Lecture 5
Matrix-Matrix Product

Based on the lecture materials of Hwu (UIUC) and Kirk (NVIDIA)
Matrix Multiplication

• Simple version first
  – illustrate basic features of memory and thread management in CUDA programs
  – Thread ID usage
  – Memory data transfer API between host and device
  – Analyze performance

• Extend to version which employs shared memory
Square Matrix Multiplication

- P = M * N of size WIDTH x WIDTH
- Without tiling:
  - One thread calculates one element of P
  - M and N are loaded WIDTH times from global memory
Memory Layout of a Matrix

This order will be important to compute the location of the element in the matrix according to thread and block indices.
Step 1: Simple Host Version

// Matrix multiplication on the (CPU) host
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;
        }
}
Step 2: Transfer Data to Device from Host

```c
void MatrixMulOnDevice(float* M, float* N, float* P, int Width) {
    int size = Width * Width * sizeof(float);
    float* Md, Nd, Pd;
    
    // 1. Allocate and Load M, N to device memory
    cudaMalloc(&Md, size);
    cudaMemcpy(Md, M, size, cudaMemcpyHostToDevice);
    cudaMalloc(&Nd, size);
    cudaMemcpy(Nd, N, size, cudaMemcpyHostToDevice);
    // Allocate P on the device
    cudaMalloc(&Pd, size);
    cudaMemcpy(Nd, N, size, cudaMemcpyHostToDevice);
    // Allocate P on the device
    cudaMalloc(&Pd, size);
```
Step 3: Output Matrix Data Transfer
(Host-side Code)

2. // Kernel invocation code – to be shown later
   ...

3. // Read P from the device
cudAmemcpy(P, Pd, size, cudAmemcpyDeviceToHost);

   // Free device matrices
cudaFree(Md); cudaFree(Nd); cudaFree(Pd);
Step 4: Kernel Function

// Matrix multiplication kernel – per thread code

__global__ void MatrixMulKernel(float* Md, float* Nd, float* Pd, int Width)
{

    // Pvalue is used to store the element of the matrix
    // that is computed by the thread
    float Pvalue = 0;

Step 4: Kernel Function (cont.)

for (int k = 0; k < Width; ++k) {
    float Melement = Md[threadIdx.y*Width+k];
    float Nelement = Nd[k*Width+threadIdx.x];
    Pvalue += Melement * Nelement;
}
Pd[threadIdx.y*Width+threadIdx.x] = Pvalue;
Step 5: Kernel Invocation
(Host-side Code)

// Setup the execution configuration
  dim3 dimGrid(1, 1);
  dim3 dimBlock(Width, Width);

// Launch the device computation threads!
MatrixMulKernel<<<dimGrid, dimBlock>>>(Md, Nd, Pd, Width);
First version: One Thread Block

- One Block of threads compute matrix $P_d$
  - Each thread computes one element of $P_d$
- Each thread
  - Loads a row of matrix $M_d$
  - Loads a column of matrix $N_d$
  - Perform one multiply and addition for each pair of $M_d$ and $N_d$ elements
  - Compute to off-chip memory access ratio close to 1:1 (not very high)
- Size of matrix limited by the number of threads allowed in a thread block
  - It is 512. So the number allowed is <23
Extend to Arbitrary Sized Square Matrices

- Use more than one block
- Have each 2D thread block to compute a \((TILE\_WIDTH)^2\) sub-matrix (tile) of the result matrix
  - Each has \((TILE\_WIDTH)^2\) threads
- Generate a 2D Grid of \((WIDTH/TILE\_WIDTH)^2\) blocks

You still need to put a loop around the kernel call for cases where \(WIDTH/TILE\_WIDTH\) is greater than max grid size (64K)!
Matrix Multiplication Using Multiple Blocks

- Break-up Pd into tiles
- Each block calculates one tile
  - Each thread calculates one element
  - Block size equal tile size
A Small Example

```
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P0,0</td>
<td>P1,0</td>
<td>P2,0</td>
<td>P3,0</td>
</tr>
<tr>
<td>P0,1</td>
<td>P1,1</td>
<td>P2,1</td>
<td>P3,1</td>
</tr>
<tr>
<td>P0,2</td>
<td>P1,2</td>
<td>P2,2</td>
<td>P3,2</td>
</tr>
<tr>
<td>P0,3</td>
<td>P1,3</td>
<td>P2,3</td>
<td>P3,3</td>
</tr>
</tbody>
</table>
```

