Advance in analysis of Whole Genome Sequencing in the field of HPC

Jarlier F [1, 2, 3,6], Fedy N [5], Sirotti L[5], Magalahes T [5], Paganiban P [5], Deloger M [1,2,3,6], Hupé P [1, 2, 3, 4, 6]

(1) Institut Curie, Paris France, (2) INSERM U900, Paris France, (3) Mines ParisTech, Paris France, (4) CNRS UMR144, Paris France, (5) Université Paris V, Paris France, (6) PSL Research University, F-75005 Paris, France.

Motivation:

The Next Generation Sequencing (NGS) technology offers new insights in cancer research and personalized medicine. Due to its large scale approach, we can detect genetic alterations with an unprecedent accuracy. Due to the decreasing cost of sequencing, whole genome sequencing becomes more widely used in research project. In a near future, it will likely become a tool for daily clinical practice. The drawbacks of such breakthrough are the volume of generated data and also the complexity of the downstream analysis. For instance, a whole genome sequencing with a 40X coverage is 400GB and a variant analysis will last for a week with current pipelines.

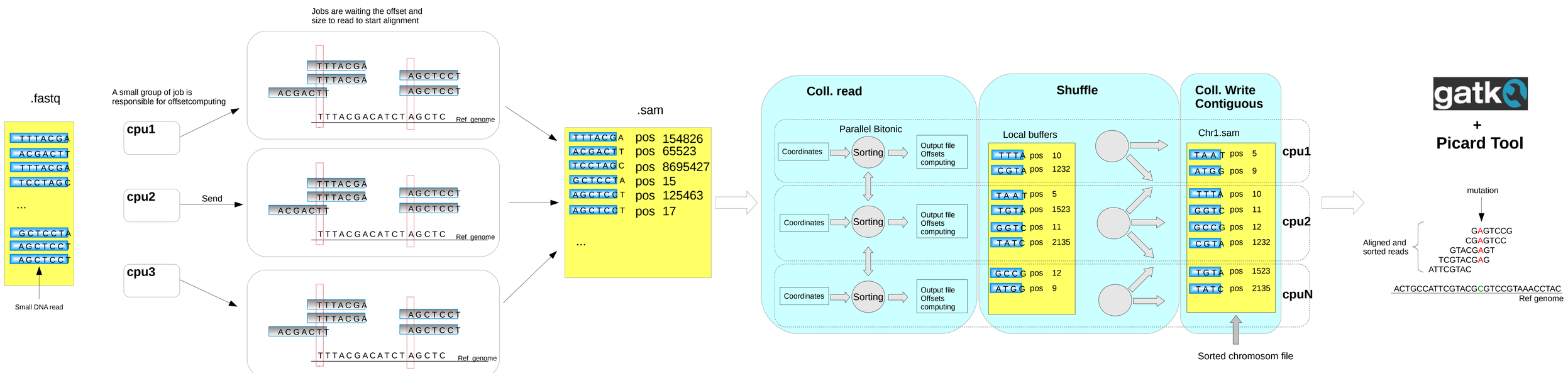
To tackle traditionnal bottlenecks, we have used the parallelization with message passing interface paradigm called MPI. This paradigm has many advantages : it transfers input-output (IO) file system latencies at network level ; MPI also provides many optimizations such as collective operations to optimize IO ; it provides communication between jobs ; it also avoids copy of data thanks to derived datatype.

Early stages of NGS pipeline consists of two steps. The first step is the alignment of small nucleotide sequences (called reads) on a referenced genome. The second step is the sorting of the alignment result according to chromosome and genomic position. These operations are essential but time consuming. Therefore we propose to optimize the two steps using MPI technology. In the first part we present an overview of the parallelized workflow that we have implemented and then we show the results we have obtained on whole human genomes an exoms samples.

1) Description of the MPI workflow

Every NGS pipeline starts with the two following operations: the alignment and the sorting. The alignment consists in finding the position of small fragments of DNA produced by the sequencers. Number of reads at a particular position is called the coverage of a sample. The deeper the coverage the better the reconstructed sequence is. Nowadays a standard coverage is around 30X, 40X and even 100X. For instance, a whole genome sequencing with 100X coverage produces 1 TB of data after alignment.

