Project 3, due on 05/01.

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. *For graduate students, implement the multiplication algorithm using shared memory.*

Use matrix of size 4096×4096, 8192×8192 and 16384×16384 to test the performance. Test your code by using 8×8 and 16×16 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 elapseTime;
cudaEventElapseTime(&elapseTime, start, stop);
printf("Time to generate: %f ms\n", elapseTime);
```

Hand-In.

1. The hardcopy of your source code (Also send the source code to me by email. Please use the email title: Project 4: your name).

2. A report which contains performance measure and a description of your algorithm using the pseudo code language.