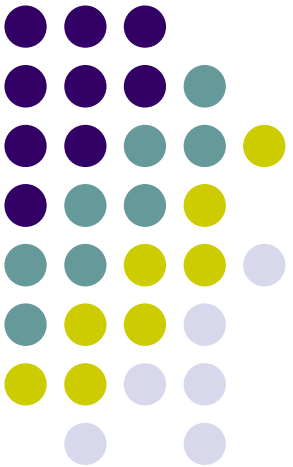
ME759
High Performance Computing for Engineering Applications

Using streams in CUDA
GPU Computing with thrust

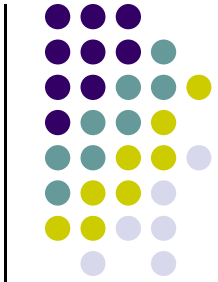
October 18, 2013



"How is education supposed to make me feel smarter? Besides, every time I learn something new, it pushes some old stuff out of my brain. Remember when I took that home winemaking course, and I forgot how to drive?"
-- Homer Simpson

© Dan Negrut, 2013

Before We Get Started…

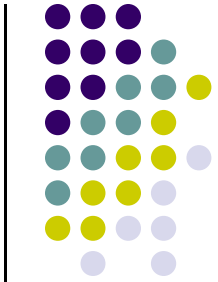


- Last time
  - Wrap up: scan operation
  - Started “streams” in CUDA: hiding data movement with useful computation

- Today:
  - Wrap up, Streams in CUDA
  - GPU computing w/ thrust

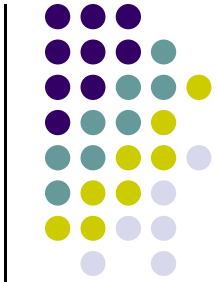
- Miscellaneous
  - HW due on Mo, October 21 at 11:59 pm
  - Start thinking about midterm projects and final projects
    - Due date for midterm project topic is Oct 23
  - Exam moved back from November 8 to November 25 at 7:15 PM (Room TBA)
    - Review session during regular class hour (show up for review only if you think it’s useful)

Example 1: Using One Stream



- Example draws on material presented in the “CUDA By Example” book
  - J. Sanders and E. Kandrot, authors
- What is the purpose of this example?
  - Shows an example of using page-locked (pinned) host memory
  - Shows one strategy that you should invoke when dealing with applications that require more memory than you can accommodate on the GPU
  - [Most importantly] Shows a strategy that you can follow to get things done on the GPU without blocking the CPU (host) – goes back to the use of `cudaMemcpyAsync`
    - While the GPU works, the CPU works too
- Remark:
  - In this example the magic happens on the host side. Focus on host code, not on the kernel executed on the GPU (the kernel code is basically irrelevant)

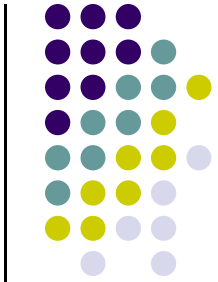
This Example's Kernel



- Computes some average, it's not important, simply something that gets done and allows us later on to gauge efficiency gains when using *multiple* streams (for now dealing with one stream only)
  - Inputs: a and b
  - Output: c

```
#include "../common/book.h"

#define N   1048576  // this is 1024*1024
#define FULL_DATA_SIZE   (N*20)

__global__ void kernel( int *a, int *b, int *c ) {
    int idx = threadIdx.x + blockIdx.x * blockDim.x;
    if (idx < N) {
        int idx1 = (idx + 1) % 256;
        int idx2 = (idx + 2) % 256;
        float   as = (a[idx] + a[idx1] + a[idx2]) / 3.0f;
        float   bs = (b[idx] + b[idx1] + b[idx2]) / 3.0f;
        c[idx] = (as + bs) / 2;
    }
}
```

The “main()” Function



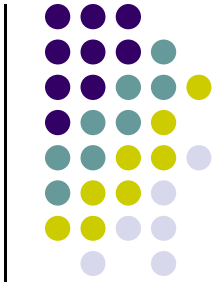
```
01| int main( void ) {
02|     cudaEvent_t     start, stop;
03|     float           elapsedTime;
04|
05|     cudaStream_t    stream;
06|     int *host_a, *host_b, *host_c;
07|     int *dev_a, *dev_b, *dev_c;
08|
09|     // start the timers
10|     HANDLE_ERROR( cudaEventCreate( &start ) );
11|     HANDLE_ERROR( cudaEventCreate( &stop ) );
12|
13|     // initialize the stream; only one stream for now…
14|     HANDLE_ERROR( cudaStreamCreate( &stream ) );
15|
16|     // allocate the memory on the GPU
17|     HANDLE_ERROR( cudaMalloc( (void**)&dev_a, N * sizeof(int) ) );
18|     HANDLE_ERROR( cudaMalloc( (void**)&dev_b, N * sizeof(int) ) );
19|     HANDLE_ERROR( cudaMalloc( (void**)&dev_c, N * sizeof(int) ) );
20|
21|     // allocate host pinned memory, used to stream
22|     HANDLE_ERROR( cudaHostAlloc( (void**)&host_a, FULL_DATA_SIZE * sizeof(int), cudaHostAllocDefault ) );
23|     HANDLE_ERROR( cudaHostAlloc( (void**)&host_b, FULL_DATA_SIZE * sizeof(int), cudaHostAllocDefault ) );
24|     HANDLE_ERROR( cudaHostAlloc( (void**)&host_c, FULL_DATA_SIZE * sizeof(int), cudaHostAllocDefault ) );
25|
26|     for (int i=0; i<FULL_DATA_SIZE; i++) {
27|         host_a[i] = rand();
28|         host_b[i] = rand();
29|     }
```

Stage 1

Stage 2

The “main()” Function

[Cntd.]



