CSCI-GA.3033-004
Graphics Processing Units (GPUs): Architecture and Programming
CUDA
Advanced Techniques 1

Mohamed Zahran (aka Z)
mzahran@cs.nyu.edu
http://www.mzahran.com
Some Refreshing Exercises!

• Suppose registers and shared memory capacities were not an issue. When is it still beneficial to put values fetched from memory into the shared memory?
Some Refreshing Exercises!

• Assume a kernel is launched with 1000 blocks. Each block has 512 threads.
  – If a variable is declared as local in the kernel, how many versions will be created throughout the lifetime of the kernel?
  – How about if the variable is created as shared?
Some Refreshing Exercises!

- A kernel contains 36 floating point operations and 7 32-bit word global memory accesses per thread. For each of the following device properties, indicate whether this kernel is compute- or memory-bound.
  - Peak FLOPS = 200 GFLOPS, Peak Memory Bandwidth = 100 GB/s
  - Peak FLOPS = 300 GFLOPS, Peak Memory Bandwidth = 250 GB/s
A Note About Performance
Defining Performance

• Which airplane has the best performance?
Performance Considerations

• There are many hardware constraints.
• Depending on the application, different constraints may dominate.
• We can improve performance of an application by trading one resource usage for another.
Performance Issue: Thread Divergence
Performance Issue:
Thread Diversion (aka Branch Divergence)

• Things are OK when all threads in a warp follow the same control-flow
• Performance loss due to thread diversion
Performance Issue: Thread Diversion

1. `__shared__ float partialSum[]`

2. `unsigned int t = threadIdx.x;
3. for (unsigned int stride = 1;
4.          stride < blockDim.x; stride *= 2)
5. {
6.     syncThreads();
7.     if (t % (2*stride) == 0)
8.         partialSum[t] += partialSum[t+stride];
9. }

Example: Sum Reduction Kernel
Performance Issue: Thread Diversion

1. __shared__ float partialSum[];
2. unsigned int t = threadIdx.x;
3. for (unsigned int stride = blockDim.x >> 1; stride > 0; stride >>= 1)
4. { __syncthreads();
5.   if (t < stride)
6.     partialSum[t] += partialSum[t+stride];
7. }
8. partialSum[0] = __sync_val_add(&partialSum[1], partialSum[0]);
9. ...

Why is this version better than the previous one?

Example: Sum Reduction Kernel
Performance Issue: Global Memory
Performance Issue: Global Memory

• Typical application: process massive amount of data within short period of time
  – From global memory
  – large amount + short period = huge bandwidth requirement

• Two main challenges regarding global memory:
  – Long latency
  – Relatively limited bandwidth
Dealing With Global Memory: TILING

• We have seen this before
• Make use of shared memory available in SMs to reduce trips to global memory
Dealing With Global Memory: Coalescing

• To more effectively move data from global memory to shared memory and registers
• For best results: can be used with tiling
• Global memory:
  – DRAM
  – Reading a bit is slow
  – So memory is implemented to read several bits in parallel

General rule:
For global memory, the more scattered the addresses are, the more reduced the throughput it.
Dealing With Global Memory: Coalescing

• If an application can make use of data from multiple consecutive locations, the DRAM can supply the data in much higher rate.

• Device memory (i.e. global memory) is accessed via 32-, 64-, or 128-byte memory transactions.
  – Those transactions must be aligned to 32-, 64, or 128-byte segment.

• Kernel must arrange its data access accordingly
Dealing With Global Memory: Coalescing

• When all threads in a warp execute a load instruction:
  – The hardware detects whether the addresses are consecutive.
  – The hardware combines (coalesces) all accesses in a consolidated access to consecutive DRAM locations

• More coalesces → less memory transactions → higher throughput → better performance
Dealing With Global Memory: Coalescing

Not coalesced

coalesced

Same warp

Thread 1

Thread 2

Linearized order in increasing address
Dealing With Global Memory: Coalescing

Favorable access direction in Kernel code of one thread.
Dealing With Global Memory: Coalescing

