CUDA Efficient Programming
Outline

1. Overview and general concepts
2. Performance Metrics
3. Memory Optimizations
4. Execution Optimization
5. Tools Overview
**Different worlds: host and device**

<table>
<thead>
<tr>
<th></th>
<th>Host</th>
<th>Device</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threading resources</td>
<td>2 threads per core (SMT), 24/32 threads per node. The thread is the atomic execution unit.</td>
<td>e.g.: 1536 (thd x sm) * 14 (sm) = 21504. The Warp (32 thd) is the atomic execution unit.</td>
</tr>
<tr>
<td>Threads</td>
<td>«Heavy» entities, context switches and resources management.</td>
<td>Extremely lightweight, managed grouped into warps, fast context switch, no resources management (statically allocated once).</td>
</tr>
<tr>
<td>Memory</td>
<td>e.g.: 48 GB / 32 thd = 1.5 GB/thd, 300 cycles lat., 6.4 GB/s band (DDR3), 3 caching levels with lots of speculation logic.</td>
<td>e.g.: 6 GB / 21504 thd = 0.3 MB/thd, 600 cycles lat*, 144 GB/s band (GDDR5)*, fake caches.</td>
</tr>
</tbody>
</table>

* coalesced
Obtaining maximum performance benefit

📍 Focus on achieving high occupancy (more on this later, for know you can translate «high occupancy» as «many many… many threads in flight»).
📍 Focus on how to exploit the SIMT (data parallel) programming model.
📍 Deeply analyze your algorithm in order to find hotspots and embarassingly parallel-enabled portions.

Furthermore, pay attention to the Amdahl’s law:

\[
S = \frac{1}{(1 - P) + P/N}
\]

**Hint**: avoid the jump-start-to-code approach: porting your serial and/or multithreaded and/or message passing CPU application to GPU is *not* in general an easy task.
CUDA Enabled GPU: compute capability

The compute capability is a kind of version tag that identifies:

- instructions and features supported by the board;
- coalescing rules;
- the board’s resources constraints;
- throughput of some instructions (hardware implementation).

The compute capability is given as a `major.dot.minor` version number (i.e: 2.0, 2.1, 3.0, 3.5).
## Compute capability: resources constraints

<table>
<thead>
<tr>
<th>Technical Specifications</th>
<th>Compute Capability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.0</td>
</tr>
<tr>
<td>Maximum dimensionality of grid of thread blocks</td>
<td>2</td>
</tr>
<tr>
<td>Maximum x-dimension of a grid of thread blocks</td>
<td></td>
</tr>
<tr>
<td>Maximum y- or z-dimension of a grid of thread blocks</td>
<td></td>
</tr>
<tr>
<td>Maximum dimensionality of thread block</td>
<td>3</td>
</tr>
<tr>
<td>Maximum x- or y-dimension of a block</td>
<td>512</td>
</tr>
<tr>
<td>Maximum z-dimension of a block</td>
<td>64</td>
</tr>
<tr>
<td>Maximum number of threads per block</td>
<td>512</td>
</tr>
<tr>
<td>Warp size</td>
<td></td>
</tr>
<tr>
<td>Maximum number of resident blocks per multiprocessor</td>
<td></td>
</tr>
<tr>
<td>Maximum number of resident warps per multiprocessor</td>
<td></td>
</tr>
<tr>
<td>Maximum number of resident threads per multiprocessor</td>
<td></td>
</tr>
<tr>
<td>Number of 32-bit registers per multiprocessor</td>
<td>8 K</td>
</tr>
<tr>
<td>Maximum number of 32-bit registers per thread</td>
<td></td>
</tr>
<tr>
<td>Maximum amount of shared memory per multiprocessor</td>
<td>16 KB</td>
</tr>
<tr>
<td>Number of shared memory banks</td>
<td></td>
</tr>
<tr>
<td>Amount of local memory per thread</td>
<td>16 KB</td>
</tr>
<tr>
<td>Constant memory size</td>
<td></td>
</tr>
<tr>
<td>Cache working set per multiprocessor for constant memory</td>
<td></td>
</tr>
<tr>
<td>Cache working set per multiprocessor for texture memory</td>
<td></td>
</tr>
<tr>
<td>Maximum width for a 1D texture reference bound to a CUDA array</td>
<td>8192</td>
</tr>
</tbody>
</table>
Performance metrics
Performance metrics

- **Wall-clock time**
  - you always want to keep that one at a minimum

- Theoretical (peak) bandwidth Vs effective bandwidth
  - that allows you to measure performance of a **memory-bound kernel**

- Theoretical (peak) FLOPS* Vs effective FLOPS**
  - that allows you to measure performance of a **compute-bound kernel**

---

*theoretical **F**Loating point **O**peration **P**er **S**econd: different kind of ops have in general different throughput. Ops throughput differs among the compute capabilities.