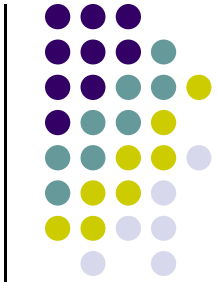
```
30|
31|    HANDLE_ERROR( cudaEventRecord( start, 0 ) );
32|    // now loop over full data, in bite-sized chunks
33|    for (int i=0; i<FULL_DATA_SIZE; i+= N) {
34|        // copy the locked memory to the device, async
35|        HANDLE_ERROR( cudaMemcpyAsync( dev_a, host_a+i, N * sizeof(int), cudaMemcpyHostToDevice, stream ) );
36|        HANDLE_ERROR( cudaMemcpyAsync( dev_b, host_b+i, N * sizeof(int), cudaMemcpyHostToDevice, stream ) );
37|
38|        kernel<<<(N+255)/256,256,0,stream>>>( dev_a, dev_b, dev_c );
39|
40|        // copy the data from device to locked memory
41|        HANDLE_ERROR( cudaMemcpyAsync( host_c+i, dev_c, N * sizeof(int), cudaMemcpyDeviceToHost, stream ) );
42|
43|    }
44|
45|    HANDLE_ERROR( cudaStreamSynchronize( stream ) );
46|
47|    HANDLE_ERROR( cudaEventRecord( stop, 0 ) );
48|
49|    HANDLE_ERROR( cudaEventSynchronize( stop ) );
50|    HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime, start, stop ) );
51|    printf( "Time taken:  %3.1f ms\n", elapsedTime );
52|
53|    // cleanup the streams and memory
54|    HANDLE_ERROR( cudaFreeHost( host_a ) );
55|    HANDLE_ERROR( cudaFreeHost( host_b ) );
56|    HANDLE_ERROR( cudaFreeHost( host_c ) );
57|    HANDLE_ERROR( cudaFree( dev_a ) );
58|    HANDLE_ERROR( cudaFree( dev_b ) );
59|    HANDLE_ERROR( cudaFree( dev_c ) );
60|    HANDLE_ERROR( cudaStreamDestroy( stream ) );
61|
62|    return 0;
63|}
```

Stage 3

Stage 4

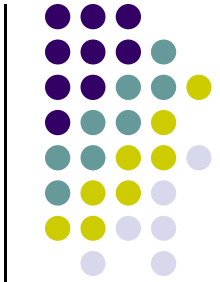
Stage 5

Example 1, Summary



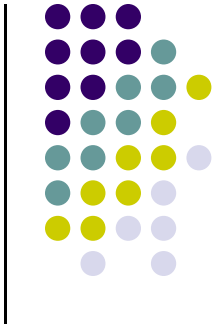
- Stage 1 sets up the events needed to time the execution of the program
- Stage 2 allocates page-locked memory on the host side so that we can fall back on asynchronous memory copy operations between host and device
- Stage 3 enques the set of GPU operations that need to be undertaken (the “chunkification”)
- Stage 4 needed for timing reporting
- Stage 5: clean up time

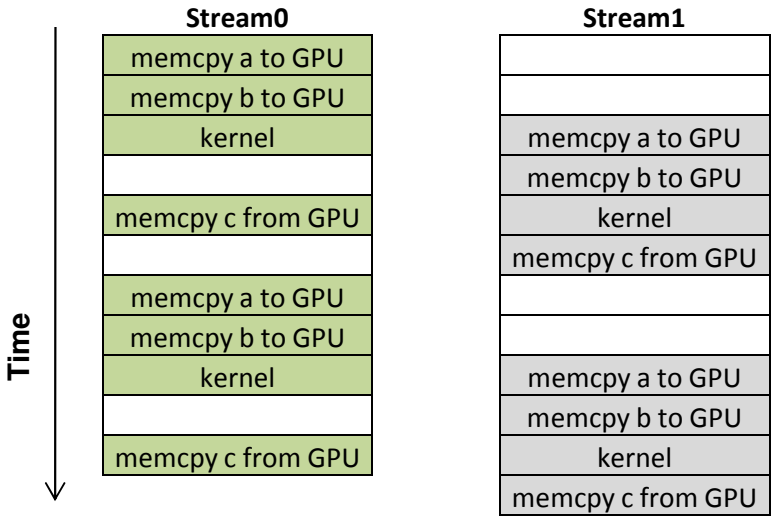
Example 2: Using Multiple Streams
[Version 2.1]



- Implement the same example but use two streams to this end
- Why would you want to use multiple streams?
  - Overlapping GPU execution with host $\leftrightarrow$ device data movement can improve overall performance
- Two ideas underlie the process
  - The idea of "chunkification" of the computation
    - Large computation is broken into pieces that are queued up for execution on the device (we already saw this in Example 1, which uses one stream)
  - The idea of overlapping execution with host $\leftrightarrow$ device data movement
  - NOTE: I didn't want to call this tiling, although it's similar to that. However, "tiling" is something that happens exclusively on the device (from global to shared memory). Here, the "chunkification" happens on the host