Access direction in Kernel code

Time Period 1

<table>
<thead>
<tr>
<th>M₀₀</th>
<th>M₁₀</th>
<th>M₀₁</th>
<th>M₂₀</th>
<th>M₃₀</th>
</tr>
</thead>
<tbody>
<tr>
<td>M₀₁</td>
<td>M₁₁</td>
<td>M₂₁</td>
<td>M₃₁</td>
<td></td>
</tr>
<tr>
<td>M₀₂</td>
<td>M₁₂</td>
<td>M₂₂</td>
<td>M₃₂</td>
<td></td>
</tr>
<tr>
<td>M₀₃</td>
<td>M₁₃</td>
<td>M₂₃</td>
<td>M₃₃</td>
<td></td>
</tr>
</tbody>
</table>

Time Period 2

<table>
<thead>
<tr>
<th>M₀₀</th>
<th>M₁₀</th>
<th>M₀₁</th>
<th>M₂₀</th>
<th>M₃₀</th>
</tr>
</thead>
<tbody>
<tr>
<td>M₀₁</td>
<td>M₁₁</td>
<td>M₂₁</td>
<td>M₃₁</td>
<td></td>
</tr>
<tr>
<td>M₀₂</td>
<td>M₁₂</td>
<td>M₂₂</td>
<td>M₃₂</td>
<td></td>
</tr>
<tr>
<td>M₀₃</td>
<td>M₁₃</td>
<td>M₂₃</td>
<td>M₃₃</td>
<td></td>
</tr>
</tbody>
</table>

...
Dealing With Global Memory: Coalescing: Programming Rules

• If you access words of size 1, 2, 4, 8, or 16 bytes and are aligned → compiler translates this to 1 global memory instruction.
  – Built-in data types fulfill this.
• If size and alignment are not fulfilled → compiles to multiple global memory access instructions.
• Less global memory instructions → more opportunities to coalesce.
• You can enforce the compiler to align using __align(x)__
  – Example: struct __align(16)__{ .... };
Dealing With Global Memory: Cache

- Fermi (and later GPUs) have cache for global memory
- Caches automatically coalesce most of kernel access patterns
Dealing With Global Memory: Prefetching

Prefetch next data elements while consuming the current data elements.
This increases the number of independent instructions between memory accesses and consumers

<table>
<thead>
<tr>
<th>A</th>
<th>Without prefetching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loop</td>
<td></td>
</tr>
<tr>
<td>Load current tile to shared memory</td>
<td></td>
</tr>
<tr>
<td>__syncthreads()</td>
<td></td>
</tr>
<tr>
<td>Compute current tile</td>
<td></td>
</tr>
<tr>
<td>__syncthreads()</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B</th>
<th>With prefetching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load first tile from global memory into registers</td>
<td></td>
</tr>
<tr>
<td>Loop</td>
<td></td>
</tr>
<tr>
<td>Deposit tile from registers to shared memory</td>
<td></td>
</tr>
<tr>
<td>__syncthreads()</td>
<td></td>
</tr>
<tr>
<td>Load next tile from global memory into registers</td>
<td></td>
</tr>
<tr>
<td>Compute current tile</td>
<td></td>
</tr>
<tr>
<td>__syncthreads()</td>
<td></td>
</tr>
</tbody>
</table>
Performance Issue: SM Resources
Performance Issue: SM Resources

- Execution resources in SM include:
  - registers
  - block slots
  - thread slots
  - shared memory

- There is an interaction among the resources that you must take into account.
**Performance Issue:**

**SM Resources**

**Example:** Assume G80 is executing the matrix multiplication with 16x16 thread blocks  
(G80 SM: 8 block slots, 768 thread slots, 8192 registers)

If a thread needs 10 registers then:

- A block needs 10x16x16 = 2560 registers  
  By using 1 extra variable the program saw a 1/3 reduction in warp parallelism
- 3 blocks -> 7680 registers (under the 8192 limit)  
- We can’t add another block (will make it 10240)  
- 3 blocks x 256 threads/block = 768 (within limit)  
  -> **performance cliff**

Assume the programmer declares one more auto var:

- 11 x 16 x 16 = 2816 registers per block  
  - 3 blocks -> 3 x 2816 = 8448 (above limit)
- SM reduces #blocks by 1 -> 5632 registers required  
- This reduces the number of threads in SM to 2 x256 -> 512
Performance Issue: SM Resources

Example: Still with G80:
• An instruction takes 4 cycles
• Assume 4 independent instructions between global memory load and its use
• Global memory latency is 200 cycles

To keep execution units fully utilized:
We need to have \( \frac{200}{(4 \times 4)} = 13 \) warps

Assume an extra register allows the programmer to use a transformation to increase independent instructions from 4 to 8, then:
• Now we need \( \frac{200}{(4 \times 8)} = 7 \) warps
• Blocks reduced from 3 to 2 \( \rightarrow \) warps reduced from 24 to 16
• Still we can fully utilize execution units

Trading thread-level parallelism with increased thread performance
Performance Issue:
Instruction Mix
Performance Issue: Instruction Mix

```
for (int k = 0; k < BLOCK_SIZE; ++k)
Pvalue += Ms[ty][k] * Ns[k][tx];
```

Is the above code efficient?