**effective **F**Loating point **O**peration **P**er **S**econd: can be difficult to count the effective number of operations that the kernel is doing during execution.*
Timing

- You can use the standard timing facilities (host side) in an almost standard way…
- …but remember: CUDA calls can be asynchronous!

```c
start = clock();
my_kernel<<<blocks, threads>>>();
cudaDeviceSynchronize();
end = clock();
```

- CUDA provides the **cudaEvents** facility. They grant you access to the GPU timer.
- Needed to time a single stream without loosing Host/Device concurrency.

```c
cudaEvent_t start, stop;
cudaEventCreate(start); cudaEventCreate(stop);
cudaEventRecord(start, 0);
My_kernel<<<block2, threads>>> ();
cudaEventRecord(stop, 0);
cudaEventSynchronize(stop);
float ElapsedTime;
cudaEventElapsedTime(&elapsedTime, start, stop);
cudaEventDestroy(start); cudaEventDestroy(stop);
```
Bandwidth

1. Get GPU main memory’s theoretical bandwidth (ECC off):

\[ B = \text{freq} \times \text{busw} \times nlin = (1.107 \text{ GHz}) \times \left( \frac{512 \times 2}{8} \right) = 141.6 \text{ GB/s} \]

2. Get kernel’s effective bandwidth:

```c
// slice of a totally memory bound kernel ahead: memcpy D2D;
// dim(mat_a)=dim(mat_b)=2048x2048
int xIdx = blockIdx.x*blockDim.x+threadIdx.x;
int yIdx = blockIdx.y*blockDim.y+threadIdx.y;
if (xIdx < 2048 && yIdx < 2048)
    mat_a[xIdx][yIdx]=mat_b[xIdx][yIdx];
```

3. Compute the effective to theoretical bandwidth ratio. Then ask:
   - Is it around 70-75% of the peak? Good job*.
   - Is it much lower than 70% of the peak? Plenty of room for memory access optimization and performance improvement*.

*once again: the bandwidth metric is valid for memory bound kernel
Memory Optimizations
Data Transfers

- Host and Device have their own address space
- GPU boards are connected to host via PCIe bus
- Low bandwidth, extremely low latency

<table>
<thead>
<tr>
<th>Technology</th>
<th>Peak Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>PClex GEN2 (16x, full duplex)</td>
<td>8 GB/s (peak)</td>
</tr>
<tr>
<td>PClex GEN3 (16x, full duplex)</td>
<td>16 GB/s (peak)</td>
</tr>
<tr>
<td>DDR3 (full duplex)</td>
<td>26 GB/s (single channel)</td>
</tr>
</tbody>
</table>

Focus on how to minimize transfers and copybacks*.  

* Try to find a good trade off!
Pinned (or page-locked memory) is a main memory area that is not pageable by the operating system;
Ensures faster transfers (the DMA engine can work without CPU intervention);
The only way to get closer to PCI peak bandwidth;
Allows CUDA asynchronous operations (including *Zero Copy*) to work correctly.

```c
// allocate page-locked memory
cudaMallocHost(&area, sizeof(double) * N);
// free page-locked memory
cudaFreeHost(area);

// allocate regular memory
area = (double*) malloc( sizeof(double) * N );
// lock area pages (CUDA >= 4.0)
ccudaHostRegister( area, sizeof(double) * N, cudaHostRegisterPortable );
// unlock area pages (CUDA >= 4.0)
ccudaHostUnregister(area);
// free regular memory
cudaFreeHost(area);
```

**Warning:** page-locked memory is a scarce resource.
Use with caution: allocating too much page-locked memory can reduce overall system performance

**Breath relief:** nVidia guys allocate up to 95% of a Linux compute node memory as ‘pinned’ memory in real world applications «without much problems» they say...
Zero Copy

CUDA allows to map a page-locked host memory area to the device’s address space;

```c
// allocate page-locked and mapped memory
cudaHostAlloc(&area, sizeof(double) * N, cudaHostAllocMapped);
// invoke retrieving device pointer for mapped area
cudaHostGetDevicePointer( &dev_area, area, 0 );
my_kernel<<< g, b >>>( dev_area );
// free page-locked and mapped memory
cudaFreeHost(area);
```

The only way to provide on-the-fly a kernel data that doesn’t fit into the device’s global memory.

