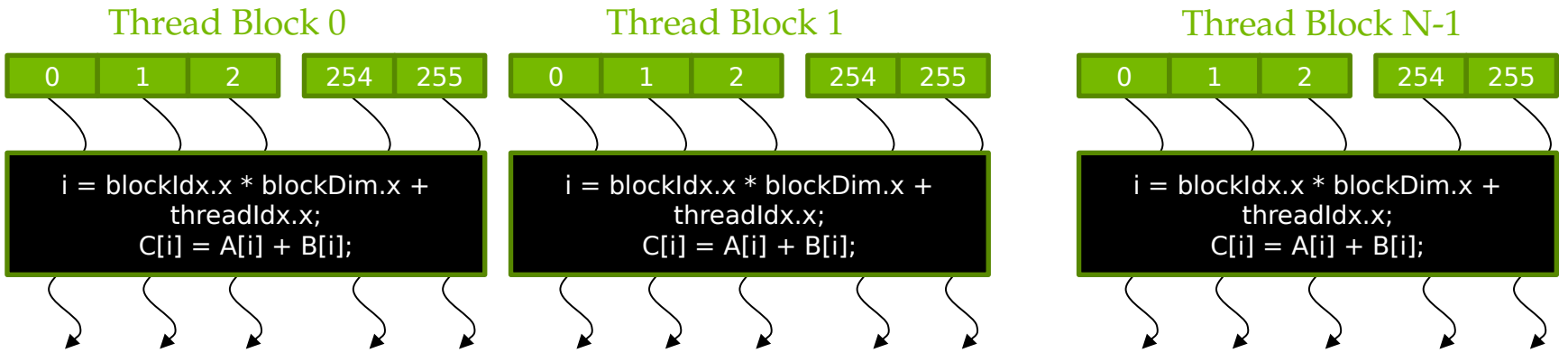
Arrays of Parallel Threads

A CUDA kernel is executed by a grid (array) of threads

- All threads in a grid run the same kernel code (Single Program Multiple Data)
- Each thread has indices that it uses to compute memory addresses and make control decisions



