Massively parallel sequence alignment using PCJ-BLAST

Marek Nowicki, Piotr Bała

With the development of Next Generation Sequencing there is a high demand in the area of computational biology to extract useful information from the databases. The most popular technique is sequence alignment based on the comparison of the DNA/RNA sequence with the ones of the known function stored in the protein or nucleotide databases. There are numerous algorithms and software packages used for these purposes. The most widely used in BLAST algorithm and its implementation provided by NCBI. Unfortunately, this implementation is limited to the single node execution which leads to long processing time which is not acceptable if used to make a diagnosis and decide on treatment. For these purposes, the analysis should be finished in hours, not days or weeks. Therefore there is growing interest in running blast on the large clusters or supercomputers. For the clusters, users would like to use new filesystems such as HDFS since it is often used to store and process genetic data.

There have been several approaches to parallelizing blast, usually trying to benefit from the parallel processing of the different queries. The example implementations are for example pyblast, mpiBLAST, pioBLAST, dedicated Blue Gene implementation or commercial one from Paracel. All of them have disadvantages which prevent efficient usage by the bioinformatics community. Most of the mentioned solutions is based on the static distribution of the input sequence and reference database. This can be done easily, but each read is processed in the different time which leads to pure load-balancing.

In this contribution, we present massively parallel execution of the blast algorithm on thousands of processors, available on supercomputers and HPC clusters. Our work is based on the optimal splitting up the set of queries running with the non-modified NCBI-blast package for sequence alignment which makes our solution flexible and extensible. The work distribution and search management are implemented in Java with the PCJ (Parallel Computing in Java) library. The PCJ-BLAST package is responsible for reading input sequence, splitting it up and start multiple NCBI-blast executables. Our implementation has mechanisms to achieve good load balancing which significantly improves scalability and reduces search time. The load balancing is dynamic and does not require previous knowledge of the time for processing individual chunks.

Our implementation uses a single instance of the searched database which significantly reduces administration overheads. This approach requires multiple reads of the database which, in the case of multimode execution becomes a bottleneck. To solve this problem we have investigated a problem of parallel I/O delivering a fast, high throughput execution of BLAST. New developments include features requested by users such as possibility to read query string from the HDFS file system. Moreover, PCJ-BLAST can be run as Spark application with minimal overhead, which allows to utilize computational power provided by the Hadoop/Spark clusters.

We have tested our solution on Cray XC40 with Aries interconnect, x86 cluster with Infiniband and Hadoop cluster. Our solution is at least 3 times faster than others and shows very good performance and efficiency up to thousands of cores. In result, we have significantly reduced time required for sequence analysis making it feasible for personalized medicine. We have also proved that PCJ library can be used as an effective tool for fast development of the scalable applications implemented in Java.

Short CV of Piotr Bała

Prof. Piotr Bała graduated in physics in 1988 and received a PhD in physics in 1993 at N. Copernicus University (Torun, Poland). He has been studying quantum effects is molecular system using high performance computing (HPC). Since 2000 he is a leader of the team at ICM University of Warsaw which developed grid tools for molecular biology. He was strongly involved in national and European grid projects, he is member of UNICORE Forum. The main focus of current research is on development of new methods for parallel and distributing computing. In particular he is leader of the team developing PCJ library for parallel computing in Java and coordinator of the HPDCJ project (CHIST-ERA). The PCJ library has received HPCC award at SC'14. Piotr Bała is an author and co-author of more than 130 scientific papers.

Full name: Piotr Bała

Job title: professor

Department: ICM

Organization: University of Warsaw

Short CV of Marek Nowicki

Dr Marek Nowicki graduated from computer science at the Nicolaus Copernicus University in Toruń, Poland. He entered a prestigious country wide program for the best Ph.D. students in math and computer science. He is also a winner of the IBM Great Minds student internships in 2013. He finished Ph.D. studies in 2014 and defended at the University of Warsaw thesis on the New programming methods for parallel programming in Java based on the PGAS (Partitioned Global Address Space) paradigm. Currently, he is working at the Faculty of Mathematics and Computer Science, Nicolaus Copernicus University in Toruń, Poland. The main research interest is on parallel computing in Java. He is involved in the HPDCJ project and takes care on the development of the PCJ library.

Full name: Marek Nowicki

Job title: adiunkt (assistant professor)

Department: Faculty of Mathematics and Computer Science

Organization: Nicloaus Copernicus University in Toruń