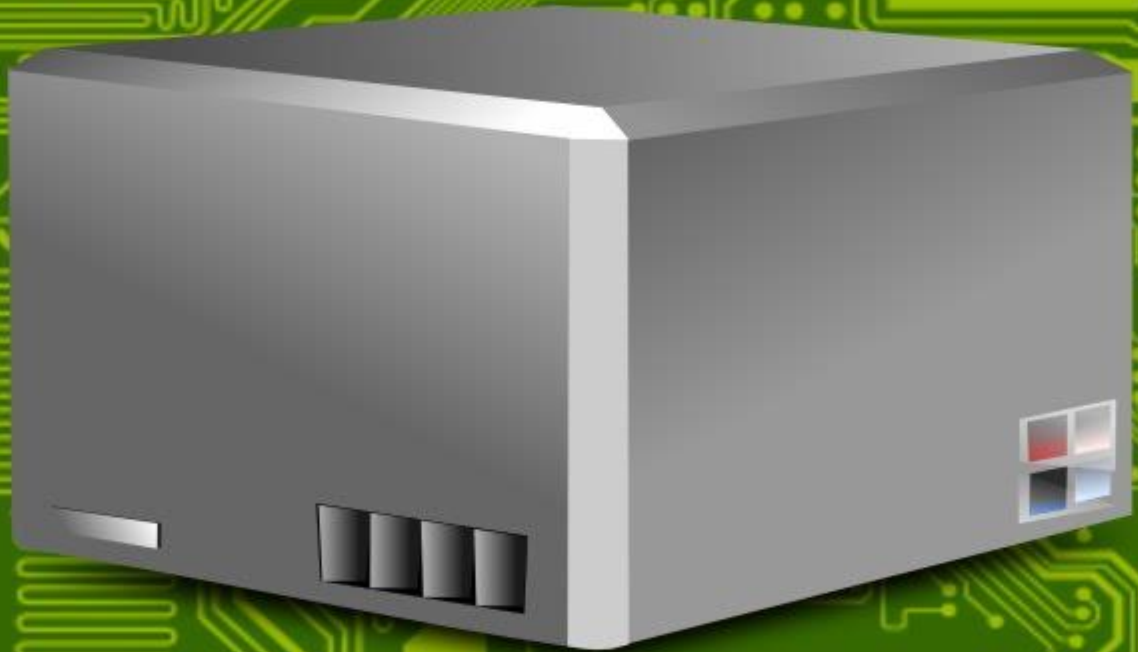
GPU Computing



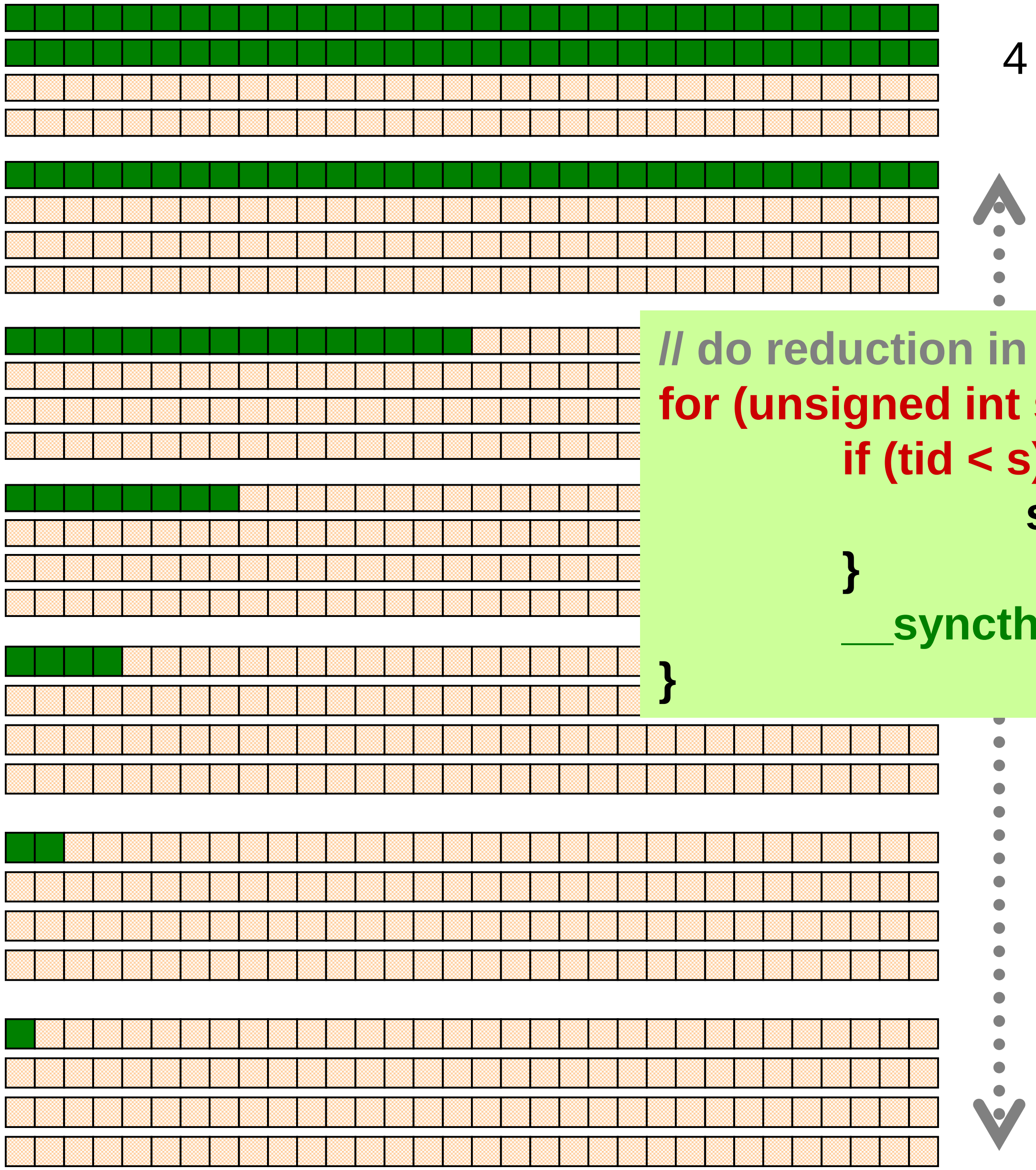
规约算法

Hui Liu
Email: hui.sc.liu@gmail.com

Reduction #4: Still way off

- Memory bandwidth is still underutilized
  - We know that reductions have low arithmetic density
- What is the potential bottleneck?
  - Loads, stores, or arithmetic for the core computation
  - Address arithmetic and loop overhead
  - Synchronization

Warp control flow



```
// do reduction in shared mem
for (unsigned int s = blockDim.x/2; s > 0; s /= 2) {
        if (tid < s) {
                sdata[tid] += sdata[tid + s];
        }
        __syncthreads();
}
```

Unrolling the last warp

- At every step the number of active threads halves
  - When s <=32 there is only one warp left
- Instructions are SIMD-synchronous within a warp
  - They all happen in lock step
  - No need to use __syncthreads()
  - We don’t need “if (tid < s)” since it does not save any work
    - All threads in a warp will “see” all instructions whether they execute them or not
- Unroll the last 6 iterations of the inner loop
  - s <= 32