Overlapping Execution and Data Transfer: A Desirable Scenario

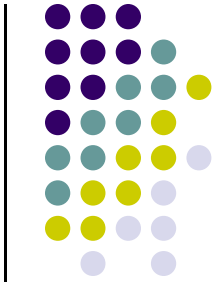




Timeline of intended application execution using two independent streams

- Observations:
  - "memcpy" actually represents an asynchronous `cudaMemcpyAsync()` memory copy call
  - White (empty) boxes represent time when one stream is waiting to execute an operation that it cannot overlap with the other stream's operation
  - The goal: keep both GPU engine types (execution and mem copy) busy
    - Note: recent hardware allows two copies to take place simultaneously: one from host to device, at the same time one goes on from device to host (you have two copy subengines)

The "main()" Function, Two Streams



```
01| int main( void ) {
02|     cudaDeviceProp  prop;
03|     int whichDevice;
04|     HANDLE_ERROR( cudaGetDevice( &whichDevice ) );
05|     HANDLE_ERROR( cudaGetDeviceProperties( &prop, whichDevice ) );
06|     if (!prop.deviceOverlap) {
07|         printf( "Device will not handle overlaps, so no speed up from streams\n" );
08|         return 0;
09|     }
```

Stage 1

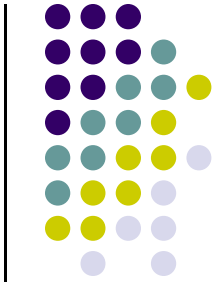
```
10|
11|     cudaEvent_t     start, stop;
12|     float           elapsedTime;
13|
14|     cudaStream_t    stream0, stream1;
15|     int *host_a, *host_b, *host_c;
16|     int *dev_a0, *dev_b0, *dev_c0;
17|     int *dev_a1, *dev_b1, *dev_c1;
18|
19|     // start the timers
20|     HANDLE_ERROR( cudaEventCreate( &start ) );
21|     HANDLE_ERROR( cudaEventCreate( &stop ) );
22|
23|     // initialize the streams
24|     HANDLE_ERROR( cudaStreamCreate( &stream0 ) );
25|     HANDLE_ERROR( cudaStreamCreate( &stream1 ) );
26|
```

Stage 2

Stage 3

```
27|     // allocate the memory on the GPU
28|     HANDLE_ERROR( cudaMalloc( (void**)&dev_a0, N * sizeof(int) ) );
29|     HANDLE_ERROR( cudaMalloc( (void**)&dev_b0, N * sizeof(int) ) );
30|     HANDLE_ERROR( cudaMalloc( (void**)&dev_c0, N * sizeof(int) ) );
31|     HANDLE_ERROR( cudaMalloc( (void**)&dev_a1, N * sizeof(int) ) );
32|     HANDLE_ERROR( cudaMalloc( (void**)&dev_b1, N * sizeof(int) ) );
33|     HANDLE_ERROR( cudaMalloc( (void**)&dev_c1, N * sizeof(int) ) );
34|
35|     // allocate host locked memory, used to stream
36|     HANDLE_ERROR( cudaHostAlloc( (void**)&host_a, FULL_DATA_SIZE * sizeof(int), cudaHostAllocDefault ) );
37|     HANDLE_ERROR( cudaHostAlloc( (void**)&host_b, FULL_DATA_SIZE * sizeof(int), cudaHostAllocDefault ) );
38|     HANDLE_ERROR( cudaHostAlloc( (void**)&host_c, FULL_DATA_SIZE * sizeof(int), cudaHostAllocDefault ) );
```

The "main()" Function, Two Streams [Cntd.]



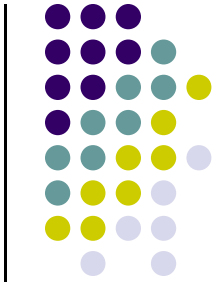
Still Stage 3

```
39|    for (int i=0; i<FULL_DATA_SIZE; i++) {
40|        host_a[i] = rand();
41|        host_b[i] = rand();
42|    }
43|
44|    HANDLE_ERROR( cudaEventRecord( start, 0 ) );
```

Stage 4

```
45|    // now loop over full data, in bite-sized chunks
46|    for (int i=0; i<FULL_DATA_SIZE; i+= N*2) {
47|        // copy the locked memory to the device, async
48|        HANDLE_ERROR( cudaMemcpyAsync( dev_a0, host_a+i, N * sizeof(int), cudaMemcpyHostToDevice, stream0 ) );
49|        HANDLE_ERROR( cudaMemcpyAsync( dev_b0, host_b+i, N * sizeof(int), cudaMemcpyHostToDevice, stream0 ) );
50|
51|        kernel<<<(N+255),256,0,stream0>>>( dev_a0, dev_b0, dev_c0 );
52|
53|        // copy the data from device to locked memory
54|        HANDLE_ERROR( cudaMemcpyAsync( host_c+i, dev_c0, N * sizeof(int), cudaMemcpyDeviceToHost, stream0 ) );
55|
56|
57|        // copy the locked memory to the device, async
58|        HANDLE_ERROR( cudaMemcpyAsync( dev_a1, host_a+i+N, N * sizeof(int), cudaMemcpyHostToDevice, stream1 ) );
59|        HANDLE_ERROR( cudaMemcpyAsync( dev_b1, host_b+i+N, N * sizeof(int), cudaMemcpyHostToDevice, stream1 ) );
60|
61|        kernel<<<N/256,256,0,stream1>>>( dev_a1, dev_b1, dev_c1 );
62|
63|        // copy the data from device to locked memory
64|        HANDLE_ERROR( cudaMemcpyAsync( host_c+i+N, dev_c1, N * sizeof(int), cudaMemcpyDeviceToHost, stream1 ) );
65| }
66|
```

The "main()" Function, Two Streams
[Cntd.]