- Extra instructions to update loop counter
- Extra instructions for conditional branch at the end of each iteration
- Using k to access matrices incurs address arithmetic instructions.
- All of the above compete with the floating-point calculations for limited instruction processing bandwidth.

<table>
<thead>
<tr>
<th>2 FP arithmetic</th>
<th>2 address arithmetic instructions</th>
<th>only 1/3 instructions are FP operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 loop branch instructions</td>
<td>1 loop increment instructions</td>
<td></td>
</tr>
</tbody>
</table>
Performance Issue: Instruction Mix

for (int k = 0; k < BLOCK_SIZE; ++k)
    Pvalue += Ms[ty][k] * Ns[k][tx];

Pvalue += Ms[ty][0] * Ns[0][tx] + ...
    Ms[ty][15] * Ns[15][tx];

Loop unrolling
Performance Issue: Thread Granularity
Performance Issue: Thread Granularity

- Algorithmic decision
- It is often advantageous to put more work into each thread and use fewer threads
  - When redundant work exists between threads
  - Example: Let a thread compute 2 tiles
    + Less redundant work
    + Potentially more independent instructions
    - More resources requirements
Performance Issue: Putting It All Together
Putting It All Together

Thread Granularity:
- normal
- merging 2 blocks
- merging 4 blocks
Until tile reaches 16x16 neither loop unrolling nor data prefetch helps.

For small tile size, global memory bandwidth severely limits performance.
Putting It All Together

Granularity adjustment can reduce global memory access.
Data prefetching becomes less beneficial as thread granularity increases.
About Algorithms for GPUs

• When designing an algorithm for GPU, the two main characteristics that determine its performance are:
  1. How much parallelism is available
  2. How much data must move through the memory hierarchy

• Then when you move from algorithm to code you need to take hardware constraints into account.
Revisiting Shared Memory
Till now

• We know that we have to use `__shared__`
• But, you leave it to the runtime system to distribute shared memory among blocks. Can you change that?
• What if you don’t know how much shared memory we need before execution?
If we know the size beforehand

```c
__global__ void staticShared(int *d, int n)
{
    __shared__ int array[64];
    int t = threadIdx.x;
    s[t] = d[t];
}
```

What if you don’t know this size beforehand?
Dynamic Shared Memory

```cpp
__global__ void DynamicShared(int *d, int n)
{
    extern __shared__ int array[];
    int t = threadIdx.x;
    s[t] = d[t];
}

main()
{
    ... 
    DynamicShared<<<100, 100, N*sizeof(int)>>>(int *, int);
}
```

Step 1:
A different declaration with no size in the kernel

Step 2:
The 3rd argument of kernel launch is the shared memory allocation size per block in bytes.
What if you need more than one array in the shared memory?

• Declare a single extern unsized array (as in the previous slide).
• Use pointers into it to divide it into multiple arrays.
• Example:

```c
extern __shared__ int s[];
int *integerData = s;   // nI ints
float *floatData = (float*)integerData[nI];  // nF floats
char *charData = (char*)floatData[nF+nI];     // nC chars
```
Potential Performance Loss: Shared Memory Bank Conflicts

• Shared memory is divided into equally sized memory modules called banks.
• Banks can be accessed simultaneously.
• Shared memory accesses that span $b$ distinct banks yield an effective bandwidth that is $b$ times as high as the bandwidth of when accesses map to the same bank.
• Exceptions are: broadcast and multicast.
Potential Performance Loss: 
**Shared Memory Bank Conflicts**

- Shared memory banks are organized such that successive 32-bit words are assigned to successive banks.
- Bandwidth is 32 bits per bank per clock cycle.
- Warp size is 32
Potential Performance Loss: Shared Memory Bank Conflicts