Reduction #5: Unrolling the last 6 iterations

```
// do reduction in shared mem
for (unsigned int s = blockDim.x/2; s > 32; s /= 2) {

        if (tid < s) {
                sdata[tid] += sdata[tid + s];
        }
        __syncthreads();
}
```

```
if (tid <32)
{
        sdata[tid] += sdata[tid + 32];
        sdata[tid] += sdata[tid + 16];
        sdata[tid] += sdata[tid + 8];
        sdata[tid] += sdata[tid + 4];
        sdata[tid] += sdata[tid + 2];
        sdata[tid] += sdata[tid + 1];
}
```

- **This saves work in all warps not just the last one**
  - **Without unrolling all warps execute the for loop and if statement**

Performance for 4M element reduction

	Time (2^{22} ints)	Bandwidth	Step Speedup	Cumulative Speedup
Kernel 1: interleaved addressing with divergent branching	8.054 ms	2.083 GB/s		
Kernel 2: interleaved addressing non-divergent branching	3.456 ms	4.854 GB/s	2.33x	2.33x
Kernel 3: sequential addressing	1.722 ms	9.741 GB/s	2.01x	4.68x
Kernel 4: first step during global load	0.965 ms	17.377 GB/s	1.78x	8.34x
Kernel 5: Unroll last warp	0.536 ms	31.289 GB/s	1.8x	15.01x

THANK YOU