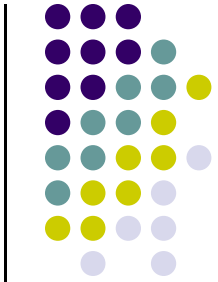


```
67|    HANDLE_ERROR( cudaStreamSynchronize( stream0 ) );
68|    HANDLE_ERROR( cudaStreamSynchronize( stream1 ) );
69|
70|    HANDLE_ERROR( cudaEventRecord( stop, 0 ) );
71|
72|    HANDLE_ERROR( cudaEventSynchronize( stop ) );
73|    HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime, start, stop ) );
74|    printf( "Time taken:  %3.1f ms\n", elapsedTime );
75|
76|    // cleanup the streams and memory
77|    HANDLE_ERROR( cudaFreeHost( host_a ) );
78|    HANDLE_ERROR( cudaFreeHost( host_b ) );
79|    HANDLE_ERROR( cudaFreeHost( host_c ) );
80|    HANDLE_ERROR( cudaFree( dev_a0 ) );
81|    HANDLE_ERROR( cudaFree( dev_b0 ) );
82|    HANDLE_ERROR( cudaFree( dev_c0 ) );
83|    HANDLE_ERROR( cudaFree( dev_a1 ) );
84|    HANDLE_ERROR( cudaFree( dev_b1 ) );
85|    HANDLE_ERROR( cudaFree( dev_c1 ) );
86|    HANDLE_ERROR( cudaStreamDestroy( stream0 ) );
87|    HANDLE_ERROR( cudaStreamDestroy( stream1 ) );
88|
89|    return 0;
90|}
```

Stage 5

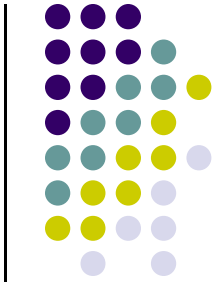
NOTE: the kernel doesn't actually change…

Example 2.1 [Version 1], Summary



- Stage 1 ensures that your device supports your attempt to overlap kernel execution with host↔device data transfer
- Stage 2 sets up the events needed to time the execution of the program
- Stage 3 allocates page-locked memory on the host side so that we can fall back on asynchronous memory copy operations between host and device and initializes data
- Stage 4 enques the set of GPU operations that need to be undertaken (the “chunkification”)
- Stage 5 takes care of timing reporting and clean up

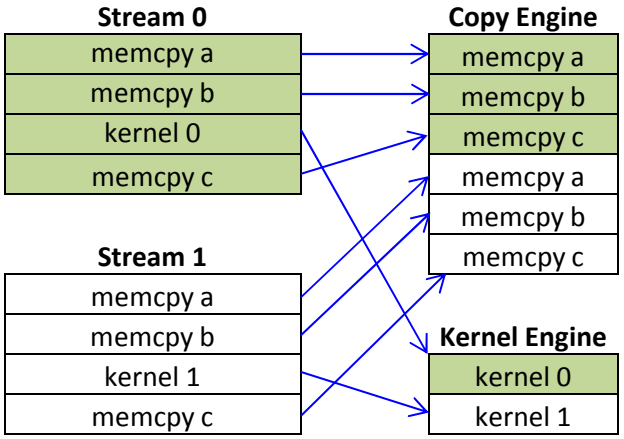
Comments, Using Two Streams
[Version 2.1]



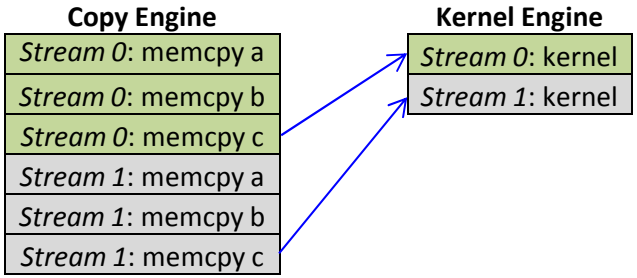
- Timing results provided by “CUDA by Example: An Introduction to General-Purpose GPU Programming,”
  - Sanders and Kandrot reported results on NVIDIA GTX285
- Using one stream (in Example 1): 62 ms
- Using two streams (this example, version 1): 61 ms
- Lackluster performance goes back to the way the two GPU engines (kernel execution and copy) are scheduled

The Two Stream Example, Version 2.1

Looking Under the Hood



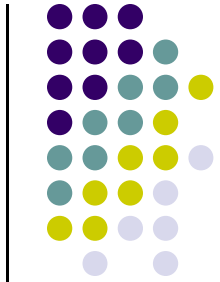
Mapping of CUDA streams onto GPU engines

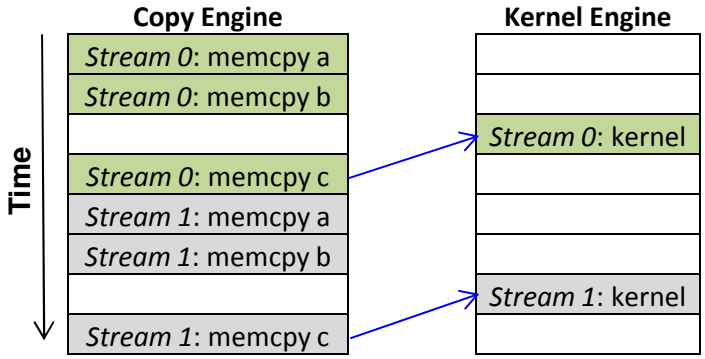


Arrows depicting the dependency of cudaMemcpyAsync () calls on kernel executions in the 2 Streams example

