```
__global__ void prescan(float *g_odata, float *g_idata, int n) {
    extern __shared__ float temp[];  // allocated on invocation
    int thid = threadIdx.x;
    int offset = 1;
    // load input into shared memory
    temp[2*thid] = g_idata[2*thid];
    temp[2*thid+1] = g_idata[2*thid+1];
    // build sum in place up the tree
    for (int d = n>>1; d > 0; d >>= 1) {
        __syncthreads();
        if (thid < d) {
            int ai = offset*(2*thid+1)-1;
            int bi = offset*(2*thid+2)-1;
            temp[bi] += temp[ai]; }
        offset *= 2; }
    // clear the last element
    if (thid == 0) { temp[n - 1] = 0; }
    // traverse down tree & build scan
    for (int d = 1; d < n; d *= 2) {
        offset >>= 1;
        __syncthreads();
        if (thid < d) {
            int ai = offset*(2*thid+1)-1;
            int bi = offset*(2*thid+2)-1;
            float t = temp[ai];
            temp[ai] = temp[bi];
            temp[bi] += t; } }
    __syncthreads();
    // write results to device memory
    g_odata[2*thid] = temp[2*thid];
    g_odata[2*thid+1] = temp[2*thid+1]; }
```

WAT

Source: Harris, Sengupta, and Owens in *GPU Gems 3*, Chapter 39