Thread Blocks: Scalable Cooperation

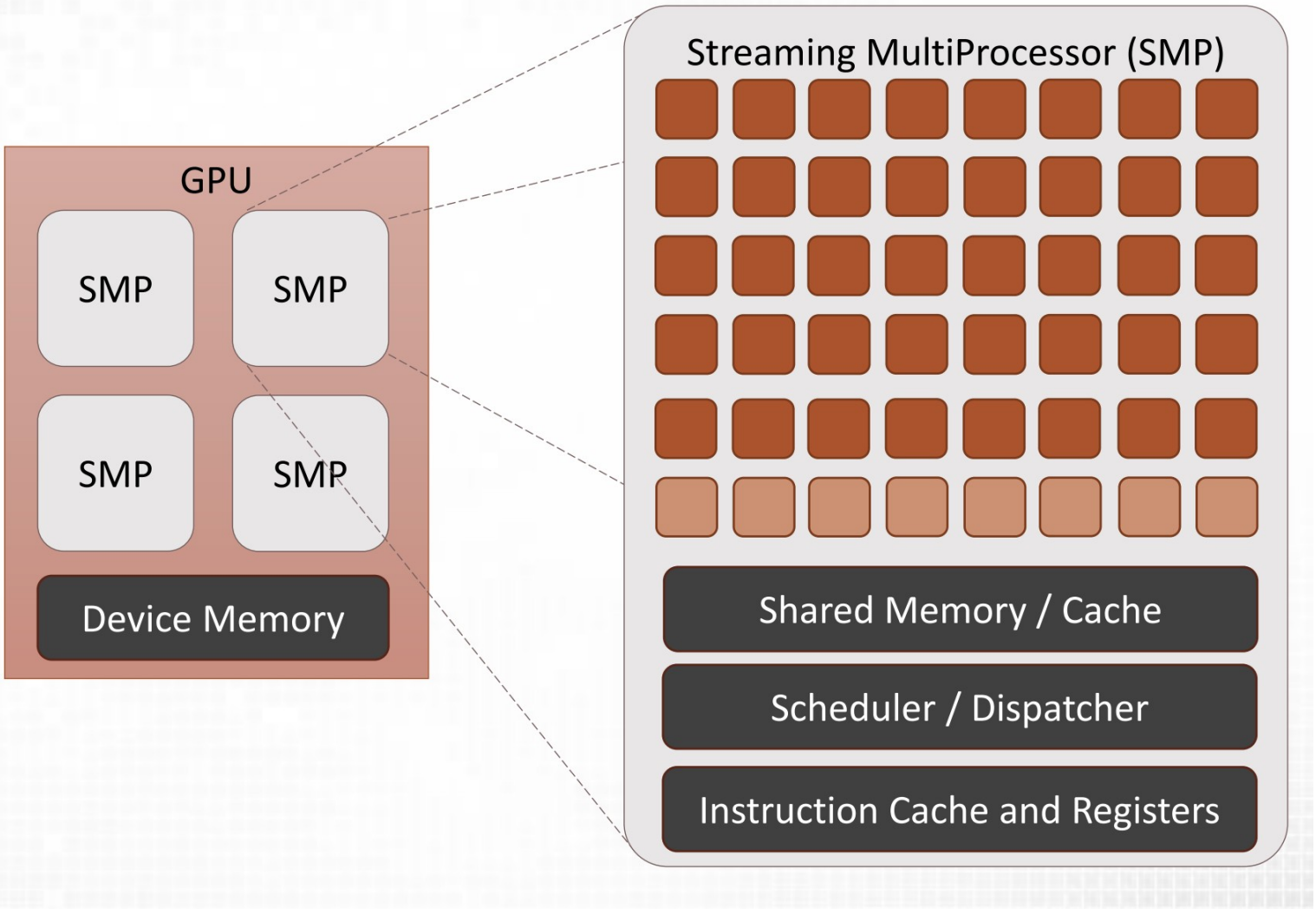


Divide thread array into multiple blocks