- At the left:
  - An illustration of how the work queued up in the streams ends up being assigned by the CUDA driver to the two GPU engines (copy and execution)
  - Important remark: FIFO is also observed in relation to scheduling the engines (not only the streams)
- At the right
  - Image shows dependency that is implicitly set up in the two streams given the way the streams were defined in the code
  - The queue in the Copy Engine, combined with the implied dependencies determines the scheduling of the Copy and Kernel Engines (see next slide)

The Two Stream Example
Looking Under the Hood



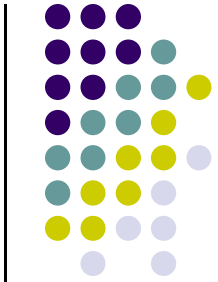


Execution timeline of the 2 Stream example (blue line shows dependency; empty boxes represent idle segments)

- Note that due to the *specific* way in which the streams were defined (depth first), basically there is no overlap of copy & execution…
  - Explains the no net-gain in efficiency compared to the one stream example
- Remedy: go breadth first, instead of depth first
  - In the current version, execution on the two engines was inadvertently blocked by the way the streams have been set up and the existing scheduling and lack of dependency checks available in the current version of CUDA

The Two Stream Example

[Version 2.2: A More Effective Implementation: Breadth First]

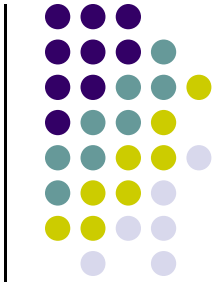


- Old way (the depth first approach):
  - Assign the copy of a, copy of b, kernel execution, and copy of c to stream0. Subsequently, do the same for stream1

- New way (the breadth first approach):
  - Add the copy of a to stream0, and then add the copy of a to stream1
  - Next, add the copy of b to stream0, and then add the copy of b to stream1
  - Next, enqueue the kernel invocation in stream0, then enqueue one in stream1.
  - Finally, enqueue the copy of c back to the host in stream0 followed by the copy of c in stream1.

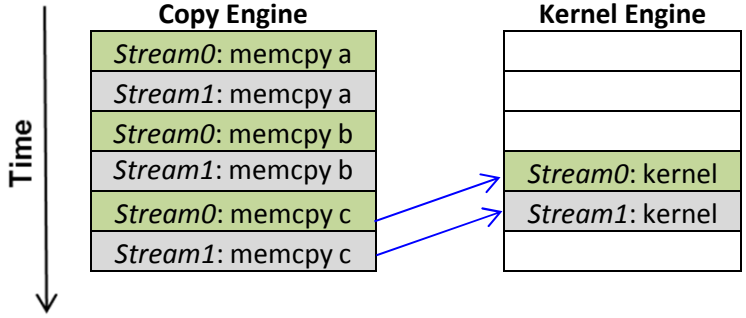
The Two Stream Example

A 20% More Effective Implementation (48 vs. 61 ms)



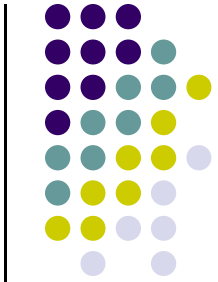
```
A|     // now loop over full data, in bite-sized chunks
B|    for (int i=0; i<FULL_DATA_SIZE; i+= N*2) {
C|       // enqueue copies of a in stream0 and stream1
D|        HANDLE_ERROR( cudaMemcpyAsync( dev_a0, host_a+i,   N * sizeof(int), cudaMemcpyHostToDevice, stream0 ) );
E|        HANDLE_ERROR( cudaMemcpyAsync( dev_a1, host_a+i+N, N * sizeof(int), cudaMemcpyHostToDevice, stream1 ) );
F|        // enqueue copies of b in stream0 and stream1
G|        HANDLE_ERROR( cudaMemcpyAsync( dev_b0, host_b+i,   N * sizeof(int), cudaMemcpyHostToDevice, stream0 ) );
H|        HANDLE_ERROR( cudaMemcpyAsync( dev_b1, host_b+i+N, N * sizeof(int), cudaMemcpyHostToDevice, stream1 ) );
I|
J|        // enqueue kernels in stream0 and stream1
K|        kernel<<<(N+255),256,0,stream0>>>( dev_a0, dev_b0, dev_c0 );
L|        kernel<<<(N+255),256,0,stream1>>>( dev_a1, dev_b1, dev_c1 );
M|
N|        // enqueue copies of c from device to locked memory
O|        HANDLE_ERROR( cudaMemcpyAsync( host_c+i, dev_c0,   N * sizeof(int), cudaMemcpyDeviceToHost, stream0 ) );
P|        HANDLE_ERROR( cudaMemcpyAsync( host_c+i+N, dev_c1, N * sizeof(int), cudaMemcpyDeviceToHost, stream1 ) );
Q|    }
```

Replaces Previous Stage 4



Execution timeline of the breadth-first approach
(blue line shows dependency)

Using Streams, Lessons Learned



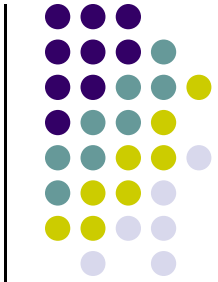
- Streams provide a basic mechanism that enables task-level parallelism in CUDA C applications

- Two requirements underpin the use of streams in CUDA C
  - `cudaHostAlloc()` should be used to allocate memory on the host so that it can be used in conjunction with a `cudaMemcpyAsync()` non-blocking copy command
    - The use of pinned (page-locked) host memory improves data transfer performance even if you only work with one stream

