Second Assignment - Report

Christian Dalvit
Student-ID: 249988
Trento, Italy
christian.dalvit@studenti.unitn.it

I. Introduction

The objective of this homework is to implement an algorithm for transposing a non-symmetric matrix and to measure and analyze various metrics of the algorithm. This report provides a description of the problem setting, algorithms, and experimental results of the implementation. The code used for this homework is made available through a public Github repository. Details on how to run the code and reproduce the results can be found in the `README.md` file of the Github repository.

II. Problem Description

For a given matrix $A \in \mathbb{R}^{n \times m}$, the transpose of the matrix $A^T \in \mathbb{R}^{m \times n}$ is defined as

$$A^T_{ij} = A_{ji}$$

In this homework, matrices have dimensions of 2^N for $N \in \mathbb{N}$, so only square matrices are considered. As a result, the implemented algorithms don't need to accommodate changes in the output matrix's shape. For the purpose of this homework, row-major memory layout of the matrix is assumed. While implementing an algorithm that computes the transpose of a matrix is straightforward, coming up with an efficient implementation is quite tricky. In general, leveraging spatial and temporal locality can improve efficiency. Because each element of the matrix is accessed only once, temporal locality cannot be exploited for computing the transposed matrix [2], so spatial locality becomes the only source for improvement. The issue with leveraging spatial locality in matrix transposition is that data is accessed along rows but written along columns, potentially leading to poor cache performance. Algorithms that respect spatial locality in their memory access pattern can benefit from quicker access to cached data. In the following, the different algorithms implemented during this homework are described and their memory access pattern is discussed.

A. CPU Algorithms

In this homework, three different algorithms for in-place matrix transposition are implemented. Pseudocode for the algorithms can be found in the appendix. The first implementation, as shown in Algorithm 1, can be directly inferred from the mathematical definition. Transposition is performed by iterating over all entries above the matrix diagonal and swapping them with the corresponding entries below the diagonal. The first implementation will serve as the baseline implementation to measure performance improvements of other implementations. The main issue with the naive implementation's memory access pattern is the disjointed access from `mat[j*size+i]`, potentially causing poor cache performance, especially with larger matrices where `mat[j*size+i]` might not be present in cache and requires loading from memory. The second

Algorithm 1: Naive implementation

```
input: Size $n$, Matrix $M$
for $i$=0 ... $n$-1 do
    for $j$=$i$+1 ... $n$-1 do
        swap($M_{i,j}$, $M_{j,i}$);
    end
end
```

implementation (Algorithm 2) tries to improve performance by prefetching the memory addresses needed in the next iteration. This should reduce cache-miss latency by moving data into the cache before it is accessed [7]. The built-in function `__builtin_prefetch` can be used to perform prefetching. `__builtin_prefetch` takes as arguments the address to be prefetched and two optional arguments *rw* and *locality*. Setting *rw* to 1 means preparing the prefetch for write access and setting *locality* to 1 means that the prefetched data has low temporal locality [7]. The second implementation is referred to as "prefetch"-implementation. The third algorithm (Algorithm

Algorithm 2: Prefetch implementation

```
input: Size $n$, Matrix $M$
for $i$=0 ... $n$-1 do
    for $j$=$i$+1 ... $n$-1 do
        swap($M_{i,j}$, $M_{j,i}$);
        builtin_prefetch($M_{i+1,j}$, 1, 1);
        builtin_prefetch($M_{j,i+1}$, 0, 1);
    end
end
```

3) implements a recursive pattern for matrix transposition. It uses the fact that

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^T = \begin{pmatrix} A^T & C^T \\ B^T & D^T \end{pmatrix}$$

