CSCE 435 Group project

0. Group number: 13

The team will communicate with discord channel

1. Group members:

1. Ravish Shardha
2. Lydia Harding
3. Brack Harmon
4. Jack Hoppe

2.Parallel sorting algorithms

2a. Brief project description (what algorithms will you be comparing and on what architectures)

All algorithms are going to be written and tested on grace cluster using MPI

- Bitonic Sort:
    - Parallel sorting algorithm that repeatedly sorts segments of a sequence of numbers into bitonic sequences. A bitonic sequence consists of a list of numbers which is first increasing, then decreasing. In the parallel version, for each i round, sorting is done with the partner that differs in the ith bit, alternating between collecting the lower half of all elements, and collecting the higher half of all elements, then sorting in order. Once all rounds have completed, the full array is sorted.
    - Bitonic sort only works on input arrays of size 2^n.
    - Bitonic sort will make the same number of comparisons for any input array, with a complexity of O(log^2n), where n is the number of elements to be sorted.
- Sample Sort:
    - Sample sort is a parallelized version of bucket sort.
    - Each processor takes a chunk of the data and sorts locally.
    - Then each processor takes s samples and those samples get combined into a buffer and sorted.
    - Global splitters or pivots are selected form the sorted samples and define endpoints for buckets.
    - each processor takes its data and filters it into each respective bucket.
    - The buckets are sorted and combined.
    - Finally the endpoints of each bucket are checked to verify the whole dataset is sorted.
    - Uses quicksort to sort partitioned buckets, so works well with random and in-order data.
- Merge Sort:
    - Divide-and-conquer sorting algorithm that splits the array into halves recursively until each sub-array contains a single element, then merges the sub-arrays to produce a sorted array.
    - Best and Worst case time complexity is O(nlogn) and space complexity is O(n).

    - For reverse and random sorted data, merge is a good choice.
    - For sorted and nearly sorted(1% random), merge sort doesn't consider the existing order so might not be the best choice.

- Radix Sort:

    - The Radix sort is a non-comparative integer sorting algorithm that iterates through each digit of the given elements starting from the least significant digit and progressing to the most significant digit. As the algorithm iterates the temporary results are placed in a "bucket" using an algorithm like the "count sort". These buckets are used to create the new ordering of the elements until all digits have been processed.
    - Through the MPI library and utilizing Grace a parallel implementation can be achieved by splitting input data into chunks then distributing them to workers. The workers will sort its chunk of data based on the current digit. After sorting, the bucket data from the count sort is redistributed across the processes. This continues until the most significant digit is reached resulting in sorted data.
    - The Radix sort requires integers or data that can be represented with integers and a fixed range of digits, i.e., (0-9)
    - Larger integers cause more passes and slows the algorithm
    - Uneven distributions of input data can lead to inefficiencies due to some buckets becoming disproportionately large.
    - Runtime: O(d*n) where d = number of digits and n = number of elements

2b. Pseudocode for each parallel algorithm

- For MPI programs, include MPI calls you will use to coordinate between processes
- **Merge Sort:**

```
INITIALIZE MPI(MPI_Init)
GET world_rank(MPI_Comm_rank)
GET world_size(MPI_Comm_size)

DIVIDE n by world_size to get chunk_size

CREATE sub_array of size chunk_size

SCATTER original_array to all processes (MPI_Scatter)
EACH PROCESS receives sub_array of size chunk_size

// Local Merge Sort
CALL mergeSort(sub_array, 0, chunk_size - 1)

// Gather sorted sub-arrays at root
GATHER sub_array at root into original_array (MPI_Gather)

IF world_rank == 0 THEN
    CALL mergeSort(original_array, 0, n - 1)  // Final merge at root
    PRINT sorted original_array

//Clean root and the rest of dynamically allocated arrays
```

```
  FREE sub_array and temp_arrays

  FINALIZE MPI
```

