The Seal suite of distributed software for high-throughput sequencing

Luca Pireddu, Simone Leo, Gianluigi Zanetti
CRS4, Polaris, Ed. 1, Pula, Italy
http://www.bioinfo.no/
http://www.uni.no/computing/units/cbu

Modern DNA sequencing machines have opened the flood gates of whole genome data; and the current processing techniques are being washed away. Medium- sized sequencing laboratories can produce Terabytes of data per week that need processing. Unfortunately, most programs available for sequence processing are not designed to scale easily to such high data rates, nor are the typical bioinformatics workflow designs. As a consequence, many sequencing operations are left struggling to cope with the high data loads, often hoping that acquiring additional hardware will solve their problems. In contrast, we believe that a change in paradigm is required to solve this problem: a shift to highly parallelized software is required the handle the parallelization that has taken place in sequencing.

In response to the growing processing requirements of the CRS4 Sequencing and Genotyping Platform (CSGP), which now houses 4 Illumina HiSeq 2000 sequencers for a total capacity of about 7000 Gbases/month, we began the development of Seal [3], a new suite of sequence processing tools based on the MapReduce [1] programming model that run on the Hadoop framework. Seal aims to replace many of the tools that are customarily used in sequencing workflows with Hadoop-based, scalable alternatives. Currently, Seal provides distributed MapReduce tools for: demultiplexing tagged reads, mapping reads to a reference (it includes a distributed version of the BWA aligner [2]), and sorting reads by alignment position. In the near future we will also be adding tools for read quality recalibration.

Seal tools have been shown to scale well in the amount of input data and the amount of computational nodes available [4]; therefore, with Seal one can increase processing throughput by simply adding more computing nodes. Moreover, thanks to the robust platform provided by Hadoop, the effort required by operators to run the analyses on a large cluster is generally reduced, since Hadoop transparently handles most hardware and transient network problems, and provides a friendly web interface to monitor job progress and logs. Finally, the Hadoop Distributed File System (HDFS) provides a scalable storage system that scales its total throughput hand in hand with the number of processing nodes. Thus, it avoids creating a bottleneck at the shared storage volume and avoids the need for an expensive high-performance parallel storage device.

Seal is currently in production use at the CRS4 Sequencing and Genotyping Platform and is being evaluated at other various sequencing centers.

References

1. J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. In OSDI '04: 6th Symposium on Operating Systems Design and Implementation, 2004.
2. Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25(14):1754—1760, 2009.
3. Luca Pireddu, Simone Leo, and Gianluigi Zanetti. Mapreducing a genomic sequencing workflow. In Proceedings of the second international workshop on MapReduce and its applications, MapReduce '11, pages 67–74, New York, NY, USA, 2011. ACM.
4. Luca Pireddu, Simone Leo, and Gianluigi Zanetti. Seal: a distributed short read mapping and duplicate removal tool. Bioinformatics, 27(15):2159–2160, 2011.

Relevant Web sites

5. http://biodoop-seal.sourceforge.net/
6. http://hadoop.apache.org/