CS 677: Parallel Programming for Many-core Processors
Lecture 4

Instructor: Philippos Mordohai
Webpage: www.cs.stevens.edu/~mordohai
E-mail: Philippos.Mordohai@stevens.edu
Logistics

• Midterm: March 26

• Project proposal presentations: March 19
  – Have to be approved by me by March 14 (during Spring break)
Project Proposal

• Problem description
  – What is the computation and why is it important?
  – Abstraction of computation: equations, graphic or pseudo-code, no more than 1 page

• Suitability for GPU acceleration
  – Amdahl’s Law: describe the inherent parallelism. Argue that it is close to 100% of computation.
  – Synchronization and Communication: Discuss what data structures may need to be protected by synchronization, or communication through host.
  – Copy Overhead: Discuss the data footprint and anticipated cost of copying to/from host memory.

• Intellectual Challenges
  – Generally, what makes this computation worthy of a project?
  – Point to any difficulties you anticipate at present in achieving high speedup
Overview

• More Performance Considerations
  – Memory Coalescing
  – Occupancy
  – Kernel Launch Overhead
  – Instruction Performance
• Summary of Performance Considerations
  – Lectures 3 and 4

• Floating-Point Considerations
Memory Coalescing (Part 2)

slides by
Jared Hoberock and David Tarjan
(Stanford CS 193G)
Consider the stride of your accesses

```c
#include <cuda_runtime.h>

__global__ void foo(int* input,
                     float3* input2)
{
    int i = blockDim.x * blockIdx.x + threadIdx.x;
    // Stride 1
    int a = input[i];
    // Stride 2, half the bandwidth is wasted
    int b = input[2*i];
    // Stride 3, 2/3 of the bandwidth wasted
    float c = input2[i].x;
}
```
Example: Array of Structures (AoS)

```c
struct record {
    int key;
    int value;
    int flag;
};

record *d_records;
cudaMalloc((void**)&d_records,
...);
```
Example: Structure of Arrays (SoA)

```c
struct SoA
{
    int * keys;
    int * values;
    int * flags;
};

SoA d_SoA_data;
cudaMalloc((void**)&d_SoA_data.keys, ...);
cudaMalloc((void**)&d_SoA_data.values, ...);
cudaMalloc((void**)&d_SoA_data.flags, ...);
```
Example: SoA vs. AoS

```c
__global__ void bar(record
 *AoS_data, SoA SoA_data)
{
    int i = blockDim.x * blockIdx.x
         + threadIdx.x;
    // AoS wastes bandwidth
    int key = AoS_data[i].key;
    // SoA efficient use of bandwidth
    int key_better = SoA_data.keys[i];
}
```
Memory Coalescing

• Structure of arrays is often better than array of structures
  – Very clear win on regular, stride 1 access patterns
  – Unpredictable or irregular access patterns are case-by-case
Occupancy

slides (mostly) by
Jared Hoberock and David Tarjan
(Stanford CS 193G)
and Joseph T. Kider Jr. (UPenn)
Reminder: Thread Scheduling

- SM implements zero-overhead warp scheduling
  - At any time, only one of the warps is executed by SM
  - Warps whose next instruction has its inputs ready for consumption are eligible for execution
  - Eligible Warps are selected for execution on a prioritized scheduling policy
  - All threads in a warp execute the same instruction when selected

- Time

TB = Thread Block, W = Warp
Thread Scheduling

• What happens if all warps are stalled?
  – No instruction issued → performance lost

• Most common reason for stalling?
  – Waiting on global memory

• If your code reads global memory every couple of instructions
  – You should try to maximize occupancy
Occupancy

- Thread instructions are executed sequentially, so executing other warps is the only way to hide latencies and keep cores busy
- **Occupancy** = number of warps running concurrently on a multiprocessor divided by maximum number of warps that can run concurrently
- Limited by resource usage:
  - Registers
  - Shared memory
Resource Limits (1)

• Pool of registers and shared memory per SM
  • Each thread block grabs registers & shared memory
  • If one or the other is fully utilized -> no more thread blocks
Resource Limits (2)

- Can only have 8 thread blocks per SM
  - If they’re too small, can’t fill up the SM
  - Need 128 threads / block (GT200), 192 threads / block (GF100)

- Higher occupancy has diminishing returns for hiding latency
Grid/Block Size Heuristics

• # of blocks > # of multiprocessors
  – So all multiprocessors have at least one block to execute

• # of blocks / # of multiprocessors > 2
  – Multiple blocks can run concurrently on a multiprocessor
  – Blocks not waiting at a __syncthreads() keep hardware busy
  – Subject to resource availability - registers, shared memory

• # of blocks > 100 to scale to future devices
Register Dependency

• Read-after-write register dependency
  – Instruction’s result can be read approximately 24 cycles later

• To completely hide latency:
  – Run at least 192 threads (6 warps) per multiprocessor
    • At least 25% occupancy for compute capability 1.0 and 1.1
    • Threads do not have to belong to the same block
Register Pressure

• Hide latency by using more threads per SM

• Limiting factors:
  – Number of registers per thread
    • 8k/16k/... per SM, partitioned among concurrent threads
  – Amount of shared memory
    • 16kB/... per SM, partitioned among concurrent blocks
Hiding Latency with more threads

Throughput, 32-bit words

GB/s

Threads Per Multiprocessor

0 0 128 256 384 512 640 768 896 1024
How do you know what you’re using?

