Here's What You Would Change in your Host Program

```c
size_t cSize = numWorkGroups * sizeof(float);
size_t abSize = NUM_ELEMENTS * sizeof(float);
float * hC = new float [ numWorkGroups ];
float * hB = new float [ NUM_ELEMENTS ];
float * hA = new float [ NUM_ELEMENTS ];
```

The Arguments to the Kernel

```c
status = clSetKernelArg( kernel, 3, sizeof(cl_mem), &dC);
status = clSetKernelArg( kernel, 2, LOCAL_SIZE * sizeof(float), NULL);
status = clSetKernelArg( kernel, 1, sizeof(cl_mem), &dB);
status = clSetKernelArg( kernel, 0, sizeof(cl_mem), &dA);
```

Recall the OpenCL Memory Model

- **Global Memory**
- **Constant Memory**
- **Local Memory**
- **Private Memory**
- **Shared Memory**

«Previous Slide
Next Slide»
Reduction Takes Place Within a Single Work-Group

Each work-item is run by a single thread

A work-group consisting of numItems work-items can be reduced to a sum in \(\log_2(\text{numItems})\) steps. In this example, numItems=8.

The reduction begins with the individual products in prods[0] .. prods[7].

The final sum will end up in prods[0], which will then be copied into dC[wgNum].

Reduction Takes Place in a Single Work-Group

Each work-item is run by a single thread

Reduction Performance

Work-Group Size = 32

Array Size (MegaNumbers)

GigaNumbers Multiplied and Reduced Per Second)