Uppsala University

Parallel and Distributed Programming

1TD070

Individual Project:
Shear Sort

Students:
Linjing Shen

Lecturer:
Prof. Roman Iakymchuk

May 31, 2024

Contents

1 Introduction

Shearsort is a sorting algorithm particularly well-suited for sorting on a two-dimensional grid, developed by Sen, Shamir, and Isaac Scherson at the University of California, Irvine [2]. The core idea is to arrange data in a 2D grid and perform alternating row and column sorting steps, gradually moving the data in the entire grid towards the final sorted state. Each row or column sorting operation is referred to as a "shear," and the overall sorting process resembles the action of shearing sheep, hence the name Shearsort. The basic idea behind Shearsort is to perform $\lceil \log(n) + 1 \rceil$ iterations to sort the values in an $n \times n$ matrix with a total of N values. In each iteration, the rows and columns are sorted. Specifically, even-indexed rows are sorted in ascending order, odd-indexed rows are sorted in descending order, and then all columns are sorted in ascending order.
Shearsort is well-suited for parallel processing, where each processing unit can handle sorting a set of rows or columns independently. Sorting of rows can be done independently, while sorting of columns can be efficiently achieved by transposing the matrix. The MPI function `MPI_Alltoall()` is particularly useful for efficient global communication and data redistribution, facilitating data exchange and coordination in parallel computing [1,3,4]. Additionally, the course materials from the Parallel Algorithm course at the University of Science and Technology of China mention various other methods for sorting two-dimensional arrays besides shear sort. For example, sorting can be performed using a divide-and-conquer strategy, continually breaking down and merging data for sorting, or employing element-swapping algorithms like odd-even transposition sorting [5].

2 Algorithms

2.1 Sequential Algorithm

In terms of data storage, projecting the obtained two-dimensional array onto a one-dimensional linear array ensures that contiguous array members are stored close to each other in memory, reducing the likelihood of cache misses. However, when shearsort begins to use vertical access in the second step, iterating over each column instead of each row, the code will attempt to access data stored in a non-sequential order. As N grows larger, cache misses become inevitable, leading to poor performance. Therefore, when performing column sorting, the code first transposes the matrix, performs row sorting on the transposed content instead of directly implementing column sorting, and then transposes it back to its normal form after completing this process.
Besides, ShearSort inherently requires local sorting of arrays. I tested the algorithm on an input matrix of size 5,000. Compared to the standard C `qsort` function, I also tried implementing quicksort and mergesort. However, both performed worse than `qsort` to varying degrees. Therefore, the code simply uses the `qsort` function.

2.2 Parallel Algorithm

For the parallel algorithm, another benefit of using a 1D array instead of a 2D array to store matrix data is that the MPI library handles one-dimensional arrays more efficiently, and it simplifies the process of data distribution and collection. Not only does this storage method adhere more closely to the principle of memory contiguity, improving access efficiency, but it also better aligns with the MPI library's parallel computing model, enhancing program performance and scalability.

In the design of parallel algorithms, considering that the size n of the 2D array may not be evenly divisible by the number of processes, a blocking method is used to allocate matrix data with unequal row counts to different processes. The size of the data block each process receives is determined by the matrix size and the number of processes, using the `sendcounts` and `displs` arrays to describe the size and offset of the data block each process receives. Starting from process 0, redundant rows are allocated to the corresponding processes one by one to avoid allocating all redundant rows to one process. This blocking method helps reduce data communication overhead and makes parallel computation more efficient. Accordingly, `MPI_Scatterv()` and `MPI_Gatherv()` are used instead of the common equal division and aggregation methods `MPI_Scatter()` and `MPI_Gather()`. These two functions are used to distribute data from the root process to other processes and to collect data from each process back to the root process, respectively.