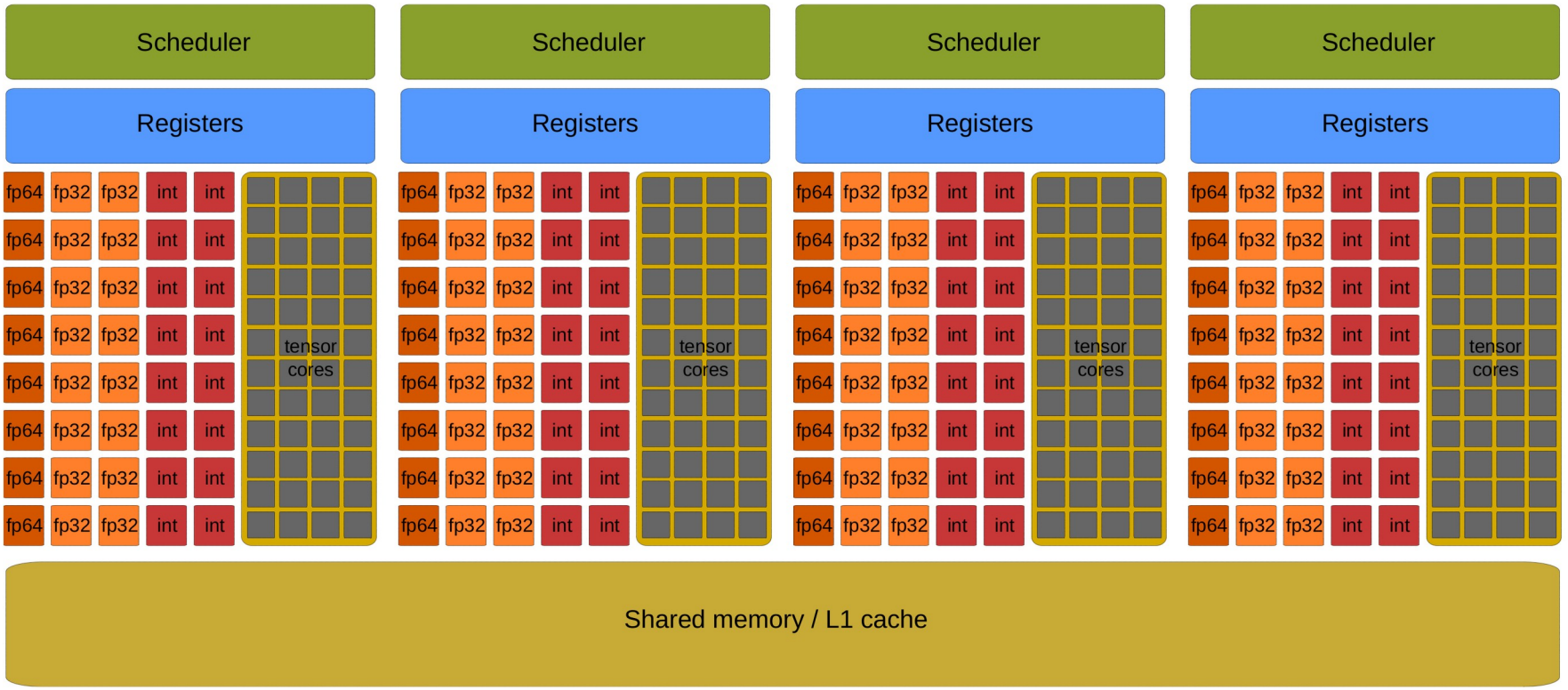
Threads within a block cooperate via shared memory, atomic operations and barrier synchronization

