Abstract Preview

[redacted]

Scaling with the flow: advantages of a MapReduce-based scalable and high-throughput sequencing workflow *L. Pireddu[1], S. Leo[1], F. Reinier[2], R. Berutti[2,3], R. Atzeni[2], G. Zanetti[1]*. 1) Distributed Computing Group, CRS4, Pula, CA, Italy; 2) Advanced Genomics Computing Technology, CRS4, Pula, CA, Italy; 3) Dept. of Biomedical Sciences, University of Sassari, Sassari, Italy.

Increasing sequence data rates can present serious problems to a growing sequencing platform, especially in its processing workflow and IT infrastructure. We discuss the advantages that the CRS4 Sequencing and Genotyping Platform (CSGP) gained by migrating its production process from its conventional workflow to its new one based on Seal (http://biodoop-seal.sf.net) and Hadoop. CSGP uses a typical BWA-based workflow that demultiplexes samples, maps reads to a reference, removes PCR duplicates and recalibrates base qualities. The original pipeline used common tools (BWA, Samtools, Picard, GATK) and parallelized computation through concurrent jobs, using a centralized file system to share data. This implementation showed weaknesses as the workload increased: low parallelism; I/O bottleneck at central storage; failure of entire analyses due to node failures or transient cluster problems. Thanks to the open-source resources currently available we developed a new custom distributed workflow that solves these problems. The new workflow is based on the open-source Seal suite, which provides a set of tools (including a distributed BWA aligner) that run on the Hadoop MapReduce framework, leveraging its functionality for genomic sequencing applications. By switching to a Seal-based workflow we have acquired computational scalability out-of-the-box. Therefore, we can now easily meet the demands imposed by the growing sequencing platform (now operating 6 Illumina sequencers) by adding more computing nodes. In addition, the much-increased parallelism has improved overall computational throughput by taking advantage of all available computing power. Moreover, the effort required by our operators to run the analyses has been reduced, since Hadoop transparently handles most hardware and transient network problems and provides a friendly web interface to monitor job progress and logs. Finally, we eliminated the need for our expensive shared parallel storage devices. Our tests reveal that Seal is efficient, achieving close to 70% of the theoretical maximum throughput per node (measured with a single-node version of the workflow on a small data set) and scales linearly at least up to 128 nodes. In summary, this case study suggests that the MapReduce programming model, Seal and Hadoop are enabling technologies that provide considerable benefits in the genomic sequencing domain. Seal now includes our new workflow as a downloadable sample application.

[redacted]