After the matrix is read from the root process and allocated to each process, the iterations of the shear sort algorithm begin. In the first step, each process performs local row-wise ascending or descending sorting. If the current iteration count is less than or equal to $\lceil \log_2(n) \rceil$, the second step begins, which is the matrix transposition algorithm. First, each process converts its data into a format usable by `MPI_Alltoallv()`, then employs `MPI_Alltoallv()` to facilitate data exchange among all processes, and finally reverts the data according to the corresponding rules within each process. This entire process achieves one transposition. Subsequently, each process performs local row-wise ascending sorting, instead of implementing column sorting of the matrix. At the end of each iteration, the matrix is transposed back using the aforementioned transposition method. Furthermore, the algorithm uses `MPI_Wtime()` to measure the program's execution time, calling it at the beginning and end of the program to calculate the runtime. `MPI_Reduce()` is then utilized to perform reduction operations on execution times across all processes, calculating global statistics and obtaining the maximum time consumed.

3 Experiments and Results

In the experimental phase, I mainly aimed to evaluate the scalability of the parallel shearsort algorithm, which includes both strong scalability experiments and weak scalability experiments. Firstly, the strong scalability experiments were conducted by keeping the input size constant and gradually increasing the number of parallel processing processes. Large input files stored in a specified directory were utilized as test data to comprehensively assess the program's performance. The execution time of the program under

different numbers of processes was recorded for each experiment, and speedup was calculated to evaluate the performance of the parallel program as the number of processors changed.
Secondly, the experimental plan also included conducting weak scalability experiments to ensure that the problem size handled by each process remained constant as the number of processes increased. This implies that as the number of processes increases, the total amount of data processed will also increase accordingly, maintaining a proportional relationship between problem size and the number of processes. This experiment required recording the execution time of the program under different problem sizes and numbers of processes and calculating speedup to assess whether the execution time of the parallel program remained stable under different problem sizes and numbers of processes.
To ensure the accuracy and reliability of the experimental results, multiple experiments were conducted for each parameter combination, and average values were obtained.

3.1 Strong scalability experiment

Table 1 and Figure 1 show the results of the strong scaling experiments, which were carried out by keeping the input size constant at 8,000 while incrementally increasing the number of parallel processing processes, using the number of processes ranging from one to sixteen. Each sub-experiment was run five times, and the average of these runs was calculated.
The formula for the acceleration ratio is as follows, where T_s is the serial execution time and T_p is the parallel execution time.

$$\text{Speedup} = \frac{T_s}{T_p}$$

Firstly, from both the figure and the table, it can be observed that as the number of processes increases, the execution time gradually decreases, which is in line with expectations. From a single process to 16 processes, the execution time decreases from 98.6729 seconds to 6.4618 seconds, indicating a significant improvement. This suggests that with the increase in the number of processes, the computational workload is more evenly distributed across multiple processes, thus accelerating the execution of the entire algorithm.
Secondly, based on the provided data, it can be seen that as the number of processes increases, the speedup gradually increases, ranging from 1.00 for a single process to 15.27 for 16 processes. Although the rate of speedup growth inevitably slows down with the increase in the number of processes, the actual speedup achieved is very close to the ideal speedup. The slowdown in speedup growth is due to the fact that as the number of processors increases, communication and synchronization overhead become more significant, thus limiting further increases in speedup.
Overall, with the increase in the number of processes, the execution time decreases, and the speedup gradually increases, indicating that the algorithm exhibits very good strong scalability. Comparing to version one's experimental results, where the algorithm was compiled with -O3 optimization for scalability experiments, while in version two,

Process Number	Execution Time (s)	Standard Deviation	Speedup
1	98.6729	0.6105	1.00
2	50.1910	0.6085	1.97
3	33.4618	0.3225	2.95
4	24.9241	0.1940	3.96
5	20.2304	0.0507	4.88
6	17.0864	0.1517	5.77
7	14.7494	0.0630	6.69
8	12.8963	0.0999	7.65
9	11.7415	0.1089	8.40
10	10.5347	0.0902	9.37
11	9.6192	0.0377	10.26
12	8.6258	0.0433	11.44
13	8.0262	0.0480	12.29
14	7.4242	0.0306	13.29
15	6.8626	0.0311	14.38
16	6.4618	0.0567	15.27

Table 1: Strong scalability experiment

compiler optimizations were disabled. In version one, at 12 processes, only half of the ideal speedup was achieved.

