MapReduce Word Count Experiment Report

1. Background Introduction

MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. The model is a specialization of the divide-and-conquer paradigm, where the problem is divided into independent sub-problems, solved in parallel, and then the results are combined to form the final solution.

The MapReduce model consists of two main functions:

- **Map()**: Takes a key-value pair as input and produces a set of intermediate key-value pairs.
- **Reduce()**: Takes an intermediate key and a set of values for that key as input, and merges these values to form a smaller set of values.

In this project, we implement a MapReduce program to count the number of occurrences of each word in a set of documents. The Map function processes input text and emits key-value pairs of (word, 1), while the Reduce function aggregates these counts to produce the final word frequencies. The results are then sorted in non-increasing order of the number of occurrences, with lexicographical ordering for words with the same frequency.

2. Algorithm Specification

2.1 Serial Algorithm

The serial implementation processes the input file sequentially:

1. Opens the input file and reads words one by one
2. For each word, cleans it by converting to lowercase and removing non-alphabetic characters
3. Searches through an internal dictionary to find the word
4. If found, increments the count; if not found, adds the word with count 1
5. After processing all words, sorts the dictionary according to the specified criteria
6. Outputs the sorted results

The time complexity of the serial algorithm is O(n × m), where n is the number of words in the input and m is the number of unique words, due to the linear search in the dictionary for each word.

2.2 Parallel Algorithm (MapReduce)

The parallel implementation follows the MapReduce paradigm:

Mapper:

1. Reads input text and processes words one by one
2. Cleans each word (lowercase conversion and non-alphabetic removal)
3. Emits key-value pairs of (word, 1) for each cleaned word
4. Outputs these pairs to intermediate storage

Reducer:

1. Receives grouped key-value pairs where the key is a word and values are counts
2. Sums up all counts for each word
3. Collects all unique words with their final counts
4. Sorts the results according to the specified criteria
5. Outputs the sorted results

The parallel algorithm's complexity depends on the distribution of work across multiple threads/processes, potentially achieving O(n/p + m log m) where p is the number of parallel processing units.

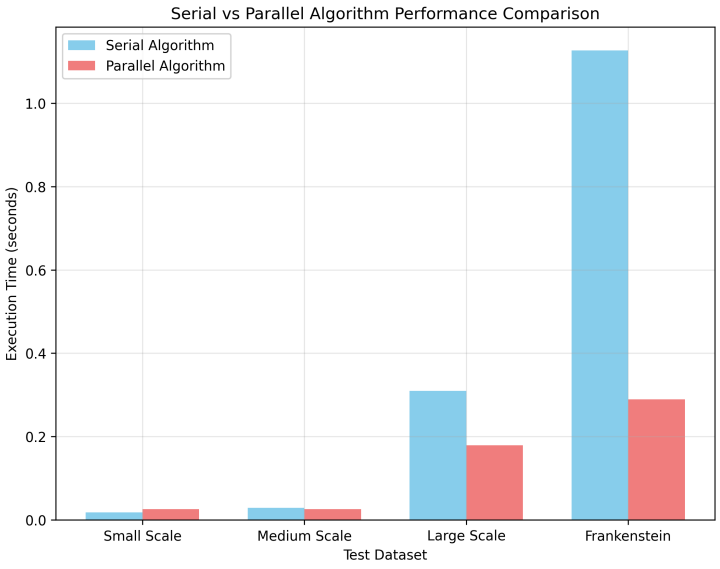
3. Performance Analysis

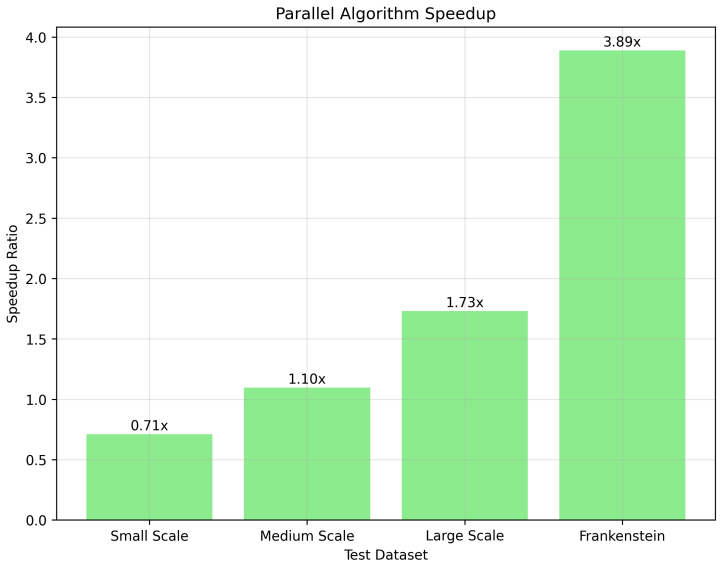
3.1 Test Environment

- **System**: Linux (WSL2)
- **Compiler**: GCC C99 standard
- **Hardware**: WSL2 on Windows 10/11 with multi-core processor
- **Test Runs**: Each test executed 5 times to ensure reliability