Where A, B, C and D are submatrices. Note that the submatrices B and C get swapped. The idea behind this algorithm

is that it splits the matrix into four sub-matrices until the submatrices fit into the cache. Then the submatrices get transposed and the quadrants get swapped. Algorithm 3 depicts pseudocode for this algorithm. For a matrix size of 128 or smaller the algorithm performs a normal transpose operation (i.e. the `swap_small_matrix`-function call in Algorithm 3). Otherwise, the matrix is split into four submatrices and the function is called recursively (i.e. the **`else`**-block in Algorithm 3). After the transposition of the submatrices, the upper-right and bottom-left quadrants need to be swapped (i.e. the `swap_quadrants`-function call in Algorithm 3).
This algorithm exploits spatial locality as it divides the matrix into sub-matrices that can fit into the cache. The third algorithm also has a reduced I/O complexity of $\mathcal{O}(\frac{N^2}{B})$ [6], where N is the size of the matrix and B the size of the blocks (i.e. the size of matrices where standard transposition is performed). It then transposes each submatrix, which can be performed more efficiently as the whole submatrix is present in cache. This algorithm also works quite well for large matrices, because they are always reduced to submatrices of sizes that fit into cache. The threshold of 128 for performing standard matrix transposition, was selected because the MacBook Air has an L1 data cache of 65KB (as described in Section III). Since each matrix element is an integer having a size of 4 bytes, a submatrix of dimension 128 needs $128 * 128 * 4 = 65536$ bytes. Therefore one submatrix fits in the L1 data cache, which should result in better performance. During the homework, the same algorithm was tested for threshold values 32, 64, 128, 256 and 512. The results indicated that the algorithm performed best with a threshold of 128 on the tested architectures.

Algorithm 3: Recursive implementation

```
input: Size $n$, Row offset $x$, Column offset $y$, Matrix $M$
if $n \leq 128$ then
    swap_small_matrix($n$, $M$, $x$, $y$);
else
    $m \leftarrow \frac{n}{2}$;
    Recursive call of this algorithm;
    transpose_block($m$, $M$, $x$, $y$);
    transpose_block($m$, $M$, $x$, $y + m$);
    transpose_block($m$, $M$, $x + m$, $y$);
    transpose_block($m$, $M$, $x + m$, $y + m$);
    Swap upper-right and bottom-left quadrants;
    swap_quadrants($m$, $M$, $x$, $y$);
end
```

B. GPU Algorithms

Graphics Processing Units (GPUs) can be used to execute parts of algorithms in parallel and hence significantly improve performance. Algorithms must be adapted for GPU execution, otherwise implementations of algorithms cannot benefit from hardware acceleration. In this homework, I present two CUDA kernels for in-place matrix transposition.

The first kernel is depicted in Algorithm 4. The kernel is a straightforward parallelization of the naive CPU algorithm shown in Algorithm 1. Instead of performing the swap of matrix elements in nested loops, each thread swaps one matrix element. Because each swap can be performed independently, no synchronization between threads in needed. Algorithm 4 first computes the coordinates of the thread in the grid, i.e., the indices that a thread should swap. Note that the implementation uses two-dimensional block and grid sizes, as they can naturally map to the two-dimensional structure of matrices. The conditions $x \leq n$ and $y \leq n$ ensure that no thread swaps indices that are outside the matrix. The condition $x \leq y$ ensures that no duplicate swaps occur, which would result in the original matrix instead of the transposed matrix. If all conditions are satisfied, a standard in-place transposition is performed by the kernel function. Block and grid dimensions are computed according to the matrix size to ensure that all matrix elements get swapped, i.e., enough threads are launched. The drawback of the Algorithm 4 is that

Algorithm 4: Naive CUDA kernel

```
input: Size $n$, Matrix $M$
$x \leftarrow blockIdx.x \cdot blockDim.x + threadIdx.x$;
$y \leftarrow blockIdx.y \cdot blockDim.y + threadIdx.y$;
if $x \leq n$ and $y \leq n$ and $x \leq y$ then
    $tmp \leftarrow M_{y,x}$;
    $M_{y,x} \leftarrow M_{x,y}$;
    $M_{x,y} \leftarrow tmp$;
end
```

every memory access in the kernel function is to the global GPU memory. An optimized kernel function could use the shared memory of the GPU, since it has higher bandwidth and lower latency [5, See Section *9.2.3. Shared Memory*]. These properties can be leveraged to improve the performance of the algorithm. This motivates the second CUDA kernel (depicted in Algorithm 5) implemented during this homework. In the second algorithm, each block transposes a tile of the matrix. The kernel function uses a two-dimensional array in the GPUs shared memory. In the first loop, data from the matrix tile is loaded from the global memory into the shared memory. After synchronization, the second loop writes the transposed data back from shared memory to global memory. Synchronizing all threads of the block is needed because the threads inside a block cannot operate independently. Calling `__syncthreads()` ensures that the whole array in shared memory has been populated with the corresponding matrix element. Only after synchronization, the threads can continue to write the data to global memory. Otherwise, threads might write data into global memory that has not been loaded yet.