• Use `nvcc --ptxas-options=-v` to get register and shared memory usage

• Plug those numbers into CUDA Occupancy Calculator
CUDA Occupancy Calculator

1. Select Compute Capability (click): 1.3
2. Enter your resource usage:
   - Threads Per Block
   - Registers Per Thread
   - Shared Memory Per Block (bytes)

3. GPU Occupancy Data is displayed here and in the graphs:
   - Active Threads per Multiprocessor
   - Active Warps per Multiprocessor
   - Active Thread Blocks per Multiprocessor
   - Occupancy of each Multiprocessor

4. Allocation Per Thread Block:
   - Warps
   - Registers
   - Shared Memory

5. Maximum Thread Blocks Per Multiprocessor: Blocks
   - Limited by Max Warps / Blocks per Multiprocessor
   - Limited by Registers per Multiprocessor
   - Limited by Shared Memory per Multiprocessor
   - Thread Block Limit Per Multiprocessor highlighted

6. CUDA Occupancy Calculator
   - Version: 2.0
   - Copyright and License

The other data points represent the range of possible block sizes, register counts, and shared memory allocation.

### Varying Block Size

<table>
<thead>
<tr>
<th>Threads Per Block</th>
<th>Occupancy</th>
<th>My Block Size 128</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>44</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>80</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>264</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

### Varying Register Count

<table>
<thead>
<tr>
<th>Registers Per Thread</th>
<th>Occupancy</th>
<th>My Register Count 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>64</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>128</td>
<td>128</td>
<td>128</td>
</tr>
</tbody>
</table>

### Varying Shared Memory Usage

<table>
<thead>
<tr>
<th>Shared Memory Per Block</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
</tr>
<tr>
<td>64</td>
</tr>
<tr>
<td>128</td>
</tr>
<tr>
<td>256</td>
</tr>
<tr>
<td>512</td>
</tr>
<tr>
<td>1024</td>
</tr>
</tbody>
</table>
CUDA GPU Occupancy Calculator

Just follow steps 1, 2, and 3 below! (or click here for help)

1.) Select Compute Capability (click):
   - 1.3

2.) Enter your resource usage:
   - Threads Per Block: 128
   - Registers Per Thread: 25
   - Shared Memory Per Block (bytes): 640

3.) GPU Occupancy Data is displayed here and in the graphs:

![Calculator Interface](image-url)
### GPU Occupancy Data

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Threads per Multiprocessor</td>
<td>512</td>
</tr>
<tr>
<td>Active Warps per Multiprocessor</td>
<td>16</td>
</tr>
<tr>
<td>Active Thread Blocks per Multiprocessor</td>
<td>4</td>
</tr>
<tr>
<td>Occupancy of each Multiprocessor</td>
<td>50%</td>
</tr>
</tbody>
</table>

### Physical Limits for GPU Compute Capability

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threads per Warp</td>
<td>32</td>
</tr>
<tr>
<td>Warps per Multiprocessor</td>
<td>32</td>
</tr>
<tr>
<td>Threads per Multiprocessor</td>
<td>1024</td>
</tr>
<tr>
<td>Thread Blocks per Multiprocessor</td>
<td>8</td>
</tr>
<tr>
<td>Total # of 32-bit registers per Multiprocessor</td>
<td>16384</td>
</tr>
<tr>
<td>Register allocation unit size</td>
<td>512</td>
</tr>
<tr>
<td>Register allocation granularity</td>
<td>block</td>
</tr>
<tr>
<td>Shared Memory per Multiprocessor (bytes)</td>
<td>16384</td>
</tr>
<tr>
<td>Shared Memory Allocation unit size</td>
<td>512</td>
</tr>
<tr>
<td>Warp allocation granularity (for register allocation)</td>
<td>2</td>
</tr>
</tbody>
</table>

### Allocation Per Thread Block

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Warps</td>
<td>4</td>
</tr>
<tr>
<td>Registers</td>
<td>3584</td>
</tr>
<tr>
<td>Shared Memory</td>
<td>1024</td>
</tr>
</tbody>
</table>

These data are used in computing the occupancy data in blue.

### Maximum Thread Blocks Per Multiprocessor

<table>
<thead>
<tr>
<th>Description</th>
<th>Blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limited by Max Warps / Blocks per Multiprocessor</td>
<td>8</td>
</tr>
<tr>
<td>Limited by Registers per Multiprocessor</td>
<td>4</td>
</tr>
<tr>
<td>Limited by Shared Memory per Multiprocessor</td>
<td>16</td>
</tr>
</tbody>
</table>

Thread Block Limit Per Multiprocessor highlighted **RED**
The other data points represent the range of possible block sizes, register counts, and shared memory allocation.