Very convenient for large data with sparse access pattern.
CUDA 4.0 introduced one (virtual) address space for all CPU and GPUs memory:

- automatically detects physical memory location from pointer value
- enables libraries to simplify their interfaces (e.g. `cudaMemcpy`)

<table>
<thead>
<tr>
<th>Pre-UVA</th>
<th>UVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each source-destination permutation has its own option</td>
<td>Same interface</td>
</tr>
<tr>
<td><code>cudaMemcpyHostToHost</code></td>
<td><code>cudaMemcpyDefault</code></td>
</tr>
<tr>
<td><code>cudaMemcpyHostToDevice</code></td>
<td></td>
</tr>
<tr>
<td><code>cudaMemcpyDeviceToDevice</code></td>
<td></td>
</tr>
<tr>
<td><code>cudaMemcpyDeviceToDevice</code></td>
<td></td>
</tr>
</tbody>
</table>

Pointers returned by `cudaHostAlloc()` can be used directly from within kernels running on UVA enabled devices (i.e. there is no need to obtain a device pointer via `cudaHostGetDevicePointer()`).
cudaDeviceCanAccessPeer(&can_access_peer_0_1, gpuid_0, gpuid_1);
cudaDeviceCanAccessPeer(&can_access_peer_1_0, gpuid_1, gpuid_0);

cudaSetDevice(gpuid_0);
cudaDeviceEnablePeerAccess(gpuid_1, 0);

cudaSetDevice(gpuid_1);
cudaDeviceEnablePeerAccess(gpuid_0, 0);

cudaMemcpy(gpu0_buf, gpu1_buf, buf_size, cudaMemcpyDefault);

- **cudaMemcpy()** knows that our buffers are on different devices (UVA), will do a P2P copy now

- Note that this will *transparently* fall back to a normal copy through the host if P2P is not available
Multi-GPUs: direct access

__global__ void SimpleKernel(float *src, float *dst) {
    const int idx = blockIdx.x * blockDim.x + threadIdx.x;
    dst[idx] = src[idx];
}

cudaDeviceCanAccessPeer(&can_access_peer_0_1, gpuid_0, gpuid_1);
cudaDeviceCanAccessPeer(&can_access_peer_1_0, gpuid_1, gpuid_0);

cudaSetDevice(gpuid_0);
cudaDeviceEnablePeerAccess(gpuid_1, 0);
cudaSetDevice(gpuid_1);
cudaDeviceEnablePeerAccess(gpuid_0, 0);

cudaSetDevice(gpuid_0);
SimpleKernel<<<blocks, threads>>> (gpu0_buf, gpu1_buf);
SimpleKernel<<<blocks, threads>>> (gpu1_buf, gpu0_buf);
cudaSetDevice(gpuid_1);
SimpleKernel<<<blocks, threads>>> (gpu0_buf, gpu1_buf);
SimpleKernel<<<blocks, threads>>> (gpu1_buf, gpu0_buf);

- After P2P initialization, this kernel can now read and write data in the memory of multiple GPUs (just dereferencing pointers!)
- UVA ensures that the kernel knows whether its argument is from local memory, another GPU or zero-copy from the host
Asynchronous CPU/GPU operations

- Asynchronous operations: control is returned to the host thread before the device has completed the requested task
  - Kernel calls are asynchronous by default
  - Memory copies from host to device of a memory block of 64 KB or less
  - Memory set function calls
  - The `cudaMemcpy()` has an asynchronous version (`cudaMemcpyAsync`)

- Remember: standard memory transfers and copybacks are *blocking*

```c
// First transfer
cudaMemcpyAsync(d_A, h_A, size, cudaMemcpyHostToDevice, 0);
// First invocation
MyKernel<<<100, 512, 0, 0>>>(d_A, size);
// Second transfer
cudaMemcpyAsync(d_B, h_B, size, cudaMemcpyHostToDevice, 0);
// Second invocation
MyKernel2<<<100, 512, 0, 0>>>(d_B, size);
// Wrapup
cudaMemcpyAsync(h_A, d_A, size, cudaMemcpyDeviceToHost, 0);
cudaMemcpyAsync(h_B, d_B, size, cudaMemcpyDeviceToHost, 0);
cudaThreadSynchronize();
```
Asynchronous GPU Operations: CUDA Stream

**A stream** is a FIFO command queue;

**Default stream (aka stream ‘0’):** Kernel launches and memory copies that do not specify any stream (or set the stream to zero) are issued to the default stream.

A stream is independent to every other active stream;

Streams are the main way to exploit concurrent execution and I/O operations

**Explicit Synchronization:**

- `cudaDeviceSynchronize()`
  - blocks host until all issued CUDA calls are complete
- `cudaStreamSynchronize(streamId)`
  - blocks host until all CUDA calls in streamId are complete
- `cudaStreamWaitEvent(streamId, event)`
  - all commands added to the stream delay their execution until the event has completed

**Implicit Synchronization:**

- any CUDA command to the default stream,
- a page-locked host memory allocation,
- a device memory set or allocation,
- ...
CUDA streams enable concurrency

Concurrency: the ability to perform multiple CUDA operations simultaneously.
Fermi architecture can simultaneously support:
- Up to 16 CUDA kernels on GPU
- 2 cudaMemcpyAsyncs (in opposite directions)
- Computation on the CPU

Requirements for Concurrency:
- CUDA operations must be in different, non-0, streams
- cudaMemcpyAsync with host from 'pinned' memory
- Sufficient resources must be available
  - cudaMemcpyAsyncs in different directions
  - Device resources (SMEM, registers, blocks, etc.)