- Now, depending on compute capability:
  - 1.x: the number of banks is 16. A shared memory request for a warp is split into one request for the first half of the warp and one request for the second half of the warp.
  - 2.x: the number of banks is 32. A shared memory request for a warp is not split.
  - 3.x: have configurable bank size (next slide)
cudaDeviceSetSharedMemConfig(i)

- "i" above can be:
  - cudaSharedMemBankSizeFourByte ← default
  - cudaSharedMemBankSizeEightByte

- This is part of the CUDA driver API, so you need to include cuda_runtime.h
Floating Points
Importance of Floating Points

- Many graphics operations are floating point operations
- GPU performance is measured in GFLOPS
Turing Award 1989 to William Kahan for design of the IEEE Floating Point Standards 754 (binary) and 854 (decimal)
Floating Point

• We need a way to represent
  - numbers with fractions, e.g., 3.1416
  - very small numbers, e.g., .000000001
  - very large numbers, e.g., 3.15576 \times 10^9

• Representation:
  - sign, exponent, mantissa: \((-1)^{\text{sign}} \times \text{mantissa} \times 2^{\text{exponent}}\)
  - more bits for mantissa gives more accuracy
  - more bits for exponent increases range

• IEEE 754 floating point standard:
  - single precision: 8 bit exponent, 23 bit mantissa
  - double precision: 11 bit exponent, 52 bit mantissa
IEEE 754 floating-point standard

- Leading “1” bit of significand is implicit (called hidden 1 technique, except when exp = -127)
- Exponent is “biased” to make sorting easier
  - all 0s is smallest exponent
  - all 1s is largest exponent
  - bias of 127 for single precision and 1023 for double precision
  - summary: $(-1)^{\text{sign}} \times (1+\text{significand}) \times 2^{\text{exponent} - \text{bias}}$

- Example:
  - decimal: -.75 = - ( $\frac{1}{2} + \frac{1}{4}$ )
  - binary: -.11 = -1.1 x 2^{-1}
  - floating point: exponent = 126 = 01111110
  - IEEE single precision: 10111111010000000000000000000000
More about IEEE floating Point Standard

Single Precision:

\[-1^{\text{sign}} \times (1 + \text{mantissa}) \times 2^{\text{exponent} - 127}\]

The variables shown in red are the numbers stored in the machine.
Floating Point Example

what is the decimal equivalent of

1  01110110  10110000...0
Based on exp
we have 3 encoding schemes

- exp ≠ 0..0 or 11...1 \(\rightarrow\) normalized encoding
- exp = 0...000 \(\rightarrow\) denormalized encoding
- exp = 1111...1 \(\rightarrow\) special value encoding
  - frac = 000...0
  - frac = something else
1. Normalized Encoding

- **Condition**: \( exp \neq 000\ldots0 \) and \( exp \neq 111\ldots1 \)

  referred to as Bias

- **Exponent is**: \( E = \text{Exp} - (2^{k-1} - 1) \), \( k \) is the \# of exponent bits
  - Single precision: \( E = \text{exp} - 127 \)
  - Double precision: \( E = \text{exp} - 1023 \)

- **Significand is**: \( M = 1.xxx\ldots x_2 \)
  - \( \text{Range}(M) = [1.0, 2.0-\varepsilon) \)
  - Get extra leading bit for free
Normalized Encoding Example

• Value: \( \text{Float } F = 15213.0; \)
  \(- 15213_{10} = 11101101101101_2 \\
  = 1.11011011011012 \times 2^{13} \)

• Significand
  \( M = \underbrace{1.11011011011012}_{\text{frac}} \)
  \( \frac{\text{frac}}{1101101101101000000000000_2} \)

• Exponent
  \( E = \exp - \text{Bias} = \exp - 127 = 13 \)
  \( \Rightarrow \exp = 140 = 10001100_2 \)

• Result:
  \[
  \begin{array}{ccc}
  \boxed{0} & 10001100 & \underbrace{110110110110101000000000000}_2 \\
  \text{s} & \exp & \text{frac}
  \end{array}
  \]
2. Denormalized Encoding

- **Condition**: $\text{exp} = 000...0$

- **Exponent value**: $E = 1 - \text{Bias}$ (instead of $E = 0 - \text{Bias}$)