**Varying Block Size**

**Varying Register Count**

**Varying Shared Memory Usage**
How to influence how many registers you use

• Pass option `-maxrregcount=X` to nvcc

• This isn’t magic, won’t get occupancy for free

• Use this very carefully when you are right on the edge
Optimizing Threads per Block

• Choose threads per block as multiple of warp size
  – Avoid wasting computation on under-populated warps
• Run as many warps as possible per SM
  – Hide latency
• SMs can run up to 8 blocks at a time

• Heuristics
  – Minimum: 64 threads per block
    • Only if multiple concurrent blocks
  – 192 or 256 threads are a better choice
    • Usually, still enough registers to compile and invoke successfully
  – This all depends on computation
Occupancy != Performance

• Increasing occupancy does not necessarily increase performance

• BUT...

• Low-occupancy SMs cannot adequately hide latency
Parameterize your Application

- Parameterization helps adaptation to different GPUs
- GPUs vary in many ways
  - # of SMs
  - Memory bandwidth
  - Shared memory size
  - Register file size
  - Max threads per block

Avoid local minima
- Try widely varying configurations
Kernel Launch Overhead

slides by
Jared Hoberock and David Tarjan
(Stanford CS 193G)
Kernel Launch Overhead

• Kernel launches aren’t free
  – A null kernel launch will take non-trivial time
  – Actual time changes with HW generations and driver software...

• Independent kernel launches are cheaper than dependent kernel launches
  – Dependent launch: Some readback to the CPU

• Launching lots of small grids comes with substantial performance loss
Kernel Launch Overheads

• If you are reading back data to the CPU for control decisions, consider doing it on the GPU

• Even though the GPU is slow at serial tasks, it can do surprising amounts of work before you used up kernel launch overhead
Instruction Performance

slides by
Joseph T. Kider Jr. (Upenn)
Instruction Performance

• Instruction cycles (per warp) is the sum of
  – Operand read cycles
  – Instruction execution cycles
  – Result update cycles

• Therefore instruction throughput depends on
  – Nominal instruction throughput
  – Memory latency
  – Memory bandwidth

• Cycle refers to the multiprocessor clock rate
  – 1.3 GHz on the Tesla C1060, for example
Maximizing Instruction Throughput

• Maximize use of high-bandwidth memory
  – Maximize use of shared memory
  – Minimize accesses to global memory
  – Maximize coalescing of global memory accesses

• Optimize performance by overlapping memory accesses with computation
  – High arithmetic intensity programs
  – Many concurrent threads
Arithmetic Instruction Throughput

• `int` and `float` add, shift, min, max and `float` mul, mad: 4 cycles per warp
  – `int` multiply is by default 32-bit
    • requires multiple cycles/warp
  – use `__mul24()` and `__umul24()` intrinsics for 4-cycle 24-bit int multiplication

• Integer division and modulo operations are costly
  – The compiler will convert literal power-of-2 divides to shifts
    • But it may miss
  – Be explicit in cases where the compiler cannot tell that the divisor is a power of 2
    • Trick: `foo % n == foo & (n-1)` if `n` is a power of 2
Loop Transformations

Mary Hall
CS6963 University of Utah
Reordering Transformations

• Analyze reuse in computation
• Apply loop reordering transformations to improve locality based on reuse
• With any loop reordering transformation, always ask
  – Safety? (doesn’t reverse dependences)
  – Profitability? (improves locality)
Loop Permutation: A Reordering Transformation

Permute the order of the loops to modify the traversal order

```
for (i=0; i<3; i++)
  for (j=0; j<6; j++)
```

```
for (j=0; j<6; j++)
  for (i=0; i<3; i++)
```

Which one is better for row-major storage?
Safety of Permutation

• **Intuition**: Cannot permute two loops $i$ and $j$ in a loop nest if doing so reverses the direction of any dependence.

```plaintext
for (i = 0; i < 3; i++)
  for (j = 0; j < 6; j++)
```

```plaintext
for (i = 0; i < 3; i++)
  for (j = 1; j < 6; j++)
```

• Ok to permute?
Tiling (Blocking): Another Loop Reordering Transformation

- Blocking reorders loop iterations to bring iterations that reuse data closer in time
Tiling Example

**For (j=1; j<M; j++)**
  **for (i=1; i<N; i++)**
  \[ D[i] = D[i] + B[j][i]; \]

Strip mine

**For (j=1; j<M; j++)**
  **for (ii=1; ii<N; ii+=s)**
    **for (i=ii; i<min(ii+s,N); i++)**
    \[ D[i] = D[i] + B[j][i]; \]

Permute

**For (ii=1; ii<N; ii+=s)**
  **for (j=1; j<M; j++)**
  **for (i=ii; i<min(ii+s,N); i++)**
  \[ D[i] = D[i] + B[j][i]; \]
Legality of Tiling

- Tiling = strip-mine and permutation
  - Strip-mine does not reorder iterations
  - Permutation must be legal
 OR
- strip size less than dependence distance
A Few Words On Tiling