Serial:
cudaMemcpyAsync(H2D) | Kernel <<< >>> | cudaMemcpyAsync(D2H)

2-way concurrency:
cudaMemcpyAsync(H2D) | K1 | K2 | K3 | K4

3-way concurrency:
HD1 | K1 | HD1 | K2 | HD2
HD2 | K1 | HD1 | K3 | HD3
HD3 | K1 | HD1 | K4 | HD4

4-way concurrency:
HD1 | K1.1 | HD1 | K2.1 | K2.2 | K2.3 | DH1 | DH2
HD2 | K1.2 | HD2 | K2.1 | K2.2 | K2.3 | DH2 | DH3
HD3 | K1.3 | HD3 | K3.1 | K3.2 | K3.3 | DH3 | DH4
HD4 | K1.4 | HD4 | K4.1 | K4.2 | K4.3 | DH4 | DH5
HD5 | K1.5 | HD5 | K5.1 | K5.2 | K5.3 | DH5 | DH6
HD6 | K1.6 | HD6 | K6.1 | K6.2 | K6.3 | DH6 | K7 on CPU
CUDA streams enable concurrency

cudaStream_t stream[3];
for (int i=0; i<3; ++i) cudaStreamCreate(&stream[i]);

float* hPtr; cudaMallocHost((void**)&hPtr, 3 * size);

for (int i=0; i<3; ++i) {
    cudaMemcpyAsync(d_inp + i*size, hPtr + i*size,
                     size, cudaMemcpyHostToDevice, stream[i]);
    MyKernel<<<100, 512, 0, stream[i]>>>(d_out+i*size, d_inp+i*size, size);
    cudaMemcpyAsync(hPtr + i*size, d_out + i*size,
                     size, cudaMemcpyDeviceToHost, stream[i]);
}
cudaDeviceSynchronize();

for (int i=0; i<3; ++i) cudaStreamDestroy(&stream[i]);
CUDA Streams: overlapping kernels execution

- Starting from capability 2.0 the board has the ability to overlap computations from multiple kernels.
  - CUDA kernels are in different streams,
  - no operations occur on the default stream,
  - the active streams are less than 16.
  - no synchronization happens between command stages,

- Threadblocks for a given kernel are scheduled if all threadblocks for preceding kernels have already been scheduled and there are SM resources available

- Concurrent execution can be limited by implicit dependencies due to hardware limitations: command issue order matters!

```
// Depth-first commands submission.
Beware: PSEUDO CODE ahead:
for each StreamId:
  do H2D data tile transfer
  launch kernel on data tile
  do D2H result data tile transfer

// Breadth-first commands submission.
Beware: PSEUDO CODE ahead:
for each StreamId:
  do H2D data tile transfer
  launch kernel on data tile
  do D2H result data tile transfer
```

**hint:** depth-first commands submission is usually better on Fermi. It’s a no-issue for Kepler K20 with HyperQ technology
Global Memory

- It is a memory area with the same purpose of the host’s main memory;
- High(er) bandwidth, high(er) latency;
- In order to exploit its bandwidth at best, all accesses must be **coalesced**, i.e. memory accesses from different threads need to be grouped together and serviced in one memory transaction.
- beware: some threads memory access patterns can be coalesced, some others cannot (coalescence rules depends on GPU compute capability)
- **FERMI** architecture introduces caching mechanisms for GMEM accesses (constant and texture are cached since 1.0)
- L1: private to thread, virtual cache implemented into shared memory
- L2: 768KB, grid-coherent, 25% better latency than DRAM

```c
// L1 = 48 KB
// SH = 16 KB
cudaFuncSetCacheConfig( kernel, cudaFuncCachePreferL1);
// L1 = 16 KB
// SH = 48 KB
cudaFuncSetCacheConfig( kernel, cudaFuncCachePreferShared );
```

**Kepler** architecture introduced some improvements:
32 KB + 32 KB partition option
Global Memory (pre-Fermi)

**Compute capability 1.0 and 1.1**

- A global memory request for a warp is split into two memory requests, one for each half-warp, that are issued independently.
- In order to exploit its bandwidth at best, all accesses must be **coalesced** (*half-warp accesses contiguously region of device memory*).
- Threads must access the words in a **strictly increasing sequence**: *the n\(^{th}\) thread in the half-warp must access the n\(^{th}\) word*.
- All 16 words must lie in the same **aligned** segment
- A coalesced memory access results in:
  - in one 64-byte memory transaction, for 4-byte words
  - in one 128-byte memory transaction, for 8-byte words
  - in two 128-byte memory transactions, for 16-byte words
Coalescing (pre-Fermi)