- **Significand is**: $M = 0.xxx...x_2$ (instead of $M=1.xxx_2$)

- **Cases**
  - $\text{exp} = 000...0$, $\text{frac} = 000...0$
    - Represents zero
    - Note distinct values: +0 and –0
  - $\text{exp} = 000...0$, $\text{frac} \neq 000...0$
    - Numbers very close to 0.0
3. Special Values Encoding

• **Condition**: $\exp = 111...1$

• **Case**: $\exp = 111...1, \frac{\text{ac}}{} = 000...0$
  - Represents value $\infty$ (infinity)
  - Operation that overflows
  - E.g., $1.0/0.0 = -1.0/-0.0 = +\infty$, $1.0/-0.0 = -\infty$

• **Case**: $\exp = 111...1, \frac{\text{ac}}{} \neq 000...0$
  - Not-a-Number (NaN)
  - Represents case when no numeric value can be determined
  - E.g., $\sqrt{-1}$, $\infty - \infty$, $\infty \times 0$
Visualization: Floating Point Encodings

-∞ - Normalized - Denorm + Denorm + Normalized +∞

NaN

-0 +0 NaN
Floating Point: IEEE 754

What is the decimal equivalent of:

10011111110100000000000000000000

- 127

So:

• Real exponent = 127 - 127 = 0
• There is hidden 1

1.1010....0
= 1.625

Final answer = -1.625
In Summary About Floating Points

<table>
<thead>
<tr>
<th>exponent</th>
<th>mantissa</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>11...1</td>
<td>≠ 0</td>
<td>NaN</td>
</tr>
<tr>
<td>11...1</td>
<td>=0</td>
<td>((-1)^s \times \infty)</td>
</tr>
<tr>
<td>00...0</td>
<td>≠0</td>
<td>denormalized</td>
</tr>
<tr>
<td>00...0</td>
<td>=0</td>
<td>0</td>
</tr>
</tbody>
</table>
Algorithm Considerations

- Non representable numbers are rounded.
- This rounding error leads to different results depending on the order of operations.
  - Non-repeatability makes debugging harder.
- A common technique to maximize floating point arithmetic accuracy is to presort data before a reduction computation.
**Example**

**Multiplying two vectors:**

\[
a = [a_1 \ a_2 \ a_3 \ a_4] \]
\[
b = [b_1 \ b_2 \ b_3 \ b_4] \]
\[
a \cdot b = a_1 b_1 + a_2 b_2 + a_3 b_3 + a_4 b_4
\]

3 methods for doing it

Serial

Parallel
Example

Multiplying two vectors:

\[ \mathbf{a} = [a_1 \ a_2 \ a_3 \ a_4] \]
\[ \mathbf{b} = [b_1 \ b_2 \ b_3 \ b_4] \]
\[ \mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 + a_3 b_3 + a_4 b_4 \]

For example, consider the vectors:

\[ \mathbf{a} = [1.907607, -0.7862027, 1.148311, 0.9604002] \]
\[ \mathbf{b} = [-0.9355000, -0.6915108, 1.724470, -0.7097529] \]

<table>
<thead>
<tr>
<th>method</th>
<th>result</th>
</tr>
</thead>
<tbody>
<tr>
<td>exact</td>
<td>.0559587528435...</td>
</tr>
<tr>
<td>serial</td>
<td>.0559588074</td>
</tr>
<tr>
<td>FMA</td>
<td>.0559587515</td>
</tr>
<tr>
<td>parallel</td>
<td>.0559587478</td>
</tr>
</tbody>
</table>

The Most accurate

Parallel is more accurate than serial
FP Support in CUDA

• From compute capability 2.0
  – single and double precision
  – fused-and-multiply in both precisions

• Four rounding modes:
  – rn  round to nearest, ties to even (default)
  – rz  round towards zero
  – ru  round towards + ∞
  – rd  round towards − ∞
FP Support in CUDA

\[ x + y \]
\[ \text{\_\_fadd\_[rn | rz | ru | rd]} (x, y) \]

\[ x \times y \]
\[ \text{\_\_fmul\_[rn | rz | ru | rd]} (x, y) \]

\[ \text{fmaf} (x, y, z) \]
\[ \text{\_\_fma\_[rn | rz | ru | rd]} (x, y, z) \]

\[ 1.0f / x \]
\[ \text{\_\_frcp\_[rn | rz | ru | rd]} (x) \]

\[ x / y \]
\[ \text{\_\_fdiv\_[rn | rz | ru | rd]} (x, y) \]

\[ \sqrt{f}(x) \]
\[ \text{\_\_fsqrt\_[rn | rz | ru | rd]} (x) \]
So..

