www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

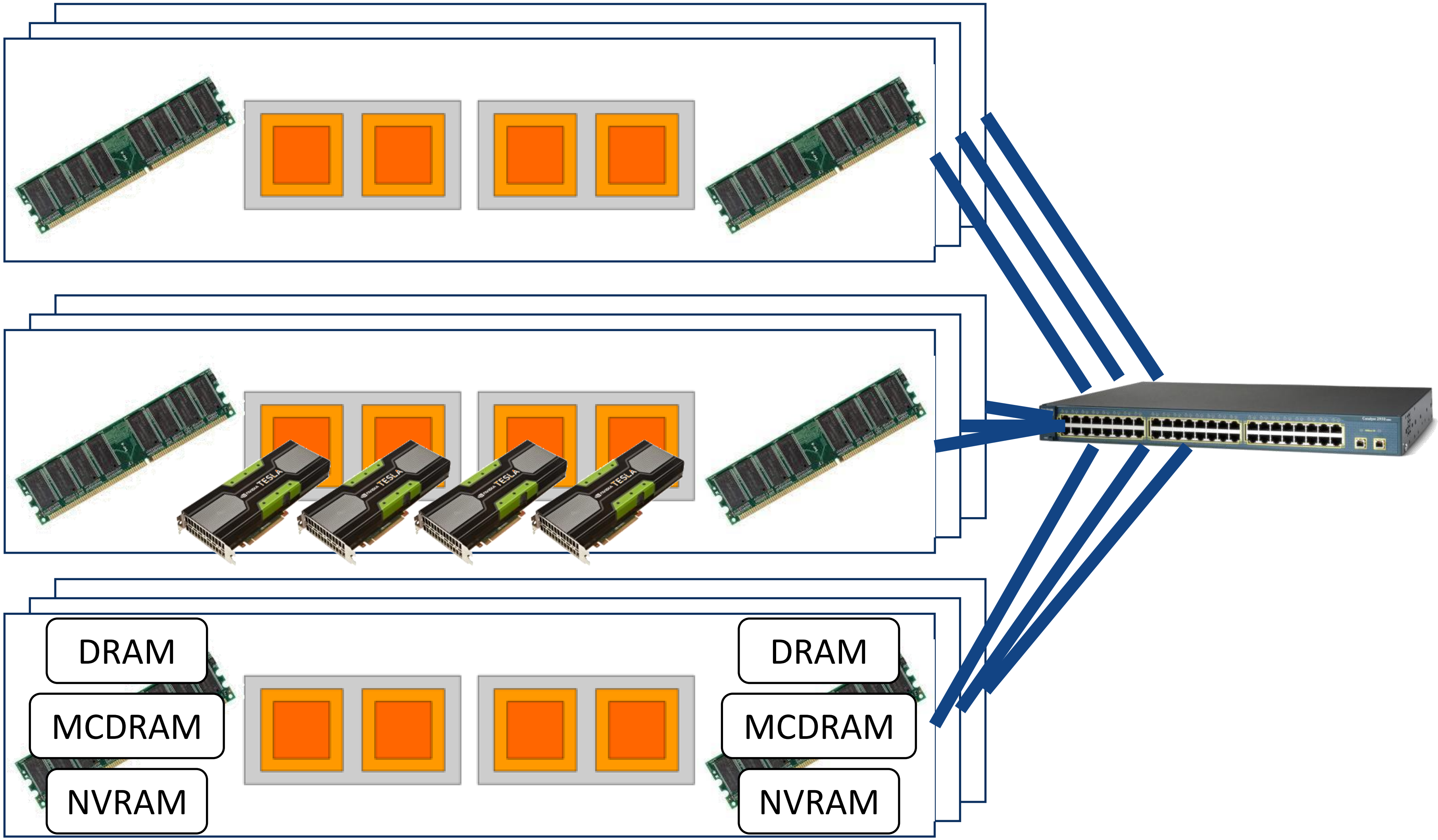
Introduction to CUDA: The Basics

Marc Jordà, Antonio J. Peña

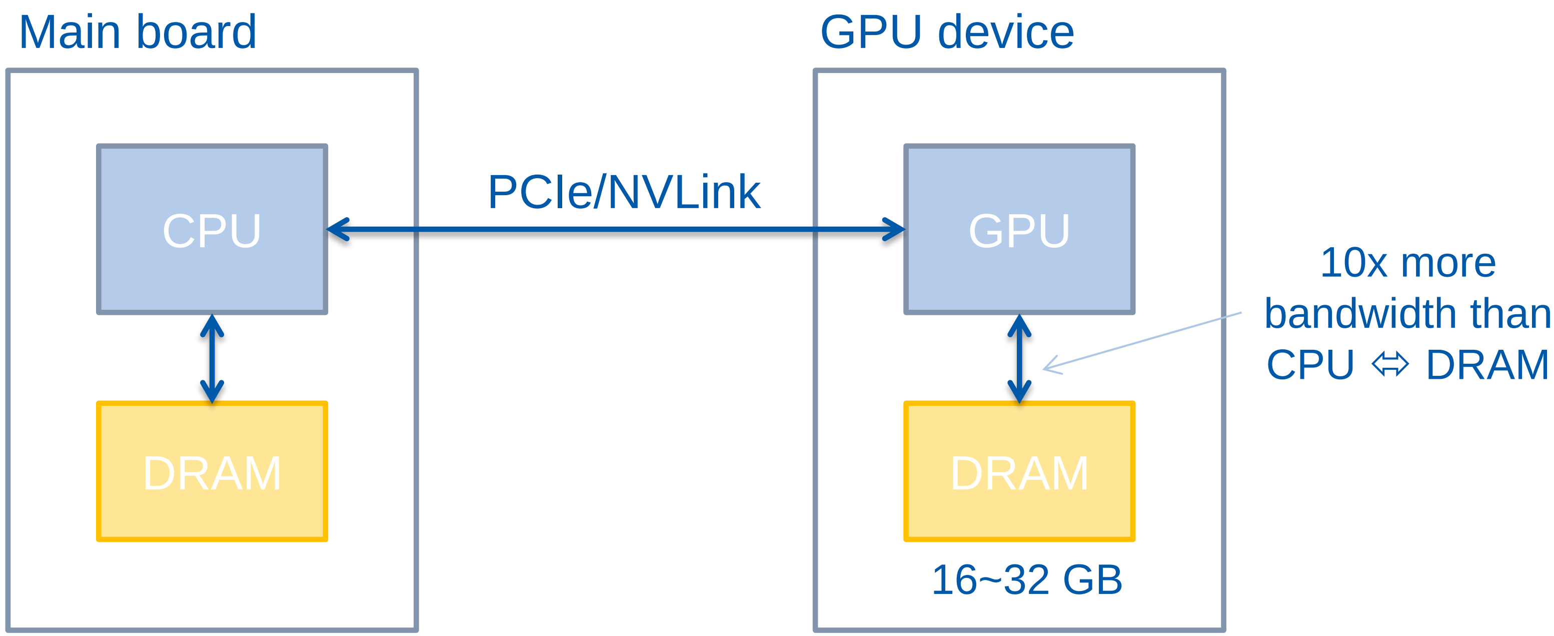
Based on material from NVIDIA’s GPU Teaching Kit

