CS 4230/6230

Programming Assignment 3

Due 11:59pm, Monday, 11/25/2024

For this assignment, various CUDA versions are to be implemented for transposed matrix-multiplication, whose sequential C code is shown below. For all the versions, use a 1D 256 x 1 thread-block. To ease the implementation effort, the codes only need to pass the correctness test for a fixed problem size of 1024. Since CHPC GPUs are under a heavy load, use CADE Lab for this assignment.

Transposed-Transposed Matrix-Matrix Multiplication ($C=A^TB^T$):

```
for (i=0;i<1024;i++)
 for (k=0;k<1024;k++)
  for (j=0;j<1024;j++)
// C[i][j] += A[k][i]*B[j][k];
   C[i*1024+j] += A[k*1024+i]*B[j*1024+k];
```

1. (20 points) Replace "FIXMEs" in **mmtt.cu** to create a version that passes the correctness test. It should use 256 x 1 thread blocks and a distinct thread to compute each output element of C and achieve coalesced access of A and B from global memory. Expected performance on CADE: around 180 GFLOPs.
2. (10 points) Reverse the mapping from the thread-grid to output data space from the above. What performance differences are observed?
3. (10 points) Starting with the version of mmtt.cu (from part 1) that achieves the expected performance, implement a version that performs 4-way unrolling along ***k*** in **mmtt_k4.cu**. Is any performance improvement achieved over mmtt?
4. (15 points) Implement a version that performs 4-way unrolling along ***j*** in **mmtt_j4.cu**. Is performance improvement achieved over mmtt?
5. (15 points) Implement a version that performs 4-way unrolling along ***i*** in **mmtt_i4.cu**. Is performance improvement achieved over mmtt?
6. (15 points) Implement a version that performs 4-way unrolling along ***i*** and ***j*** in **mmtt_i4j4.cu**. Is performance improvement achieved over previous versions?
7. (15 points) Implement a version that uses shared memory to buffer elements of A and B in **mmtt_sm.cu**. No loop unrolling is required. Is any performance improvement achieved over mmtt?

Upload all your CUDA files and report.pdf file on Gradescope. The report.pdf file should summarize performance trends with the various optimized versions and include traces or screenshots of execution.