Bitonic Sort:

```
////////////////////
// MAIN
// MPI_Init,
// MPI_Comm_size(num_procs)
// MPI_Comm_rank(rank)

// Generate the input array on each processor (rand, in-order, reverse-order, or
perturbed)

// MPI_Barrier

// BITONIC SORT
// dimensions = log2(num_proc)
// For i = 0 to dimensions - 1:
    // For j = i down to 0:
        if (i + 1)st bit of rank == jth bit of rank then
            COMP EXCHANGE MIN (j)
        else
            COMP EXCHANGE MAX (j)
        // MPI_Barrier to ensure steps are in sync


// MPI_Barrier to ensure all sorting is complete

// VERIFY SORT
// Check if array is sorted locally
// Check if array end is less than start of neighbor process's array
// If both true for all processors, array is sorted.

// free array

// MPI Finalize
// RETURN
// END MAIN

////////////////////
// HELPER FUNCTIONS

////////////////////
// COMP EXCHANGE MIN (j)
// partner process = rank XOR (1 << j)
// MPI_Sendrecv array with partner, store in buffer_receive

// Concatenate array and buffer receive, store as temp buffer
// Sort temp
// Set array to be the lower half of temp
```

```
// free buffers
// RETURN
// END COMP EXCHANGE MIN
////////////////////

////////////////////
// COMP EXCHANGE MAX (j)
// partner process = rank XOR (1 << j)
// MPI_Sendrecv array with partner, store in buffer_receive

// Concatenate array and buffer receive, store as temp
// Sort temp
// Set array to be the higher half of temp

// free buffers
// RETURN
// END COMP EXCHANGE MAX
////////////////////
```

Sample Sort:

```
    // Starting MPI commands
    MPI_Init
    MPI_Comm_rank
    MPI_Comm_world
    etc.

    // Set Number of elements per processor
    data_size = total_size / num_processors

    // Fill local data with random integers
    for each element in data from rank * data_size to (rank + 1) * data_size:
    data[i] = random int from 1 to 999

    // Here is where we start timing the Samplesort algorithm

    // Step 1: Sort the local data
    sort local_data

    // Step 2: Select local data samples
    set samples_list empty
    samples_list[i] = sample an index from local_data

    // Step 3: Gather all the samples at rank 0 (rank 0 handles the samples)
    call MPI_Gather with sample data to fill gathered_samples


    // Step 4: Rank 0 sorts the samples and selects splitters
    if(rank == 0)
        sort gathered_samples
        for each sample
```

```
            splitters[i] = sample from gathered_samples


    // Step 5: Broadcast the splitters to all processes
    call MPI_Bcast with splitters and num_samples

    // Step 6: Partition the local data based on the splitters

    create send_counts, send_offsets, and partitioned_data arrays
    populate partitioned_data with info from data based on offsets and send counts

    // Step 7: Send partitioned data to respective processor buckets
    use MPI_Alltoall with send_counts and recv_counts
    create recv_data array and populate with received information using
MPI_Alltoallv

    // Step 8: Sort the received data locally
    sort recv_data

    // Here is where we end timing the Samplesort algorithm

    // print sorted data (optional)

    // Ending MPI commands (MPI_Finalize)
```

Radix Sort:

```
1. Initialize MPI, get rank, and size
INITIALIZE MPI(MPI_Init)
GET world_rank(MPI_Comm_rank)
GET world_size(MPI_Comm_size)

2. Generate different types of inputs listed in 2c.
    Create input arrays of different sizes etc
    Define length of data

3. Distribute data to processes with MPI_Scatter
    rank = 0:
        divide n by world size to calculate chunck size
        create arrays with chunk size and then fill
        send the chunks out to the processes
        MPI_Scatter(chunks)

4. Radix sort for worker
    -iterate through each digit:
        -each process performs local count sort:

        -initialize count array representing numbers 0-9
         array count[10] = {0};

        -Count occurence of each digit from local chunk
         Extract current digit with modulo (%)
```

```
           count[current digit]++;

          -Gather count data from each process and combine
           MPI_Allreduce(total_count)

          -Caluclate Cumulative count
          cumulative[i] = cumulative[i-1] + total_count[i]

          -Redistribute the elements based on the combined count
          Then place into processes depending on that calculation
          MPI_Alltoall

5. Use MPI_Gather to collect sorted data
    After processing all digits gather data
    MPI_Gather(sorted data)

6. Finalize MPI
MPI_Finalize()
```