- Tiling can be used hierarchically to compute partial results on a block of data wherever there are capacity limitations
  - Between grids if total data exceeds global memory capacity
  - Across thread blocks if shared data exceeds shared memory capacity (also to partition computation across blocks and threads)
  - Within threads if data in constant cache exceeds cache capacity or data in registers exceeds register capacity or (as in example) data in shared memory for block still exceeds shared memory capacity
Summary of Performance Considerations
Summary of Performance Considerations

- Thread Execution and Divergence
- Communication Through Memory
- Instruction Level Parallelism and Thread Level Parallelism
- Memory Coalescing
- Shared Memory Bank Conflicts
- Parallel Reduction
- Prefetching
- Loop Unrolling and Transformations
- Occupancy
- Kernel Launch Overhead
- Instruction Performance
Thread Execution and Divergence

- Instructions are issued per 32 threads (warp)
- Divergent branches:
  - Threads within a single warp take different paths
    - if-else, ...
  - Different execution paths within a warp are serialized
- Different warps can execute different code with no impact on performance
An Example

// is this barrier divergent?
for (int offset = blockDim.x / 2;
     offset > 0;
     offset >>= 1)
{
    ...
    __syncthreads();
}
A Second Example

// what about this one?
__global__ void do_i_halt(int *input)
{
    int i = ...
    if(input[i])
    {
        ...
        __syncthreads(); // a divergent barrier
    // hangs the machine
Communication Through Memory

• Carefully partition data according to access patterns
• Read-only $\Rightarrow$ __constant__ memory (fast)
• R/W & shared within block $\Rightarrow$ __shared__ memory (fast)
• R/W within each thread $\Rightarrow$ registers (fast)
• Indexed R/W within each thread $\Rightarrow$ local memory (slow)
• R/W inputs/results $\Rightarrow$ cudaMalloc‘ed global memory (slow)

© David Kirk/NVIDIA and Wen-mei W. Hwu, 2007-2010
ECE 408, University of Illinois, Urbana-Champaign
Communication Through Memory

• Question:

```c
__global__ void race(void)
{
    __shared__ int my_shared_variable;
    my_shared_variable = threadIdx.x;

    // what is the value of
    // my_shared_variable?
}
```
Instruction Level Parallelism and Thread Level Parallelism

• Dynamic partitioning gives more flexibility to compilers/programmers
  – One can run a smaller number of threads that require many registers each or a large number of threads that require few registers each
  • This allows for finer grain threading than traditional CPU threading models
  – The compiler can tradeoff between instruction-level parallelism and thread level parallelism
Memory Coalescing

• When accessing global memory, peak performance utilization occurs when all threads in a half warp access continuous memory locations.
Memory Layout of a Matrix in C

Access direction in Kernel code

Time Period 1
T₁ T₂ T₃ T₄

Time Period 2
T₁ T₂ T₃ T₄

...
Memory Layout of a Matrix in C

Access direction in Kernel code

M_{0,0} M_{1,0} M_{2,0} M_{3,0}
M_{0,1} M_{1,1} M_{2,1} M_{3,1}
M_{0,2} M_{1,2} M_{2,2} M_{3,2}
M_{0,3} M_{1,3} M_{2,3} M_{3,3}

Time Period 1

Time Period 2

\[ M_{0,0} \quad M_{1,0} \quad M_{2,0} \quad M_{3,0} \quad M_{0,1} \quad M_{1,1} \quad M_{2,1} \quad M_{3,1} \quad M_{0,2} \quad M_{1,2} \quad M_{2,2} \quad M_{3,2} \quad M_{0,3} \quad M_{1,3} \quad M_{2,3} \quad M_{3,3} \]
Example: SoA vs. AoS

```c
__global__ void bar(record *AoS_data, SoA SoA_data)
{
    int i = blockDim.x * blockIdx.x
             + threadIdx.x;
    // AoS wastes bandwidth
    int key = AoS_data[i].key;
    // SoA efficient use of bandwidth
    int key_better = SoA_data.keys[i];
}
```
Shared Memory Bank Conflicts

• Shared memory is as fast as registers \textit{if there are no bank conflicts}
• Bank conflicts are less of an issue in newer versions of CUDA
Parallel Reduction:
No Divergence until <= 16 sub-sums
Prefetching

- One could **double buffer** the computation, getting better instruction mix within each thread
  - This is classic software pipelining in ILP compilers

```
Loop {
  Load current tile to shared memory
  syncthreads()
  Compute current tile
  syncthreads()
}

Load next tile from global memory

Loop {
  Deposit current tile to shared memory
  syncthreads()
  Load next tile from global memory
  Compute current tile
  syncthreads()
  }
```
Instruction Mix Considerations: Loop Unrolling

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

There are very few mul/add between branches and address calculation

Loop unrolling can help. (Beware that any local arrays used after unrolling will be dumped into Local Memory)

```c
Pvalue += Ms[ty][k] * Ns[k][tx] + ...
    Ms[ty][k+15] * Ns[k+15][tx];
```
Occupyancy