Montevideo, 21-25 October 2019

Heterogeneous Parallel Computing

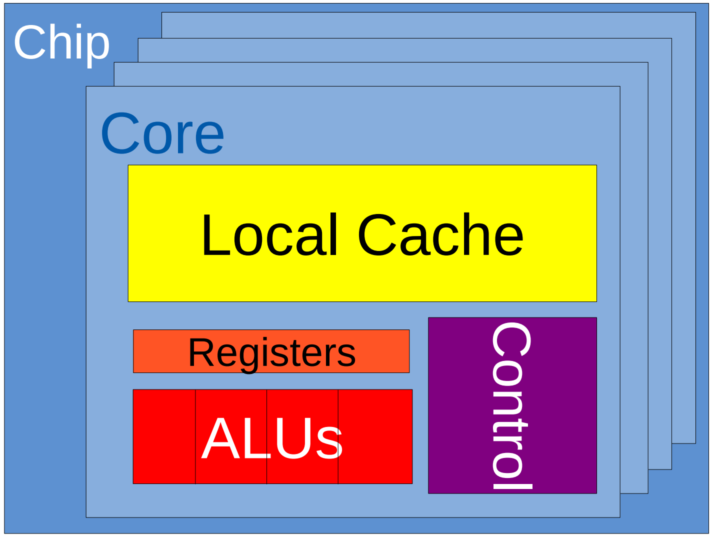


Heterogeneous Node

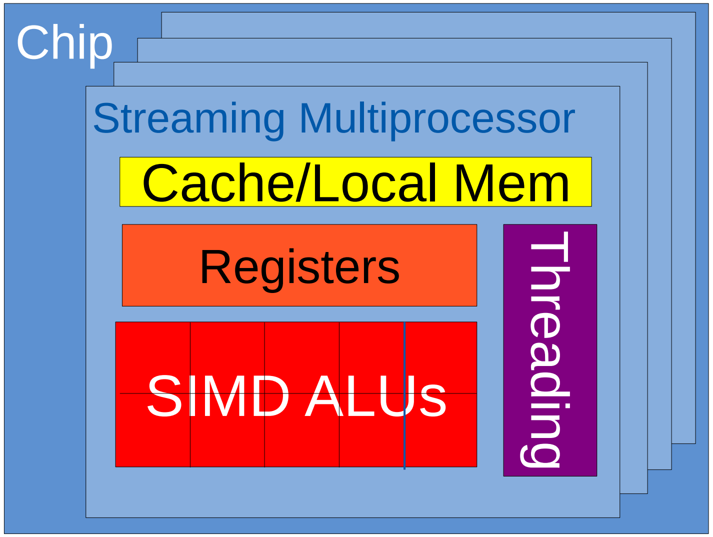


CPU and GPU are designed very differently

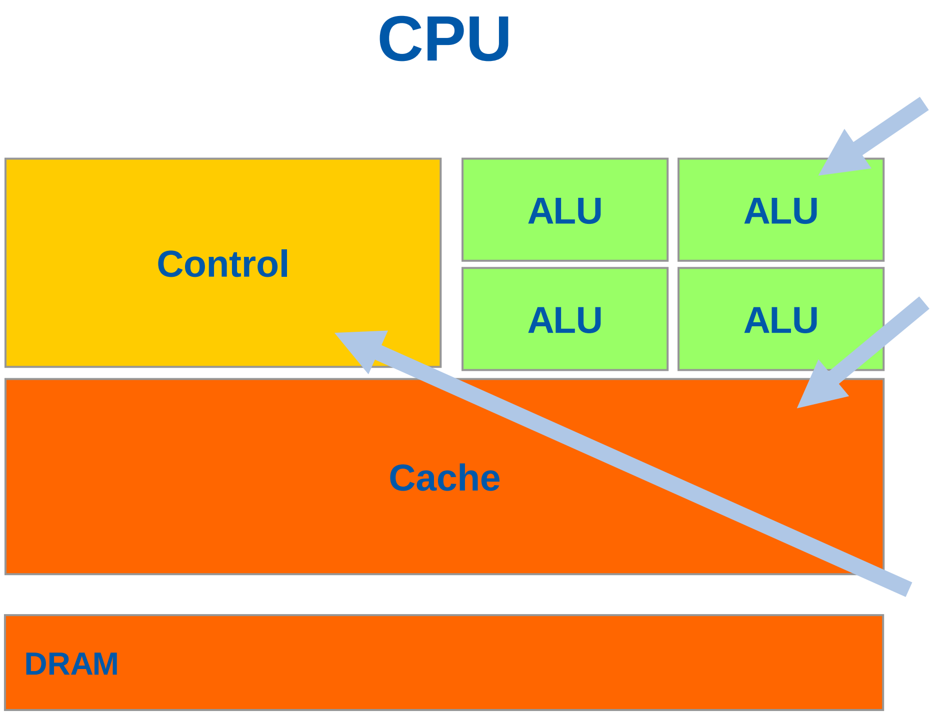
CPU
Latency Oriented Cores



GPU
Throughput Oriented Cores

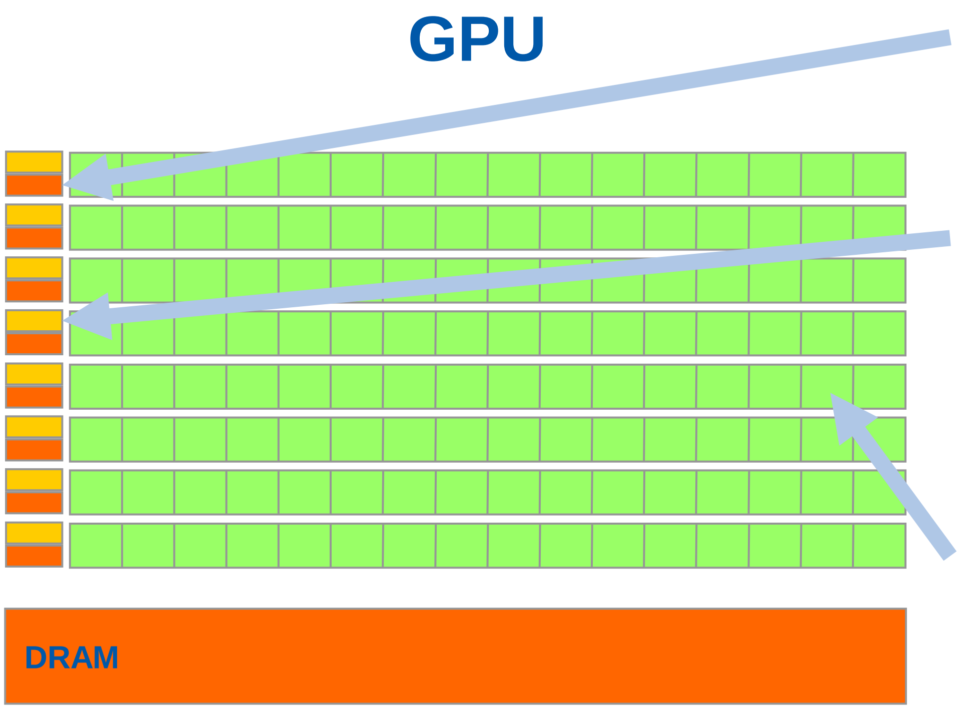


CPUs: Latency Oriented Design



- Powerful ALUs
  - Short pipeline, reduced operation latency
- Large caches
  - Convert long latency memory accesses to short latency cache accesses
- Sophisticated control
  - Branch prediction and return value prediction, speculative execution, etc.
  - Data forwarding for reduced data latency

GPUs: Throughput Oriented Design



- Small caches
  - To boost memory throughput
- Simple control
  - No branch prediction
  - No speculative execution
  - No data forwarding
- SIMD ALUs
  - Vector units (similar to AVX)
  - Many, long latency but heavily pipelined for high throughput
- Can have many active threads
  - Throughput oriented
  - Helps tolerate latencies

Applications should Use Both CPU and GPU

- CPUs for sequential parts where latency matters
  - CPUs can be 10X+ faster than GPUs for sequential code

- GPUs for parallel parts where throughput wins
  - GPUs can be 10X+ faster than CPUs for parallel code

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

PROGRAMMING ALTERNATIVES TO USE THE GPU

3 Ways to Accelerate Applications

Applications

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

Libraries: Easy, High-Quality Acceleration

- **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

