Parallel Sorting Algorithm report

Introduction:

Sorting is a fundamental operation in computer science, essential for optimizing the efficiency of various applications. We have a lot of different algorithm out there such as QuickSort, MergeSort and HeapSort. Those are well known and have been studied for a long period.

But with the advent of multicore systems, it becomes crucial to leverage parallelism to enhance sorting performance on large datasets.

The goal of the project is to compare: **Sequential Sorting and Parallel Sorting** on large datasets and see if we effectively gain an increase in performance:

Definition:

- **Sequential time:** This is the time used by the **sequential sorting algorithm** to process an array of n elements:
- **Parallel time:** This is the time used by the **parallel sorting algorithm** to process an array of n elements:
- **Speedup:** This is the quotient of the sequential time and the parallel time. It's used to measure the effectiveness of the parallel sorting algorithm over the sequential version.

Key Concepts:

- **Divide and Conquer**: Many parallel sorting algorithms are based on the divide-and-conquer paradigm. The dataset is divided into smaller parts, each of which is sorted independently and in parallel. The results are then combined to produce the final sorted array.
- **Data Partitioning**: Efficient partitioning of data is crucial for parallel sorting. The data must be divided in a way that balances the load across all processing units, ensuring that no single unit becomes a bottleneck.
- **Synchronization and Communication**: Managing synchronization and communication between parallel tasks is vital. Efficient algorithms minimize the overhead associated with these operations to maximize the speedup gained from parallelism.

Approach:

- **Initialization:** First of all initialize two arrays. We'll incrementally increase their size and check the speedup time to draw a conclusion

```java
// Initialize an ArrayList to hold integers
ArrayList<Integer> array = new ArrayList<>();
// Fill the array with random integers
fillarray(array);
List<Integer> array2 = array.subList(0, array.size());
```

- **Sequential Time:** We try to evaluate the required time for the sequential operation. We should note that behind the scenes the **Collections.sort** is using a synchronous Merge Sort.

```java
//Sorting and timing using sequential sorting
Long startSequentialTime = System.currentTimeMillis();
Collections.sort(array2);
Long sequentialTime = System.currentTimeMillis() - startSequentialTime;
```

- **Parallel Time:** We try to evaluate the required time for the parallel operation. The ForkJoinPool is gonna be used to manage the Recursive Task that we are creating. It's gonna control and manage threads allocation

```java
// Create a ForkJoinPool to manage parallel tasks
ForkJoinPool pool = new ForkJoinPool();

// Create a MergeSortTask to sort the array
MergeSortTask task = new MergeSortTask(array,  start: 0, array.size());

// Record the start time for performance measurement
Long startTime = System.currentTimeMillis();

// Invoke the MergeSortTask using the ForkJoinPool
List<Integer> sortedList = pool.invoke(task);

// Record the end time for performance measurement
Long parallelTime = System.currentTimeMillis() - startTime;
```

- **MergeSortTask:** We defined an instance of RecursiveTask. Indeed, due to the recursive nature of the merge sort algorithm. We use an instance of RecursiveTask to easily handle the depth of recursion.

```java
public class MergeSortTask extends RecursiveTask<List<Integer>> {
    // Threshold for switching to sequential sort
    1 usage
    private static final int THRESHOLD = 15;

    // List of numbers to be sorted
    4 usages
    private List<Integer> numbers;

    // Start and end indices for the current task
    5 usages
    private int start, end;

    /**
     * Constructor to initialize the MergeSortTask.
     *
     * @param numbers the list of integers to sort
     * @param start the starting index of the sublist
     * @param end the ending index of the sublist
     */
    3 usages   atakoutene
    public MergeSortTask(List<Integer> numbers, int start, int end) {
        this.numbers = numbers;
        this.start = start;
        this.end = end;
    }
```

The leftTask is gonna be executed asyncrhonously on another thread while the right one will be executed on the current thread. They will be merged together after completion using the merge function. It should be noted that for task dealing with less than **THRESHOLD(15)** elements, we use a simple sequential merge sort algorithm.

```java
@Override
protected List<Integer> compute() {
    // If the sublist size is below the threshold, sort it directly
    if (end - start ≤ THRESHOLD) {
        List<Integer> nums = new ArrayList<>(numbers.subList(start, end));
        Collections.sort(nums);
        return nums;
    }

    // Find the middle index to split the list
    int mid = (start + end) / 2;

    // Create subtasks for the left and right halves
    MergeSortTask leftTask = new MergeSortTask(numbers, start, mid);
    MergeSortTask rightTask = new MergeSortTask(numbers, mid, end);

    // Execute the left task asynchronously
    leftTask.fork();

    // Execute the right task directly
    List<Integer> right = rightTask.compute();

    // Wait for the left task to complete and get the result
    List<Integer> left = leftTask.join();

    // Merge the results of the left and right tasks
    return merge(right, left);
}
```

- **Evaluation**

```
For 10000 elements.
Elapsed Time for sequential sort: 10ms
Elapsed Time for parallel sort: 19ms
The speedup is: 0.5263158
```

```
For 100000 elements.
Elapsed Time for sequential sort: 49ms
Elapsed Time for parallel sort: 54ms
The speedup is: 0.9074074
```

```
For 1000 elements.
Elapsed Time for sequential sort: 1ms
Elapsed Time for parallel sort: 4ms
The speedup is: 0.25
```

```
For 1000000 elements.
Elapsed Time for sequential sort: 478ms
Elapsed Time for parallel sort: 325ms
The speedup is: 1.4707693
```

```
For 100000000 elements.
Elapsed Time for sequential sort: 27795ms
Elapsed Time for parallel sort: 19510ms
The speedup is: 1.424654
```

Analysis:

1. **1,000 Elements**:
   - Sequential Sort: 1 ms
   - Parallel Sort: 4 ms
   - Speedup: 0.25

   For small data sizes (1000 elements), parallel sorting performs worse than sequential sorting. This is likely due to the overhead of parallel processing outweighing the benefits.
2. **10,000 Elements**:
   - Sequential Sort: 10 ms
   - Parallel Sort: 19 ms
   - Speedup: 0.5263158

   For 10000 elements, parallel sorting still performs worse than sequential sorting, with the speedup being less than 1. The overhead continues to have a significant impact.
3. **100,000 Elements**:
   - Sequential Sort: 49 ms
   - Parallel Sort: 54 ms
   - Speedup: 0.9074074

   As the data size increases to 100000 elements, the performance of parallel sorting starts to get closer to sequential sorting, though it is still slightly slower. The speedup is approaching 1, indicating that the benefits of parallel processing are starting to be realized.
4. **1,000,000 Elements**:
   - Sequential Sort: 478 ms
   - Parallel Sort: 325 ms
   - Speedup: 1.4707693

   For large data sizes (1000000 elements), parallel sorting outperforms sequential sorting, with a speedup greater than 1. This shows that parallel sorting is more efficient for larger datasets, as the overhead becomes negligible compared to the sorting work itself.

Conclusion

- **Small Data Sizes (1000 to 10000 elements)**: Parallel sorting is less efficient due to the overhead of managing parallel tasks.
- **Medium Data Sizes (100000 elements)**: Parallel sorting becomes nearly as efficient as sequential sorting, with overhead starting to be compensated by the benefits of parallel processing.
- **Large Data Sizes (1000000 elements)**: Parallel sorting is significantly more efficient than sequential sorting, with a clear performance improvement and speedup greater than 1.

Overall, parallel sorting demonstrates its strength in handling large datasets effectively, whereas sequential sorting remains more efficient for smaller datasets due to lower overhead.