2c. Evaluation plan - what and how will you measure and compare

- Input sizes, Input types:
  - The input will consist of multiple arrays of numerical data, with sizes increasing by powers of 2. These arrays will be tested on varying numbers of processors, which will also increase by powers of 2. Furthermore, there would be four main input types: sorted, nearly sorted, random, and reverse sorted.
  - These would be the inputs:
    - For input_size's:
      - 2^16, 2^18, 2^20, 2^22, 2^24, 2^26, 2^28
    - For input_type's:
      - Sorted, Random, Reverse sorted, 1%perturbed
    - MPI: num_procs:
      - 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024
- Strong scaling (same problem size, increasing number of processors/nodes)
  - We will keep the problem size fixed (e.g., a 2^18-sized array) and increase the number of processors/nodes. For each configuration, we will graph the computation time. This analysis will be done for all four input types: sorted, nearly sorted, random, and reverse sorted. Overall, there would be a total of 7 plots, one for each input size with 4 lines per plot representing each input type.
- Weak scaling (increase problem size, increase number of processors)
  - We will increase the problem size while also increasing the number of processors (e.g., 2, 4, 8, etc.). The computation time for each array size will be graphed across varying numbers of processors. This analysis will also be conducted for all four input types: sorted, nearly sorted, random, and reverse sorted. There would be four plots for each input type for weak scaling.

3a. Caliper instrumentation

Please use the caliper build `/scratch/group/csce435-f24/Caliper/caliper/share/cmake/caliper` (same as lab2 build.sh) to collect caliper files for each experiment you run.

Your Caliper annotations should result in the following calltree (use `Thicket.tree()` to see the calltree):

```
main
|_ data_init_X      # X = runtime OR io
|_ comm
|    |_ comm_small
|    |_ comm_large
|_ comp
|    |_ comp_small
|    |_ comp_large
|_ correctness_check
```

Required region annotations:

- `main` - top-level main function.
  - `data_init_X` - the function where input data is generated or read in from file. Use *data_init_runtime* if you are generating the data during the program, and *data_init_io* if you are reading the data from a file.
  - `correctness_check` - function for checking the correctness of the algorithm output (e.g., checking if the resulting data is sorted).
  - `comm` - All communication-related functions in your algorithm should be nested under the `comm` region.
    - Inside the `comm` region, you should create regions to indicate how much data you are communicating (i.e., `comm_small` if you are sending or broadcasting a few values, `comm_large` if you are sending all of your local values).
    - Notice that auxillary functions like MPI_init are not under here.
  - `comp` - All computation functions within your algorithm should be nested under the `comp` region.
    - Inside the `comp` region, you should create regions to indicate how much data you are computing on (i.e., `comp_small` if you are sorting a few values like the splitters, `comp_large` if you are sorting values in the array).
    - Notice that auxillary functions like data_init are not under here.
  - `MPI_X` - You will also see MPI regions in the calltree if using the appropriate MPI profiling configuration (see **Builds/**). Examples shown below.

All functions will be called from `main` and most will be grouped under either `comm` or `comp` regions, representing communication and computation, respectively. You should be timing as many significant functions in your code as possible. **Do not** time print statements or other insignificant operations that may skew the performance measurements.

