GenArch: High Performance Computer Architectures for Large-Scale Population Genome Data Analysis

Description

The potential of precision medicine is tremendous in the transition towards future digital healthcare system. Genome data analysis is a critical component in the success of precision medicine. Sequencing and analysing the whole genome of a patient will allow developing personalised therapies, anticipating health problems and enabling preventive therapies. Intensive international research is promoting population-wide sequencing iniciatives. With the advent of second-generation sequencing systems, the amount of genomics data available to researchers has doubled every 7 months, a much faster rate than computation power. If this trend continues, in the next decade we will be able to sequence billions of whole human genomes every year, generating exabytes of raw genomics data and enabling large-scale population genomics analysis. Furthermore, emergent third-generation technologies are expected to be widely used in the near future, producing 10x-100x longersequences, further increasing throughput and reducing costs.

However, genomics data requires complex analyses that are extremely compute and data-intensive. A single whole-genome analysis typically requires hundreds of CPU hours of computation, suffering from very inefficient execution due to poor memory utilization. As a result, the performance bottleneck in genome analysis is quickly moving from the sequencing side (as it used to be for the past decade) towards the computing side. This poses a huge compute challenge for genome data analysis applications. Simply relying on new and faster processors will not be enough to make precision medicine based on genome data analysis a reality, especially now that we areclose to the end of Moores Law. Thus, a novel approach is required to close the gap between genomics data production and current computing power. Recently, the research team introduced the novel WaveFront Alignment (WFA) algorithm for accelerated sequence comparison, which delivers significant speedups over classical methods and better scalability with the increasingly longer sequences.

Barcelona Supercomputing Center - Centro Nacional de Supercomputación