Compute capability 1.0 and 1.1

- stricter access requirements
- memory accesses serviced on a half-warp (16 threads) basis
- not all threads need to participate but their memory accesses must be aligned and in order:
  - k-th thread must access k-th word in the segment

![Diagram showing coalescing transactions and errors](image)
Coalescing (pre-Fermi)

Compute capability 1.2 and 1.3

- The memory controller is much improved
FERMI (Compute Capability 2.x) GMEM Operations

- Two types of loads:
  - **Caching**
    - default mode
    - attempts to hit in L1, then L2, then GMEM
    - load granularity is **128-byte line**
  - **Non-caching**
    - compile with `-Xptxas -dlcm=cg`
    - attempts to hit in L2, then GMEM does not hit in L1.
    - load granularity is **32-bytes**

- **Stores:**
  - Invalidate L1, write-back for L2
Global Memory Load Operation (Fermi)

- Memory operations are issued per warp (32 threads)
  - just like all other instructions

- Operation:
  - Threads in a warp provide memory addresses
  - Determine which lines/segments are needed
  - Request the needed lines/segments

<table>
<thead>
<tr>
<th>Warp requests 32 aligned, consecutive 4-byte words (128 bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Caching Load</strong></td>
</tr>
<tr>
<td>Addresses fall within 1 cache-line</td>
</tr>
<tr>
<td>128 bytes move across the bus</td>
</tr>
<tr>
<td>Bus utilization: <strong>100%</strong></td>
</tr>
</tbody>
</table>

Addresses from a warp

Memory addresses

0 32 64 96 128 160 192 224 256 288 320 352 384 416 448
## Global Memory Load Operation (Fermi)

### Warp requests 32 aligned, permuted 4-byte words (128 bytes)

<table>
<thead>
<tr>
<th></th>
<th>Caching Load</th>
<th>Non-caching Load</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addresses fall within</td>
<td>1 cache-line</td>
<td>4 segments</td>
</tr>
<tr>
<td></td>
<td>128 bytes move across the bus</td>
<td>128 bytes move across the bus</td>
</tr>
<tr>
<td>Bus utilization:</td>
<td><strong>100%</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

### Warp requests 32 misaligned, consecutive 4-byte words (128 bytes)

<table>
<thead>
<tr>
<th></th>
<th>Caching Load</th>
<th>Non-caching Load</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addresses fall within</td>
<td>2 cache-lines</td>
<td>at most 5 segments</td>
</tr>
<tr>
<td></td>
<td>256 bytes move across the bus</td>
<td>160 bytes move across the bus</td>
</tr>
<tr>
<td>Bus utilization:</td>
<td><strong>50%</strong></td>
<td><strong>at least 80%</strong></td>
</tr>
</tbody>
</table>
Global Memory Load Operation (Fermi)

### All threads in a warp request the same 4-byte word (4 bytes)

<table>
<thead>
<tr>
<th>Caching Load</th>
<th>Non-caching Load</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addresses fall within 1 cache-line</td>
<td>Addresses fall within 1 segments</td>
</tr>
<tr>
<td>128 bytes move across the bus</td>
<td>32 bytes move across the bus</td>
</tr>
<tr>
<td>Bus utilization: <strong>3.125%</strong></td>
<td>Bus utilization: <strong>12.5%</strong></td>
</tr>
</tbody>
</table>

### Warp requests 32 scattered 4-byte words (128 bytes)

<table>
<thead>
<tr>
<th>Caching Load</th>
<th>Non-caching Load</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addresses fall within N cache-lines</td>
<td>Addresses fall within N segments</td>
</tr>
<tr>
<td>N*128 bytes move across the bus</td>
<td>N*32 bytes move across the bus</td>
</tr>
</tbody>
</table>
Shared memory

A sort of *explicit* cache (i.e. under programmer control)

Resides on the chip so it is *much* faster than the on-board memory

Divided into equally-sized memory modules (banks) which can be accessed simultaneously (32 banks can be accessed simultaneously by the same warp)

48KB on Fermi by default*

*Kepler* architecture introduced some improvements:
- ability to switch from 4B to 8B banks
- (2x bandwidth for double precision codes)

**Uses:**
- Inter-thread communication within a block
- Cache data to reduce redundant global memory accesses
- To improve global memory access patterns

**Organization:**
- 32 banks, 4-byte wide banks
- Successive 4-byte words belong to different banks
- Each bank has 32-bit per cycle bandwidth.
If at least two threads belonging to the same half-warp (whole warp for capability 1.0) access the same shared memory bank, there is a bank conflict and the accesses are serialized (groups transactions in conflict-free accesses);