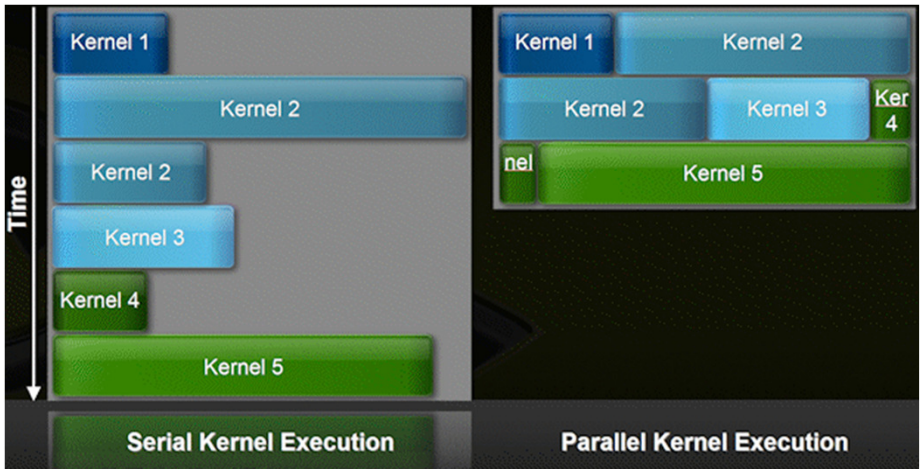
  - Effective latency hiding of kernel execution with memory copy operations requires a breadth-first approach to enqueuing operations in different streams
    - This is a consequence of the two engine setup associated with a GPU

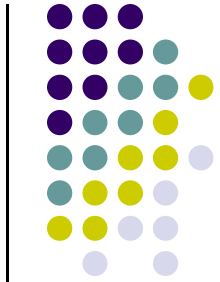
Concurrent Kernel Execution

[one slide detour]



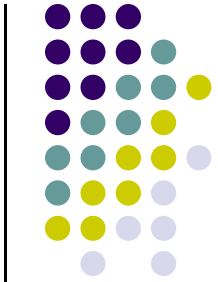
- Fermi: up to 16 kernels can be run on the device at the same time
- When is this useful?
  - Devices of compute capability 2.x are pretty wide (large number of SMs)
  - Sometimes you launch kernels whose execution configuration is smaller than the GPU's "width"
  - Then, two or three independent kernels can be "squeezed" on the GPU at the same time
- Represents the GPU's attempt to look like a MIMD architecture
  - Requires use of multiple streams to stand a chance of concurrent kernel execution





GPU Computing using `thrust`

3 Ways to Accelerate on GPU



Application

Libraries

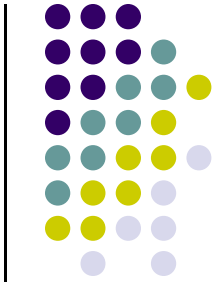
Directives

Programming Languages

Direction of increased performance (and effort)

NVIDIA [C. Woolley]→

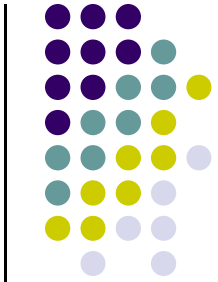
Acknowledgments



- The `thrust` slides include material provided by Nathan Bell of NVIDIA

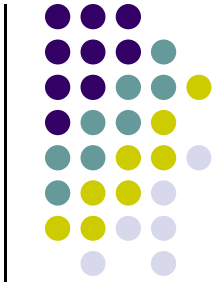
- Slightly modified, assuming responsibility for any mistakes

Design Philosophy, `thrust`



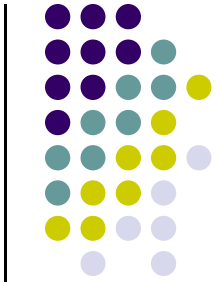
- Increase programmer productivity
  - Build complex applications quickly
- Encourage generic programming
  - Leverage parallel primitives
- Should run fast
  - Efficient mapping to hardware

What is `thrust`?



- A template library for CUDA
  - Mimics the C++ STL
- Containers
  - On host and device
- Algorithms
  - Sorting, reduction, scan, etc.

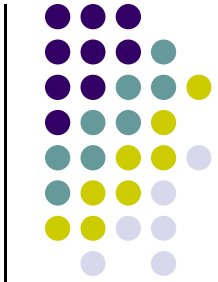
What is `thrust`? [Cntd.]



- `thrust` is a header library – all the functionality is accessed by `#include`-ing the appropriate `thrust` header file
- Program is compiled with `nvcc` as per usual, no special tools are required
- Lots of C++ syntax, related to high-level host-side code that you write
  - The concept of execution configuration, shared memory, etc. : it's all gone

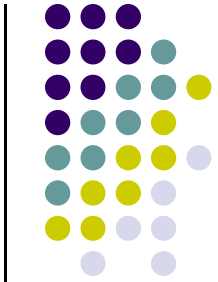
NVIDIA [N. Bell]→

Why Should One Use `thrust`?