• Thread instructions are executed sequentially, so executing other warps is the only way to hide latencies and keep memory busy

• **Occupyancy** = number of warps running concurrently on a multiprocessor divided by maximum number of warps that can run concurrently

• Limited by resource usage:
  – Registers
  – Shared memory
Optimizing Threads per Block

- Choose threads per block as multiple of warp size
  - Avoid wasting computation on under-populated warps
- Run as many warps as possible per SM
  - Hide latency
- SMs can run up to 8 blocks at a time

- Heuristics
  - Minimum: 64 threads per block
    - Only if multiple concurrent blocks
  - 192 or 256 threads are a better choice
    - Usually, still enough registers to compile and invoke successfully
  - This all depends on computation
Kernel Launch Overhead

• Kernel launches aren’t free
  – A null kernel launch will take non-trivial time
  – Actual time changes with HW generations and driver software...

• Independent kernel launches are cheaper than dependent kernel launches
  – Dependent launch: Some readback to the cpu

• Launching lots of small grids comes with substantial performance loss
Compute Capabilities

• Reminder: do not take various constants, such as size of shared memory etc., for granted since they continuously change

• Check CUDA programming guide for the features of the compute capability of your GPU
Floating-Point Considerations
Objective

• To understand the fundamentals of floating-point representation
• To know the IEEE-754 Floating Point Standard
• GeForce 8800 CUDA Floating-point speed, accuracy and precision
  – Deviations from IEEE-754
  – Accuracy of device runtime functions
  – -fastmath compiler option
  – Future performance considerations
What is IEEE floating-point format?

• A floating point binary number consists of three parts:
  – sign (S), exponent (E), and mantissa (M).
  – Each (S, E, M) pattern uniquely identifies a floating point number.

• For each bit pattern, its IEEE floating-point value is derived as:
  
  – value = \((-1)^S \times M \times 2^E\), where \(1.0 \leq M < 10.0\)

• The interpretation of S is simple: S=0 results in a positive number and S=1 a negative number.
Normalized Representation

• Specifying that $1.0_B \leq M < 10.0_B$ makes the mantissa value for each floating point number unique.
  – For example, the only one mantissa value allowed for $0.5_D$ is $M = 1.0$
    • $0.5_D = 1.0_B * 2^{-1}$
  – Neither $10.0_B * 2^{-2}$ nor $0.1_B * 2^0$ qualifies

• Because all mantissa values are of the form $1.XX...$, one can omit the “1.” part in the representation.
  – The mantissa value of $0.5_D$ in a 2-bit mantissa is 00, which is derived by omitting “1.” from 1.00.
Note: Two’s Complement

• The negative value of a number can be derived by:
  • Complementing every bit
  • Adding 1

• Example: -3
  • $3 = 011_B$
  • Complement every bit: 100
  • Add 1: 101
Exponent Representation

- In an n-bits exponent representation, $2^{n-1}-1$ is added to its 2's complement representation to form its excess representation.
  - See Table for a 3-bit exponent representation
- A simple unsigned integer comparator can be used to compare the magnitude of two FP numbers
- Symmetric range for +/- exponents (111 reserved)

<table>
<thead>
<tr>
<th>2's complement</th>
<th>Actual decimal</th>
<th>Excess-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>0</td>
<td>011</td>
</tr>
<tr>
<td>001</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>010</td>
<td>2</td>
<td>101</td>
</tr>
<tr>
<td>011</td>
<td>3</td>
<td>110</td>
</tr>
<tr>
<td>100 (reserved pattern)</td>
<td>(reserved pattern)</td>
<td>111</td>
</tr>
<tr>
<td>101</td>
<td>-3</td>
<td>000</td>
</tr>
<tr>
<td>110</td>
<td>-2</td>
<td>001</td>
</tr>
<tr>
<td>111</td>
<td>-1</td>
<td>010</td>
</tr>
</tbody>
</table>
A simple, hypothetical 6-bit FP format

- Assume 1-bit S, 3-bit E, and 2-bit M
  - $0.5_D = 1.00_B \times 2^{-1}$
  - $0.5_D = 001000$, where $S = 0$, $E = 010$, and $M = (1.)00$

<table>
<thead>
<tr>
<th>2’s complement</th>
<th>Actual decimal</th>
<th>Excess-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>0</td>
<td>011</td>
</tr>
<tr>
<td>001</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>010</td>
<td>2</td>
<td>101</td>
</tr>
<tr>
<td>011</td>
<td>3</td>
<td>110</td>
</tr>
<tr>
<td><strong>100</strong></td>
<td><strong>(reserved pattern)</strong></td>
<td><strong>111</strong></td>
</tr>
<tr>
<td>101</td>
<td>-3</td>
<td>000</td>
</tr>
<tr>
<td>110</td>
<td>-2</td>
<td>001</td>
</tr>
<tr>
<td>111</td>
<td>-1</td>
<td>010</td>
</tr>
</tbody>
</table>
Representable Numbers