Note that the size of the shared memory tile is `TILE_DIM` $\times$ `TILE_DIM+1`. This size is selected in order to avoid memory bank conflicts. Shared memory bank conflicts occur

Algorithm 5: Tiled CUDA kernel

```
input: Size $n$, Matrix $M$
*Shared memory tile of size*
  $TILE\_DIM \times TILE\_DIM + 1$;
*__shared__* $T$;
$x_t \leftarrow threadIdx.x$;
$y_t \leftarrow threadIdx.y$;
$x \leftarrow blockIdx.x \cdot blockDim.x + x_t$;
$y \leftarrow blockIdx.y \cdot blockDim.y + y_t$;
for $i = 0;\ i \leq TILE\_DIM;\ i \leftarrow i + blockDim.y$ do
    if $x \leq n$ and $y + i \leq n$ then
        $T_{y_t+i,x_t} \leftarrow M_{y+i,x}$;
    end
end
*Synchronize threads of the block*;
syncthreads();
for $i = 0;\ i \leq TILE\_DIM;\ i \leftarrow i + blockDim.y$ do
    if $x \leq n$ and $y + i \leq n$ then
        $M_{y+i,x} \leftarrow T_{x_t,y_t+i}$;
    end
end
```

if two threads want to access the same shared memory bank at the same time. This results in serialized access, where the requests are processed one after another instead of in parallel, reducing performance. By offsetting each row with padding, the memory accesses are spread across different banks, reducing the chances of bank conflicts during shared memory access. This should result in better performance.

III. CPU Experiments

A. Setup

The three algorithms were implemented in C. Each algorithm was compiled with optimization levels `-O0`, `-O1`, `-O2`, `-O3`. Each resulting binary was evaluated 50 times for matrix sizes between 2^8 and 2^{14}. Cache data was collected using Valgrind [3]. Each binary was benchmarked with the Cachegrind tool for matrix size 2^{14}. The experiments were conducted on two different architectures

- MacBook Air (2020) having a M1 chip with 3,2 GHz and 8 cores, 8GB of RAM, 131KB L1 instruction cache, 65KB L1 data cache, 4.2 MB L2 cache and a cacheline size of 128 byte.
- iMac (2011) having a Intel Core i5 with 2.5GHz and 4 cores, 8GB of RAM, 64KB L1 cache, 1MB L2 cache, 6MB L3 cache and a cacheline size of 64 byte.

Unfortunately Valgrind is not officially supported for ARM-based Apple computers [4] and Open-Source projects working on compatibility for M1 processors are still in the experimental phase [1]. Therefore it is not possible to provide cache performance data for the MacBook Air experiments.

B. Results

The execution times of the different algorithms are compared in Figure 1. The third algorithm (`oblivious128` in Figure 1) achieved the highest speedup with optimization flags. It showed a speedup of 7.81 on the MacBook Air and 5.7 on the iMac when comparing non-optimized code to code with `-O3` enabled for matrix size 2^{14}. On the MacBook Air, the third algorithm gained an additional speedup of 1.46 for matrix size 2^{14} using `-O2` and `-O3` optimizations compared to `-O1` optimization. The second algorithm (`naive_prefetch` in Figure 1) performs slightly better for large matrices with some optimization enabled, resulting in a speedup of 2.11 on the MacBook Air and 1.73 on the iMac for matrix size 2^{14}. Enabling optimization flags can be disadvantageous for execution time performance (see `naive` and `naive_prefetch` implementations for matrix size 2^{10}).

