Distributed DNA alignment, a stream based approach

Wolf Beat
Department of Computer Science
University of applied sciences Fribourg

January 22, 2013

1 Introduction

The field of bioinformatics research becomes becomes more and more popular. Thanks to recent advances in techniques of DNA sequencing, a growing number of genetic data is digitalized and analyzed. The collected data is used for diverse purposes, such as diagnostics of genetic diseases, or the mapping of the evolutionary tree of different species. The increasing interest in this technology allowed to decrease the time spent on sequencing a single human genome from about the 13 years achieved by the Human Genome project [5], to as low as 6 hours today. This speed increase has been achieved in only 10 years, which is rarely seen in any technical domain. As a comparison, it is notably faster than the Moore's law, to which computer science is subject to. However this performance improvement created severe processing power problems, which could only partially be resolved by algorithmic advances.

2 DNA sequencing

To understand the problem that needs to be solved, we have to understand the kind of data that is generated by the DNA sequencing and how it is processed. The DNA consists in several chromosomes, which are long sequences of nucleotide bases. The four nucleotide bases are labelled A, C, T and G. The human DNA consists of over 3 billion of nucleotide base pairs. The bases encode the genes of an human organism, and the order of those bases is the information searched by the DNA sequencing process. This process cannot be achieved, for a given human DNA, in one shot. Instead, the DNA is read in many small pieces of 20 to few thousand bases long. These sequences are then aligned against a reference sequence, to find the most likely place each sequence was in the original DNA. It has to be noted that the sequences are rarely identical to the reference sequence and there are multiple regions in the reference sequence which are identical making the alignement process a complex task. Eventually it is the differences in the sequenced DNA compared to the reference sequence which are of interest because they can indicate a disease causing mutation.

3 Distributed alignment

To overcome the increasing gap between the DNA sequencing speed and the speed of the alignment process, new computing approaches need to be explored. As the millions of sequences that need to be aligned are independent tasks, a possible approach is to realize the alignment process on a massively parallel computing infrastructure. Many different distributed architectures have been used to solve this problem, from a simple single multicore machine, to the use of graphics cards (GPUs) through networks of thousands of processing units (Grid). Recently the concept of cloud computing has became a very popular trend. While it removes the burden of individual laboratories to maintain a powerful computing infrastructure, it also raises new questions and problems. One of the major problem arising with the different cloud based solutions for DNA sequence alignment, is the data transfer time from the local network to the cloud, and back again. Tools like Cloudburst [2] or Crossbow [6] need that the data to be analyzed is entirely transfered to the cloud prior to start the computation. This leads to a time overhead of up to several hours before any calculation can be started. In two recent publications, the ETHZ ([3], [4]), has investigated the possibility of using a

stream processing approach to overcome the problem of the data transfer time overhead. With this approach, calculations can be started as soon as the first data reaches the computing infrastructure, the data is processed sequence by sequence as and when they arrive.

4 Streaming approach validation

The publication of the ETHZ was based on technologies, that were not necessarily meant to be used in a cloud or streaming based environment. Data conversions had to be made between several incompatible applications which is not good for the overall performance. To validate the concept of streamed alignment, the GensearchNGS aligner, which was created during the master thesis "Analysis and visualization of DNA sequences using cloud computing" [1] and in collaboration with Phenosystems SA, was ported to use a local area network grid through the use of the RMI technology. To test the implementation, a Core 2 Duo laptop clocked at 2.5 GHz was connected to a Phenom X6 1090T clocked at 3.2GHz over a 1Gb/s switch. The laptop has 2 CPU cores and the Phenom X6 has 6 cores. A small dataset of 50k sequences was aligned against the chromosome 1 which is 247 Mega bases long.

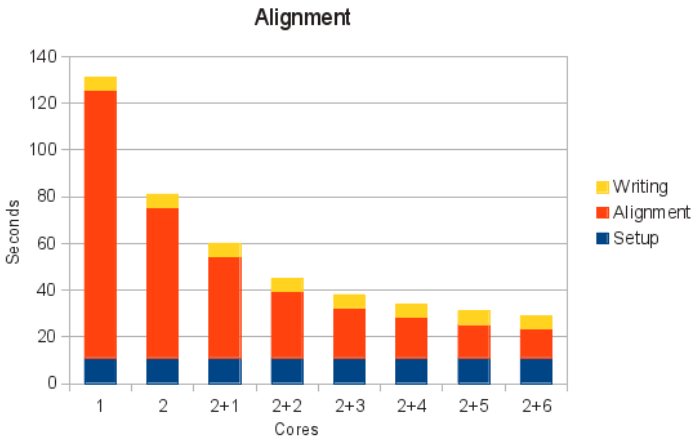

Figure 1: Time needed to align 50k sequences

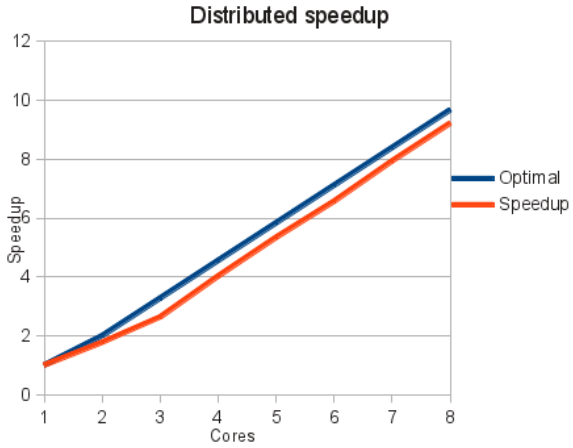

Figure 2: Adjusted speedup over two computers

The results are promising. Even using a technology like RMI that might not be best adapted for the job, a linear speedup was achieved (Figure 2). The main computer for the benchmark was the laptop, the second computer joined as soon as 3 threads were used. The speedup values were adjusted to the power of the processors.

5 Future work

Now that the concept has been validated, more flexible solutions will be explored. Streaming frameworks like DSPE [7], will be looked at because they remove the need to handle implementation complexity of a streaming algorithm and allow the deployment on more diverse architectures. Also looked at will be an implementation of the algorithm using the POP [8] technology, which will enable nearly effortless distributed calculations. Using those technologies will require some improvements. DSPE was developed to run on multicore and GPU accelerated systems. A project is currently beeing done that combines POP-C++ and DSPE, allowing DSPE applications to run on a POP-C++ powered grid. Further work will need to be done to bring this environment to the cloud.

References

[1] B. Wolf, Analysis and visualization of DNA sequences using cloud computing, 2011

[2] M. C. Schatz, CloudBurst: Highly Sensitive Read Mapping with MapReduce, Bioinformatics, vol. 25, no. 11, 2009.

[3] R. Kienzler, A. Ranganathan, and N. Tatbul, Large-scale DNA Sequence Analysis in the Cloud: A Stream-based Approach, 2011.

[4] R. Kienzler, R. Bruggmann, and A. Ranganathan, Stream As You Go: The Case for Incremental Data Access and Processing in the Cloud, 2012.

[5] Human genome project
http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

[6] Crossbow
http://bowtie-bio.sourceforge.net/crossbow/

[7] Domain specific language for parallel real-time stream processing
http://www.systemdesigner.ch/?page_id=31

[8] POP-C++ and POP-Java
http://gridgroup.hefr.ch/popc/doku.php/main_page