

## Project 4. Due on 05/02/2015

### Implementing Matrix vector multiplication.

The purpose of this assignment is to implement float type matrix-vector multiplication for matrix and vector of arbitrary sizes. Use the 2D block grid and 2D thread block to implement the algorithm.

Optional: *Implement the multiplication algorithm using shared memory.*

Use matrix of size  $4096 \times 4096$ ,  $8192 \times 8192$  and  $16384 \times 16384$  to test the performance. Test your code by using  $8 \times 8$  and  $16 \times 16$  thread blocks, respectively.

To measure the performance of the GPU kernel execution, use the following code:

```
cudaEvent_t start, stop;
cudaEventCreate(&start);
cudaEventCreate(&stop);
cudaEventRecord(start, 0);

/// your kernel call here

cudaEventRecord(stop, 0);
cudaEventSynchronize(stop);

float elapsedTime;
cudaEventElapsedTime(&elapsedTime, start, stop);
printf("Time to generate: %f ms\n", elapsedTime);
```

### Hand-In.

1. Send the source code and report to me by email. Please use the email title: acms40212S16-Proj4-your-ND-ID.
2. Report needs to contain performance measure and a description of your algorithm using the pseudo code language.