The effective bandwidth of the different algorithms is compared in Figure 2. It's evident across all graphs that the effective bandwidth decreases as the matrix size increases. This trend arises because the matrix size grows exponentially with a base of 2, while the execution time grows exponentially with a base of roughly 10, as illustrated in Figure 1. Consequently, the execution time increases more rapidly than the matrix size, leading to a decline in effective bandwidth.

Cache data, as depicted in Figure I in the appendix, reveals distinctive behaviors among implementations. Specifically, the `naive` and `naive_prefetch` implementations show similar cache patterns, while the `oblivious128` implementation stands out with different cache behavior. Comparing cache data across implementations, it's evident that the `oblivious128` implementation requires approximately twice as many instructions, data reads, and data writes as the `naive` implementation. The variance could be a result of the extra instructions required for transposing submatrices and copying quadrants in the `oblivious128` algorithm, leading to a greater overall instruction count. Another notable finding is that the last-level data cache read misses are significantly lower for the `oblivious128` implementation. This may be due to the improved fit of submatrices utilized in the `oblivious128` algorithm within the cache.

TABLE I
Summarized cache data for different CPU implementations with `-O3` optimization

Name	Ir	Dr	DLmr	Dw	D1mw	DLmw
`naive`	20.6	5.1	0.14	2.4	0.016	0.016
`oblivious128`	41.1	10.1	0.041	4.7	0.033	0.033

[a]Values are given in billions

IV. GPU Experiments

A. Setup

For benchmarking the GPU implementations, the following setup was used. Both GPU algorithms were compiled with `nvcc` of the CUDA 12.5 module provided on the Marzola cluster of the University of Trento. To measure the speedup

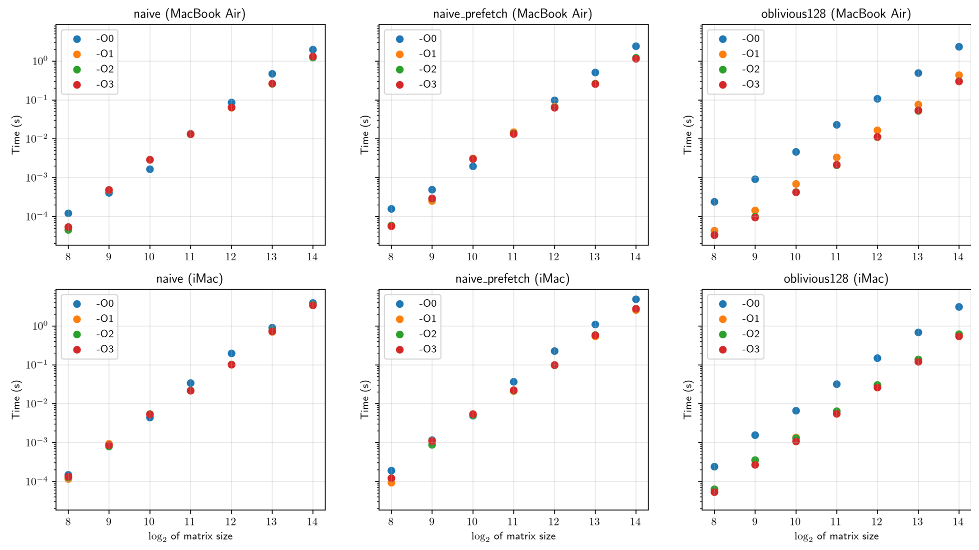

Fig. 1. Comparison of the mean execution time of various algorithms for matrix sizes ranging from 2^8 to 2^{14} for the MacBook Air architecture (upper row) and iMac architecture (bottom row). The y-axis denotes the execution time in seconds and is depicted on a logarithmic scale.