TILE_WIDTH = 2
A Small Example: Multiplication
Revised Matrix Multiplication Kernel using Multiple Blocks

```c
__global__ void MatrixMulKernel(float* Md, float* Nd, float* Pd, int Width)
{
    // Calculate the row index of the Pd element and M
    int Row = blockIdx.y*TILE_WIDTH + threadIdx.y;
    // Calculate the column index of Pd and N
    int Col = blockIdx.x*TILE_WIDTH + threadIdx.x;

    float Pvalue = 0;
    // each thread computes one element of the block sub-matrix
    for (int k = 0; k < Width; ++k)
        Pvalue += Md[Row*Width+k] * Nd[k*Width+Col];

    Pd[Row*Width+Col] = Pvalue;
}
```

Note how the Row and Column indices are computed.
Analysis of this version

• Each thread loads $2 \cdot \text{Width}$ elements and from global memory and does that many floating point computations
  – So this version does one flop per 4 byte memory load
• On a G80 bandwidth of memory transfer from global memory is $\sim 86$ GB/sec, and so we are limited to $\sim 21$ G floating point loads
  – Flop rate is also limited to this number.
• But the 8800 GTX is supposed to achieve $\sim 340$ Gflops
  – Need to use shared memory and do more computations per global memory access.
Hardware Implementation: Memory Architecture

- The local, global, constant, and texture spaces are regions of device memory
- 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
More 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

<table>
<thead>
<tr>
<th>Memory</th>
<th>Location</th>
<th>Cached</th>
<th>Access</th>
<th>Who</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local</td>
<td>Off-chip</td>
<td>No</td>
<td>Read/write</td>
<td>One thread</td>
</tr>
<tr>
<td>Shared</td>
<td>On-chip</td>
<td>N/A - resident</td>
<td>Read/write</td>
<td>All threads in a block</td>
</tr>
<tr>
<td>Global</td>
<td>Off-chip</td>
<td>No</td>
<td>Read/write</td>
<td>All threads + host</td>
</tr>
<tr>
<td>Constant</td>
<td>Off-chip</td>
<td>Yes</td>
<td>Read</td>
<td>All threads + host</td>
</tr>
<tr>
<td>Texture</td>
<td>Off-chip</td>
<td>Yes</td>
<td>Read</td>
<td>All threads + host</td>
</tr>
</tbody>
</table>
Access Times

- Register – dedicated HW - single cycle
- Shared Memory – dedicated HW - single cycle
- 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
Language Extensions:
Variable Type Qualifiers

| __device__ __local__ | int LocalVar; | local | thread | thread |
| __device__ __shared__ | int SharedVar; | shared | block | block |
| __device__ | int GlobalVar; | global | grid | application |
| __device__ __constant__ | int ConstantVar; | constant | grid | application |

- __device__ is optional when used with __local__, __shared__, or __constant__

- **Automatic variables** without any qualifier reside in a register
  - Except arrays that reside in local memory
Variable Type Restrictions

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

- Global memory resides in device memory (DRAM) - much slower access than shared memory
- So, a profitable way of performing computation on the device is to tile data to take advantage of fast shared memory:
  - Partition data into subsets that fit into shared memory
  - Handle each data subset with one thread block by:
    - Loading the subset from global memory to shared memory, using multiple threads to exploit memory-level parallelism
    - Performing the computation on the subset from shared memory; each thread can efficiently multi-pass over any data element
    - Copying results from shared memory to global memory
A Common Programming Strategy (Cont.)

• Constant memory also resides in device memory (DRAM) - much slower access than shared memory
  – But… cached!
  – Highly efficient access for read-only data

• Carefully divide data according to access patterns
  – R/Only  → constant memory (very fast if in cache)
  – R/W shared within Block  → shared memory (very fast)
  – R/W within each thread  → registers (very fast)
  – R/W inputs/results  → global memory (very slow)

Texture memory -- later
Idea: Use Shared Memory to reuse global memory data

• Each input element is read by Width threads.
• Load each element into Shared Memory and have several threads use the local version to reduce the memory bandwidth
  – Tiled algorithms
Tiled Multiply

- Break up the execution of the kernel into phases so that the data accesses in each phase is focused on one subset (tile) of Md and Nd.
Breaking Md and Nd into Tiles
Each phase of a Thread Block uses one tile from \(\text{Md} \) and one from \(\text{Nd} \)

