GPU Recap

```
__managed__ byte filter[NF] = {...};

int device; cudaGetDevice(&device);
cudaMemPrefetchAsync(filter, NF, device);
void *dimage = cudaMalloc(nbytes);
cudaMemcpyAsync(dimage, readImage(file), nbytes,
                          cudaMemcpyDefault, 0);

cudaEvent_t event1, event2;
cudaEventCreate(&event1); cudaEventCreate(&event2);
cudaEventRecord(event1, 0);

convolve<<<nbytes/512, 512, datasize, 0>>> (dimage);

// cudaDeviceSynchronize();
cudaEventRecord(event2, 0);
cudaEventSynchronize(event2);
cudaEventElapsedTime(&ms, event1, event2);

cudaMemcpy(himage, dimage, nbytes, cudaMemcpyDefault);
```

```
extern __shared__ float data[];

__global__ void convolve (float *image) {
    __shared__ byte sfilter[NF];
    if(threadIdx.x < F)
        sfilter[threadIdx.x] = filter[threadIdx.x];
    __syncthreads();
    size_t gtid = threadIdx.x +
                  blockIdx.x * blockDim.x;
    image[gtid] = filterit(..);
}
```

Subodh Kumar