If all the threads access the same address, a broadcast is performed;

If part of the half-warp accesses the same address, a multicast is performed (capability >= 2.0);
Lunch break

The second part will start at 14:30. Please, try to be on time 😊
Texture Memory

- **Read only**, must be set by the host;
- Load requests are cached (dedicated cache);
- specifically, texture memories and caches are designed for graphics applications where memory access patterns exhibit a great deal of spatial locality;
- Dedicated texture cache hardware provides:
  - Out-of-bounds index handling (clamp or wrap-around)
  - Optional interpolation (on-the-fly interpolation)
  - Optional format conversion
- could bring benefits if the threads within the same block access memory using regular 2D patterns, but you need appropriate binding;

For typical linear patterns, global memory (if coalesced) is faster.
// allocate array and copy image data
cudaChannelFormatDesc channelDesc =
    cudaCreateChannelDesc(32, 0, 0, 0, cudaChannelFormatKindFloat);
cudaArray* cu_array;
cudaMallocArray( &cu_array, &channelDesc, width, height );
cudaMemcpyToArray( cu_array, 0, 0, h_data, size, cudaMemcpyHostToDevice);
// set texture parameters
tex.addressMode[0] = cudaAddressModeWrap;
tex.addressMode[1] = cudaAddressModeWrap;
tex.filterMode = cudaFilterModeLinear;
tex.normalized = true;  // access with normalized texture coordinates
// Bind the array to the texture
cudaBindTextureToArray( tex, cu_array, channelDesc);

// declare texture reference for 2D float texture
texture<float, 2, cudaReadModeElementType> tex;

__global__ void transformKernel( float* g_odata, int width, int height, float theta)
{
    // calculate normalized texture coordinates
    unsigned int x = blockIdx.x*blockDim.x + threadIdx.x;
    unsigned int y = blockIdx.y*blockDim.y + threadIdx.y;
    float u = x / (float) width;
    float v = y / (float) height;
    // transform coordinates
    u -= 0.5f;
    v -= 0.5f;
    float tu = u*cosf(theta) - v*sinf(theta) + 0.5f;
    float tv = v*cosf(theta) + u*sinf(theta) + 0.5f;
    // read from texture and write to global memory
    g_odata[y*width + x] = tex2D(tex, tu, tv);
}
Kepler global loads through texture

The compiler (LLVM) can detect texture-compliant loads and map them to the new «global load through texture» PTX instruction:

- global loads are going to pass through texture pipeline;
- dedicated cache (no L1 pressure) and memory pipe, relaxed coalescing;
- automatically generated by compiler (no texture map needed) for accesses through compliant pointers (constant and restricted);
- useful for bandwidth-limited kernels
  - global memory bandwidth and texture memory bandwidth stack up.
Constant Memory

- Extremely fast on-board memory area
- **Read only**, must be set by the host
- 64 KB, cached reads in a dedicated L1 (register space)
- Coalesced access if all threads of a warp read the same address (serialized otherwise)
- **__constant__** qualifier in declarations
- Useful:
  - To off-load long argument lists from shared memory (compute capability 1.x)
  - For coefficients and other data that is read uniformly by warps

```c
__device__ __constant__ parameters_t_t args;
__host__ void copy_params(const parameters_t* const host_args) {
    cudaMemcpyToSymbol("args", host_args, sizeof(parameters_t_t));
}
```
Registers

Just like CPU registers, access has no latency;
used for scalar data local to a thread;
taken by the compiler from the Streaming Multiprocessor (SM) pool and statically allocated to each thread;

- each SM of a Fermi GPU has a 32KB register file, 64KB for a Kepler GPU

*register pressure one of the most dangerous occupancy limiting factors.*
Registers

Some tips to reduce register pressure:

- try to offload data to shared memory;
- use launch bounds to force the number of resident blocks;
- limit register usage via compiler option.

```c
#define MAX_THREADS_PER_BLOCK 256
#define MIN_BLOCKS_PER_MP 2

__global__ void
__launch_bounds__( MAX_THREADS_PER_BLOCK, MIN_BLOCKS_PER_MP )
my_kernel( int* inArr, int* outArr ) { ... }
```

```bash
# nvcc -Xptas -v mykernel.cu
ptxas info : Compiling entry function '_Z12my_kernelP9domain_t_' for 'sm_20'
ptxas info : Used 13 registers, 8+16 bytes smem

# nvcc --maxrregcount 10 -Xptas -v mykernel.cu
ptxas info : Compiling entry function '_Z12my_kernelP9domain_t_' for 'sm_20'
ptxas info : Used 10 registers, 12+0 bytes lmem, 8+16 bytes smem
```
Local memory

“Local” because it’s private on a per-thread basis;

it’s actually a global memory area used to spill out data when the SM runs out of register resources;

arrays declared inside a kernel go to LMEM;

local memory accesses are cached (just like global memory).