3.2 Weak scalability experiment

In the weak scalability experiment, the experiment guarantees that each process processes a 1-D array of data with a total number of 1,000,000, and calculates the time-consumption and efficiency when the number of processes ranges from 1 to 64. The results are shown in Table 2 and Figure 2. Each configuration is run five times, and the average is calculated. The efficiency calculation formula is as follows, where p denotes the number of processes, T_s is the serial execution time and T_p is the parallel execution time.

$$\mathrm{E} = \frac{T_s}{p * T_p}$$

From both the figure and the table, it can be observed that as the size of the input matrix increases, there is a trend of gradually increasing execution time, and correspondingly, the efficiency decreases. For instance, from an input size of 1,000 to 3,000 and from 1 process to 9 processes, the execution time and efficiency can maintain relatively stable. However, when the input size increases to 4,000 and the number of parallel processes reaches 16, the efficiency of the code decreases, although it remains within a reasonable range.

Furthermore, when the problem size is greater than or equal to 5,000 and the number of parallel processes is greater than or equal to 25, despite using more parallel processing units, both the efficiency and execution time cannot be well-maintained. It is evident that

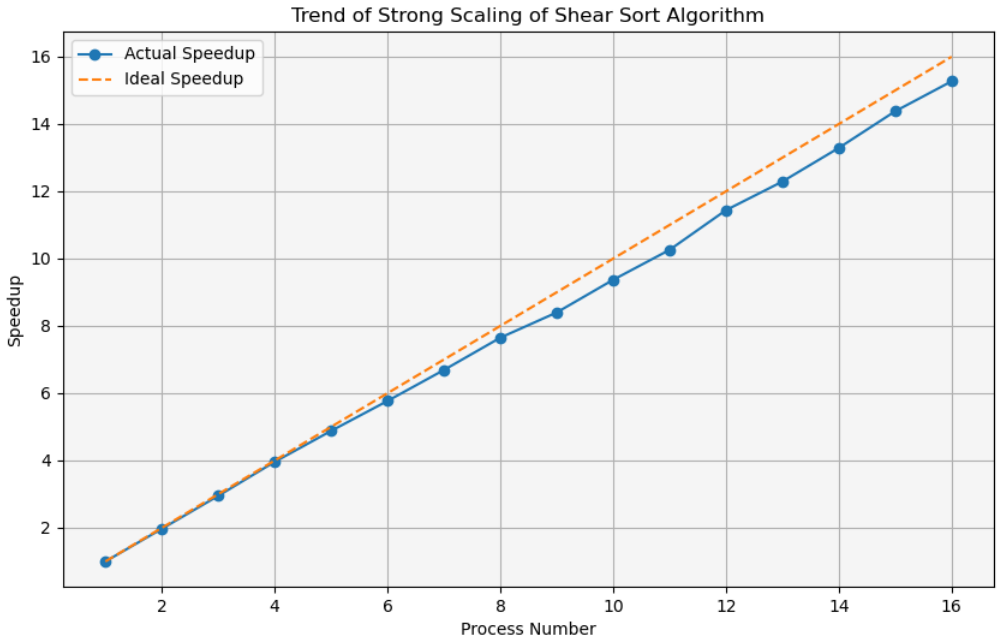

Figure 1: Strong scalability experiment

Process Number	Input Size	Execution Time (s)	Efficiency
1	1,000	1.1392	1.00
4	2,000	1.0693	1.05
9	3,000	1.2742	1.02
16	4,000	1.5550	0.86
25	5,000	2.1270	0.68
36	6,000	2.6773	0.55
49	7,000	3.4746	0.43
64	8,000	4.6105	0.33

Table 2: Weak scalability experiment

the efficiency gradually decreases with the increase in the number of processes, which may be due to the corresponding increase in communication and synchronization overhead. Larger problem sizes imply more communication between processes, and additionally, larger problem sizes also consume more memory space.
Overall, the algorithm exhibits acceptable weak scalability, and it is reasonable for efficiency to decrease when the number of processes is too large. Additionally, I conducted experiments where each process handled larger-scale data, such as one process handling 4,000,000 or 16,000,000 data points, compared to each process handling 1,000,000 data points. These experiments showed better weak scalability.