When doing floating-point operations in parallel you have to decide:

• How much accuracy is good enough?
• Do you need single-precision or double precision?
• Can you tolerate presorting overhead, if you care about rounding errors?
Asynchronous Execution
Asynchronous Execution

- Asynchronous = returns to host right-away and does not wait for device
- This includes (but not limited to):
  - Kernel launches;
  - Memory copies between two addresses to the same device memory;
  - Memory copies from host to device of a memory block of 64 KB or less;
  - Memory copies performed by functions that are suffixed with Async;
Asynchronous Execution

• Some CUDA API calls and all kernel launches are asynchronous with respect to the host code.
• This means error-reporting is also asynchronous.

```c
cudaMemcpyAsync(a_d, a_h, size, cudaMemcpyHostToDevice, 0);
kernel<<<grid, block>>>(a_d);
cpuFunction();
```

stream
Streams
Streams

- A sequence of operations that execute on the device in the order in which they are issued by the host code
- Operations in different streams can be interleaved and, when possible, they can even run concurrently.
- A stream can be sequence of kernel launches and host-device memory copies
- Can have several open streams to the same device at once
- Need GPUs with concurrent transfer/execution capability
- Potential performance improvement: can overlap transfer and computation
**Streams**

- By default all transfers and kernel launches are assigned to stream 0
  - This means they are executed in order
Example: Default Stream

```c
cudaMemcpy(d_a, a, numBytes, cudaMemcpyHostToDevice);
increment<<<1,N>>>(d_a);
cudaMemcpy(a, d_a, numBytes, cudaMemcpyDeviceToHost);
```

- In the code above, from the perspective of the device, all three operations are issued to the same (default) stream and will execute in the order that they were issued.
- From the perspective of the host:
  - data transfers are blocking or synchronous transfers
  - kernel launch is asynchronous.

```c
anyCPUfunction();
cudaMemcpy(a, d_a, numBytes, cudaMemcpyDeviceToHost);
```
Example: Non-Default Stream

Non-default streams in CUDA C/C++ are declared, created, and destroyed in host code as follows:

```c
cudaStream_t stream1;
cudaError_t result;
result = cudaStreamCreate(&stream1);
result = cudaStreamDestroy(stream1);
```

To issue data transfer to non-default stream (non-blocking):

```c
result = cudaMemcpyAsync(d_a, a, N, cudaMemcpyHostToDevice, stream1);
```

To launch a kernel to non-default stream:

```c
increment<<<1,N,0,stream1>>>(d_a);
```
Important

• All operations to non-default streams are non-blocking with respect to the host code.
• Sometimes you need to synchronize the host code with operations in a stream.
• You have several options:
  – cudaDeviceSynchronize() → blocks host
  – cudaStreamSynchronize(stream) → blocks host
  – cudaStreamQuery(stream) → does not block host
Streams

- The amount of overlap execution between two streams depends on:
  - Device supports overlap transfer and kernel execution (compute capability 1.1 and higher)
  - Devices supports concurrent kernel execution (compute capability 2.x and higher)
  - Device supports concurrent data transfer (compute capability 2.x and higher)
  - The order on which commands are issued to each stream
Using streams to overlap device execution with data transfer

• Conditions to be satisfied first:
  – The device must be capable of *concurrent copy and execution*.
  – The kernel execution and the data transfer to be overlapped must both occur in *different, non-default streams*.
  – The host memory involved in the data transfer must be *pinned memory* (more on this later).
Using streams to overlap device execution with data transfer

for (int i = 0; i < nStreams; ++i) {

    int offset = i * streamSize;

    cudaMemcpyAsync(&d_a[offset], &a[offset],
                    streamBytes,
                    cudaMemcpyHostToDevice,
                    stream[i]);

    kernel<<<......>>>(d_a, offset);

    cudaMemcpyAsync(&a[offset], &d_a[offset],
                    streamBytes,
                    cudaMemcpyDeviceToHost,
                    stream[i]);

}
So..

