cs378: Concurrency Honors: GPU K-Means Lab 4 Writeup Template

You
Department of Computer Science
UT Austin

August 23, 2018

1 Graph

Using the random-n2048-d16-c16.txt, random-n16384-d24-c16.txt, and random-n65536-d32-c16 sample inputs, create graphs of speedup over your sequential CPU kmeans finding 16 centroids (see lab for other parameters). Include your best performing multi-threaded CPU k-means, basic CUDA, CUDA with shared memory, and any additional extra credit optimizations you might have explored (successful or not). Be sure to label the graph clearly. In this sample writeup, all graphs are combined into one grouped bar graph: it is fine to do it this way, or to provide separate graphs.

Questions to answer:

- Report the GPU hardware details, CPU hardware details, and OS version on the machine where you did your measurements.
- Which of your implementations is fastest? Does it match your expectations of which should be fastest? Estimate the best-case performance speedup your CUDA implementations should have based on the number of threads in your program and the number of processing contexts actually supported by your hardware. How far of that prediction is your best-case performance?
- Which of the parallel implementations is slowest, and does it match your expectations? Why or why not?
- What fraction of the end-to-end runtime in your CUDA versions is spent in data transfer?
- What implementation approach did you take in Step 4? Did it improve performance or scalability? How much? Why/why not?
- How much time did you spend on the lab?

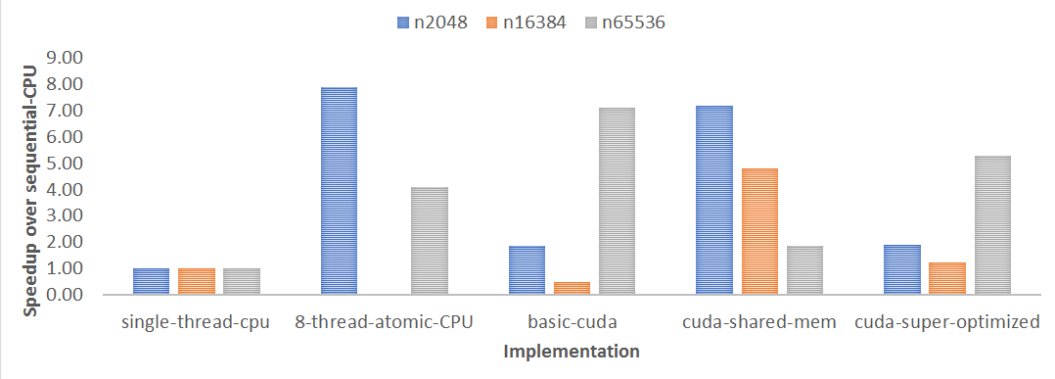

Figure 1: Speedup over sequential-CPU for different GPU and CPU implementations of K-Means. *NOTE: the data in this graph are random. If your data are different that's GOOD.*