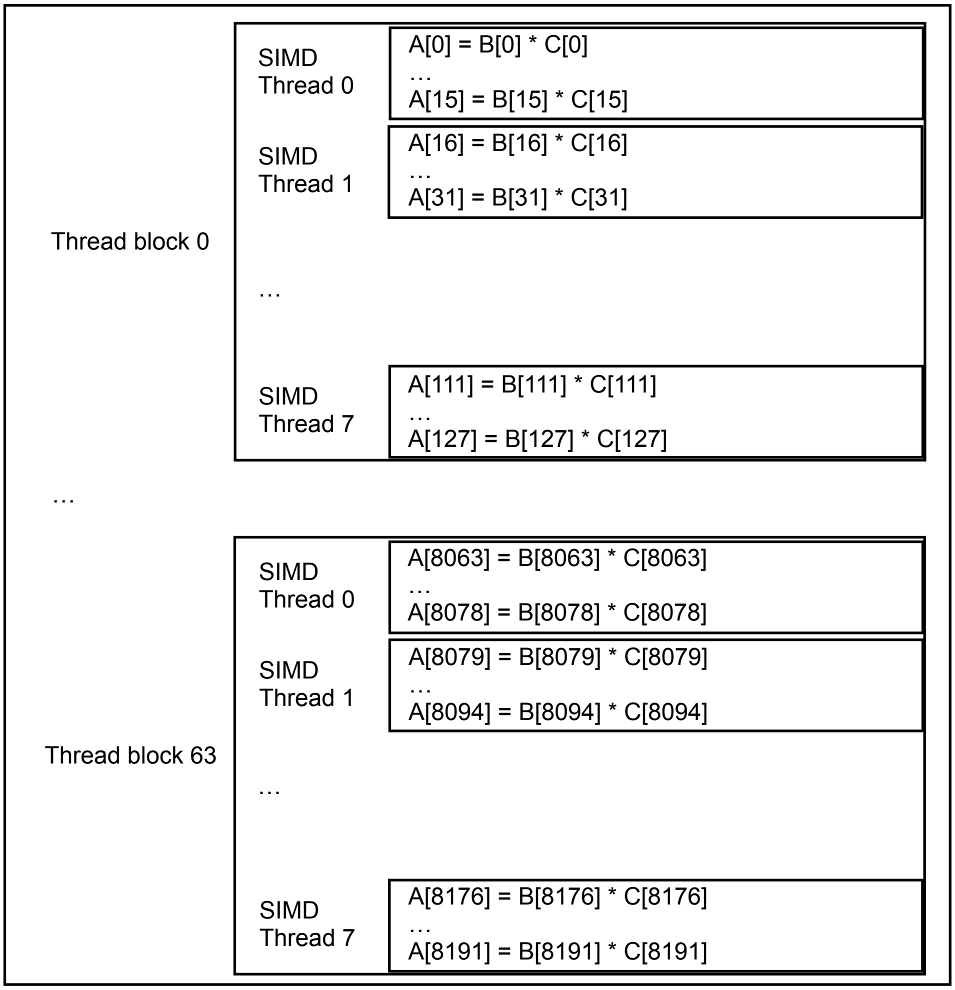
ECEn 528
Study Guide – GPUs

- Read Section 4.4 of H&P
  - Things to focus on
    - The terminology is really a mess! For the first four terms in Figure 4.12, look at both H&P and NVIDIA's terms. For the rest of them, focus on H&P's terminology.
    - Think about how individual elements in a single data-parallel loop get mapped to SIMD processors, threads of SIMD instructions, and SIMD instructions.

  - Clarifications
    - I'd like to strangle the guy at NVIDIA who thought it would be cute to name things "CUDA threads". No, I don't care if you've got funny support for branches making it look like different lanes are on different PC's, **they are not really threads**.
    - Similarly, calling individual collections of functional units (the equivalent of vector lanes) "thread processors" should earn this individual a mention in Dante's Inferno.

  - Answer the following questions:

1. Redraw Figure 4.13 for thread blocks with 8 SIMD threads and threads of SIMD instructions which calculate 16 elements per instruction.



Thread block 0

SIMD Thread 0
A[0] = B[0] * C[0]
…
A[15] = B[15] * C[15]

SIMD Thread 1
A[16] = B[16] * C[16]
…
A[31] = B[31] * C[31]

…

SIMD Thread 7
A[111] = B[111] * C[111]
…
A[127] = B[127] * C[127]

…

Thread block 63

SIMD Thread 0
A[8063] = B[8063] * C[8063]
…
A[8078] = B[8078] * C[8078]

SIMD Thread 1
A[8079] = B[8079] * C[8079]
…
A[8094] = B[8094] * C[8094]

…

SIMD Thread 7
A[8176] = B[8176] * C[8176]
…
A[8191] = B[8191] * C[8191]

2. Problem 4.16

1.5GHz * 16 SIMD * 16 FP = 384 GLOPS/s
This would require 384 GFLOPS/s * (12 bytes / FLOP) = 4.608 TB/s of bandwidth, which isn’t sustainable

3. Problem 4.13

a. 1.5 GHz x 0.8 x 0.85 x 0.70 x 10 SIMD x (32 results / 4 cycles) = 57.12 GFLOPS/s
b. 1) speedup of 2
   2) speedup of 1.5
   3) speedup of 0.95/0.85 = 1.12

4. Problem 4.11 (c)

```
unsigned int tid = threadIdx.x;
for (unsigned int s = blockDim.x/2; s > 0; s >>= 1) {
    if (tid < s) {
        parts[tid] += parts[tid+s];
    }
    __syncthreads();
}
```