Threads in different blocks do not interact

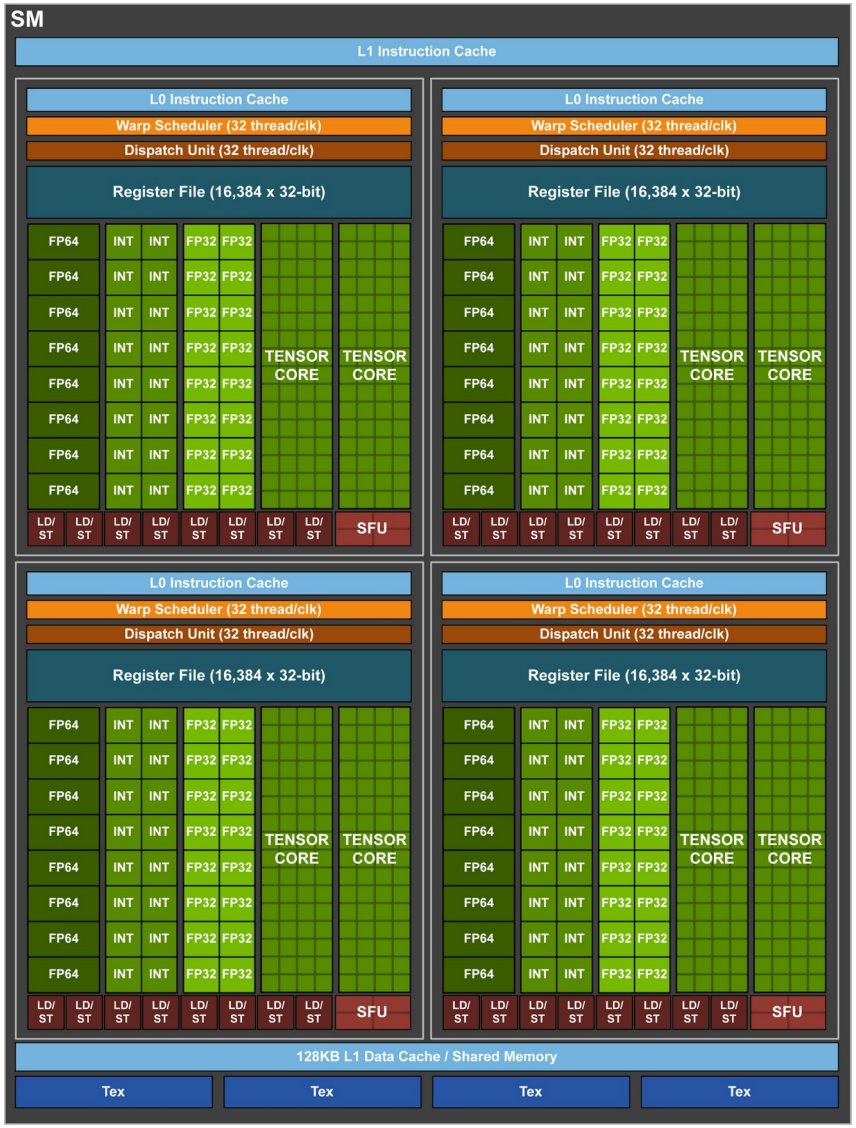
GPU Hardware



Inside Streaming Multiprocesor



Inside Volta SMP



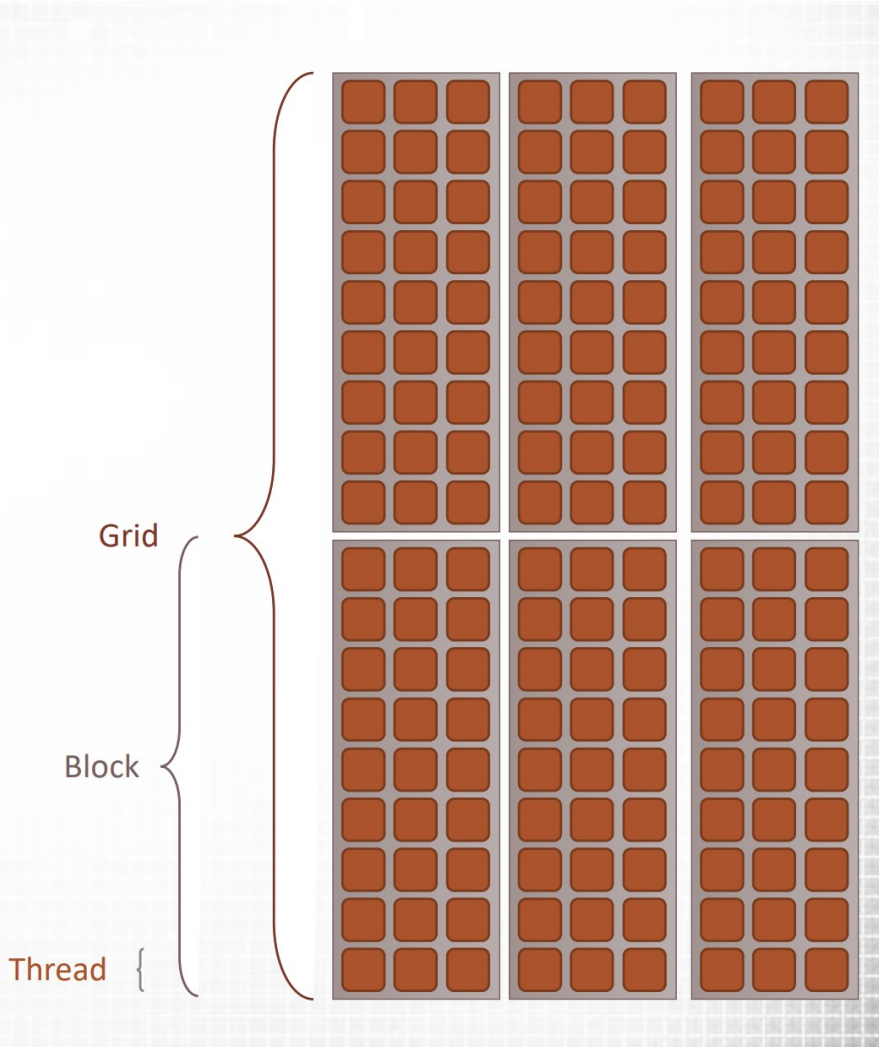
GPU Software Model

Hardware abstracted as a Grid of Thread Blocks

- Blocks map to SMPs
- Each thread maps to a CUDA core

No need to know HW

Code is portable across GPUs



Built-in Variables

threadIdx

The location of a thread within a block

blockIdx

The location of a block within a grid

blockDim

The dimensions of the blocks (how many threads in each block)

gridDim

The dimensions of the grid (how many blocks)