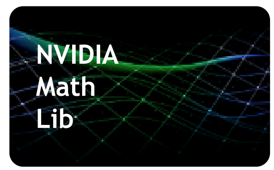
Linear Algebra
FFT, BLAS,
SPARSE, Matrix

CULA|tools

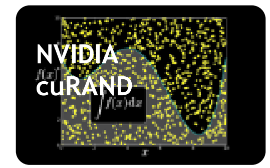
MAGMA

CUSP

Numerical & Math
RAND, Statistics

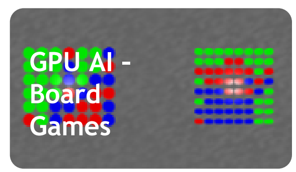


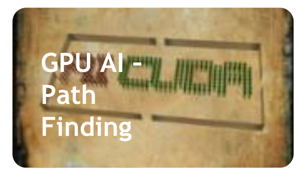
ArrayFire



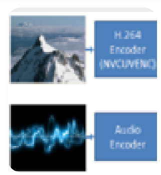
Data Struct. & AI
Sort, Scan, Zero Sum

Thrust





Visual Processing
Image & Video



Sundog™
Software

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Vector Addition in Thrust

```
thrust::device_vector<float> deviceInput1(inputLength);
thrust::device_vector<float> deviceInput2(inputLength);
thrust::device_vector<float> deviceOutput(inputLength);

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

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

Compiler Directives: Easy, Portable Acceleration

- **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: Most Performance and Flexible Acceleration

- **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

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

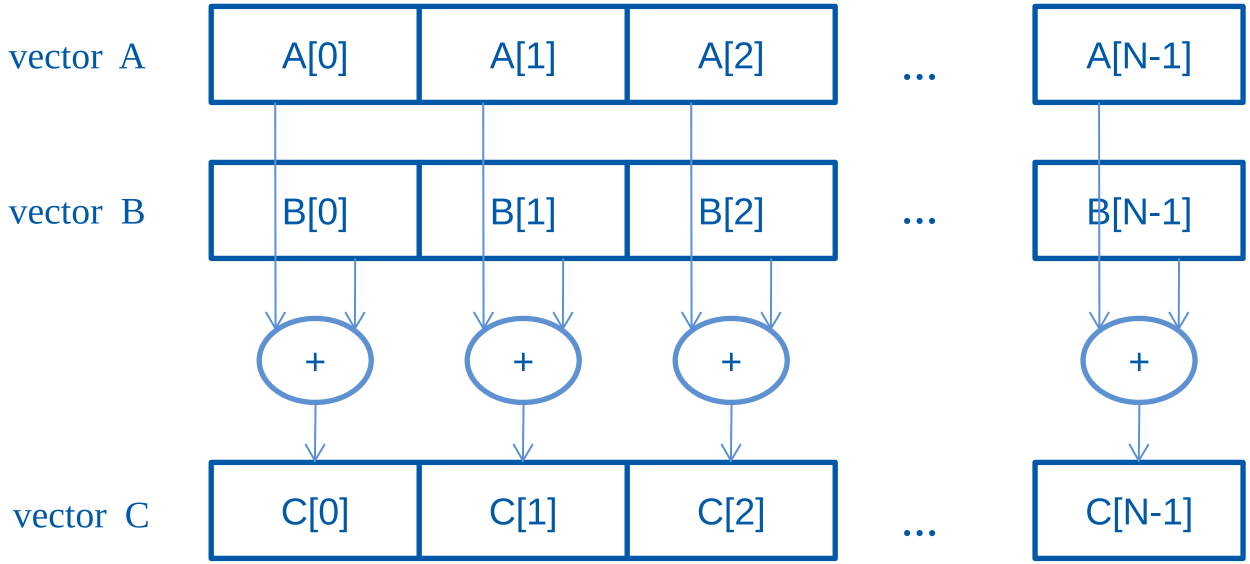
GPU Programming Languages

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

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MEMORY ALLOCATION AND DATA MOVEMENT API FUNCTIONS

Data Parallelism - Vector Addition Example

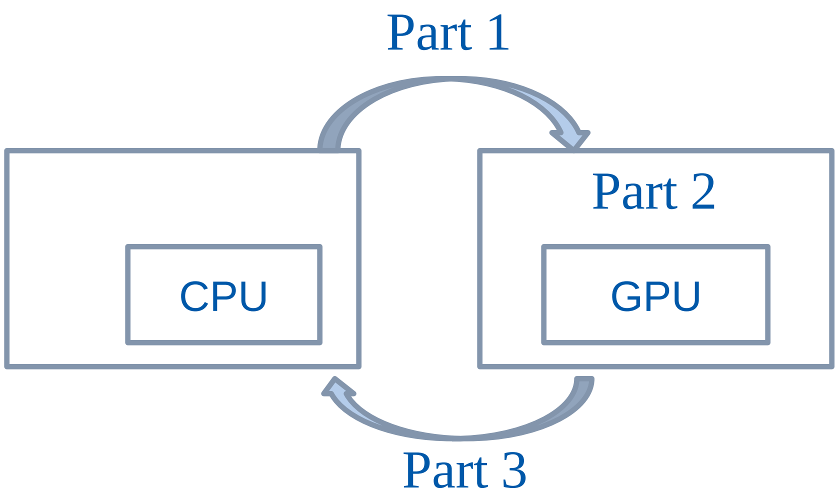


Vector Addition – Traditional C Code

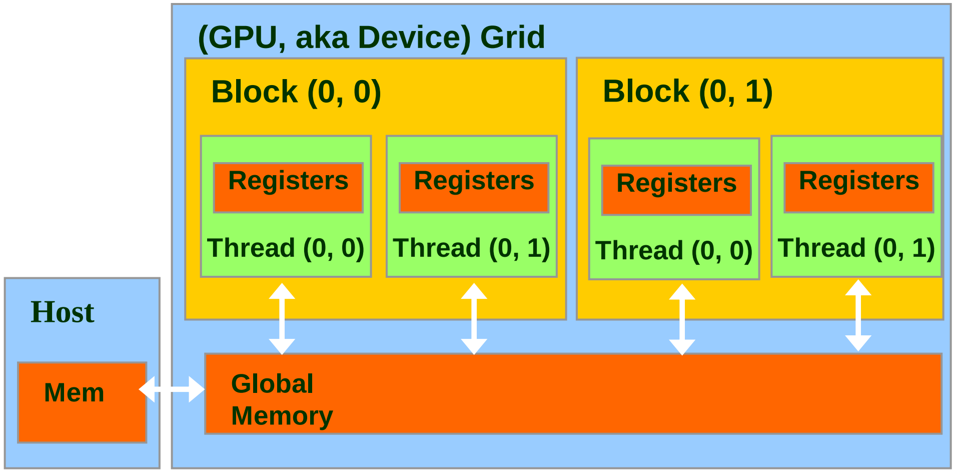
```
// Compute vector sum C = A + B
void vecAdd(float *h_A, float *h_B, float *h_C, int n)
{
    int i;
    for (i = 0; i<n; i++) h_C[i] = h_A[i] + h_B[i];
}

