HIGH PERFORMANCE PARALLEL PROGRAMMING (CS61064)

Soumyajit Dey
CSE, IIT Kharagpur

Recap

- The Host-Kernel Model for CPU-GPU Systems
- The CUDA programming language
- Mapping multi-dimensional kernels to multi-dimensional data

Recap

- Querying device properties
- The concept of scheduling warps
- Performance bottlenecks
  - Branch Divergence
  - Global memory accesses

Parallel Patterns

- Matrix Multiplication (Gather Operation)
- Convolution (Stencil Operation)
- **Reduction**

Reduction

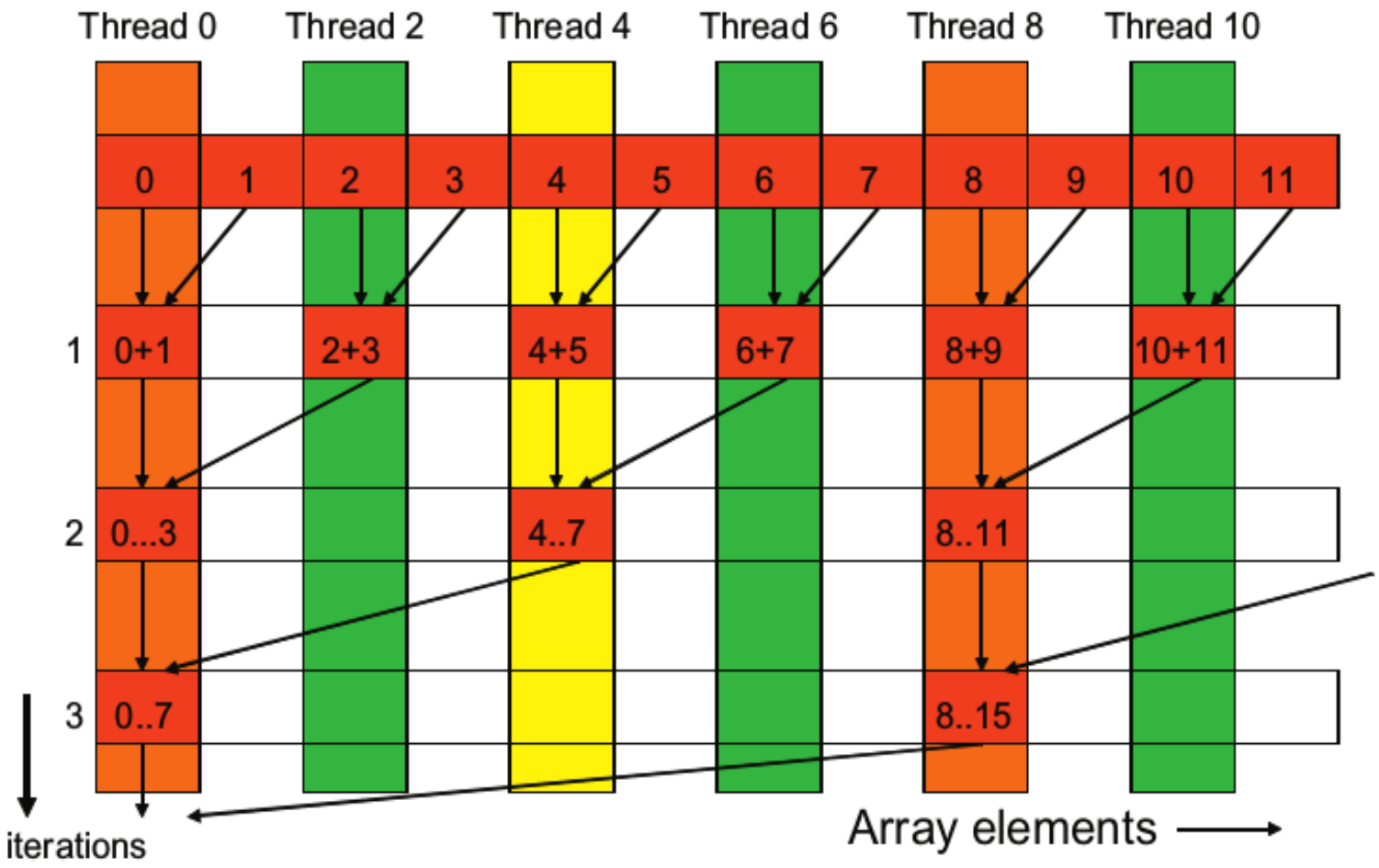

Figure: Reducing an array of 12 elements

Reduction Kernel

```
__global__
void reduce0(int *g_idata, int *g_odata){
extern __shared__ int sdata[];
unsigned int tid = threadIdx.x;
unsigned int i = blockIdx.x*blockDim.x +
    threadIdx.x;
sdata[tid] = g_idata[i];
__syncthreads();
for(unsigned int s=1;s<blockDim.x;s*= 2)
    {
     if (tid %(2*s)==0)
          sdata[tid]+=sdata[tid+s];
__syncthreads();
}
if (tid==0)
     g_odata[blockIdx.x] = sdata[0];
}
```

