Overview of Lecture

- Where data can be stored
  - And how to get it there
- Some guidelines for where to store data
  - Who needs to access it?
  - Read only vs. Read/Write
  - Footprint of data
- High level description of how to write code to optimize for memory hierarchy
  - More details Wednesday and next week
- Reading:
  - Chapter 5, Kirk and Hwu book
  - Or, http://courses.ece.illinois.edu/ece498/al/textbook/Chapter4-CudaMemoryModel.pdf

Administrative

- Next assignment due Friday, 5 PM
  - Use handin program on CADE machines
    - "handin CS6235 lab1 <probfile>"
- TA: Preethi Kotari
  - Email: preethik@cs.utah.edu
  - Office hours: Tu-Th, 2-3PM, MEB 3423
- Mailing list
  - CS6235s12-discussion@list.eng.utah.edu
  - Please use for all questions suitable for the whole class
  - Feel free to answer your classmates questions!

Targets of Memory Hierarchy Optimizations

- Reduce memory latency
  - The latency of a memory access is the time (usually in cycles) between a memory request and its completion
- Maximize memory bandwidth
  - Bandwidth is the amount of useful data that can be retrieved over a time interval
- Manage overhead
  - Cost of performing optimization (e.g., copying) should be less than anticipated gain
Optimizing the Memory Hierarchy on GPUs, Overview

Device memory access times non-uniform so data placement significantly affects performance.
- But controlling data placement may require additional copying, so consider overhead.
- Optimizations to increase memory bandwidth. Idea: maximize utility of each memory access.
  - Coalesce global memory accesses
  - Avoid memory bank conflicts to increase memory access parallelism
  - Align data structures to address boundaries

Hardware Implementation: Memory Architecture

- The local, global, constant, and texture spaces are regions of device memory (DRAM)
- Each multiprocessor has:
  - A set of 32-bit registers per processor
  - On-chip shared memory
    - Where the shared memory space resides
  - A read-only constant cache
    - To speed up access to the constant memory space
  - A read-only texture cache
    - To speed up access to the texture memory space
  - Data cache (Fermi only)

Terminology Review

- device = GPU = set of multiprocessors
- Multiprocessor = set of processors & shared memory
- Kernel = GPU program
- Grid = array of thread blocks that execute a kernel
- Thread block = group of SIMD threads that execute a kernel and can communicate via shared memory

Memory Location Cached Access Who
Local Off-chip No Read/Write One thread
Shared On-chip N/A - resident Read/Write All threads in a block
Global Off-chip No Read/Write All threads + host
Constant Off-chip Yes Read All threads + host
Texture Off-chip Yes Read All threads + host

Reuse and Locality

- Consider how data is accessed
  - Data reuse:
    - Same data used multiple times
    - Intrinsic in computation
  - Data locality:
    - Data is reused and is present in “fast memory”
    - Same data or same data transfer
- If a computation has reuse, what can we do to get locality?
  - Appropriate data placement and layout
  - Code reordering transformations
Access Times

- Register – dedicated HW - single cycle
- Constant and Texture caches – possibly single cycle, proportional to addresses accessed by warp
- Shared Memory – dedicated HW - single cycle if no "bank conflicts"
- Local Memory – DRAM, no cache - "slow"
- Global Memory – DRAM, no cache - "slow"
- Constant Memory – DRAM, cached, 1...10s...100s of cycles, depending on cache locality
- Texture Memory – DRAM, cached, 1...10s...100s of cycles, depending on cache locality
- Instruction Memory (invisible) – DRAM, cached

Data Placement: Conceptual

- Copies from host to device go to some part of global memory (possibly, constant or texture memory)
- How to use SP shared memory
  - Must construct or be copied from global memory by kernel program
  - How to use constant or texture cache
    - Read-only "reused" data can be placed in constant & texture memory by host
- Also, how to use registers
  - Most locally-allocated data is placed directly in registers
  - Even array variables can use registers if compiler understands access patterns
  - Can allocate "superwords" to registers, e.g., float4
  - Excessive use of registers will "spill" data to local memory
- Local memory
  - Deals with capacity limitations of registers and shared memory
  - Eliminated worried about race conditions
  - ... but SLOW

Data Placement: Syntax

- Through type qualifiers
  - __constant__, __shared__, __local__, __device__
- Through cudaMemcpy calls
  - Flavor of call and symbolic constant designate where to copy
- Implicit default behavior
  - Device memory without qualifier is global memory
  - Host by default copies to global memory
  - Thread-local variables go into registers unless capacity exceeded, then local memory

Language Extensions: Variable Type Qualifiers

<table>
<thead>
<tr>
<th>Memory</th>
<th>Scope</th>
<th>Lifetime</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>device</strong></td>
<td>local</td>
<td>thread</td>
</tr>
<tr>
<td><strong>local</strong></td>
<td>shared</td>
<td>block</td>
</tr>
<tr>
<td><strong>shared</strong></td>
<td>shared</td>
<td>block</td>
</tr>
<tr>
<td><strong>constant</strong></td>
<td>global</td>
<td>grid</td>
</tr>
</tbody>
</table>

__device__ is optional when used with __local__, __shared__, or __constant__
Variable Type Restrictions

- **Pointers** can only point to memory allocated or declared in global memory:
  - Allocated in the host and passed to the kernel:
    ```cpp
    __global__ void KernelFunc(float* ptr)
    ```
  - Obtained as the address of a global variable:
    ```cpp
    float* ptr = &GlobalVar;
    ```

Now Let’s Look at Shared Memory

- **Common Programming Pattern** (5.3 of CUDA 4 manual)
  - Load data into shared memory
  - Synchronize (if necessary)
  - Operate on data in shared memory
  - Synchronize (if necessary)
  - Write intermediate results to global memory
  - Repeat until done

