CUDA Parallel Floyd Warshall Project

ofo5108 – Ömer Faruk Özdemir

Algorithm

Can be found in the project description.

Blocking Factor

32, I chose a blocking factor for 2 reasons.

1. To have 32B aligned access to the global memory we need to select block size as multiple of 32.
2. The best speedup experimentally.

Optimizations

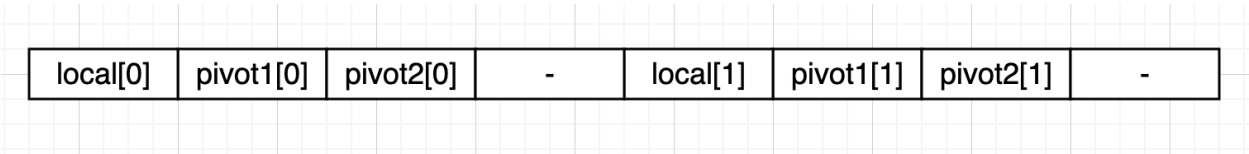
1. Global memory accesses are aligned to 32B to make it coalescing.
2. Data is represented as 1B(uint8_t) to minimize memory usage.
3. In each warp, every thread writes into a different bank.
4. Only the upper triangle of the matrix is calculated, to decrease memory access and usage. Accesses into lower triangle are mapped into upper triangle.

Unhelpful optimizations

1. Doing write operations with the ordering of 32 thread chunks did not change the performance. I thought it would decrease bank conflicts with thread synchronization, but it seems it does not help.

Data Layout

The data is padded with global memory with infinities, if the given data is not multiple of the blocking factor.
CPU and GPU global memory is contiguous memory region: `uint8_t *gdata`
Shared memory layout is below,
Each cell is a byte.



Since we want to minimize bank conflicts, each local memory is put into different words.
It is safe to put pivots in the same banks with locals, because they are read only.
Reads and writes are separated with a syncthreads call, thus there is no read-write bank conflict, hence reads are all broadcasts to threads.

Experiment and Analysis

PSU CSE machines

gcc (GCC) 4.8.5 20150623 (Red Hat 4.8.5-44)

nvcc: NVIDIA (R) Cuda compiler driver
Copyright (c) 2005-2023 NVIDIA Corporation
Built on Mon_Apr__3_17:16:06_PDT_2023
Cuda compilation tools, release 12.1, V12.1.105
Build cuda_12.1.r12.1/compiler.32688072_0

Correctness

diff on ./dataset/50-1225:

diff on ./dataset/100-4000:

diff on ./dataset/500-40000:

diff on ./dataset/1000-400000:

diff on ./dataset/2000-1200000:

Performance

Dataset: 2000
Sequential: real 0m5.5410s
CUDA 32 Blocking Factor: real 0m1.656s
Speedup: 3.34x