

Using cloud computing for parallel analysis of genome-wide datasets

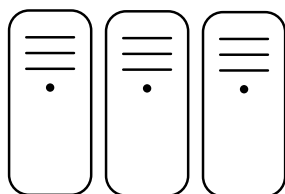
Introduction

Analysis of today's genome-wide datasets poses ever-increasing demands for computational capacity. For example, imputation, empirical p-values, epistatic analysis or QTL analysis of a large number of traits may easily take days or even months on a single computer.

For many analysis tasks, a linear increase in performance can be achieved by parallelization – partitioning the data by markers, subjects, or traits. Parallelization requires utilizing multiple calculation servers. However, acquiring and maintaining a computational cluster is expensive. As an alternative to costly calculation clusters, the emergence of cloud computing offers a new way of acquiring more computational power without initial investments and maintenance costs. With cloud computing, virtual calculation nodes can be added or removed on demand.

Local cluster/GRID

- + Very high bandwidth
- High initial investment
- Maintenance
- Fixed capacity



Cloud computing

- + Very flexible and scalable
- + Ecological
- Relatively low bandwidth

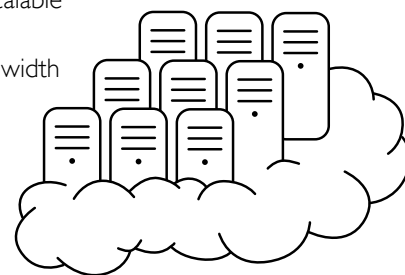


Figure 1: Benefits and disadvantages of local clusters vs. cloud computing

Overhead costs in parallelization

Even though parallelization sounds like a good solution to the question of how to increase performance in data analysis, some obstacles need to be resolved. One of the biggest obstacles in parallelization performance is the overhead cost generated by data preparation, data transfer and data analysis input file creation.

Even though cloud computing offers an inexpensive way to rent calculation capacity on demand, it poses some challenges. Compared to locally installed clusters, data transport speeds to a remote cloud environment are typically much slower and there is some initial overhead of installing the necessary software for each new node.

Experimental analysis

We performed an experimental analysis to measure the importance of minimizing the overhead costs, and to analyze

feasibility of cloud computing for distributed genetic analyses; in particular genome-wide quantitative trait (QTL) analysis of a very large number of traits using PLINK¹.

In our experimental analysis we used the following analysis setting:

Genome-wide multi QTL analysis using PLINK¹

- 960 quantitative traits
- 2.5 million imputed markers
- 1000 subjects
- 2 covariates

We tested the following server configurations

- PLINK¹ on command line - one processor core
- Two local 4-core calculation servers
- 1, 2, 4, 5, 10 and 20 4-core servers in EC2, 3.5 GB RAM/core

For our results, we measured the parallelization efficiency, time and costs.

Cloud environment

For the calculations we used Amazon EC2² cloud computing service. Amazon has secure server farms both in Europe and the US.

BC|SNPmax – infrastructure for parallelization

In our experimental analysis, we used Biocomputing Platforms' BC|SNPmax system to carry out the experiment and as the infrastructure for parallelization process. The system provides an easy-to-configure queue system, which allows connecting an arbitrary amount of calculation nodes with little overhead and which can automatically distribute analyses over all the available nodes and collect the results to a central repository.

Techniques to minimize overhead costs in BC|SNPmax

To minimize the overhead costs, we used the following techniques, in-built in the BC|SNPmax platform:

- Main database server only feeds the cloud and distributes all other work to the calculation servers
- Use of very high data compression rate in data transfer (BCD data format)
- Caching the transferred compressed binary file in the cloud
- Generation of the input-files for the analysis programs in the cloud

Results

The following graph presents parallelization performance index PPI. We have calculated PPI in the following way:

$$\text{PPI} = \frac{\text{Time of analysis with single core}}{\text{Total time of analysis} * \# \text{ of cores used}}$$