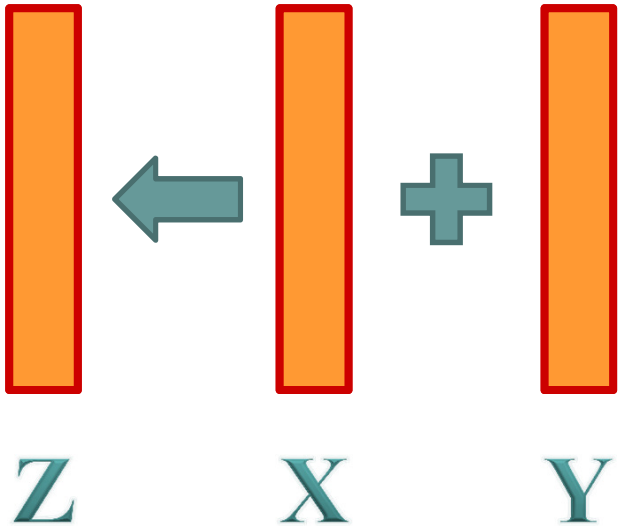


- Extensively tested
- Open Source
  - Permissive License (Apache v2)
- Active community

Example: Vector Addition

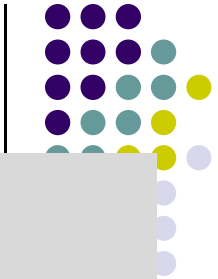


```
for (int i = 0; i < N; i++)
    Z[i] = X[i] + Y[i];
```



NVIDIA [N. Bell]→

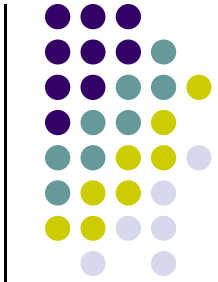
Example, Vector Addition



```
#include <thrust/device_vector.h>
#include <thrust/transform.h>
#include <thrust/functional.h>
#include <iostream>

int main(void) {
  thrust::device_vector<float> X(3);
  thrust::device_vector<float> Y(3);
  thrust::device_vector<float> Z(3);

  X[0] = 10; X[1] = 20; X[2] = 30;
  Y[0] = 15; Y[1] = 35; Y[2] = 10;

  thrust::transform(X.begin(), X.end(),
                    Y.begin(),
                    Z.begin(),
                    thrust::plus<float>());

  for (size_t i = 0; i < Z.size(); i++)
    std::cout << "Z[" << i << "] = " << Z[i] << "\n";

  return 0;
}
```


NVIDIA [N. Bell]→

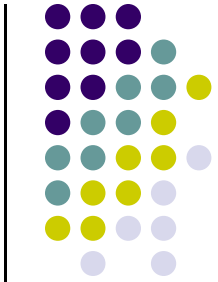
Example, Vector Addition



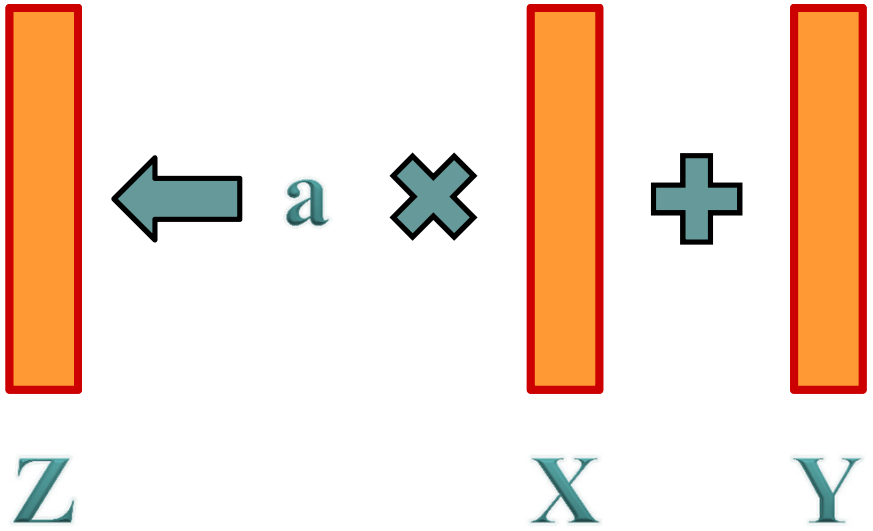
```
[negrut@euler01 CodeBits]$ nvcc --version
nvcc: NVIDIA (R) Cuda compiler driver
Copyright (c) 2005-2011 NVIDIA Corporation
Built on Thu_Jan_12_14:41:45_PST_2012
Cuda compilation tools, release 4.1, V0.2.1221
[negrut@euler01 CodeBits]$ nvcc -O2 exThrust.cu -o exThrust.exe
[negrut@euler01 CodeBits]$ ./exThrust.exe
Z[0] = 25
Z[1] = 55
Z[2] = 40
[negrut@euler01 CodeBits]$
```

- Note: file extension should be **`.cu`**

Example: SAXPY



```
for (int i = 0; i < N; i++)
    Z[i] = a * X[i] + Y[i];
```



NVIDIA [N. Bell]→

SAXPY

functor

```
struct saxpy
{
  float a;

  saxpy(float a) : a(a) {}

  __host__ __device__
  float operator()(float x, float y)
  {
    return a * x + y;
  }
};
```

state

constructor

call operator

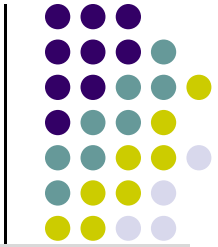
```
int main(void)
{
  thrust::device_vector<float> X(3), Y(3), Z(3);

  X[0] = 10; X[1] = 20; X[2] = 30;
  Y[0] = 15; Y[1] = 35; Y[2] = 10;

  float aVal = 2.0f;

  thrust::transform(X.begin(), X.end(),
                    Y.begin(),
                    Z.begin(),
                    saxpy(aVal));

  for (size_t i = 0; i < Z.size(); i++)
    std::cout << "Z[" << i << "] = " << Z[i] << "\n";

  return 0;
}
```