Built-in Variables in 1-D

threadIdx.x -- "thread index" within block in "x" dimension

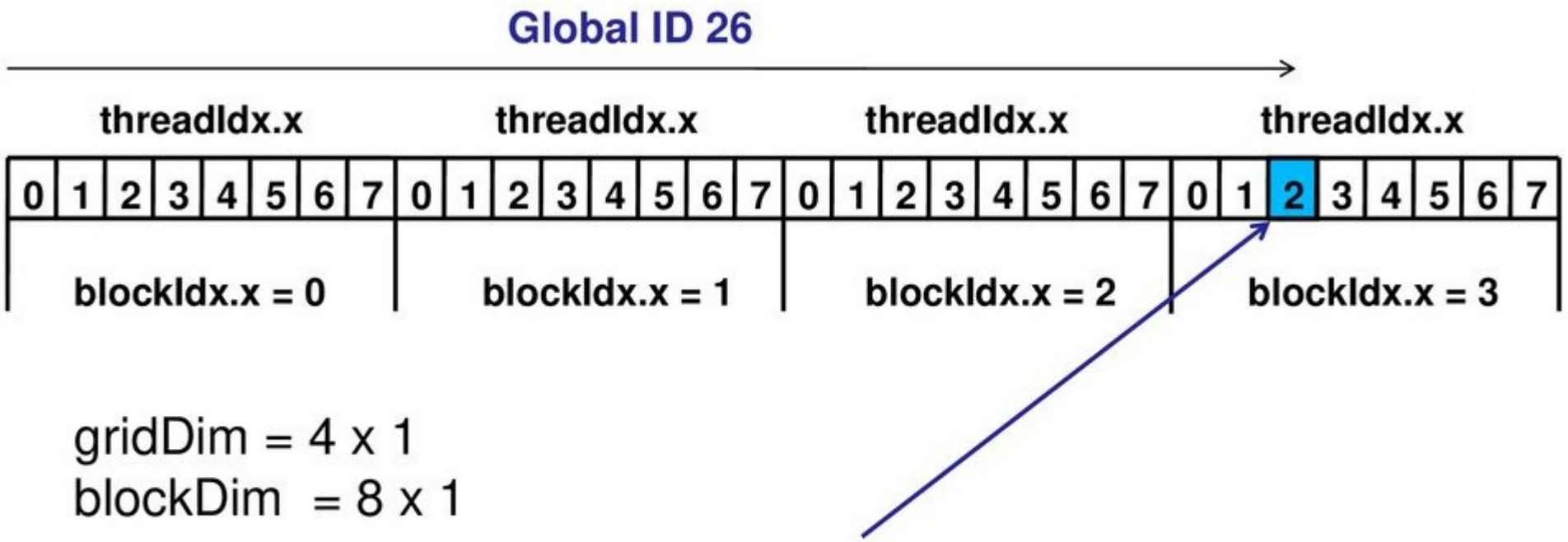
blockIdx.x -- "block index" within grid in "x" dimension

blockDim.x -- "block dimension" in "x" dimension
(i.e. number of threads in block in x dimension)

Full global thread ID in x dimension can be computed by:

```
x = blockIdx.x * blockDim.x + threadIdx.x;
```

Grid, Block, Thread



gridDim = 4 x 1
blockDim = 8 x 1

Global thread ID = **blockIdx.x * blockDim.x + threadIdx.x**
= 3 * 8 + 2 = thread 26 with linear global addressing

Vector Addition Kernel Function – Device Code

```
// Compute vector sum C = A + B
// Each thread performs one addition

__global__
void vecAddKernel(float* A, float* B, float* C, int n)
{
  int i = threadIdx.x+blockDim.x*blockIdx.x;
  if(i<n) C[i] = A[i] + B[i];
}
```

Vector Addition Kernel Function – Device Code

```
// Compute vector sum C = A + B
// each thread calculates two (adjacent) output elements of a vector addition

__global__
void vecAddKernel(float* A, float* B, float* C, int n)
{
  int i = (threadIdx.x+blockDim.x*blockIdx.x)*2;
    if(i<n) C[i] = A[i] + B[i];
  int j = (threadIdx.x+blockDim.x*blockIdx.x)*2 + 1;
    if(j<n) C[j] = A[j] + B[j];
}
```

CUDA C Keywords for Function Declarations

	Executed on the:	Only callable from the:
`__device__ float DeviceFunc()`	device	device
`__global__ void KernelFunc()`	device	host
`__host__ float HostFunc()`	host	host

__global__ defines a kernel function

- Each "__" consists of two underscore characters
- A kernel function must return void

