

Introduction

The role of high performance computing is getting more and more relevant in biological and medical scientific research due to the increasing quantity of data produced by the high throughput analysis techniques emerging in these fields. The use of distributed architecture environments can be an appropriate solution both from the computational point of view and for the data management, but it may appear still too complex for the end users: submitting jobs, monitoring their status and retrieving the results can be challenging when working on huge quantity of data. That's the case of Genetic Linkage Analysis, a statistical method for detecting genetic linkage between disease loci and markers of known locations by following their inheritance in families through the generations. This is a NP-hard problem and the computational cost and memory requirements of the major algorithms proposed in literature grows exponentially with pedigree size and markers' number. Implementations of the mentioned algorithms reflect these limits making analyses of medium/large data sets very hard on a single CPU.

The aim of the present work is to enable the use of high performance computing infrastructures, such as Clusters and Grid Infrastructures, for the execution of linkage analysis workflows on large data sets, especially for SNPs, dense biallelic markers (from 10.000 per chip to more than 1 million).

Methods

Many of the most used linkage analysis software have been ported into our Cluster and the Grid environments (gLite EGEE) and a suitable and customizable workflow has been designed and implemented in order to realize a parallel approach for the linkage analysis. The parallelism exploits independent comparisons of markers' genotypes against pedigree data, splitting inputs to achieve the best threshold for computational time/memory cost on the basis of the specific algorithm adopted by the linkage program.

At a higher level, all required steps are parameterized and monitored with the support of a web interface, designed to simplify and speed up the whole process of linkage analysis. Users can create the analysis workflow arranging customizable modules, representing workflow steps, into a visual blackboard; each module can be customized choosing different kinds of input files, introducing data pre-processing steps, selecting the favorite algorithm with the proper parameters and choosing the HPC environment to be used.

At a lower level the workflow engine splits the workload into small jobs and distributes analysis tasks over the available resources, the Cluster's nodes or the Grid computing elements; on the Cluster side the distribution of the jobs is natively managed by the workflow engine, while on the Grid side this is achieved by a reliable, scalable and secure grid system facility, called VNAS, a software layer built on top of the grid middleware which monitors each single submitted job and ensures its computation completes successfully by managing the resubmission of failed jobs automatically.

When all tasks are computed the results are retrieved, merged and made available for download through the web interface. *Figure 1* shows this pipeline.