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 vecAdd CUDA Host Code

```
#include <cuda.h>
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;
    // Part 1
    // Allocate device memory for A, B, and C
    // copy A and B to device memory

    // Part 2
    // Kernel launch code – the device performs the actual vector addition

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



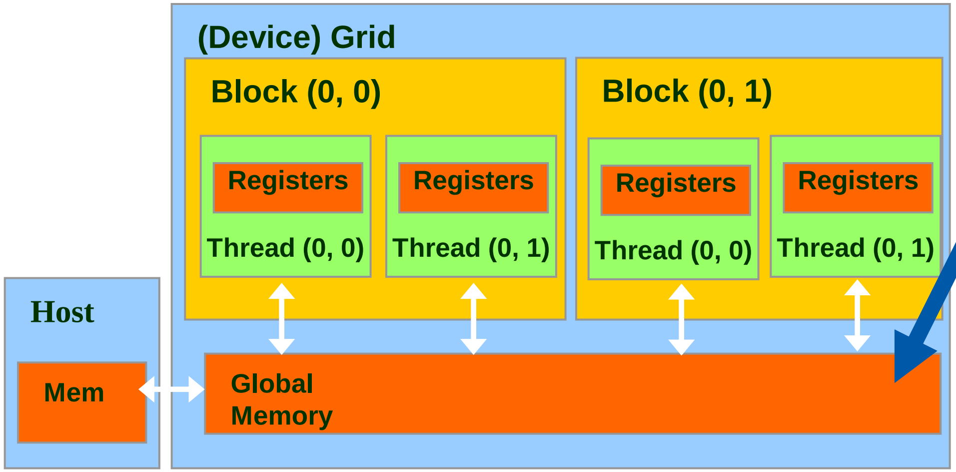
Partial Overview of CUDA Memories



- GPU threads
  - Grouped in thread blocks to form the thread grid
- Device code can:
  - R/W per-thread registers
  - R/W all-shared global memory
- Host code can
  - Transfer data between global memory and host memory

We will cover more memory types and more sophisticated memory models later.

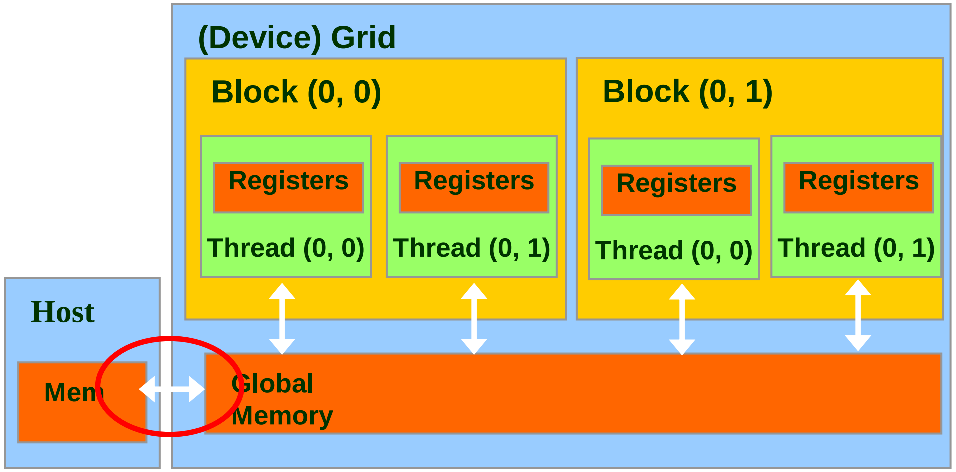
CUDA Device Memory Management API functions



- cudaMalloc()
  - Allocates an object in the device global memory
  - Two parameters
    - **Address of a pointe**r to the allocated object
    - **Size of** allocated object in terms of bytes
  - Regular C/C++ pointer, only valid in GPU code and CUDA copy functions

- cudaFree()
  - Frees object from device global memory
  - One parameter
    - **Pointer** to freed object

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

  - Transfer to device is asynchronous

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);
}
```