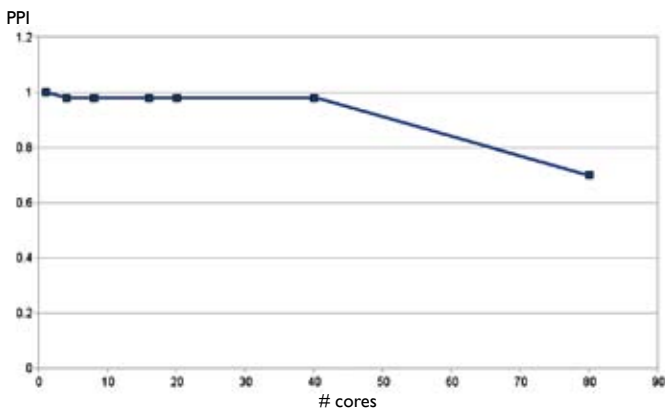


Figure 2: PPI

Our results indicate that when overhead costs are low, like when using BC|SNPmax database platform, the PPI remains unchanged when running the analysis from 4 to 40 cores. However, with 80 cores PPI drops dramatically. This is because

the example task was divided to 80 segments, losing the effect of the caching system, resulting in cores waiting for the data and thus causing the dramatic drop in the PPI.

Costs of using parallel calculation cores

The following graph shows that there is only around \$50 difference in the total analysis costs when using 4 cores or 40 cores. This cost difference comes from the Amazon EC2 hourly billing policy. When using 20-40 cores, the entire hour was not needed to perform analysis.

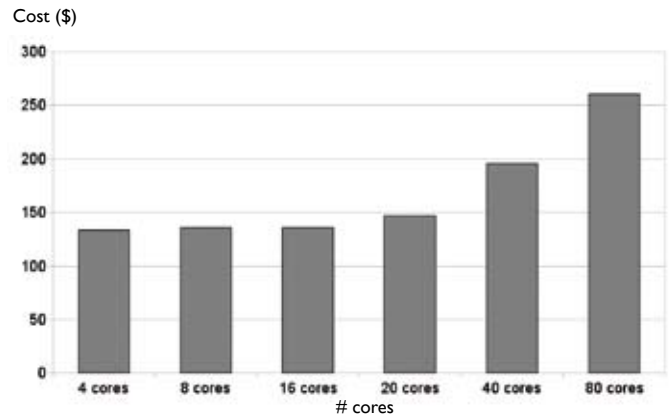


Figure 3: Costs of adding more cores

Conclusions

Our results indicate that distributing analysis to multiple servers clearly improves performance of many types of complex genetic analyses. However, our results also underline the importance of the efficient and robust parallelization infrastructure, such as the BC|SNPmax platform, in achieving good parallelization performance by minimizing overhead costs.

Cloud computing enables researchers to access a large number of calculation servers without significant investment. Our findings indicate that cloud computing offers major cost savings compared to a local cluster, especially when analysis needs are sporadic.

Additional information

For more information, please visit www.bcplatforms.com or contact us by email at info@bcplatforms.com

References:

- 1) Purcell S. et al. (2007), PLINK: a toolset for whole-genome association and population-based linkage analysis, *American journal of Human Genetics* 81
- 2) <http://aws.amazon.com/ec2/>

BC Platforms is an IBM® Advanced Business Partner™, an Affymetrix GeneChip-compatible™ company and belongs to the Illumina® Connect™ partnership program.

Contact us

Headquarters
Biocomputing Platforms Ltd
Innopoli 2, Tekniikantie 14,
FI-02150 Espoo, Finland
Tel +358 9 2517 7340
Fax +358 10 296 1288

US office (Silicon Valley)
Biocomputing Platforms Ltd
560 S. Winchester Blvd., Suite 500
San Jose, CA 95128
Tel (408) 572-5570
Fax (408) 572-5571

Email and web
For general enquiries:
info@bcplatforms.com
Personnel:
firstname.lastname@bcplatforms.com
www.bcplatforms.com

BC Platforms