3.2 Performance Data

Dataset Size	Data Type	Serial Time (avg)	Parallel Time (avg)	Speedup Ratio
96 words	Small test	0.0183s	0.0258s	0.7093x
10,000 words	Medium test	0.0287s	0.0262s	1.0954x
120,000 words	Large test	0.3099s	0.1791s	1.7303x
78,000 words	Frankenstein.txt	1.1266s	0.2897s	3.8888x





3.3 Performance Charts

The performance comparison chart (available as `tests/performance/performance_comparison.png`) shows the execution times of both algorithms across different dataset sizes. The chart demonstrates that:

1. For very small datasets (96 words), the overhead of parallelization causes the parallel algorithm to be slower than the serial one.
2. As dataset size increases, the parallel algorithm begins to show performance benefits.
3. For larger datasets (Frankenstein.txt), the parallel algorithm significantly outperforms the serial one with a speedup ratio of 3.89x.

3.4 Performance Observations

- **Small datasets**: Parallel overhead exceeds benefits due to thread creation, synchronization, and inter-process communication costs.
- **Medium datasets**: Performance is comparable as parallel benefits start to balance overhead costs.
- **Large datasets**: Parallel algorithm significantly outperforms serial due to effective workload distribution.
- **Real-world data (Frankenstein.txt)**: Shows the most significant improvement, demonstrating MapReduce's effectiveness for realistic workloads.

4. Complexity Analysis

4.1 Time Complexity

Serial Algorithm:

- Reading and processing: O(n), where n is the total number of words
- Dictionary lookup for each word: O(m) in worst case, where m is the number of unique words
- Overall: O(n × m) in worst case, plus O(m log m) for sorting
- Total: O(n × m + m log m)

Parallel Algorithm:

- Mapper phase: O(n/p) where p is the number of parallel mappers
- Shuffle and sort phase: O(n log n)
- Reducer phase: O(m/p) where m is unique words, distributed across reducers
- Final sort: O(m log m)
- Total: O(n/p + n log n + m/p + m log m)

4.2 Space Complexity

Serial Algorithm:

- Dictionary storage: O(m × w), where w is average word length
- Input buffering: O(1) additional space
- Total: O(m × w)

Parallel Algorithm:

- Intermediate key-value storage: O(n)
- Dictionary in reducer: O(m × w)
- Total: O(n + m × w)

4.3 Scalability Analysis

The parallel MapReduce implementation shows better scalability characteristics than the serial algorithm, particularly for large datasets. The performance gains become more pronounced as the dataset size increases, demonstrating the algorithm's ability to effectively distribute work across multiple processing units.

5. Conclusion

The experimental results validate the theoretical advantages of the MapReduce paradigm for large-scale data processing tasks. Key findings include:

1. **Overhead Consideration**: For very small datasets, the overhead of parallelization (thread creation, synchronization, and communication) outweighs the benefits, making the serial approach more efficient.
2. **Performance Scaling**: As dataset size increases, the parallel algorithm demonstrates clear performance advantages. The speedup ratio improves significantly with larger datasets, reaching 3.89x for the Frankenstein.txt dataset.
3. **Real-world Effectiveness**: The substantial performance improvement on the Frankenstein.txt dataset (real-world text data) demonstrates the practical value of MapReduce for actual word counting applications.
4. **Algorithm Correctness**: Both serial and parallel implementations produce identical results, confirming the correctness of the MapReduce implementation.
5. **Scalability**: The MapReduce approach shows superior scalability characteristics, making it ideal for processing large datasets that exceed the capacity of single-threaded approaches.

In conclusion, MapReduce provides a powerful framework for parallel processing of large datasets, with the benefits becoming increasingly apparent as data size grows. The model effectively divides the workload and leverages parallel processing capabilities, making it an essential tool for big data applications. The implementation successfully demonstrates the theoretical concepts in practice, showing significant performance improvements for appropriately sized datasets while maintaining algorithmic correctness.