Nesting Code Regions Example - all computation code regions should be nested in the "comp" parent code region as following:

```
CALI_MARK_BEGIN("comp");
CALI_MARK_BEGIN("comp_small");
sort_pivots(pivot_arr);
CALI_MARK_END("comp_small");
CALI_MARK_END("comp");
```

```
# Other non-computation code
...

CALI_MARK_BEGIN("comp");
CALI_MARK_BEGIN("comp_large");
sort_values(arr);
CALI_MARK_END("comp_large");
CALI_MARK_END("comp");
```

Calltree Example:

```
# MPI Mergesort
4.695 main
├─ 0.001 MPI_Comm_dup
├─ 0.000 MPI_Finalize
├─ 0.000 MPI_Finalized
├─ 0.000 MPI_Init
├─ 0.000 MPI_Initialized
├─ 2.599 comm
│  ├─ 2.572 MPI_Barrier
│  └─ 0.027 comm_large
│     ├─ 0.011 MPI_Gather
│     └─ 0.016 MPI_Scatter
├─ 0.910 comp
│  └─ 0.909 comp_large
├─ 0.201 data_init_runtime
└─ 0.440 correctness_check
```

I) Calltree For Merge Sort:

```
1.552 main
├─ 0.000 MPI_Init
├─ 0.006 data_init_runtime
├─ 0.899 comm
│  ├─ 0.432 comm_large
│  │  ├─ 0.431 MPI_Scatter
│  │  └─ 0.001 MPI_Gather
│  └─ 0.467 MPI_Barrier
├─ 0.011 comp
│  ├─ 0.006 comp_small
│  └─ 0.036 comp_large
├─ 0.000 MPI_Finalize
├─ 0.001 correctness_check
├─ 0.000 MPI_Initialized
├─ 0.000 MPI_Finalized
└─ 0.004 MPI_Comm_dup
```

II) Calltree For Sample Sort:

```
In [4]: print(tk.tree(metric_column="Avg time/rank"))
```

```
 _____ _     _       _          _
|_   _| |__ (_) ___ | | _____| |_
  | | | '_ \| |/ __| |/ / _ \ __|
  | | | | | | | (__|   <  __/ |_
  |_| |_| |_|_|\___|_|\_\___|\__|  v2024.1.0

0.652 main
├─ 0.000 MPI_Init
├─ 0.000 MPI_Comm
├─ 0.002 data_init_runtime
├─ 0.051 comp
│  ├─ 0.051 comp_small
│  └─ 0.000 comp_large
├─ 0.020 comm
│  ├─ 0.007 comm_large
│  │  ├─ 0.002 MPI_Gather
│  │  └─ 0.005 MPI_Gatherv
│  └─ 0.013 comm_small
│     ├─ 0.001 MPI_Bcast
│     ├─ 0.000 MPI_Alltoall
│     ├─ 0.001 MPI_Alltoallv
│     └─ 0.011 MPI_Gather
├─ 0.000 correctness_check
├─ 0.000 MPI_Finalize
├─ 0.000 MPI_Initialized
├─ 0.000 MPI_Finalized
└─ 0.175 MPI_Comm_dup

Legend (Metric: Avg time/rank Min: 0.00 Max: 0.65 indices: {'profile': 1806359314})
0.59 - 0.65
0.46 - 0.59
0.33 - 0.46
0.20 - 0.33
0.07 - 0.20
0.00 - 0.07

name User code    ◀ Only in left graph    ▶ Only in right graph
```

III) Calltree for Bitonic Sort:

```
  _____ _     _       _          _
 |_   _| |__ (_) ___ | | _____| |_
   | | | '_ \| |/ __| |/ / _ \ __|
   | | | | | | | (__|   <  __/ |_
   |_| |_| |_|_|\___|_|\_\___|\__|  v2024.1.0

4.466 main
├─ 0.000 MPI_Init
├─ 0.001 data_init_runtime
├─ 2.856 comm
│  ├─ 0.141 comm_large
│  │  ├─ 0.135 MPI_Scatter
│  │  └─ 0.006 MPI_Gather
│  ├─ 2.567 MPI_Barrier
│  └─ 0.147 comm_small
│     └─ 0.146 MPI_Sendrecv
├─ 0.002 comp
│  ├─ 0.000 comp_small
│  └─ 0.001 comp_large
├─ 0.000 MPI_Finalize
├─ 0.000 correctness_check
├─ 0.000 MPI_Initialized
├─ 0.000 MPI_Finalized
└─ 0.113 MPI_Comm_dup
```

