**Code Performance:**

**First version:**

**Compiled as nvcc first.cu**

Number of threads: 1024 (32x32)

Number of blocks: 4096 (64x64)

Time to calculate results on CPU: 68763.898438 ms -- result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on CPU\_Opt (single thread): 32394.253906 ms -- result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on CPU\_Opt (32 threads): 32398.128906 ms -- result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on GPU: 52.637344 ms

Effective performance: 303.967 GFlop

Effective bandwith: 0.956 GB

result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on GPU (optimised): 77.231682 ms

Effective performance: 207.169 GFlop

Effective bandwith: 0.607 GB

result[10] = 24576.000000 -- result[1000] = 2052096.000000

**Compiled as nvcc first.cu -O3**

Number of threads: 1024 (32x32)

Number of blocks: 4096 (64x64)

Time to calculate results on CPU: 30975.724609 ms -- result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on CPU\_Opt (single thread): 11007.387695 ms -- result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on CPU\_Opt (32 threads): 11011.590820 ms -- result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on GPU: 62.923073 ms

Effective performance: 254.279 GFlop

Effective bandwith: 0.800 GB

result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on GPU (optimised): 100.004478 ms

Effective performance: 159.993 GFlop

Effective bandwith: 0.469 GB

result[10] = 24576.000000 -- result[1000] = 2052096.000000

**Compiled as nvcc first.cu -Xcompiler -fopenmp -Xcompiler -O3 -O3**

Number of threads: 1024 (32x32)

Number of blocks: 4096 (64x64)

Time to calculate results on CPU: 40542.781250 ms -- result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on CPU\_Opt (single thread): 10864.010742 ms -- result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on CPU\_Opt (32 threads): 1527.979248 ms -- result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on GPU: 51.901951 ms

Effective performance: 308.274 GFlop

Effective bandwith: 0.970 GB

result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on GPU (optimised): 89.232964 ms

Effective performance: 179.306 GFlop

Effective bandwith: 0.525 GB

result[10] = 24576.000000 -- result[1000] = 2052096.000000

**Compiled as nvcc first.cu -Xcompiler -fopenmp -O3 -Xcompiler -O3**

Number of threads: 1024 (32x32)

Number of blocks: 4096 (64x64)

Time to calculate results on CPU: 27698.773438 ms -- result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on CPU\_Opt (single thread): 10822.304688 ms -- result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on CPU\_Opt (32 threads): 1536.970825 ms -- result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on GPU: 62.511936 ms

Effective performance: 255.951 GFlop

Effective bandwith: 0.805 GB

result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on GPU (optimised): 87.386238 ms

Effective performance: 183.095 GFlop

Effective bandwith: 0.536 GB

result[10] = 24576.000000 -- result[1000] = 2052096.000000

**Second version:**

**Compiled as nvcc second.cu -Xcompiler -fopenmp -O3 -Xcompiler -O3**

Number of threads: 1024 (32x32)

Number of blocks: 4096 (64x64)

Time to calculate results on CPU: 43967.253906 ms -- result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on CPU\_Opt (single thread): 10846.297852 ms -- result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on CPU\_Opt (32 threads): 1504.871948 ms -- result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on GPU: 46.512001 ms

Effective performance: 343.997 GFlop

Effective bandwith: 1.082 GB

result[10] = 24576.000000 -- result[1000] = 2052096.000000

Time to calculate results on GPU (optimised): 19.650175 ms

Effective performance: 814.242 GFlop

Effective bandwith: 2.385 GB

result[10] = 24576.000000 -- result[1000] = 2052096.000000

From the results it can be seen that both versions of code are calculating matrix multiplication correctly.

Comparison on CPU:

SERIAL: The first version of the code is 1.58 times slower than the second version on CPU.

Burda problem row major access b’ye erişirken k width kadar atlama yapıyor ve cache miss artıyor. Çözüm b’nin transpose’unu almak.

W11 Lec 2 35. Dk

Memory bank confict

0 0

1 0

2 0

Diye gidiyor

Iki yerde var birincisi s\_a[threadIdx.x][threadIdx.y]

ikincisi s\_b[threadIdx.x][threadIdx.y]

Aynı warp içinde aynı warp‘dan k farklı kelime okuyorsan buna k-way bank conflict diyoruz. Burada 32-way bank conflict var.

s\_a[i][threadIdx.y]’da sorun yok çünkü bu broadcasting aynı değere erişiyor.

s\_b[threadIx.x][i] sıkıntılı çünkü aynı memory bank farklı kelimeler.

Bu ikincisi üstekilerden daha sıkıntılı çünkü thread başı 32 kere oluyor.

İkinci versiyonda iki memory bank conflict de çözülüyor.