DISCLAIMER: local memory is not a GPU resource you want to use: It used by the compiler as needed. Its use can hardly hit your kernel performance too: variables that you think are in registers are instead stored in the device global memory.
Execution Optimization
Occupancy

The board’s occupancy is the ratio of active warps to the maximum number of warps supported on a multiprocessor.

Keeping the hardware busy helps the warp scheduler to hide latencies.
Occupancy: constraints

Every board’s resource can become an occupancy limiting factor:

- shared memory allocated per block,
- registers allocated per thread,
- block size
  - (max threads (warp) per SM/max blocks per SM)

Given an actual kernel configuration, is possible to predict the maximum *theoretical occupancy* allowed.
Occupancy: block sizing tips

Some experimentation is required.

However there are some heuristic rules:
- threads per block should be a multiple of warp size;
- a minimum of 64 threads per block should be used;
- 128-256 threads per block is universally known to be a good starting point for further experimentation;
- prefer to split very large blocks into smaller blocks.
Kepler: dynamic parallelism

One of the biggest CUDA limitations is the need to fit a single grid configuration for the whole kernel.

Kepler K20 (in addition to CUDA 5.x) introduced Dynamic Parallelism

It enables a global kernel to be called from within another kernel

The child grid can be dynamically sized and optionally synchronized

---

Parent-Child Launch Nesting

```c
__global__ ChildKernel(void* data){
    //Operate on data
}
__global__ ParentKernel(void *data){
    ChildKernel<<<16, 1>>>(data);
}
// In Host Code:
ParentKernel<<<256, 64>>>(data);
```
Instructions throughput

Arithmetic ops:
- prefer integer shift operators instead of division and modulo (would be less useful with LLVM);
- beware of (implicit) casts (very expensive);
- use intrinsics for transcendental functions where possible;
- try the fast math implementation.
## Capability: instruction throughput

<table>
<thead>
<tr>
<th>Compute Capability</th>
<th>1.0</th>
<th>1.1</th>
<th>1.2</th>
<th>1.3</th>
<th>2.0</th>
<th>2.1</th>
<th>3.0</th>
<th>3.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>32-bit floating-point add, multiply, multiply-add</td>
<td>8</td>
<td>8</td>
<td>32</td>
<td>48</td>
<td>192</td>
<td>192</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64-bit floating-point add, multiply, multiply-add</td>
<td>1</td>
<td>1</td>
<td>16(*)</td>
<td>4</td>
<td>8</td>
<td>64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32-bit integer add</td>
<td>10</td>
<td>10</td>
<td>32</td>
<td>48</td>
<td>160</td>
<td>160</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32-bit integer compare</td>
<td>10</td>
<td>10</td>
<td>32</td>
<td>48</td>
<td>160</td>
<td>160</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32-bit integer shift</td>
<td>8</td>
<td>8</td>
<td>16</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logical operations</td>
<td>8</td>
<td>8</td>
<td>32</td>
<td>48</td>
<td>160</td>
<td>160</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32-bit integer multiply, multiply-add, sum of absolute difference</td>
<td>Multiple instructions</td>
<td>Multiple instructions</td>
<td>16</td>
<td>16</td>
<td>32</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-bit integer multiply (umul24)</td>
<td>8</td>
<td>8</td>
<td>Multiple instructions</td>
<td>Multiple instructions</td>
<td>Multiple instructions</td>
<td>Multiple instructions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32-bit floating-point reciprocal, reciprocal square root, base-2 logarithm (log2f), base 2 exponential (exp2f), sine (sinf), cosine (cosf)</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>32</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type conversions from 8-bit and 16-bit integer to 32-bit types</td>
<td>8</td>
<td>8</td>
<td>16</td>
<td>16</td>
<td>128</td>
<td>128</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type conversions from and to 64-bit types</td>
<td>Multiple instructions</td>
<td>1</td>
<td>16(*)</td>
<td>4</td>
<td>8</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All other type conversions</td>
<td>8</td>
<td>8</td>
<td>16</td>
<td>16</td>
<td>32</td>
<td>32</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(*) Throughput is lower for GeForce GPUs.
Control Flow

Different execution paths inside the same warp are managed by the predication mechanism and lead to thread divergence.