In Practice, Check for API Errors in Host Code

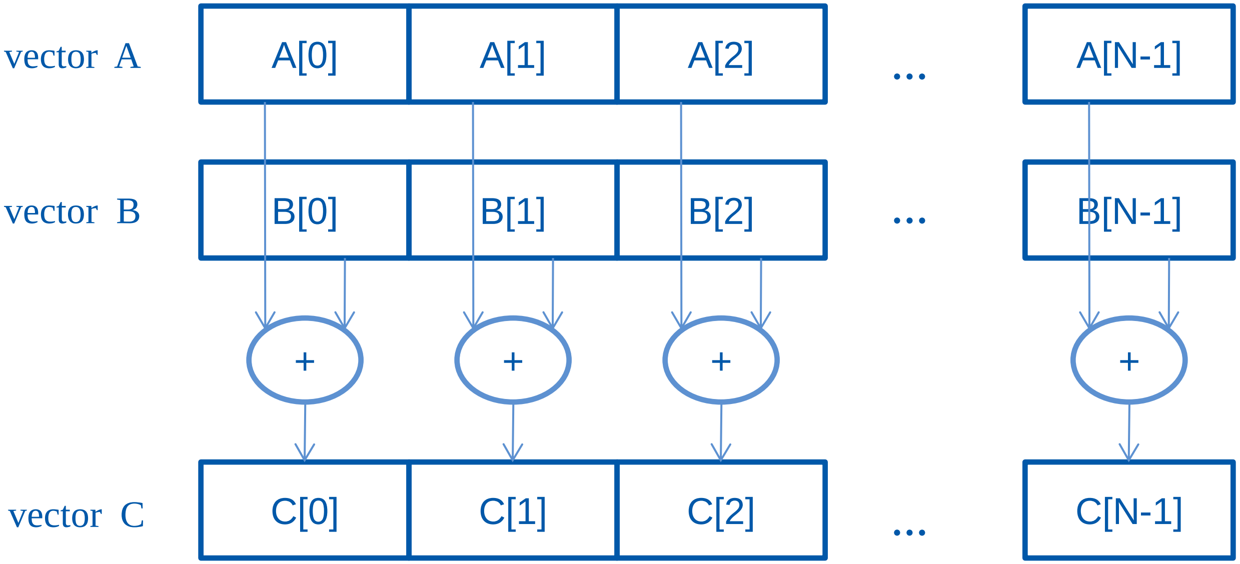
```
cudaError_t err = cudaMalloc((void **) &d_A, size);

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

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

THREADS AND KERNEL FUNCTIONS

Data Parallelism - Vector Addition Example



CUDA Execution Model

- Heterogeneous host (CPU) + device (GPU) application C program
  - Serial parts in **host** C code
  - Parallel parts in **device** SPMD (Single Program, Multiple Data) kernel code

Serial Code (host)

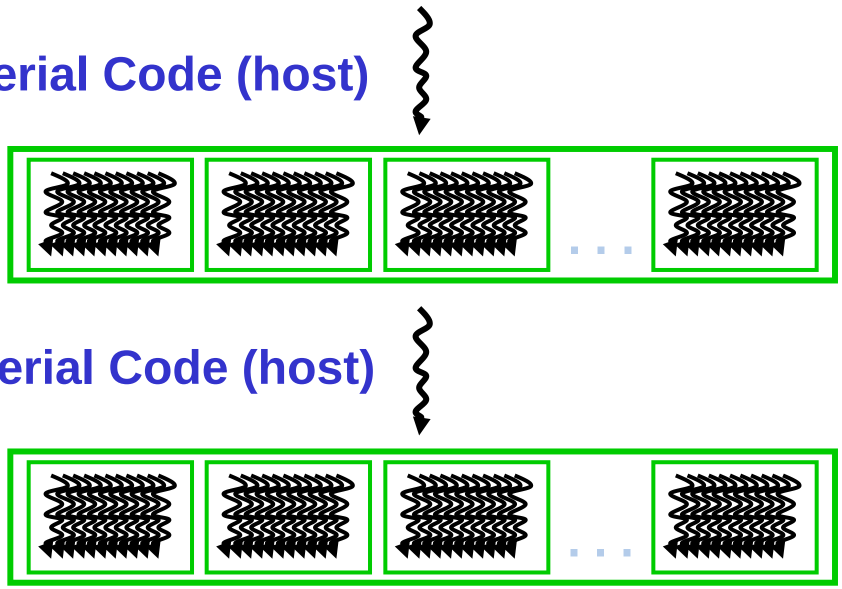
Parallel Kernel (device)

KernelA<<< nBlk, nTid >>>(args);

Serial Code (host)

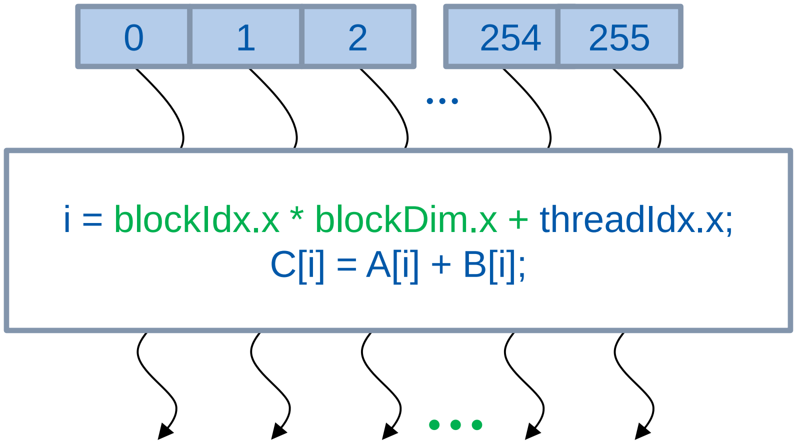
Parallel Kernel (device)

KernelB<<< nBlk, nTid >>>(args);

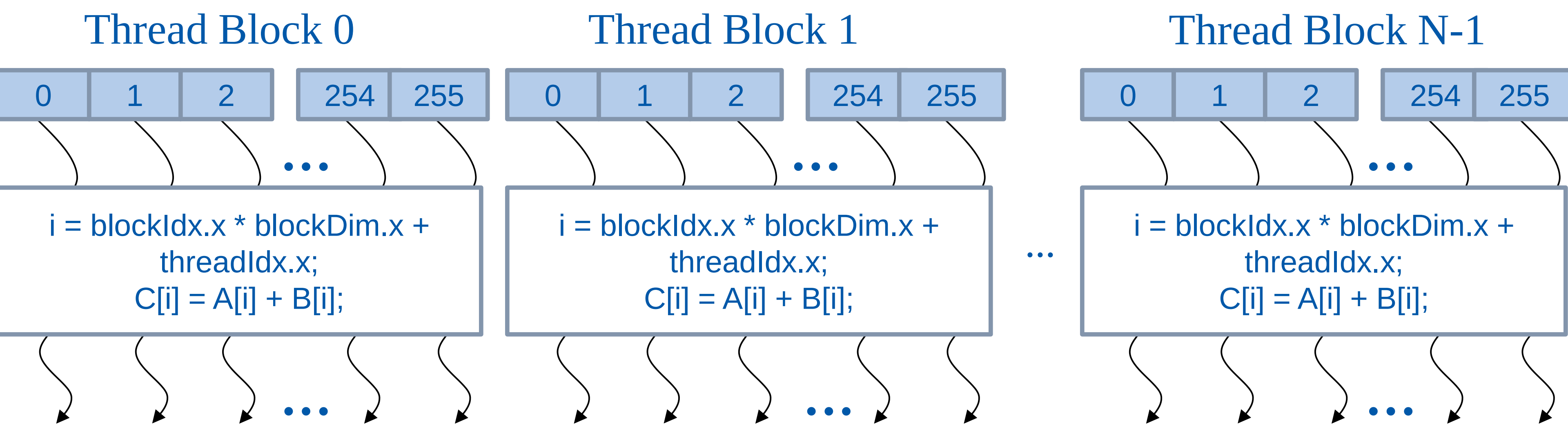


Arrays of Parallel Threads

- 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 indexes that it uses to compute memory addresses and make control decisions



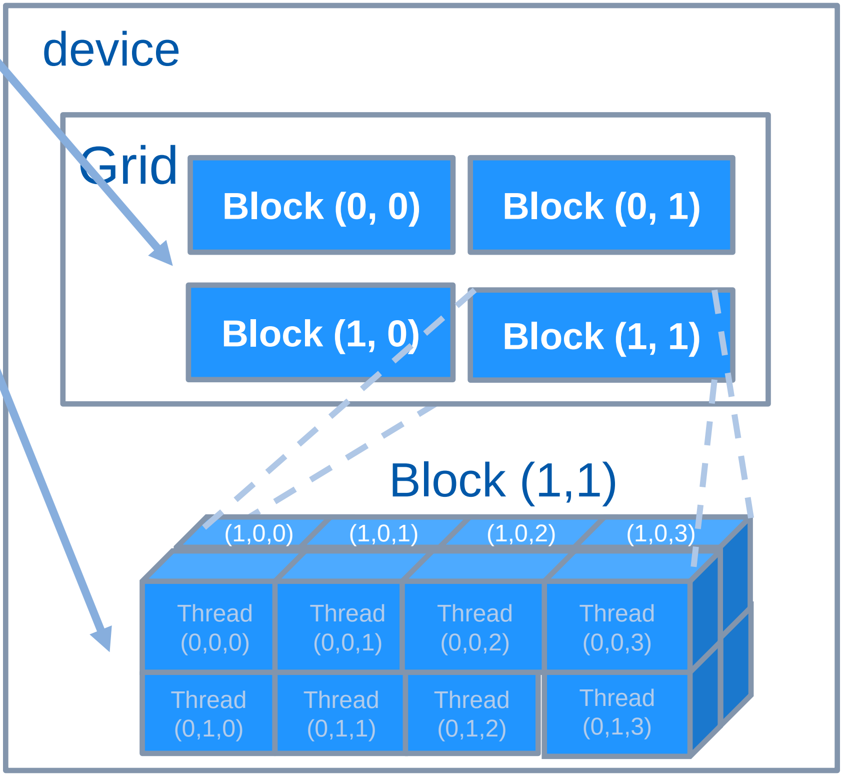
Thread Blocks: Scalable Cooperation



- 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

blockIdx and threadIdx

- 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
  - …



INTRODUCTION TO THE CUDA TOOLKIT

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

Hello World! with Device Code

