EGI User Forum 2011

Contribution ID: **148** Type: **Poster**

Large DNA re-sequencing experiments on the Dutch BiG Grid infrastructure

Monday, 11 April 2011 09:00 (8 hours)

Conclusions

The integrated workflow for next generation sequence analysis avoids manual work. Based on our experiences the obtained speed up is around 2-5 fold per experiment, however the optimal splitting size needs to be determined. Several experiments can be analyzed at the same time, thus the speed up could be higher. In the period August-December 2010 42 TB of disk space and 200 CPU days were used on the worker nodes by our institute.
The workflow needs improvements regarding fault tolerance and data management. For example. if one of the alignments fails, the entire workflow fails. Therefore a procedure that will re-submit failed jobs will be developed, avoiding re-submission of successful jobs. Furthermore, much network traffic is generated in our current setup, this requires optimization of data positioning near or on nodes where the tasks will run. In addition, we need to build in mechanisms to clean up temporary data, especially when extremely large experiments are performed.

Overview

In 2010 we presented a pilot implementing two popular sequence alignment programs, Blast and Blat, on the Dutch grid. We implemented these programs as workflow components and we analyzed 2.5 GB of next generation sequencing data with a 30x decrease of analysis time. Since then larger experiments with another DNA sequencer have been performed where the throughput was increased to 8-35 GB per experiment. These large scale projects include whole exome, transcriptome, genome and small RNA sequencing experiments.
This growth in scale required a revision of the approach initially adopted, including the porting of a new alignment software, data splitting for higher parallelism, improved resource selection for storage and computing, and more controlled access to data. The platform (eBioInfra) has also been upgraded. There was a new MOTEUR release and more resources became available on the Dutch grid. We present the adopted solutions and comment on the current results.

Impact

The improved infrastructure and the new workflow language allowed us to analyze large scale sequencing experiments more quickly. Our modular design allows easy implementation of new bioinformatics software, which we can combine with the already implemented components. We use the workflows on a daily basis in our institute.
This methodology will be used in the Genome of the Netherlands project (GoNL). For this project the genomes of 750 individuals are sequenced (30 TB of input data) which needs to be analyzed. The goal of the project is to get a full characterization of common variants in the Dutch population and insight into de novo mutations. The estimated computing time is 8-13 CPU years, depending on the splitting size.

Description of the work

The ABI Solid DNA sequencer produces shorter sequences than the Roche sequencer, but with a larger throughput. The programs BLAST and BLAT are not suitable for the shorter sequences. Therefore we implemented the BWA sequence aligner as a workflow component.
There was also an urgency to protect the data from unauthorized use. The access control settings can be set with the glite-commands and we have also implemented workflow components that set these permissions automatically. Additionally we made use of Virtual Organization (VO) subgroups to further limit the access to the data.
To speed up the analysis time we split the input data at the start of the workflow. A new component writes the split data directly from the worker node to the Logical File Catalog (LFC).
Each fraction of the data is aligned in parallel against the human genome with BWA, downloading the input data and the entire human genome database to the worker node and returning the alignment results.
At the end of the workflow the alignment results are merged. The merge-component reads the directory with the alignment results, copies them to the worker node, merges them, and transfers the results to the LFC. Since this component uses much local disk space we selected resources with sufficient free disk space available.
During last year we migrated from the SCUFL to the GWENDIA workflow language. GWENDIA is an array-based data flow language that gave us more power to implement these workflows. More specifically, this allowed us to implement better split and merge components. All steps, including post-alignment components were combined into one workflow.

URL

http://www.bioinformaticslaboratory.nl/

Primary author: Ms VAN SCHAIK, Barbera (Academic Medical Center)

Co-authors: Prof. VAN KAMPEN, Antoine (Academic Medical Center); Prof. BAAS, Frank (Academic Medical Center); Mr WILLEMSEN, Marcel (Academic Medical Center); Mr SANTCROOS, Mark (Academic Medical Center); Dr OLABARRIAGA, Silvia (Academic Medical Center)

Presenter: Ms VAN SCHAIK, Barbera (Academic Medical Center)

Session Classification: Posters

Track Classification: Poster