NVIDIA [N. Bell]→

SAXPY Example [without functor]

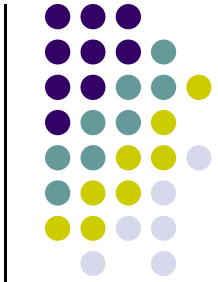
```
#include <thrust/device_vector.h>
#include <thrust/transform.h>
#include <thrust/functional.h>
#include <iostream>

using namespace thrust::placeholders;

int main(void) {
  thrust::device_vector<float> X(3), Y(3), Z(3);

  X[0] = 10; X[1] = 20; X[2] = 30;
  Y[0] = 15; Y[1] = 35; Y[2] = 10;

  float a = 2.0f;

  thrust::transform(X.begin(), X.end(),
                    Y.begin(),
                    Z.begin(),
                    a * _1 + _2);

  for (size_t i = 0; i < Z.size(); i++)
    std::cout << "Z[" << i << "] = " << Z[i] << "\n";

  return 0;
}
```

Diving In



```
#include <thrust/host_vector.h>
#include <thrust/device_vector.h>
#include <thrust/sort.h>

int main(void) {
    // generate 32M random numbers on the host
    thrust::host_vector<int> h_vec(32 << 20);
    thrust::generate(h_vec.begin(), h_vec.end(), rand);

    // transfer data to the device
    thrust::device_vector<int> d_vec = h_vec;

    // sort data on the device (846M keys per sec on GeForce GTX 480)
    thrust::sort(d_vec.begin(), d_vec.end());

    // transfer data back to host
    thrust::copy(d_vec.begin(), d_vec.end(), h_vec.begin());

    return 0;
}
```

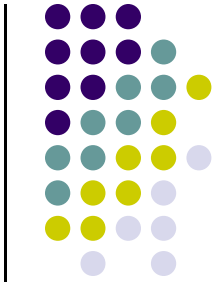
Containers



- Concise and readable code
  - Avoids common memory management errors
    - e.g.: Vectors automatically release memory when they go out of scope

```
// allocate host vector with two elements
thrust::host_vector<int> h_vec(2);

// copy host vector to device
thrust::device_vector<int> d_vec = h_vec;

// write device values from the host
d_vec[0] = 13;
d_vec[1] = 27;

// read device values from the host
std::cout << "sum: " << d_vec[0] + d_vec[1] << std::endl;
```

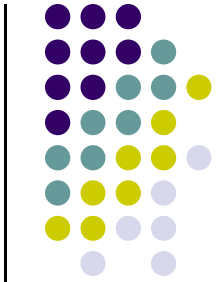
Containers



- Compatible with STL containers

```
// list container on host
std::list<int> h_list;
h_list.push_back(13);
h_list.push_back(27);

// copy list to device vector
thrust::device_vector<int> d_vec(h_list.size());
thrust::copy(h_list.begin(), h_list.end(), d_vec.begin());

// alternative method using vector constructor
thrust::device_vector<int> d_vec2(h_list.begin(), h_list.end());
```

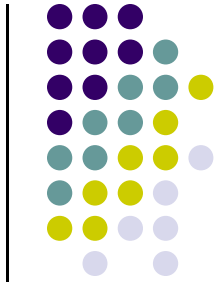
Namespaces



- Avoid name collisions

```
// allocate host memory
thrust::host_vector<int> h_vec(10);

// call STL sort
std::sort(h_vec.begin(), h_vec.end());

// call Thrust sort
thrust::sort(h_vec.begin(), h_vec.end());

// for brevity
using namespace thrust;

// without namespace
int sum = reduce(h_vec.begin(), h_vec.end());
```

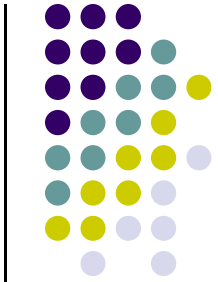
Iterators



- A pair of iterators defines a “range”

```
// allocate device memory
device_vector<int> d_vec(10);

// declare iterator variables
device_vector<int>::iterator begin  = d_vec.begin();
device_vector<int>::iterator end    = d_vec.end();
device_vector<int>::iterator middle = begin + d_vec.size()/2;
// sum first and second halves
int sum_half1 = reduce(begin, middle);
int sum_half2 = reduce(middle, end);

// empty range
int empty = reduce(begin, begin);
```

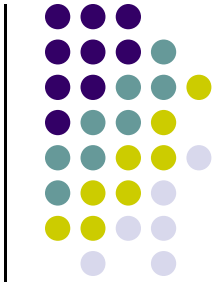
Iterators



- Iterators act like pointers

```
// declare iterator variables
device_vector<int>::iterator begin = d_vec.begin();
device_vector<int>::iterator end   = d_vec.end();

// pointer arithmetic
begin++;

// dereference device iterators from the host
int a = *begin;
int b = begin[3];

// compute size of range [begin,end)
int size = end - begin;
```

Iterators



- Encode memory location
  - Automatic algorithm selection

```
// initialize random values on host
host_vector<int> h_vec(100);
thrust::generate(h_vec.begin(), h_vec.end(), rand);

// copy values to device
device_vector<int> d_vec = h_vec;

// compute sum on host
int h_sum = thrust::reduce(h_vec.begin(), h_vec.end());

// compute sum on device
int d_sum = thrust::reduce(d_vec.begin(), d_vec.end());
```

NVIDIA [N. Bell]→