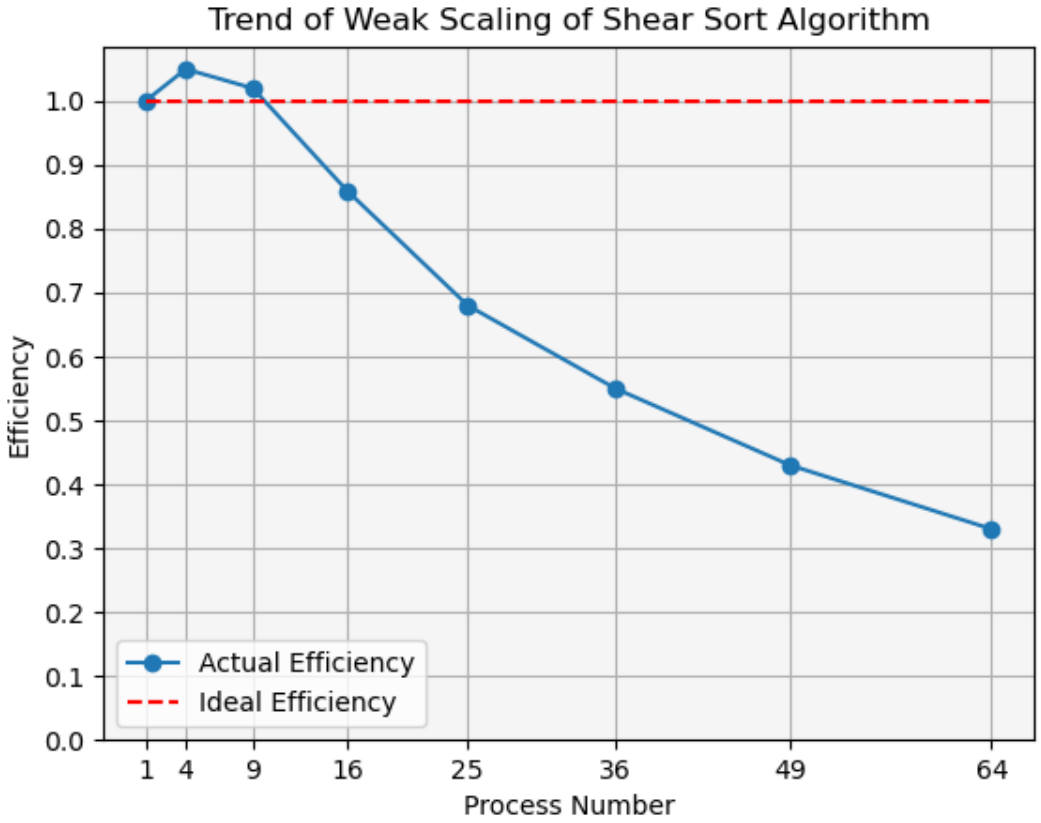

Figure 2: Weak scalability experiment

4 Discussion

Based on the data and analysis from the scalability experiments, it can be concluded that the algorithm achieves ideal strong scalability and acceptable weak scalability. This is because as the problem size or the number of processes increases, the frequency of inter-process communication and synchronization operations also increases, resulting in overhead, which becomes particularly evident when dealing with large-scale problems.