Multiple Kernel Invocations

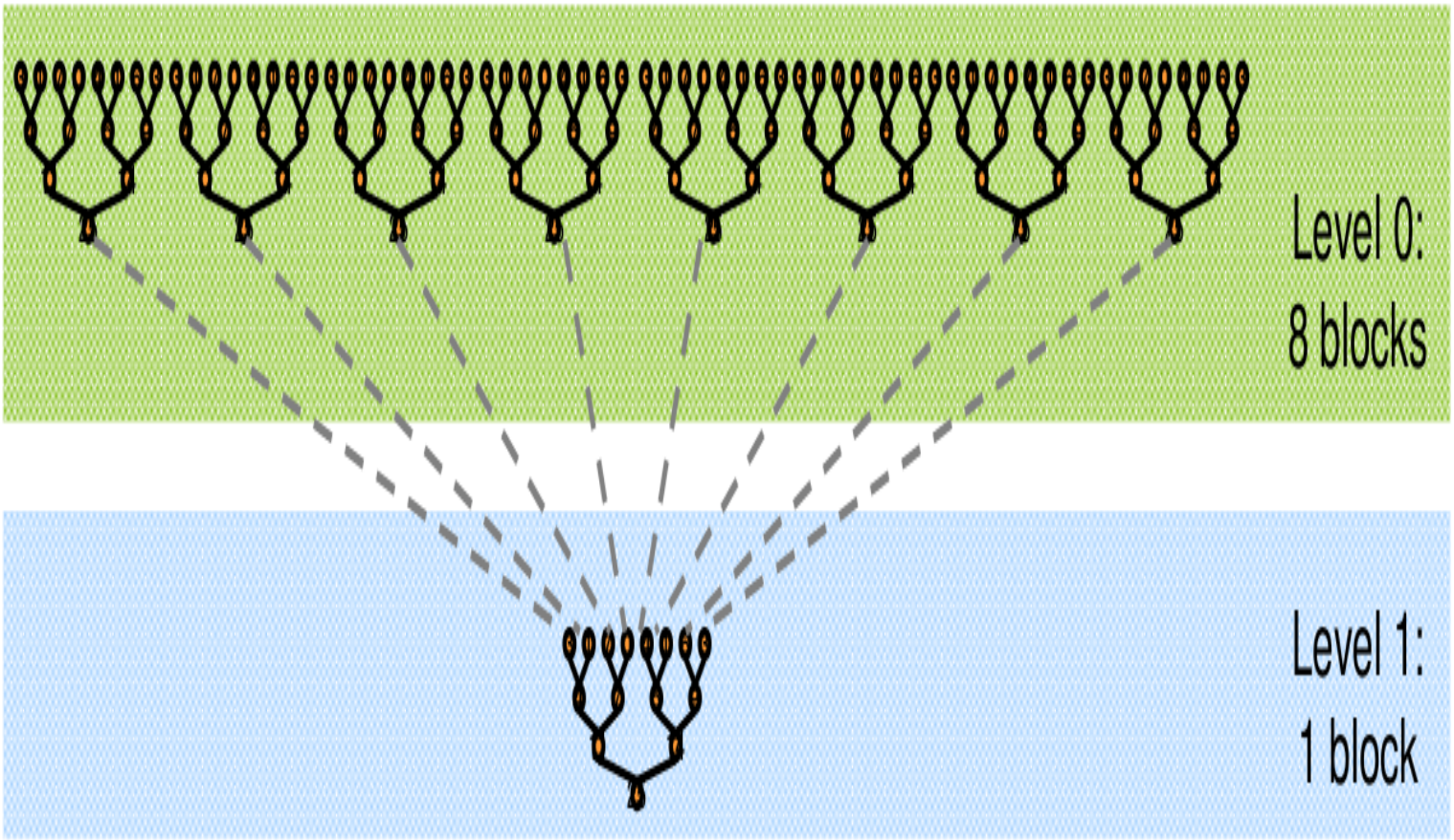

Figure: Host Side execution of kernels