Contribution ID: **36** Type: **Presentation**

TopHat on the Grid: an automatic workflow for sequence alignment exploiting EGI/IGI grid infrastructure

Friday, 21 September 2012 13:52 (22 minutes)

Description of the work

The contribution presents an execution of a complex workflow based on a reverse engineering of TopHat over the EGI/IGI grid infrastructure.It aligns RNA-Seq reads to human genomes using the ultra high-throughput short read aligner Bowtie. The workflow then analyzes the mapping results to identify splice junctions between exons.The job submission is executed by means of an already developed service called JST (https://indico.egi.eu/indico/contributionI
It handles the execution requests and deals with the real grid job submission, monitoring and resubmission. In order to provide a reliable workflow the exit status of each step is checked and each calculation could be re-executed in case of failure. Data transfers are executed using a grid Storage Element as temporary buffer. This tool could be used to exploit both standard grid resources and WNoDeS (http://web.infn.it/wnodes/index.php/wnodes) enabled cloud resources. A particular attention will be devoted to explain how we address the problem of transferring input and output data that usually exceed the 3GB size for each job. This implementation provides an improvement of algorithm of TopHat making it parallelizable. Three main blocks have been identified in TopHat,each of which is composed by several segments that can be executed independently on the Grid. Each segment is analyzed by Bowtie. Its aim is to map each short read segment onto the human genome reference.The main advantage of using the cloud solution based on WNoDeS is the possibility of deploying all the computational steps of the workflow in the cloud environment and not only the most CPU intensive ones as in the grid environment.
Indeed for few steps of the workflow we need a dedicated environment that is hard to replicate over a standard grid infrastructure.

Link for further information

www.computer.org/portal/web/csdl/doi/10.1109/ISMS.2012.76
bioinformatics.oxfordjournals.org/content/25/9/1105.full.pdf+html

Wider impact of this work

The described activities are the first prototype implementation of a mixed workflow that makes use of local dedicated machines, grid worker nodes and cloud resources and provide a good example of flexible and dynamic resource allocation.This work is useful to all the researchers that could not easily deploy their analysis on the EGI/IGI grid infrastructure only. Our framework allows to dispatch over the grid infrastructures only steps that are really CPU consuming with a minimal impact and modification on the already working workflow. Biotechnological laboratories are producing more and more sequencing data that open new horizons and new challenges to both computing infrastructures and software engineering. Our approach allows to reduce the elaboration time in the selected blocks; for example it is easy to obtain a big speedup factor in specific steps of the workflow.

Printable Summary

We will present the activity related to the use of TopHat, a fast splice junction mapper for RNA-Seq reads over a grid distributed infrastructure.TopHat manages data for Next Generation Sequencing technology that allows a more accurate analysis as: detection of new isoforms, differential gene expression analysis and detection of aberrant mutations.This technology allows to obtain from the molecules of DNA/RNA smaller fragments,

called read, which can be sequenced in parallel.The workflow is executed partially over a dedicated infrastructure hosted in the Istituto Superiore Mario Boella, while the steps that are very CPU consuming are executed dynamically on the grid.The processing framework is able to submit jobs to the grid infrastructure by means of Job Submission Tool.The jobs could be executed both on a standard grid infrastructure or on a cloud infrastructure based on WNoDeS like the IGI one.

Primary author: Dr DONVITO, Giacinto (INFN)

Co-authors: ACQUAVIVA, Andrea (Politecnico di Torino); CESINI, Daniele (INFN); Dr MOSSUCCA, Lorenzo (Istituto Superiore Mario Boella); GAIDO, Luciano (INFN); TERZO, Olivier (Istituto Superiore Mario Boella); RUIU, Pietro (Istituto Superiore Mario Boella)

Presenter: Dr DONVITO, Giacinto (INFN)

Session Classification: Workflow community workshop

Track Classification: Virtual Research Environments (Gergely Sipos: track leader)