__device__ declares device function, which is called and executed on device

__host__ is optional if used alone

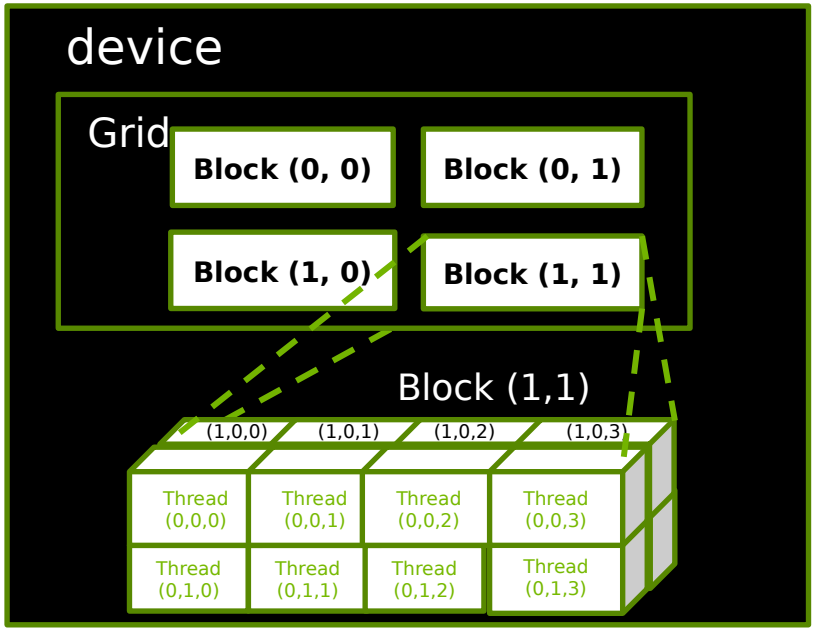
Indexing

Each thread uses indices to decide what data to work on

- blockIdx: 1D, 2D, or 3D
- threadIdx: 1D, 2D, or 3D

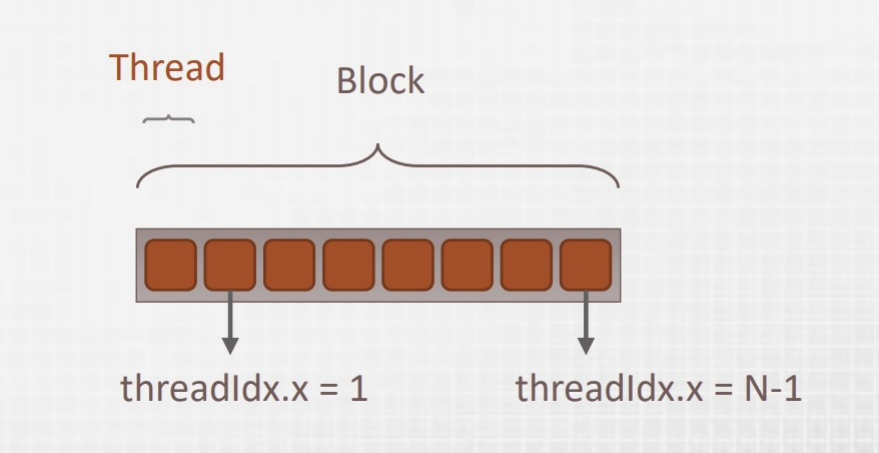
Simplifies memory addressing when processing multidimensional data

- Image processing
- Solving PDEs on volumes
- …



Kernel Function Invocation

```
dim3 blocksPerGrid(1,1,1); //use only one block
dim3 threadsPerBlock(N,1,1); //use N threads in the block
myKernel<<<blocksPerGrid, threadsPerBlock>>>(result);
```



Vector Addition Kernel Launch - Host Code

```
void vecAdd(float* h_A, float* h_B, float* h_C, int n)
{
 // d_A, d_B, d_C allocations and copies omitted
 // Run ceil(n/256.0) blocks of 256 threads each
  vecAddKernel<<<ceil(n/256.0),256>>>(d_A, d_B, d_C, n);
}
```

Only one block will give poor performance, since a block gets allocated to a single SMP