Minimize/avoid the number of execution branches inside a threads warp;
make the compiler’s life easier by unrolling loops (hand-coded, pragma or option);
use signed counters for loops (relaxed semantic in respect to the unsigned int: it allows more aggressive loop optimizations);
Exploiting Multi-GPUs

CUDA >= 4.0 introduced the N-to-N bound feature:

1. Every **host** thread can be bound to any board
2. Every board can be bound to an arbitrary number of **host** threads
3. Multi-GPU can be exploited through your favourite multi-threading paradigm (OpenMP, pthreads, etc…)

```c
#pragma omp parallel
#pragma omp sections
{
#pragma omp section
{
  cutilSafeCall(cudaSetDevice(0));
  cudaMemcpy(device_data_1, host_data_1, size, cudaMemcpyHostToDevice);
  my_kernel<<< grid, block >>>(device_data_1);
  // ...
}
#pragma omp section
{
  cutilSafeCall(cudaSetDevice(1));
  cudaMemcpy(device_data_2, host_data_2, size, cudaMemcpyHostToDevice);
  my_kernel<<< grid, block >>>(device_data_2);
  // ...
}
}
```
Tools Overview
Development tools

- **Common**
  - Memory Checker
  - Built-in profiler
  - Visual Profiler

- **Linux**
  - CUDA GDB
  - Parallel Nsight for Eclipse

- **Windows**
  - Parallel Nsight for VisualStudio
Profiling tools: built-in

The CUDA runtime provides a useful profiling facility without the need of external tools.

```
export CUDA_PROFILE=1
export CUDA_PROFILE_CONFIG=$HOME/.config
```

// Contents of config

gld_coherent

gld_incoherent

gst_coherent

gst_incoherent

gld_incoherent: Number of non-coalesced global memory loads

gld_coherent: Number of coalesced global memory loads

gst_incoherent: Number of non-coalesced global memory stores

gst_coherent: Number of coalesced global memory stores

local_load: Number of local memory loads

local_store: Number of local memory stores

branch: Number of branch events taken by threads

divergent_branch: Number of divergent branches within a warp

instructions: instruction count

warp_serialize: Number of threads in a warp that serialize based on address conflicts to shared or constant memory

cta_launched: executed thread blocks

```text
method,gputime,cputime,occupancy,gld_incoherent,gld_coherent,gst_incoherent,gst_coherent
method=[  memcopy ] gputime=[ 438.432 ]
gld_incoherent=[ 0 ] gld_coherent=[ 1952 ] gst_incoherent=[ 62464 ] gst_coherent=[ 0 ]
method=[  memcopy ] gputime=[ 349.344 ]
```
Profiling: Visual Profiler

- Traces execution at host, driver and kernel levels (unified timeline)
- Supports automated analysis (hardware counters)
Well-known tool enhanced with CUDA extensions
Works well on single-gpu systems (OS graphics disabled)
Can be run under GDB-targeted tools and GUIs (multi-gpu systems)

(cuda-gdb) info cuda threads
BlockIdx ThreadIdx To BlockIdx ThreadIdx Count Virtual PC Filename Line
Kernel 0* (0,0,0) (0,0,0) (0,0,0) (255,0,0) 256 0x0000000000866400 bitreverse.cu 9
(cuda-gdb) thread
[Current thread is 1 (process 16738)]
(cuda-gdb) thread 1
[Switching to thread 1 (process 16738)]
#0 0x000019d5 in main () at bitreverse.cu:34
34 bitreverse<<<1, N, N*sizeof(int)>>>(d);
(cuda-gdb) backtrace
#0 0x000019d5 in main () at bitreverse.cu:34
(cuda-gdb) info cuda kernels
Kernel Dev Grid SMs Mask GridDim BlockDim Name Args
0 0 1 0x00000001 (1,1,1) (256,1,1) bitreverse data=0x110000
Debugging: CUDA-MEMCHECK

- It’s able to detect buffer overflows, misaligned global memory accesses and leaks
- Device-side allocations are supported
- Standalone or fully integrated in CUDA-GDB

```
$ cuda-memcheck --continue ./memcheck_demo
============== CUDA-MEMCHECK
Mallocing memory
Running unaligned_kernel
Ran unaligned_kernel: no error
Sync: no error
Running out_of_bounds_kernel
Ran out_of_bounds_kernel: no error
Sync: no error
============== Invalid __global__ write of size 4
============== at 0x00000038 in memcheck_demo.cu:5:unaligned_kernel
============== by thread (0,0,0) in block (0,0,0)
============== Address 0x200200001 is misaligned
==============
============== Invalid __global__ write of size 4
============== at 0x00000030 in memcheck_demo.cu:10:out_of_bounds_kernel
============== by thread (0,0,0) in block (0,0,0)
============== Address 0x87654320 is out of bounds
==============
============== ERROR SUMMARY: 2 errors
```
Parallel NSight

- Plug-in for major IDEs (Eclipse and VisualStudio)
- Aggregates all external functionalities:
  - Debugger (fully integrated)
  - Visual Profiler
  - Memory correctness checker
- As a plug-in, it extends all the convenience of IDEs to CUDA

On Windows systems:
- Now works on a single GPU
- Supports remote debugging and profiling
- Latest version (2.2) introduced live PTX assembly view, warp inspector and expression lamination
Parallel NSight
Parallel NSight