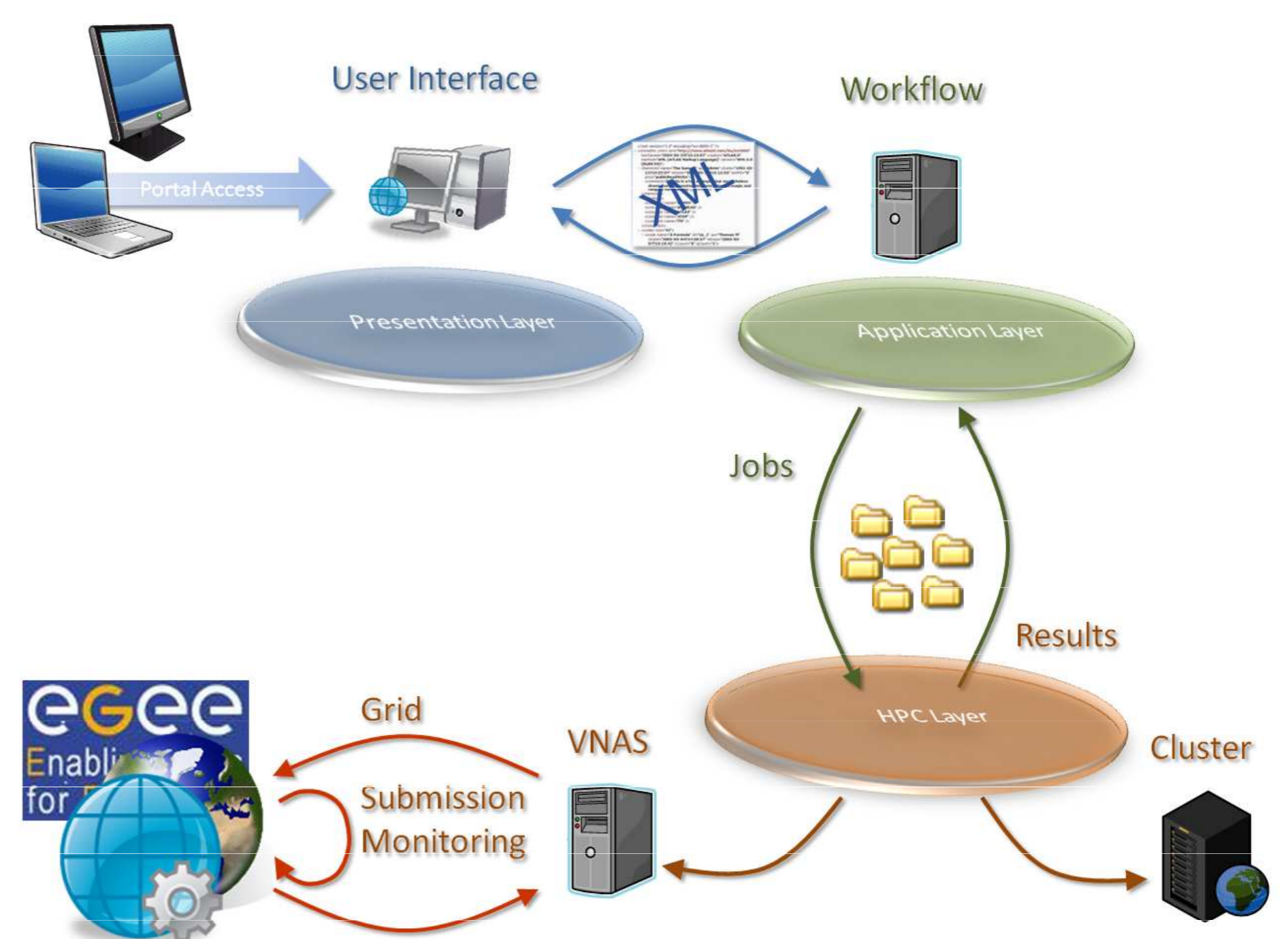


Figure 1: The pipeline of the linkage analysis managed by our system.

Results

Efficiency tests and simulations run on the framework developed, *Figure 2* and *Table 1*, show that distributed analysis pipelines with size of data (linkage variables) close to the computational limits for a mid-range workstation achieved significant in computational time compared to dual-core 2 GHz single CPU execution; increasing the number of individuals of the Pedigree tree some computations were still performed on the distributed infrastructure, but resulted infeasible on a desktop PC due to memory overflow.

Test results also show that this approach is most useful in high-end challenges, because the infrastructure overheads (like queues, latency, net congestion) are less affecting the overall execution time, while small challenges may still show higher efficiency when run on a single CPU workstation.