```
__global__ void mykernel(void) {
}

int main(void) {
   mykernel<<<1,1>>>();
   printf("Hello World!\n");
   return 0;
}
```

Output:

```
$ nvcc main.cu
$ ./a.out
Hello World!
```

- Notes
  - `mykernel` does nothing
  - nvcc only parses .cu files for CUDA

Developer Tools - Debuggers

NSIGHT | **CUDA-GDB** | **CUDA MEMCHECK**

NVIDIA Provided

allinea DDT | TotalView®

3rd Party

https://developer.nvidia.com/debugging-solutions

Compiler Flags

- There are two compilers being used
  - NVCC: Device code
  - Host Compiler: C/C++ code

- NVCC supports some host compiler flags
  - If flag is unsupported, use –Xcompiler to forward to host
    - e.g. –Xcompiler –fopenmp

- Debugging Flags
  - -g: Include host debugging symbols
  - -G: Include device debugging symbols and disables optimization of kernel code
  - -lineinfo: Include line information with symbols

CUDA-MEMCHECK

- Memory debugging tool
  - No recompilation necessary

  $ cuda-memcheck --tool <memcheck|racecheck|synccheck|initcheck> ./cuda_program

- Can detect the following errors
  - Memory leaks
  - Memory errors (OOB, misaligned access, illegal instruction, etc)
  - Race conditions
  - Illegal Barriers
  - Uninitialized Memory

- For line numbers use the following compiler flags:
  - -G (disables device code optimization)
  - -lineinfo -Xcompiler -rdynamic

http://docs.nvidia.com/cuda/cuda-memcheck

CUDA-GDB

- *cuda-gdb* is an extension of GDB
  - Provides seamless debugging of CUDA and CPU code

- Works on Linux and Macintosh
  - For a Windows debugger use NSIGHT Visual Studio Edition

http://docs.nvidia.com/cuda/cuda-gdb

Example: cuda-gdb

%> cuda-gdb --args ./a.out

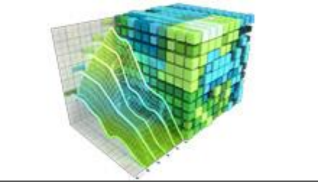
```
(cuda-gdb) b main        //set break point at main
(cuda-gdb) r                            //run application
(cuda-gdb) l                            //print line context
(cuda-gdb) b foo                    //break at kernel foo
(cuda-gdb) c                            //continue
(cuda-gdb) cuda thread              //print current thread
(cuda-gdb) cuda thread 10           //switch to thread 10
(cuda-gdb) cuda block               //print current block
(cuda-gdb) cuda block 1             //switch to block 1
(cuda-gdb) d                  //delete all break points
(cuda-gdb) set cuda memcheck on    //turn on cuda memcheck
(cuda-gdb) r                  //run from the beginning
```

http://docs.nvidia.com/cuda/cuda-gdb

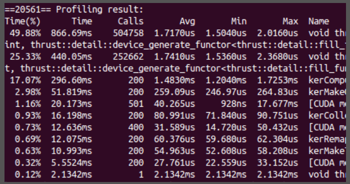
Developer Tools - Profilers

NSIGHT

NVVP



NVPROF