Additionally, larger problem sizes require more memory to store data, which may lead to memory access bottlenecks, thereby limiting the improvement in parallel efficiency. Moreover, when the problem size is large, load imbalance issues may arise, where some processes may need to handle more workload than others, leading to decreased efficiency.
Apart from this version of the code, I also attempted other implementations of shearsort, such as using `MPI_Scatterv()` and `MPI_Gatherv()` to transpose matrices instead of `MPI_Alltoallv()`. However, test results showed significant improvements in both strong and weak scalability when using `MPI_Alltoallv()` compared to multiple uses of `MPI_Scatterv()` and `MPI_Gatherv()` to transpose matrices. When using `MPI_Scatterv()` and `MPI_Gatherv()`, parallelizing with 16 processes on a 5,000-sized input matrix yielded less than a speedup of 2, possibly due to the increased communication overhead caused by gathering and scattering within iterations, leading to communication imbalance, while `MPI_Alltoallv()` patterns such as full-to-full communication exchange the same amount of data among all processes, avoiding single-point bottlenecks.
Although I compared quicksort and mergesort with `qsort`, there are many other local sorting methods, and some might have better scalability and shorter runtime than `qsort`'s implementation. In my previous tests, for a 5,000-sized input matrix, quicksort performed better in terms of execution time than `qsort` up to 20 processes, but its overall strong and weak scalability was about 5% worse. Mergesort, on the other hand, had longer runtime compared to `qsort`.
To achieve better performance, optimizing memory usage could be considered, such as reusing already allocated memory space to avoid frequent allocations and deallocations, such as within the `temp` variable loop. Additionally, although I accounted for cases where the number of processes does not evenly divide the matrix size, load imbalance issues may still arise. Dynamic adjustment of task allocation to ensure roughly equal loads among processes could be attempted.

5 Peer Reviewing

Yangmei Lin:
This report provides a detailed exploration of the Shear Sort algorithm in a parallel and distributed programming environment. The structure is well-organized, with clear sections covering the introduction, algorithms, experiments, results, and discussions as required. The comprehensive coverage enhances the reader's understanding of the algorithm and its application in parallel processing. The use of tables and figures to present experimental data enhances the clarity and visual appeal of the report.
The discussion of data storage is impressive, providing a good explanation of why a 1D array is used instead of a 2D array. The implementation utilizes `MPI_Alltoallv()` for data exchange after comparing it with other collective communication functions. There is also a comprehensive examination of local sort algorithms.
The discussion section focuses on the issues encountered in the project, including attempts and solutions to address them. For the poor weak scalability, the author performs several examinations and provides the best explanation.
Regarding the efficiency of weak scalability, the numbers do not match the given formu-

lation. Additionally, there is a zigzag pattern observed in the weak scalability figure, and it would be beneficial to include an explanation for this phenomenon.
The code is well-structured and organized, making efficient use of MPI functions, especially `MPI_Alltoallv()`. The code correctly transposes the matrix by reordering the local data before and after the Alltoallv communication, which is a crucial step in the Shear Sort algorithm.
It is better to consider creating functions to reduce code duplication, such as for the reordering of local data.
Overall, it is an insightful and valuable research on Shear Sort in parallel processing.

References

[1] Norm Matloff. Programming on parallel machines. *University of California, Davis*, 39319, 2011.

[2] Sandeep Sen, Isaac D. Scherson, and Adi Shamir. Shear sort: A true two-dimensional sorting techniques for vlsi networks. In *International Conference on Parallel Processing*, 1986. https://api.semanticscholar.org/CorpusID:40018053.

[3] Taendyr. Transpose matrix with mpi alltoallv, 2021. https://stackoverflow.com/questions/65665462/transpose-matrix-with-mpi-alltoallv.

[4] velenos14. C mpi: How to transpose a pseudo 2d array already scattered across processes, 2023. https://stackoverflow.com/questions/76851433/c-mpi-how-to-transpose-a-pseudo-2d-array-already-scattered-across-processes.

[5] Yun Xu. Sorting and selecting in synchronous. Teaching Resources, Feb.–Jun. 2024. https://wenku.baidu.com/view/83fd796858fafab069dc02ec?fr=xiongzhanghao&bfetype=new&_wkts_=1715902069200&needWelcomeRecommand=1.