Progress of Thesis Work

and discussion on next steps

Chengxin Ma

Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

September 20, 2019

Table of Contents

Table of Contents

Scope of the Thesis Work

The scope of the thesis work is the sorting stage of the genomic data processing pipeline.

- Input: SAM file(s) (output of the alignment stage, containing mapping between reads and reference genome)
- Output: BAM file(s) (containing sorted reads according to coordinates)

Goal of the Thesis Work

The goal of the thesis work is to improve the performance of the sorting stage, especially when the volume of data is of large scale, which is beyond the capability of a single computing node.

Evaluation Criteria

- Scalability
- Performance
- Fault tolerance
- Which has the highest priority?

Challenges

A cluster of multiple nodes needs to be designed to handle the sorting task. This brings the following challenges:

- How to coordinate the nodes? (Master/slave architecture or P2P architecture?)
- How to enable fast and reliable communication among the nodes?
- On each node, how to accelerate the data processing?

Table of Contents

Architecture

Master/slave.
One node acts as the coordinator, and the rest acts as the slaves.
Note that complexity of communication among the nodes is $O(N^2)$.

Load balancing

- In *SparkGA*: *static* LB based on chromosome sizes and further *dynamic* LB based on number of reads
- In *FuxiSort* (2015 winner of Sort Benchmark), they sample the input to determine the boundaries of range partition
- Experiments need to be performed to see which is the most suitable LB technique in our case

Applicable Techniques & Technologies

For fast communication among the nodes: RDMA
For improving performance on each node:

- use multithreading (native in picard SortSam, click here to read source code)
- improve I/O performance of large files by memory-mapped file or Arrow

Table of Contents

Next Steps

Comments and suggestions?