Rest of Today’s Lecture

- **Mechanics of how to place data in shared memory and constant memory**
- **Tiling transformation to reuse data within**
  - Shared memory
  - Data cache (Fermi only)

Mechanics of Using Shared Memory

- `__shared__` type qualifier required
- Must be allocated from global/device function, or as “extern”
- **Examples**
  ```cpp
  extern __shared__ float d_s_array[M];
  // create or copy from global memory
  d_s_array[i] = ...;
  // synchronize threads before use
  __syncthreads();
  // now can use any element
  ... = d_s_array[x];
  // more synchronization needed if updated
  ```
  ```cpp
  __global__ void compute2()
  {
    __shared__ float d_s_array[M];
    // create or copy from global memory
    d_s_array[j] = ...
    // synchronize threads before use
    __syncthreads();
    // may write result back to global memory
    d_g_array[j] = d_s_array[j];
  }
  ```
Reuse and Locality

- Consider how data is accessed
  - **Data reuse:**
    - Same data used multiple times
    - Intrinsic in computation
  - **Data locality:**
    - Data is reused and is present in "fast memory"
    - Same data or same data transfer
  - If a computation has reuse, what can we do to get locality?
    - Appropriate data placement and layout
    - Code reordering transformations

Temporal Reuse in Sequential Code

- Same data used in distinct iterations I and I'

```c
for (i=1; i<N; i++)
  for (j=1; j<N; j++)
```

- A[j] has self-temporal reuse in loop i

Spatial Reuse (Ignore for now)

- Same data transfer (usually cache line) used in distinct iterations I and I'

```c
for (i=1; i<N; i++)
  for (j=1; j<N; j++)
```

- A[j] has self-spatial reuse in loop j
- **Multi-dimensional array note:** C arrays are stored in row-major order

Group Reuse

- Same data used by distinct references

```c
for (i=1; i<N; i++)
  for (j=1; j<N; j++)
```

- A[j], A[j+1] and A[j-1] have group reuse (spatial and temporal) in loop j
Tiling (Blocking): Another Loop Reordering Transformation

- Tiling 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-1,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-1,N); i++)
D[i] = D[i] + B[j][i];
```

Legality of Tiling

- Tiling is safe only if it does not change the order in which memory locations are read/written
  - We'll talk about correctness after memory hierarchies
- Tiling can conceptually be used to perform the decomposition into threads and blocks
  - We'll show this later, too

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
Matrix Multiplication
A Simple Host Version in C

```c
void MatrixMulOnHost(float* M, float* N, float* P, int Width)
{
    for (int i = 0; i < Width; ++i)
        for (int j = 0; j < Width; ++j) {
            double sum = 0;
            for (int k = 0; k < Width; ++k) {
                double a = M[i * width + k];
                double b = N[k * width + j];
                sum += a * b;
            }
            P[i * Width + j] = sum;
        }
}
```

CUDA Code – Kernel Execution Configuration

```c
// Setup the execution configuration
dim3 dimBlock(BLOCK_SIZE, BLOCK_SIZE);
dim3 dimGrid(M.height / dimBlock.y, M.width / dimBlock.x);
```

For very large N and M dimensions, one will need to add another level of blocking and execute the second-level blocks sequentially.

CUDA Code – Kernel Overview

```c
// Block index
int bx = blockIdx.x;
int by = blockIdx.y;
// Thread index
int tx = threadIdx.x;
int ty = threadIdx.y;

// Pvalue stores the element of the block sub-matrix
// that is computed by the thread
float Pvalue = 0;

// Loop over all the sub-matrices of M and N
// required to compute the block sub-matrix
for (int m = 0; m < M.width/BLOCK_SIZE; ++m) {
    code from the next few slides;
}```
CUDA Code - Load Data to Shared Memory

// Get a pointer to the current sub-matrix Msub of M
Matrix Msub = GetSubMatrix(M, m, by);
// Get a pointer to the current sub-matrix Nsub of N
Matrix Nsub = GetSubMatrix(N, bx, m);
__shared__ float Ms[BLOCK_SIZE][BLOCK_SIZE];
__shared__ float Ns[BLOCK_SIZE][BLOCK_SIZE];
// each thread loads one element of the sub-matrix
Ms[ty][tx] = GetMatrixElement(Msub, tx, ty);
// each thread loads one element of the sub-matrix
Ns[ty][tx] = GetMatrixElement(Nsub, tx, ty);

CUDA Code - Compute Result

// Synchronize to make sure the sub-matrices are loaded
// before starting the computation
__syncthreads();

// each thread computes one element of the block sub-matrix
for (int k = 0; k < BLOCK_SIZE; ++k)
    Pvalue += Ms[ty][k] * Ns[k][tx];

// Synchronize to make sure that the preceding
// computation is done before loading two new
// sub-matrices of M and N in the next iteration
__syncthreads();

CUDA Code - Save Result

// Get a pointer to the block sub-matrix of P
Matrix Psub = GetSubMatrix(P, bx, by);

// Write the block sub-matrix to device memory;
// each thread writes one element
SetMatrixElement(Psub, tx, ty, Pvalue);

This code should run at about 150 Gflops on a GTX or Tesla.
State-of-the-art mapping (in CUBLAS 3.2 on C2050) yields just above 600 Gflops. Higher on GTX480.

Matrix Multiply in CUDA

- Imagine you want to compute extremely large matrices.
  - That don’t fit in global memory
- This is where an additional level of tiling could be used, between grids
Summary of Lecture

- How to place data in constant memory and shared memory
- Introduction to Tiling transformation
- Matrix multiply example

Next Time

- Complete this example
  - Also, registers and texture memory
- Reasoning about reuse and locality
- Introduction to bandwidth optimization