After intensive study of the alignment and sorting algorithms we have noticed that a major part of bottlenecks are in the IO file access where all the latency is. To solve that problem we have decided to transfer IO constraints to the network level. Other optimizations have been implemented such as : collective operation, shared memory, indexing, derived datatype. It turns out that these optimizations have reduced drastically the time and the memory consumption.



Step 1 :
- Offset computing

Step 2 :
- Alignment (BWA MEM)
- Shared memory

Step 3 :
- MPI Collective write

Step 4 :
- Collective read and bufferisation
- Indexing
- Parallel Bitonic sort of position

Step 5 :
- Dispatch of the reads Bruck algorithm (AlltoAll)

Step 6 :
- MPI Collective write

Step 7 :
- Variant calling

Overview of the alignment algorithm

Overview of the sorting algorithm

Variant calling (Non MPI)

2) Description of the COBALT architecture at TGCC

For fast reading and writing, collective operations are mandatory but these optimizations can only be achieved upon a custom distributed file system.

COBALT configuration

CPUs: 2x 14-cores Intel Broadwell@2.4GHz (AVX2)

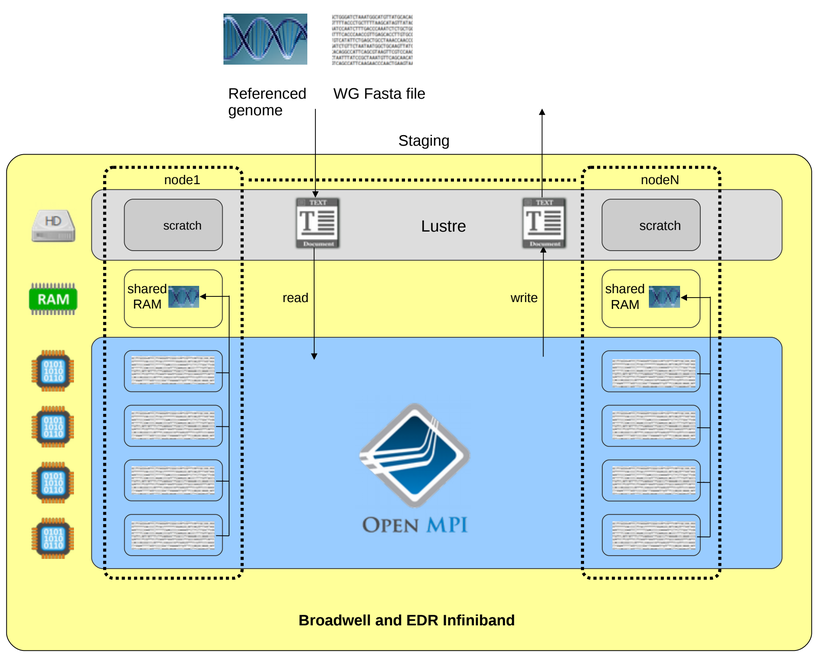
Cores/Node: 28

Cores: 39816

RAM/Core: 4,5GB

Interconnect : EDR InfiniBand 100Gb

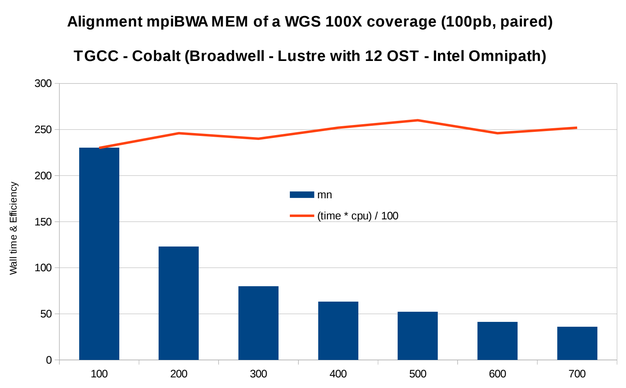
Lustre: 2.7.19.9



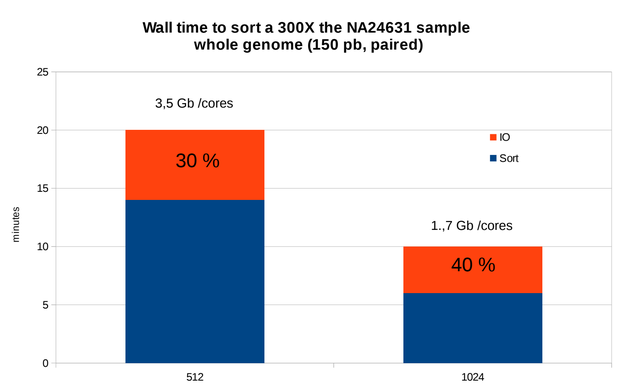
Cluster architecture at TGCC (Très Grand Centre de Calcul) CEA – Bruyères-Le-Châtel (France)

3) Results

In this section we present results we have obtained on TGCC cluster (CEA, Bruyères-Le-Châtel, France).



Wall time for aligning 100X human whole genome.
1.2 billion reads occupy 700GB on the disk. The efficiency is due to transfert of IO bounds to network bounds.



300X SAM files contains 6,5 billion reads and is 1,3TB large.
The total memory usage is 1.7 TB.

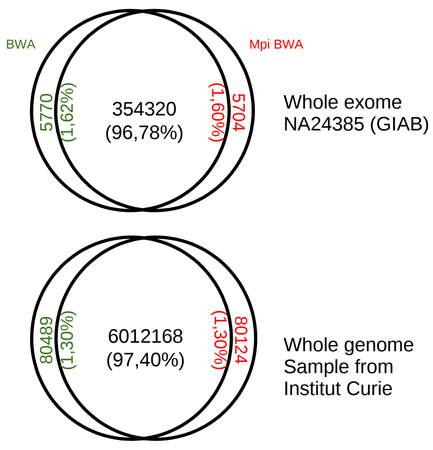
MPI effect on reproducibility of the alignment