- The representable numbers of a given format is the set of all numbers that can be exactly represented in the format.
- See Table for representable numbers of an **unsigned 3-bit integer format**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>0</td>
</tr>
<tr>
<td>001</td>
<td>1</td>
</tr>
<tr>
<td>010</td>
<td>2</td>
</tr>
<tr>
<td>011</td>
<td>3</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
</tr>
<tr>
<td>101</td>
<td>5</td>
</tr>
<tr>
<td>110</td>
<td>6</td>
</tr>
<tr>
<td>111</td>
<td>7</td>
</tr>
</tbody>
</table>
### Representable Numbers of a 5-bit Hypothetical IEEE Format

<table>
<thead>
<tr>
<th>E</th>
<th>M</th>
<th>S=0</th>
<th>S=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>00</td>
<td>$2^{-1}$</td>
<td>$-(2^{-1})$</td>
</tr>
<tr>
<td></td>
<td>01</td>
<td>$2^{-1}+1 \times 2^{-3}$</td>
<td>$-(2^{-1}+1 \times 2^{-3})$</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>$2^{-1}+2 \times 2^{-3}$</td>
<td>$-(2^{-1}+2 \times 2^{-3})$</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>$2^{-1}+3 \times 2^{-3}$</td>
<td>$-(2^{-1}+3 \times 2^{-3})$</td>
</tr>
<tr>
<td>01</td>
<td>00</td>
<td>$2^0$</td>
<td>$-(2^0)$</td>
</tr>
<tr>
<td></td>
<td>01</td>
<td>$2^0+1 \times 2^{-2}$</td>
<td>$-(2^0+1 \times 2^{-2})$</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>$2^0+2 \times 2^{-2}$</td>
<td>$-(2^0+2 \times 2^{-2})$</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>$2^0+3 \times 2^{-2}$</td>
<td>$-(2^0+3 \times 2^{-2})$</td>
</tr>
<tr>
<td>10</td>
<td>00</td>
<td>$2^1$</td>
<td>$-(2^1)$</td>
</tr>
<tr>
<td></td>
<td>01</td>
<td>$2^1+1 \times 2^{-1}$</td>
<td>$-(2^1+1 \times 2^{-1})$</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>$2^1+2 \times 2^{-1}$</td>
<td>$-(2^1+2 \times 2^{-1})$</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>$2^1+3 \times 2^{-1}$</td>
<td>$-(2^1+3 \times 2^{-1})$</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>Reserved pattern</td>
<td></td>
</tr>
</tbody>
</table>

© David Kirk/NVIDIA and Wen-mei W. Hwu, 2007-2010
ECE408, University of Illinois, Urbana-Champaign
Representable Numbers of a 5-bit Hypothetical IEEE Format

<table>
<thead>
<tr>
<th>E</th>
<th>M</th>
<th>S=0</th>
<th>S=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>00</td>
<td>$2^{-1}$</td>
<td>$-(2^{-1})$</td>
</tr>
<tr>
<td></td>
<td>01</td>
<td>$2^{-1}+1*2^{-3}$</td>
<td>$-(2^{-1}+1*2^{-3})$</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>$2^{-1}+2*2^{-3}$</td>
<td>$-(2^{-1}+2*2^{-3})$</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>$2^{-1}+3*2^{-3}$</td>
<td>$-(2^{-1}+3*2^{-3})$</td>
</tr>
<tr>
<td>01</td>
<td>00</td>
<td>$2^{0}$</td>
<td>$-(2^{0})$</td>
</tr>
<tr>
<td></td>
<td>01</td>
<td>$2^{0}+1*2^{-2}$</td>
<td>$-(2^{0}+1*2^{-2})$</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>$2^{0}+2*2^{-2}$</td>
<td>$-(2^{0}+2*2^{-2})$</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>$2^{0}+3*2^{-2}$</td>
<td>$-(2^{0}+3*2^{-2})$</td>
</tr>
<tr>
<td>10</td>
<td>00</td>
<td>$2^{1}$</td>
<td>$-(2^{1})$</td>
</tr>
<tr>
<td></td>
<td>01</td>
<td>$2^{1}+1*2^{-1}$</td>
<td>$-(2^{1}+1*2^{-1})$</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>$2^{1}+2*2^{-1}$</td>
<td>$-(2^{1}+2*2^{-1})$</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>$2^{1}+3*2^{-1}$</td>
<td>$-(2^{1}+3*2^{-1})$</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td>Reserved pattern</td>
</tr>
</tbody>
</table>

Cannot represent Zero!
Representable Numbers of a 5-bit Hypothetical IEEE Format

- Exponent bits define major intervals of representable numbers
- Mantissa bits define the number of representable numbers in each interval
  - With N mantissa bits, $2^N$ representable numbers per interval
- Representable numbers come closer to each other in the neighborhood of 0
  - Desirable property
- There is a gap around 0
  - Significantly larger error between 0 and 0.5 than 0.5 and 1
Flush to Zero (Abrupt Underflow)

- Treat all bit patterns with E=0 as 0.0
  - This takes away several representable numbers near zero and lumps them all into 0.0
  - For a representation with large M, a large number of representable numbers will be removed
## Flush to Zero