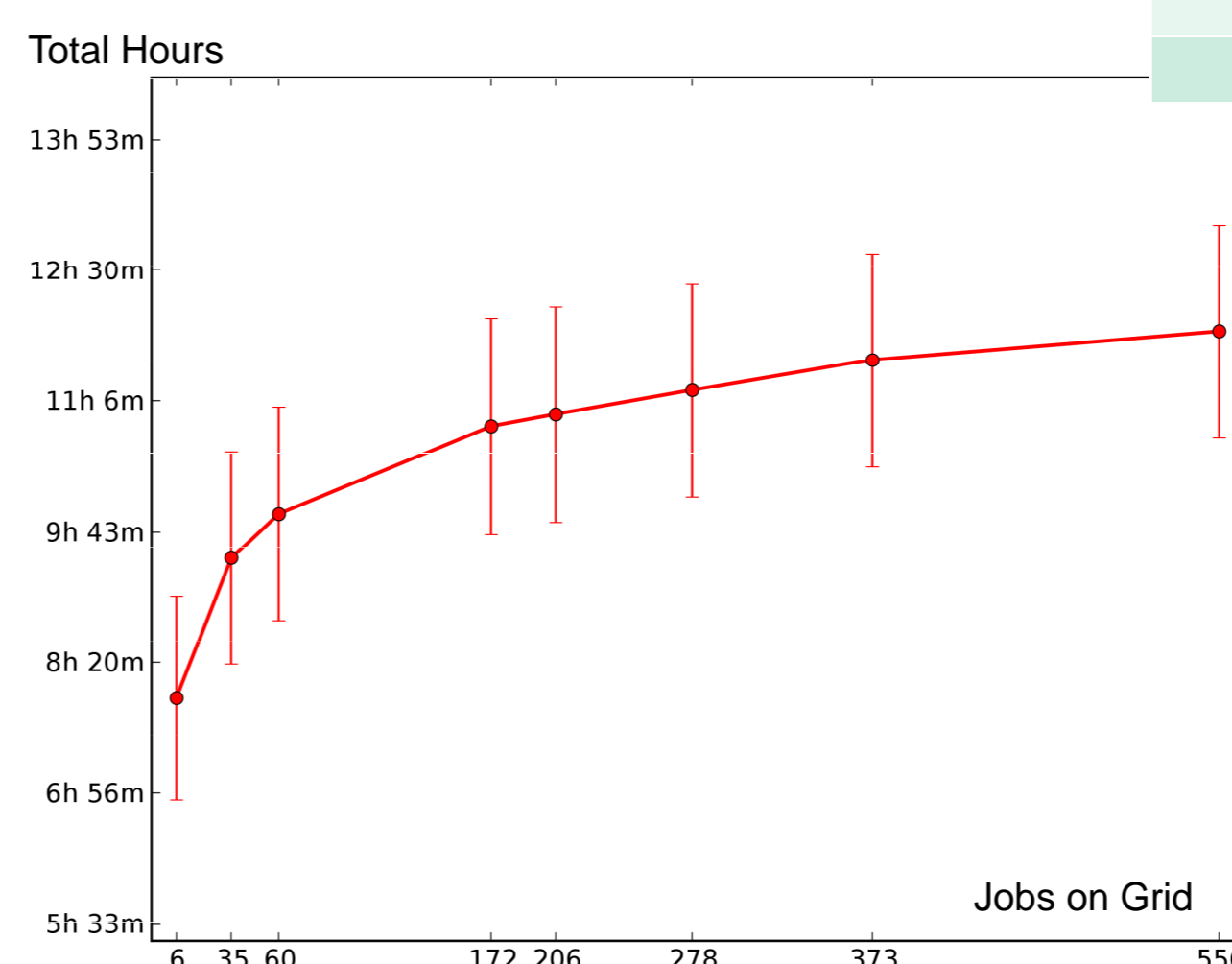
Conclusions

The developed application enables the user to launch genetic linkage analysis calculations for medium/large challenges over a distributed computational infrastructure like Clusters and the EGEE Grid. It offers a customizable workflow to achieve parallel processing of the pipeline tasks, a web user interface that provides an easier approach to linkage analysis software and a reliable software layer that manages low-level interactions with the distributed computing elements. This approach grants a user friendly access to high performance and distributed computation technologies with very basic informatics knowledge requirements.

Supplementary Information

The CNR-BIOINFORMATICS web site is available at <http://www.cnr-bioinformatics.it>

Figure 2: Grid tests and simulations executed on the framework show benefits using grid environment; confidence intervals are due to grid overheads.



Illumina Chip	# Runs [50 SNP]	# Jobs [6 h]	Comput. Cost (time)	
			Single CPU	Grid
10 k	200	6	33 h	~ 8 h 10 m
66 k	1,320	35	220 h	~ 9 h 30 m
100 k	2,000	60	333 h	~ 9 h 50 m
317 k	6,340	172	1056 h	~ 10 h 50 m
370 k	7,400	206	1233 h	~ 11 h 05 m
500 k	10,000	278	1665 h	~ 11 h 15 m
670 k	13,400	373	2233 h	~ 11 h 30 m
1 M	20,000	556	3332 h	~ 11 h 57 m

Table 1: Tests performed with SNP Illumina Chips (from 10k to 1M SNPs) comparing computational cost for Single CPU against Grid; these tests and simulations produce N runs (execution of linkage analysis software, i.e. GeneHunter, each run takes 10 mins of CPU), merged in jobs of 6 hours.