- Streams are a good way to overlap execution and transfer, hardware permits.
- Don’t confuse kernels, threads, and streams.
Pinned Pages

• Allocate page(s) from system RAM (\texttt{cudaMallocHost()} or \texttt{cudaHostAlloc()})
  – Accessible by device (but wait till next slides)
  – Cannot be paged out
  – Enables highest memory copy performance (\texttt{cudaMemcpyAsync()})
  – Don’t forget \texttt{cudaFreeHost();}

• If too much pinned pages, overall system performance may greatly suffer.
cudaHostAlloc(void ** ptr,
       size_t size,
       unsigned int flags)

The Flags:

- **cudaHostAllocDefault**: causes cudaHostAlloc() to emulate cudaMallocHost().
- **cudaHostAllocPortable**: The memory returned by this call will be considered as pinned memory by all CUDA contexts, not just the one that performed the allocation.
- **cudaHostAllocMapped**: Maps the allocation into the CUDA address space. The device pointer to the memory may be obtained by calling cudaHostGetDevicePointer().
- **cudaHostAllocWriteCombined**: Allocates the memory as write-combined (WC).
  - WC memory can be transferred across the PCI Express bus more quickly on some system configurations.
  - Cannot be read efficiently by most CPUs.
Host page accessible by the device??

• The pointer to the host memory is not directly transferable to device, except with:
  – `cudaHostGetDevicePointer(void ** pDevice, void * pHost, unsigned int flags)`
  – flags are 0 for now

• Accessing host memory from device without explicit copy is called zero-copy mechanism.
Steps for Zero-Copy

1. `cudaHostAlloc (void ** ptr, size_t size, unsigned int flags)`
   – flag here: `cudaHostAllocMapped`

2. `cudaHostGetDevicePointer()`

3. Then use the pointer in your kernel on device as if it is in the GPU memory
#include <stdio.h>
#include <cuda.h>
#include <stdlib.h>
#define N 32         // size of vectors

__global__ void add(int *a, int *b, int *c) {
    int tid = blockIdx.x * blockDim.x + threadIdx.x;
    if(tid < N) c[tid] = a[tid] + b[tid];
}

int main(int argc, char *argv[]) {
    int T = 32, B = 1;            // threads per block and blocks per grid
    int *a, *b, *c; // host pointers
    int *dev_a, *dev_b, *dev_c; // device pointers to host memory
    int size = N * sizeof(int);

    cudaMemcpy(a, dev_a, size, cudaMemcpyHostToDevice);
    cudaMemcpy(b, dev_b, size, cudaMemcpyHostToDevice);
    cudaMemcpy(c, dev_c, size, cudaMemcpyHostToDevice);

    // load arrays with some numbers
    cudaMemcpy(dev_a, a, size, cudaMemcpyDeviceToHost);
    cudaMemcpy(dev_b, b, size, cudaMemcpyDeviceToHost);
    cudaMemcpy(dev_c, c, size, cudaMemcpyDeviceToHost);

    add<<<B, T>>>(dev_a, dev_b, dev_c);

    cudaFreeHost(a);
    cudaFreeHost(b);
    cudaFreeHost(c);

    return 0;
}
So..

- If the CPU program requires a lot of memory, then pinned pages is not a good idea.
Other Sources of Concurrency

- Some devices of compute capability 2.x and higher can **execute multiple kernels** concurrently.
- The maximum number of kernel launches that a device can execute concurrently is 32 on devices of compute capability 3.5 and 16 on devices of lower compute capability.
- A kernel from one CUDA context cannot execute concurrently with a kernel from another CUDA context.
- Kernels that use many textures or a large amount of local memory are less likely to execute concurrently with other kernels.
- Some devices of compute capability 2.x and higher can perform a **copy from page-locked host memory to device memory** concurrently with a copy from device memory to page locked host memory.
Conclusions

- As we program GPUs we need to pay attention to several performance bottlenecks:
  - Branch diversion
  - Global memory latency
  - Global memory bandwidth
  - Shared memory bank conflicts
  - Communication
  - Limited resources
- We have several techniques in our arsenal to enhance performance
  - Try to make threads in the same warp follow the same control flow
  - Tiling
  - Coalescing
  - Loop unrolling
  - Increase thread granularity
  - Trade one resource for another
  - Memory access pattern
  - Streams
- Pay attention to interaction among techniques