**Assignment4 Report**

**CSE436, Summer 2016**

**Kazumi Malhan**

**07/01/2016**

1. Backgrounds and Motivation

In recent years, the graphics processing unit (GPU) is getting more attention for parallel programming, as it has high throughput and optimized for parallel execution with its large amount of cores. However, GPU programming involves new overhead such as data movement between CPU and GPU memory. In assignment 3, matrix multiplication function is implemented in five different types: serial, OpenMP, CUDA with global memory, CUDA with shared memory, and CUDA library. The performance of these functions and time consumption distribution are analyzed to understand the effectiveness and cost of GPU programming.

2. Function Implementation Description

3. Performance Report

4. Conclusion

References