<table>
<thead>
<tr>
<th>E</th>
<th>M</th>
<th>S=0</th>
<th>S=1</th>
<th>S=0</th>
<th>S=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>00</td>
<td>2(^{-1})</td>
<td>-(2(^{-1}))</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>01</td>
<td>2(^{-1})+1*2(^{-3})</td>
<td>-(2(^{-1})+1*2(^{-3}))</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>2(^{-1})+2*2(^{-3})</td>
<td>-(2(^{-1})+2*2(^{-3}))</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>2(^{-1})+3*2(^{-3})</td>
<td>-(2(^{-1})+3*2(^{-3}))</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>01</td>
<td>00</td>
<td>2(^{0})</td>
<td>-(2(^{0}))</td>
<td>2(^{0})</td>
<td>-(2(^{0}))</td>
</tr>
<tr>
<td></td>
<td>01</td>
<td>2(^{0})+1*2(^{-2})</td>
<td>-(2(^{0})+1*2(^{-2}))</td>
<td>2(^{0})+1*2(^{-2})</td>
<td>-(2(^{0})+1*2(^{-2}))</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>2(^{0})+2*2(^{-2})</td>
<td>-(2(^{0})+2*2(^{-2}))</td>
<td>2(^{0})+2*2(^{-2})</td>
<td>-(2(^{0})+2*2(^{-2}))</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>2(^{0})+3*2(^{-2})</td>
<td>-(2(^{0})+3*2(^{-2}))</td>
<td>2(^{0})+3*2(^{-2})</td>
<td>-(2(^{0})+3*2(^{-2}))</td>
</tr>
<tr>
<td>10</td>
<td>00</td>
<td>2(^{1})</td>
<td>-(2(^{1}))</td>
<td>2(^{1})</td>
<td>-(2(^{1}))</td>
</tr>
<tr>
<td></td>
<td>01</td>
<td>2(^{1})+1*2(^{-1})</td>
<td>-(2(^{1})+1*2(^{-1}))</td>
<td>2(^{1})+1*2(^{-1})</td>
<td>-(2(^{1})+1*2(^{-1}))</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>2(^{1})+2*2(^{-1})</td>
<td>-(2(^{1})+2*2(^{-1}))</td>
<td>2(^{1})+2*2(^{-1})</td>
<td>-(2(^{1})+2*2(^{-1}))</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>2(^{1})+3*2(^{-1})</td>
<td>-(2(^{1})+3*2(^{-1}))</td>
<td>2(^{1})+3*2(^{-1})</td>
<td>-(2(^{1})+3*2(^{-1}))</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>Reserved pattern</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Denormalized Numbers

• The actual method adopted by the IEEE standard is called denormalized numbers or gradual underflow.
  – The method relaxes the normalization requirement for numbers very close to 0
  – Whenever E=0, the mantissa is no longer assumed to be of the form 1.XX. Rather, it is assumed to be 0.XX. In general, if the n-bit exponent is 0, the value is

\[ 0.M \times 2^{-2^{(n-1)} + 2} \]
## Denormalization

<table>
<thead>
<tr>
<th>E</th>
<th>M</th>
<th>( S=0 )</th>
<th>( S=1 )</th>
<th>( S=0 )</th>
<th>( S=1 )</th>
<th>( S=0 )</th>
<th>( S=1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>00</td>
<td>( 2^{-1} )</td>
<td>(-2^{-1})</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>01</td>
<td>01</td>
<td>( 2^{-1}+1*2^{-3} )</td>
<td>(-2^{-1}+1*2^{-3})</td>
<td>0</td>
<td>0</td>
<td>(1*2^{-2})</td>
<td>(-1*2^{-2})</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>( 2^{-1}+2*2^{-3} )</td>
<td>(-2^{-1}+2*2^{-3})</td>
<td>0</td>
<td>0</td>
<td>(2*2^{-2})</td>
<td>(-2*2^{-2})</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>( 2^{-1}+3*2^{-3} )</td>
<td>(-2^{-1}+3*2^{-3})</td>
<td>0</td>
<td>0</td>
<td>(3*2^{-2})</td>
<td>(-3*2^{-2})</td>
</tr>
<tr>
<td>01</td>
<td>00</td>
<td>( 2^0 )</td>
<td>(-2^0)</td>
<td>(2^0)</td>
<td>(-2^0)</td>
<td>(2^0)</td>
<td>(-2^0)</td>
</tr>
<tr>
<td>01</td>
<td>01</td>
<td>( 2^0+1*2^{-2} )</td>
<td>(-2^0+1*2^{-2})</td>
<td>(2^0+1*2^{-2})</td>
<td>(-2^0+1*2^{-2})</td>
<td>(2^0+1*2^{-2})</td>
<td>(-2^0+1*2^{-2})</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>( 2^0+2*2^{-2} )</td>
<td>(-2^0+2*2^{-2})</td>
<td>(2^0+2*2^{-2})</td>
<td>(-2^0+2*2^{-2})</td>
<td>(2^0+2*2^{-2})</td>
<td>(-2^0+2*2^{-2})</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>( 2^0+3*2^{-2} )</td>
<td>(-2^0+3*2^{-2})</td>
<td>(2^0+3*2^{-2})</td>
<td>(-2^0+3*2^{-2})</td>
<td>(2^0+3*2^{-2})</td>
<td>(-2^0+3*2^{-2})</td>
</tr>
<tr>
<td>10</td>
<td>00</td>
<td>( 2^1 )</td>
<td>(-2^1)</td>
<td>(2^1)</td>
<td>(-2^1)</td>
<td>(2^1)</td>
<td>(-2^1)</td>
</tr>
<tr>
<td>01</td>
<td>01</td>
<td>( 2^1+1*2^{-1} )</td>
<td>(-2^1+1*2^{-1})</td>
<td>(2^1+1*2^{-1})</td>
<td>(-2^1+1*2^{-1})</td>
<td>(2^1+1*2^{-1})</td>
<td>(-2^1+1*2^{-1})</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>( 2^1+2*2^{-1} )</td>
<td>(-2^1+2*2^{-1})</td>
<td>(2^1+2*2^{-1})</td>
<td>(-2^1+2*2^{-1})</td>
<td>(2^1+2*2^{-1})</td>
<td>(-2^1+2*2^{-1})</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>( 2^1+3*2^{-1} )</td>
<td>(-2^1+3*2^{-1})</td>
<td>(2^1+3*2^{-1})</td>
<td>(-2^1+3*2^{-1})</td>
<td>(2^1+3*2^{-1})</td>
<td>(-2^1+3*2^{-1})</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Arithmetic Instruction Throughput