```
==20561== Profiling result:
Time(%)      Time     Calls       Avg       Min       Max  Name
 49.88%  866.69ms    504758  1.7170us  1.5040us  2.0160us  void thr
int, thrust::detail::device_generate_functor<thrust::detail::fill_
 25.33%  440.05ms    252662  1.7410us  1.5360us  2.3680us  void thr
t, thrust::detail::device_generate_functor<thrust::detail::fill_fu
 17.07%  296.60ms       200  1.4830ms  1.2040ms  1.7253ms  kerComp
  2.98%  51.819ms       200  259.09us  246.97us  264.83us  kerMake
  1.16%  20.173ms       501  40.265us     928ns  17.677ms  [CUDA m
  0.93%  16.198ms       200  80.991us  71.840us  90.751us  kerColl
  0.73%  12.636ms       400  31.589us  14.720us  50.432us  [CUDA m
  0.69%  12.075ms       200  60.376us  59.680us  62.304us  kerRema
  0.63%  10.993ms       200  54.963us  52.608us  58.208us  kerMake
  0.32%  5.5524ms       200  27.761us  22.559us  33.152us  [CUDA m
  0.12%  2.1342ms         1  2.1342ms  2.1342ms  2.1342ms  void thr
```

NVIDIA Provided

TAU

Tuning and Analysis Utilities

VampirTrace

3rd Party

https://developer.nvidia.com/performance-analysis-tools

NVPROF

Command Line Profiler

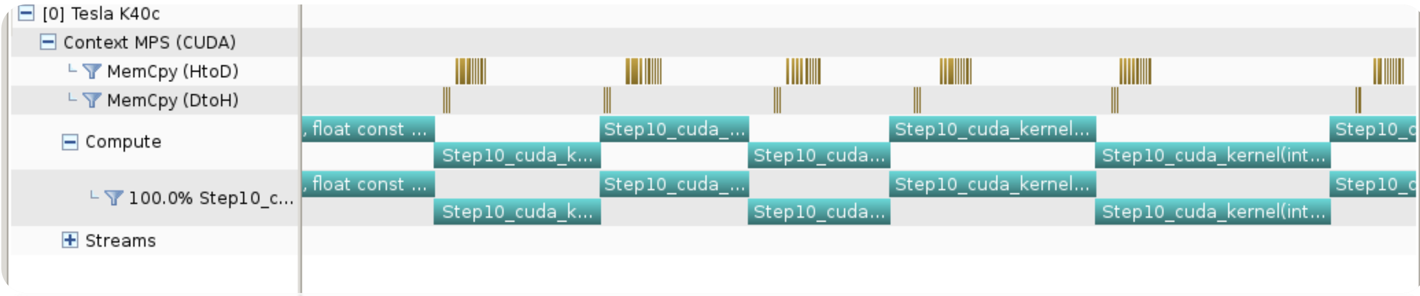
- Compute time in each kernel
- Compute memory transfer time
- Collect metrics and events
- Support complex process hierarchy's
- Collect profiles for NVIDIA Visual Profiler
- No need to recompile

Example: nvprof

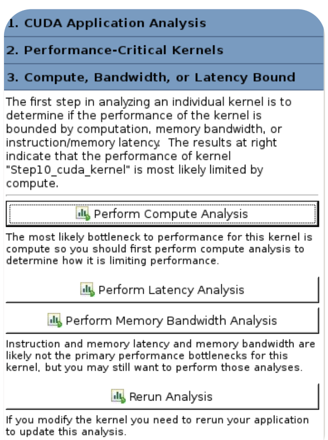
1. Collect profile information
   %> nvprof ./a.out
2. View available metrics
   %> nvprof --query-metrics
3. View global load/store efficiency
   %> nvprof --metrics gld_efficiency,gst_efficiency ./a.out
4. Store a timeline to load in NVVP
   %> nvprof –o profile.timeline ./a.out
5. Store analysis metrics to load in NVVP
   %> nvprof –o profile.metrics --analysis-metrics ./a.out

NVIDIA's Visual Profiler (NVVP)

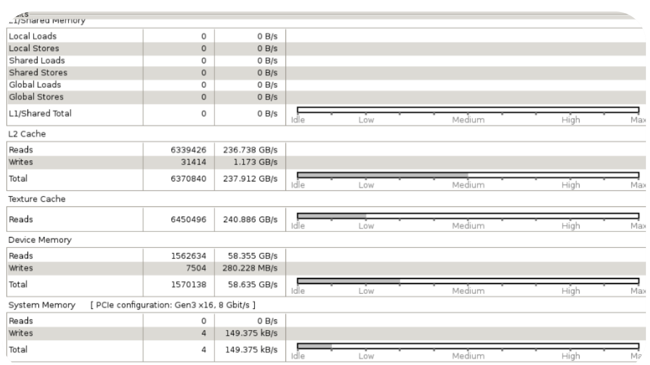
Timeline



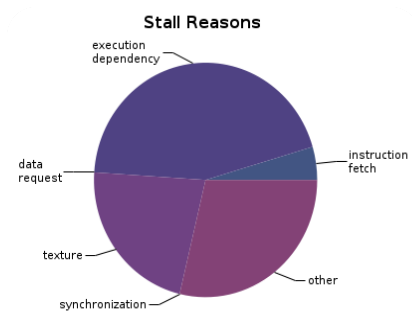
Guided System



Analysis



L1/Shared Memory			
Local Loads	0	0 B/s	
Local Stores	0	0 B/s	
Shared Loads	0	0 B/s	
Shared Stores	0	0 B/s	
Global Loads	0	0 B/s	
Global Stores	0	0 B/s	
L1/Shared Total	0	0 B/s	Idle Low Medium High Max
L2 Cache			
Reads	6339426	236.738 GB/s	
Writes	31414	1.173 GB/s	
Total	6370840	237.912 GB/s	Idle Low Medium High Max
Texture Cache			
Reads	6450496	240.886 GB/s	Idle Low Medium High Max
Device Memory			
Reads	1562634	58.355 GB/s	
Writes	7504	280.228 MB/s	
Total	1570138	58.635 GB/s	Idle Low Medium High Max
System Memory [PCIe configuration: Gen3 x16, 8 Gbit/s]			
Reads	0	0 B/s	
Writes	4	149.375 kB/s	
Total	4	149.375 kB/s	Idle Low Medium High Ma

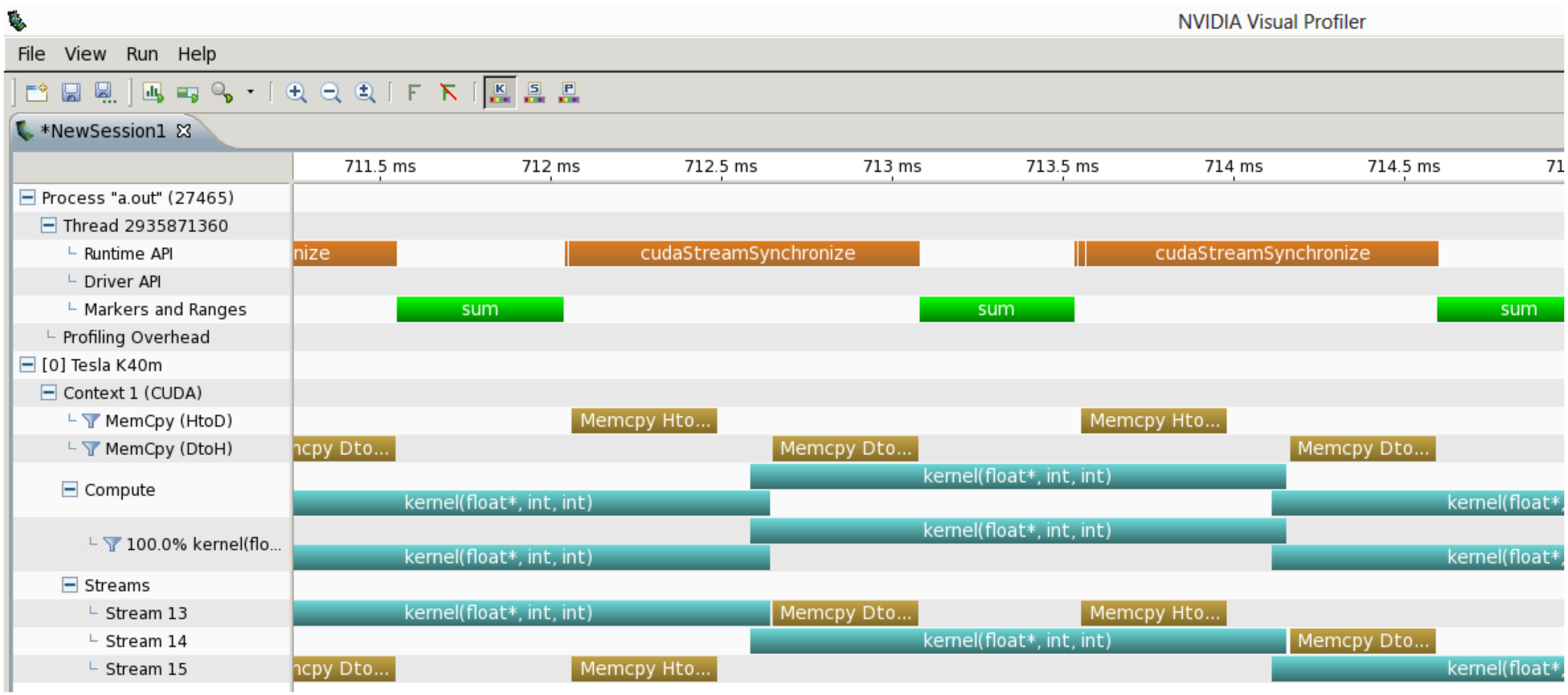


NVTX

- Our current tools only profile API calls on the host
  - What if we want to understand better what the host is doing?
- The NVTX library allows us to annotate profiles with ranges
  - Add: #include <nvToolsExt.h>
  - Link with: -lnvToolsExt
- Mark the start of a range
  - nvtxRangePushA("description");
- Mark the end of a range
