A GPU-based High Performance Computing Infrastructure for specialized NGS analyses.

Andrea Manconi[1], Marco Moscatelli[1], Matteo Gnocchi[1], Giuliano Armano[2], Luciano Milanesi[1]
[1] Institute for Biomedical Technologies, National Research Council, Segrate (MI), Italy
[2] Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari (CA), Italy

Corresponding Author:
Andrea Manconi[1]
Via F.lli Cervi 93, Segrate, MI, 20090, Italy
Email address: andrea.manconi@itb.cnr.it

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2175v1 | CC BY 4.0 Open Access | rec: 28 Jun 2016, publ: 28 Jun 2016

Motivation

Recent advances in genome sequencing and biological data analysis technologies used in bioinformatics have led to a fast and continuous increase in biological data. The difficulty of managing the huge amounts of data currently available to researchers and the need to have results within a reasonable time have led to the use of distributed and parallel computing infrastructures for their analysis.

Recently, bioinformatics is exploring new approaches based on the use of hardware accelerators as GPUs. From an architectural perspective, GPUs are very different from traditional CPUs. Indeed, the latter are devices composed of few cores with lots of cache memory able to handle a few software threads at a time. Conversely, the former are devices equipped with hundreds of cores able to handle thousands of threads simultaneously, so that a very high level of parallelism can be reached. Use of GPUs over the last years has resulted in significant increases in the performance of certain applications.

Despite GPUs are increasingly used in bioinformatics most laboratories do not have access to a GPU cluster or server. In this context, it is very important to provide useful services to use these tools.

Methods

A web-based platform has been implemented with the aim to enable researchers to perform their analysis through dedicated GPU-based computing resources. To this end, a GPU cluster equipped with 16 NVIDIA Tesla k20c cards has been configured.

The infrastructure has been built upon the Galaxy technology [1]. Galaxy is an open web-based scientific workflow system for data intensive biomedical research accessible to researchers that do not have programming experience. Let us recall that Galaxy provides a public server, but it does not provide support to GPU-computing.

By default, Galaxy is designed to run jobs on local systems. However, it can also be configured to run jobs on a cluster. The front-end Galaxy application runs on a single server, but tools are run on cluster nodes instead. To this end, Galaxy supports different distributed resource managers with the aim to enable different clusters.

For the specific case, in our opinion SLURM [2] represents the most suitable workload manager to manage and control jobs. SLURM is a highly configurable workload and resource manager and it is currently used on six of the ten most powerful computers in the world including the Piz Daint, utilizing over 5000 NVIDIA Tesla K20 GPUs.

Results

GPU-based tools [3] devised by our group for quality control of NGS data have been used to test the infrastructure. Initially, this activity required to make changes to the tools with the aim to optimize the parallelization on the cluster according to the adopted workload manager. Successively, the tools have been converted into web-based services accessible through the Galaxy portal. Currently, we are working to optimize the workload manager configuration. As for future work, we planned to share through Galaxy other GPU-based tools for NGS analyses released by our group [4][5] as well as specialized workflows created using these and other validated tools imported from the Galaxy ToolShed repository. These activities will be carried-out through the European ELIXIR project.

Supplementary Information

The work has been supported by the Italian Ministry of Education and Research through the Flagship InterOmics (PB05), and the European MIMOmics (305280) and ELIXIR (https://www.elixir-europe.org/) projects.

References

1. Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. *"Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences"*. ***Genome Biol***. 2010 Aug 25;11(8):R86.
2. http://slurm.schedmd.com/
3. Manconi, A et al. *"G-CNV: a GPU-based tool for preparing data to detect CNVs with read-depth methods"*. Frontiers in Bioengineering and Biotechnology, 2015, 3.
4. Manconi, A et al. *"GPU-BSM: A GPU-based tool to map bisulfite-treated reads"*. PLoS ONE 9(5): e97277. doi:10.1371/journal.pone.0097277, 2014.
5. Manconi, A et al. *"A tool for mapping single nucleotide polymorphisms using graphics processing units"*. BMC Bioinformatics, *15*(Suppl 1), S10, 2014.