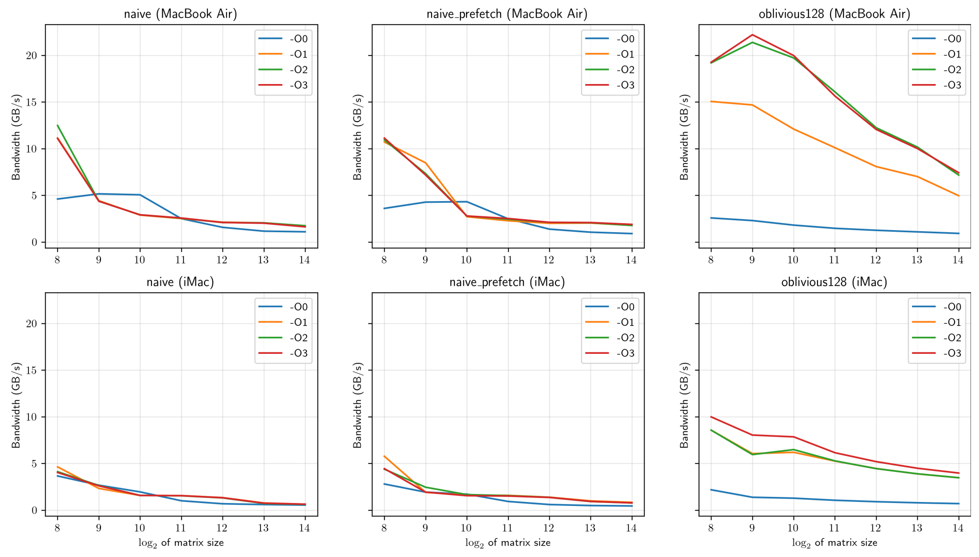

Fig. 2. Comparison of the mean effective bandwidth of different algorithms for matrix sizes between 2^8 and 2^{14} for the MacBook Air architecture (upper row) and iMac architecture (bottom row).

compared to the CPU implementations, also Algorithm 1 and Algorithm 3 were compiled using `-O3` optimizations. Each resulting binary was evaluated 50 times for matrix sizes between 2^8 and 2^{14}. CUDA algorithms were also benchmarked with block sizes of 4, 8, 16 and 32. The grid size was computed using the matrix and block size. Time was only measured for the kernel execution, so the time for moving data from and to the GPU device is not contained in the measurements. The experiments were only conducted on the Marzola cluster of the University of Trento, since there was no other possible access to GPUs during the homework. On the cluster an Nvidia A30 GPU, with a theoretical peak bandwidth of 933GB/s was used to benchmark the implementations.

B. *Results*

The results obtained during the experiments are summarized in Table II and depicted in Figure 3 and Figure 4. Note that the label `gpu_tiled_16` stands for the tiled GPU algorithm (Algorithm 5) with a block size of 16. One can make multiple observations from the data. The most obvious observation is that GPU algorithms perform significantly better than the CPU algorithms. In every scenario, the GPU algorithms have a higher effective bandwidth, compared to the CPU algorithms. In the best case, the effective bandwidth of the `gpu_tiled_16` algorithm is 94 times higher compared to the CPU algorithm. The second observation is that while the effective bandwidth decreases as matrix size increases for CPU algorithms, the effective bandwidth for GPU algorithms

TABLE II

Comparison of the mean effective bandwidth for different algorithms by matrix size

Name	2^8	2^9	2^{10}	2^{11}	2^{12}	2^{13}	2^{14}
`cpu_oblivious128`	15.77	14.36	11.99	9.05	6.98	6.18	5.49
`gpu_naive_4`	17.33	24.17	49.58	66.96	75.05	77.02	76.25
`gpu_naive_16`	23.16	31.60	98.01	225.78	209.64	228.44	307.66
`gpu_naive_32`	19.27	27.57	72.19	120.98	137.47	148.41	142.33
`gpu_tiled_4`	17.69	24.60	52.08	71.55	76.06	76.67	77.14
`gpu_tiled_16`	23.74	39.72	128.99	311.84	437.74	500.53	517.83
`gpu_tiled_32`	22.65	35.09	115.23	309.69	403.84	451.99	471.01

[a]Values are given in GB/s

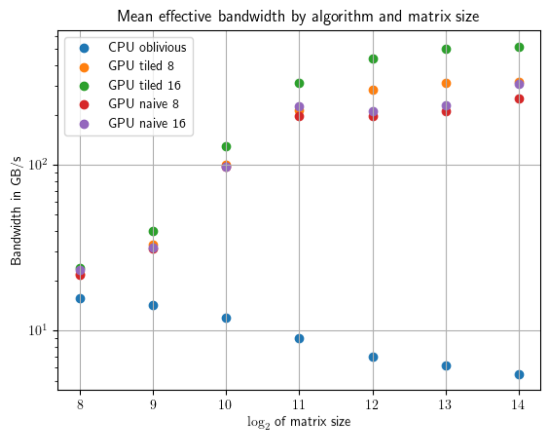

Fig. 3. Comparison of mean effective bandwidth for different algorithms and matrix sizes