  - nvtxRangePop();
- Ranges are allowed to overlap

http://devblogs.nvidia.com/parallelforall/cuda-pro-tip-generate-custom-application-profile-timelines-nvtx/

NVTX Profile



NSIGHT

- CUDA enabled Integrated Development Environment
  - Source code editor: syntax highlighting, code refactoring, etc
  - Build Manger
  - Visual Debugger
  - Visual Profiler
- Linux/Macintosh
  - Editor = Eclipse
  - Debugger = cuda-gdb with a visual wrapper
  - Profiler = NVVP
- Windows
  - Integrates directly into Visual Studio
  - Profiler is NSIGHT VSE

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

QUIZ

Question 1

- If we want to allocate an array of v integer elements in CUDA device global memory, what would be an appropriate expression for the second argument of the cudaMalloc() call?

  a) n
  b) v
  c) n * sizeof(int)
  d) v * sizeof(int)

Question 1

- If we want to allocate an array of *v* integer elements in CUDA device global memory, what would be an appropriate expression for the second argument of the cudaMalloc() call?

a) n
b) v
c) n * sizeof(int)
d) v * sizeof(int)

Question 2

- If we want to allocate an array of *n* floating-point elements and have a floating-point pointer variable *d_A* to point to the allocated memory, what would be an appropriate expression for the first argument of the *cudaMalloc()* call?

  a) n
  b) (void *) d_A
  c) *d_A
  d) (void **) &d_A

Question 2 - Answer

- If we want to allocate an array of *n* floating-point elements and have a floating-point pointer variable *d_A* to point to the allocated memory, what would be an appropriate expression for the first argument of the *cudaMalloc()* call?

  a) n
  b) (void *) d_A
  c) *d_A
  d) (void **) &d_A

Explanation: *&d_A* is pointer to a pointer of *float*. To convert it to a generic pointer required by *cudaMalloc()* should use *(void **)* to cast it to a generic double-level pointer.

Question 3

- If we want to copy 3,000 bytes of data from host array *h_A* (*h_A* is a pointer to element 0 of the source array) to device array *d_A* (*d_A* is a pointer to element 0 of the destination array), what would be an appropriate API call for this in CUDA?

a) cudaMemcpy(3000, h_A, d_A, cudaMemcpyHostToDevice);
b) cudaMemcpy(h_A, d_A, 3000, cudaMemcpyDeviceTHost);
c) cudaMemcpy(d_A, h_A, 3000, cudaMemcpyHostToDevice);
d) cudaMemcpy(3000, d_A, h_A, cudaMemcpyHostToDevice);

Question 3 - Answer

- If we want to copy 3000 bytes of data from host array h_A (h_A is a pointer to element 0 of the source array) to device array d_A (d_A is a pointer to element 0 of the destination array), what would be an appropriate API call for this in CUDA?

a) cudaMemcpy(3000, h_A, d_A, cudaMemcpyHostToDevice);
b) cudaMemcpy(h_A, d_A, 3000, cudaMemcpyDeviceTHost);
c) cudaMemcpy(d_A, h_A, 3000, cudaMemcpyHostToDevice);
d) cudaMemcpy(3000, d_A, h_A, cudaMemcpyHostToDevice);

www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Thank you!

For further information please contact

marc.jorda@bsc.es, antonio.pena@bsc.es