<table>
<thead>
<tr>
<th>Phase 1</th>
<th>Phase 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(T_{0,0})</td>
<td>(\textbf{Md}_{0,0})</td>
</tr>
<tr>
<td>(\downarrow)</td>
<td>(\text{Mds}_{0,0})</td>
</tr>
<tr>
<td>(T_{1,0})</td>
<td>(\textbf{Md}_{1,0})</td>
</tr>
<tr>
<td>(\downarrow)</td>
<td>(\text{Mds}_{1,0})</td>
</tr>
<tr>
<td>(T_{0,1})</td>
<td>(\textbf{Md}_{0,1})</td>
</tr>
<tr>
<td>(\downarrow)</td>
<td>(\text{Mds}_{0,1})</td>
</tr>
<tr>
<td>(T_{1,1})</td>
<td>(\textbf{Md}_{1,1})</td>
</tr>
<tr>
<td>(\downarrow)</td>
<td>(\text{Mds}_{1,1})</td>
</tr>
</tbody>
</table>

*Note: \(\text{PValue}\) and \(\text{PdValue}\) are placeholders for the actual values.*
Threads, Warps, Blocks

• There are (up to) 32 threads in a Warp
  – Only <32 when there are fewer than 32 total threads
• There are (up to) 16 Warps in a Block
• Each Block (and thus, each Warp) executes on a single SM
• G80 has 16 SMs
• At least 16 Blocks required to “fill” the device
• More is better
  – If resources (registers, thread space, shared memory) allow, more than 1 Block can occupy each SM
First-order Size Considerations in G80

- Each **thread block** should have many threads
  - TILE_WIDTH of 16 gives $16 \times 16 = 256$ threads

- There should be many thread blocks
  - A 1024*1024 Pd gives $64 \times 64 = 4096$ Thread Blocks

- Each thread block perform $2 \times 256 = 512$ float loads from global memory for $256 \times (2 \times 16) = 8,192$ mul/add operations.
  - Memory bandwidth no longer a limiting factor
// Setup the execution configuration

dim3 dimBlock(TILE_WIDTH, TILE_WIDTH);
dim3 dimGrid(Width / TILE_WIDTH,
             Width / TILE_WIDTH);
Tiled Matrix Multiplication Kernel

__global__ void MatrixMulKernel(float* Md, float* Nd, float* Pd, int Width)
{
    __shared__ float Mds[TILE_WIDTH][TILE_WIDTH];
    __shared__ float Nds[TILE_WIDTH][TILE_WIDTH];

    int bx = blockIdx.x;  int by = blockIdx.y;
    int tx = threadIdx.x; int ty = threadIdx.y;

    int Row = by * TILE_WIDTH + ty;
    int Col = bx * TILE_WIDTH + tx;

    float Pvalue = 0;

    for (int m = 0; m < Width/TILE_WIDTH; ++m) {
        Mds[ty][tx] = Md[Row*Width + (m*TILE_WIDTH + tx)];
        Nds[ty][tx] = Nd[Col + (m*TILE_WIDTH + ty)*Width];
        __syncthreads();

        for (int k = 0; k < TILE_WIDTH; ++k)
            Pvalue += Mds[ty][k] * Nds[k][tx];
        __syncthreads();
    }

    Pd[Row*Width+Col] = Pvalue;
}

Tiled Multiply

- Each block computes one square sub-matrix $P_{d_{sub}}$ of size $TILE\_WIDTH$
- Each thread computes one element of $P_{d_{sub}}$
G80 Shared Memory and Threading

• Each SM in G80 has 16KB shared memory
  – SM size is implementation dependent!
  – For TILE_WIDTH = 16, each thread block uses 2*256*4B = 2KB of shared memory.
  – Can potentially have up to 8 Thread Blocks actively executing
    • This allows up to 8*512 = 4,096 pending loads. (2 per thread, 256 threads per block)
  – The next TILE_WIDTH 32 would lead to 2*32*32*4B = 8KB shared memory usage per thread block, allowing only up to two thread blocks active at the same time

• Using 16x16 tiling, we reduce the accesses to the global memory by a factor of 16
  – The 86.4B/s bandwidth can now support (86.4/4)*16 = 347.6 GFLOPS!
Tiling Size Effects

![Bar chart showing performance differences for tiled and unrolled operations with various tile sizes.](chart.png)