3b. Collect Metadata

Have the following code in your programs to collect metadata:

```
adiak::init(NULL);
adiak::launchdate();    // launch date of the job
adiak::libraries();     // Libraries used
adiak::cmdline();       // Command line used to launch the job
adiak::clustername();   // Name of the cluster
adiak::value("algorithm", algorithm); // The name of the algorithm you are using
(e.g., "merge", "bitonic")
adiak::value("programming_model", programming_model); // e.g. "mpi"
adiak::value("data_type", data_type); // The datatype of input elements (e.g.,
double, int, float)
adiak::value("size_of_data_type", size_of_data_type); // sizeof(datatype) of input
elements in bytes (e.g., 1, 2, 4)
adiak::value("input_size", input_size); // The number of elements in input dataset
(1000)
adiak::value("input_type", input_type); // For sorting, this would be choices:
("Sorted", "ReverseSorted", "Random", "1_perc_perturbed")
adiak::value("num_procs", num_procs); // The number of processors (MPI ranks)
adiak::value("scalability", scalability); // The scalability of your algorithm.
choices: ("strong", "weak")
adiak::value("group_num", group_number); // The number of your group (integer,
e.g., 1, 10)
adiak::value("implementation_source", implementation_source); // Where you got the
source code of your algorithm. choices: ("online", "ai", "handwritten").
```

They will show up in the `Thicket.metadata` if the caliper file is read into Thicket.

See the `Builds/` directory to find the correct Caliper configurations to get the performance metrics. They will show up in the `Thicket.dataframe` when the Caliper file is read into Thicket.

We have included the metadata code mentioned in 3b in our algorithms. The values that we are getting in the metadata are the following:

- Launch date of the job
- Libraries used in our algorithm
- Name of the cluster
- Name of the algorithm you are using
- The programming model used for communication which in our case is MPI
- The datatype of input elements, which in our case is int
- The sizeof(datatype) of input(int) elements in bytes
- The number of elements in the input dataset/n - 2^16, 2^18, 2^20, 2^22, 2^24, 2^26, 2^28
- The input type used for sorting("Sorted", "ReverseSorted", "Random", "1_perc_perturbed").
- The number of processors used(MPI ranks) - 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024
- The scalability of your algorithm. Either if we are doing strong or weak
- Our group number, which is 13
- The place where we got the source code of your algorithm

4. Performance evaluation

Include detailed analysis of computation performance, communication performance. Include figures and explanation of your analysis.

4a. Vary the following parameters

For input_size's:

- 2^16, 2^18, 2^20, 2^22, 2^24, 2^26, 2^28

For input_type's:

- Sorted, Random, Reverse sorted, 1%perturbed

MPI: num_procs:

- 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024

This should result in 4x7x10=280 Caliper files for your MPI experiments.

4b. Hints for performance analysis

To automate running a set of experiments, parameterize your program.

- input_type: "Sorted" could generate a sorted input to pass into your algorithms
- algorithm: You can have a switch statement that calls the different algorithms and sets the Adiak variables accordingly

- num_procs: How many MPI ranks you are using

When your program works with these parameters, you can write a shell script that will run a for loop over the parameters above (e.g., on 64 processors, perform runs that invoke algorithm2 for Sorted, ReverseSorted, and Random data).

4c. You should measure the following performance metrics

- `Time`
  - Min time/rank
  - Max time/rank
  - Avg time/rank
  - Total time
  - Variance time/rank

5. Presentation

Plots for the presentation should be as follows:

- For each implementation:
  - For each of comp_large, comm, and main:
    - Strong scaling plots for each input_size with lines for input_type (7 plots - 4 lines each)
    - Strong scaling speedup plot for each input_type (4 plots)
    - Weak scaling plots for each input_type (4 plots)

Analyze these plots and choose a subset to present and explain in your presentation.

6. Final Report

Submit a zip named `TeamX.zip` where `X` is your team number. The zip should contain the following files:

- Algorithms: Directory of source code of your algorithms.
- Data: All `.cali` files used to generate the plots seperated by algorithm/implementation.
- Jupyter notebook: The Jupyter notebook(s) used to generate the plots for the report.
- Report.md