- int and float add, shift, min, max and float mul, mad: 4 cycles per warp
  - int multiply (*) is by default 32-bit
    - requires multiple cycles / warp
  - Use __mul24() / __umul24() intrinsics for 4-cycle 24-bit int multiply

- Integer divide and modulo are expensive
  - Compiler will convert literal power-of-2 divides to shifts
  - Be explicit in cases where compiler can’t tell that divisor is a power of 2!
Arithmetic Instruction Throughput

- Reciprocal, reciprocal square root, sin/cos, log, exp: 16 cycles per warp
  - These are the versions prefixed with “__”
  - Examples: __rcp(), __sin(), __exp()

- Other functions are combinations of the above
  - \( y / x = rcp(x) \times y \) == 20 cycles per warp
  - \( sqrt(x) = rcp(rsqrt(x)) \) == 32 cycles per warp
Runtime Math Library

- There are two types of runtime math operations
  - \texttt{\_\_func():} direct mapping to hardware ISA
    - Fast but low accuracy (see prog. guide for details)
    - Examples: \texttt{\_\_sin(x), \_\_exp(x), \_\_pow(x,y)}
  - \texttt{func():} compile to multiple instructions
    - Slower but higher accuracy: error of 5 ulp or less
    - (ulp: units in the last place or units of least precision)
    - Examples: \texttt{sin(x), exp(x), pow(x,y)}

- The \texttt{-use\_fast\_math} compiler option forces every \texttt{func()} to compile to \texttt{\_\_func()}
Make your program float-safe!

- Hardware now has double precision support
  - G80 is single-precision only
  - Double precision has additional performance cost
  - Careless use of double or undeclared types may run more slowly on G80+
- Important to be float-safe (be explicit whenever you want single precision) to avoid using double precision where it is not needed
  - Add ‘f’ specifier on float literals:
    - foo = bar * 0.123; // double assumed
    - foo = bar * 0.123f; // float explicit
  - Use float version of standard library functions
    - foo = sin(bar); // double assumed
    - foo = sinf(bar); // single precision explicit
Deviation from IEEE-754

- Addition and Multiplication are IEEE 754 compliant
  - Maximum 0.5 ulp (units in the last place) error
- However, often combined into multiply-add (FMAD)
  - Intermediate result is truncated
- Division is non-compliant (2 ulp)
- Not all rounding modes are supported
- Denormalized numbers were not supported
  - Fully supported in single and double precision by compute capability 2.0 or better
Floating-Point Calculation Results Can Depend on Execution Order

Order 1
\[1.00 \times 2^0 + 1.00 \times 2^0 + 1.00 \times 2^{-2} + 1.00 \times 2^{-2}\]
\[= 1.00 \times 2^1 + 1.00 \times 2^{-2} + 1.00 \times 2^{-2}\]
\[= 1.00 \times 2^1 + 1.00 \times 2^{-2}\]
\[= 1.00 \times 2^1\]

Order 2
\[(1.00 \times 2^0 + 1.00 \times 2^0) + (1.00 \times 2^{-2} + 1.00 \times 2^{-2})\]
\[= 1.00 \times 2^1 + 1.00 \times 2^{-1}\]
\[= 1.01 \times 2^1\]

Pre-sorting is often used to increase stability of floating point results.
Special Bit Patterns in the IEEE Standard Format

<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>