One thread per block is worse due to warp execution

Number of blocks ~ Multiple of Sms in GPU

More on Kernel Launch – Host Code

```
void vecAdd(float* h_A, float* h_B, float* h_C, int n)
{
  dim3 DimGrid(ceil(n/256.0), 1, 1);
  dim3 DimBlock(256, 1, 1);
  vecAddKernel<<<DimGrid,DimBlock>>>(d_A, d_B, d_C, n);
}
```

Vector Addition with 8000 Elements

Each thread calculates one output element

Thread block size is 256 threads

The programmer configures the kernel launch to have a minimal number of thread blocks to cover all output elements

How many threads will be in the grid?

ceil(8000/256) * 256 = 32 * 256 = 8192

(the minimal multiple of 256 to cover 8000 is 256*32 = 8192)

Device Synchronization

Kernel calls are non-blocking

- Host continues after kernel launch
- Overlaps CPU and GPU execution

cudaDeviceSynchronize()

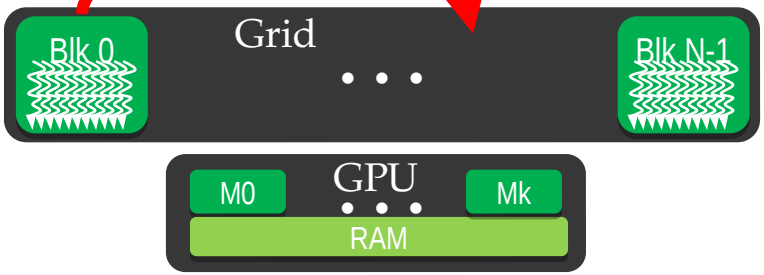
- Call from the host to block until GPU kernels have completed

Standard cudaMemcpy calls are blocking

Kernel Execution

```
__host__
void vecAdd(…)
{
  dim3 DimGrid(ceil(n/256.0),1,1);
  dim3 DimBlock(256,1,1);
vecAddKernel<<<DimGrid,DimBlock>>>(d_A,d_B,d_C,n);
}
```

```
__global__
void vecAddKernel(float *A,
     float *B, float *C, int n)
{
   int i = blockIdx.x * blockDim.x
     + threadIdx.x;

   if( i<n ) C[i] = A[i]+B[i];
}
```



Important CUDA Syntax Extensions

Declaration specifiers

__global__ void foo(...); // kernel entry point (runs on GPU)

Syntax for kernel launch

foo<<<500, 128>>>(...); // 500 thread blocks, 128 threads each

Built in variables for thread identification

dim3 threadIdx; dim3 blockIdx; dim3 blockDim; dim3 gridDim;

Example: Original C Code

```
void saxpy_serial(int n, float a, float *x, float *y)
{
    for (int i = 0; i < n; ++i)
    y[i] = a*x[i] + y[i];
}
int main()
{
    // omitted: allocate and initialize memory
     saxpy_serial(n, 2.0, x, y); // Invoke serial SAXPY kernel
   // omitted: using result
}
```

CUDA Code

```
__global__ void saxpy(int n, float a, float *x, float *y) {
        int i = blockIdx.x*blockDim.x + threadIdx.x;
        if(i<n) y[i]=a*x[i]+y[i];
}
int main() {
        // omitted: allocate and initialize memory
        int nblocks = (n + 255) / 256;
        cudaMalloc((void**) &d_x, n);
        cudaMalloc((void**) &d_y, n);
        cudaMemcpy(d_x,h_x,n*sizeof(float),cudaMemcpyHostToDevice);
        cudaMemcpy(d_y,h_y,n*sizeof(float),cudaMemcpyHostToDevice);
        saxpy<<<nblocks, 256>>>(n, 2.0, d_x, d_y);
        cudaMemcpy(h_y,d_y,n*sizeof(float),cudaMemcpyDeviceToHost);
        // omitted: using result
}
```

nvcc Compiler

NVIDIA provides a CUDA-C compiler

nvcc

NVCC compiles device code then forwards code on to the host compiler (e.g. g++)

Can be used to compile & link host only applications

References

Chapter 2

(Programming Massively Parallel Processors : A Hands-on Approach, David B. Kirk, Wen-Mei W. Hwu, Morgan Kaufmann Publishers, 3rd edition)