	BWA t10_run1	BWA t10_run2	BWA t10_run3	BWA t20	BWA t30	mpiBWA 1node_t1_n10	mpiBWA 2nodes_t1_n10	mpiBWA 3nodes_t1_n10	mpiBWA 3nodes_t10_n30
BWA t10_run1	100 %	100	100	99,9	99,9	96,7	96,7	96,7	96,7
BWA t10_run2		100	100	99,9	99,9	96,7	96,7	96,7	96,7
BWA t10_run3			100	99,9	99,9	96,7	96,7	96,7	96,7
BWA t20				100	99,9	96,7	96,7	96,7	96,7
BWA t30					100	96,7	96,7	96,7	96,7
MpiBWA 1node_t1_n10						100	100	100	96,8
MpiBWA 2nodes_t1_n10							100	100	96,8
MpiBWA 3nodes_t1_n10								100	96,8
MpiBWA 3nodes_t10_n30									100

(t = threads, n = cores)

The drift is due to the randomization when multi-hits happens

MPI effect on Variant calling



3) Conclusion

Our MPI program is able to tackle current bottleneck of whole genome sequencing analysis pipeline. In this study we have decided to use parallelization for the alignment and the sorting of NGS data. We have drastically reduced the alignement time and even allows the sorting of reads that was not feasible before. From our results MPI technology is an efficient candidate and performs very well on the cluster architecture we have tested.

The cpu, memory usage, the network and file systems are crucial for a good scalability. MPI adresses efficiently these different aspects as we have shown here. MPI among other optimization technics will definitely help bioinformatics developpers to cross the barrier of the Big Data.

References : www.open-mpi.org ; Fast and accurate short read alignment with Burrows–Wheeler transform (Li H. et al. 2009) ; The Sequence Alignment/Map format and SAMtools (Li H et al. 2009) ; Bruck et al. (1997) Efficient Algorithms for All-to-All Communications in Multiport Message-Passing Systems. EEE Transactions on Parallel and Distributed Systems, 8(11):1143–1156, 1997.

Acknowledgements : We thank François Prud'homme, Maxime Chevailliot, Eric Viara from the Institut Curie IT. We thank Claude Scarpelli for providing us with TGCC cluster access. Many thanks to Gabriele Pacuicci and Kristina Kermanshahche form Intel Company.