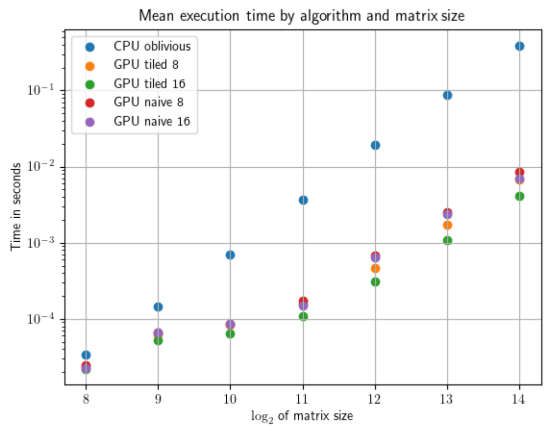

Fig. 4. Comparison of mean execution time for different algorithms and matrix sizes

increases for bigger matrices. The third observation is that the difference in performance between `gpu_naive` (Algorithm 4) and `gpu_tiled` (Algorithm 5) is quite small for smaller matrices, but grows as the matrix size increases. That means that for larger matrices, the usage of shared memory has a bigger effect on performance. The fourth observation is that the block size does affect the performance. In the experiments the optimal performance was achieved by using a block size of 16. Increasing the block size to 32 caused a drop in effective bandwidth for all GPU algorithms and matrix sizes.

V. Conclusion

After analyzing various algorithms and metrics, it's evident that utilizing blocks can enhance the performance of matrix transposition algorithms. This was demonstrated through the better execution time and effective bandwidth of compiler-optimized versions of the `oblivious128` implementation, which incorporates a form of blocking. The third algorithm also presents promising directions for parallelization, as each submatrix can be processed independently, offering straightforward potential for parallel execution.

The main observation made during the homework was the significant increase in performance by using GPU acceleration. Using GPUs for parallelization can result in large improvements with respect to effective bandwidth and execution time. It was also observed that a careful design of the GPU algorithm is needed. Different implementations and parameter choices can result in significant performance differences, even if the algorithm is executed on a GPU. The presented CUDA kernels could be further improved by using more advanced features of the CUDA platform, such as multiple streaming processors. Another approach could be to finetune the parameter choices. In the presented experiments, a block size of 16 was optimal, but further parameter choices could be investigated. One could also optimize the CUDA implementation to run on multiple GPUs.

References

[1] Louis Brunner. *Valgrind for macOS*. 2024. URL: https://github.com/LouisBrunner/valgrind-macos/issues/56#issuecomment-1971933471 (visited on 04/27/2024).

[2] Siddhartha Chatterjee and Sandeep Sen. "Cache-efficient matrix transposition". In: *Proceedings Sixth International Symposium on High-Performance Computer Architecture. HPCA-6 (Cat. No. PR00550)*. IEEE. 2000, pp. 195–205.

[3] Valgrind Developers. *Valgrind*. 2024. URL: https://valgrind.org/ (visited on 04/27/2024).

[4] Valgrind Developers. *Valgrind Documentation*. 2024. URL: https://valgrind.org/docs/manual/dist.news.html (visited on 04/27/2024).

[5] Nvidia. *CUDA C++ Best Practices Guide*. 2024. URL: https://docs.nvidia.com/cuda/cuda-c-best-practices-guide/contents.html (visited on 06/05/2024).

[6] Sergey Slotin. *Algorithmica*. 2022. URL: https://en.algorithmica.org/hpc/external-memory/oblivious/ (visited on 04/27/2024).

[7] GCC Team. *GCC Documentation*. 2023. URL: https://gcc.gnu.org/onlinedocs/gcc/Other-Builtins.html (visited on 04/27/2024).