Contribution ID: **66** Type: **Oral contribution**

BioDCV: a grid-enabled complete validation setup for functional profiling

Wednesday, 1 March 2006 14:45 (15 minutes)

Abstract
BioDCV is a distributed computing system for the complete validation of gene profiles. The system is composed of a suite of software modules that allows the definition, management and analysis of a complete experiment on DNA microarray data. The BioDCV system is grid-enabled on LCG/EGEE middleware in order to build predictive
classification models and to extract the most important genes on large scale molecular oncology studies. Performances are evaluated on a set of 6 cancer microarray datasets of different sizes and complexity, and then compared with results
obtained on a standard Linux cluster facility.

Introduction
The scientific objective of BioDCV is a large scale comparison of prognostic gene signatures from cancer microarray datasets realized by a complete validation system and run in Grid. The models will constitute a reference experimental landscape for new studies. Outcomes of BioDCV consist of a predictive model, the straighforward evaluation of its accuracy, the lists of genes ranked for importance, the identification of patient subtypes. Molecular oncologists from medical research centers and collaborating bioinformaticians are currently the target end-users of BioDCV. The comparisons presented in this paper demonstrate the factibility of this approach on public data as well as on original microarray data from IFOM-Firc. The complete validation schema developed in our system involves an intensive replication of a basic classification task on resampled versions of the dataset. About 5x105 base
models are developed, which may become 2x106 if the experiment is replicated with randomized output labels. The scheme must ensure that no selection bias effect is contaminating the experiment. The cost of this caution is high computational complexity.

Porting to the Grid
To guarantee fast, slim and robust code, and relational access to data and a model descriptions, BioDCV was written in C and interfaced with SQLite (http://www.sqlite.org), a database engine which supports concurrent access and transactions useful in a distributed environment where a dateset may be replicated for up to a few million models. In this paper, we present the porting of our application to grid systems, namely the Egrid (http://www.egrid.it) computational grids. The Egrid infrastructure is based on Globus/EDG/LCG2 middleware and is integrated as an independent virtual organization within Grid.it, the INFN production
grid. The porting requires just two wrappers, one shell script to submit jobs and one
C MPI program. When the user submits a BioDCV job to the grid, the grid middleware looks for the CE (Computing Element: where user tasks are delivered) and the WNs

(Worker Nodes: machines where the grid user programs are actually executed) require to run the parallel program. As soon as the resources (CPUs in WNs) are available, the shell script wrapper is executed on the assigned CE. This script distributes the microarray dataset from the SE (Storage Element stores user data in the grid) to all the involved WNs. It then starts the C MPI wrapper which spawns several instances of the BioDCV program itself. When all BioDCV instances are completed, the wrapper copies all outputs including model and diagnostic data from the WNs to the starting SE. Finally, the process outputs are returned, thus allowing the reconstruction of a complete data archive for the study.

Experiments and results
Two experiments were designed to measure the performance of the BioDVC parallel application in two different computing available environments: a standard Linux cluster and a computational grid.
In Benchmark 1, we study the scalability of our application as a function of the number of CPUs. The benchmark is executed on a Linux clusters formed by 8 Xeon 3.0 CPUs and on the EGEE grid infrastructure ranging from 1 to 64 Xeon CPUs. Two DNA microarray datasets are considered: LiverCanc (213 samples, ATAC-PCR, 1993 genes) and
PedLeuk (327 samples, Affymetrix, 12625 genes). On both dataset we obtain a speed-up curve very close to linear. The speed-up factor for n CPUs is defined as the user time for one CPU divided by the user time for n CPUs.
In Benchmark 2, we characterize the BioDCV application different d (number of features) and N (number of samples) values for a complete validation experiment, and we execute a task for each dataset on the EGEE grid infrastructure using a fixed number of CPUs. The benchmark was run on a suite of six microarray datasets: LiverCanc, PedLeuk, BRCA (62 samples, cDNA, 4000 genes), Sarcoma (35 samples, cDNA, 7143 genes), Wang (286 samples, Affymetrix, 17816 genes), Chang (295 samples, cDNA, 25000 genes). It can be observed that effective execution time (total execution time without queueing time at grid site) increases linearly with the dataset footprint, i.e. the product of number of genes and number of samples. The performance penalty payed with respect to a standard parallel run performed on local cluster is limited and it is mainly due to data transfer from user machine to grid site and between WNs.

Discussion and Conclusions
The two experiments, which sum up to 139 CPU days within the Egrid infrastructure, implicate that general behavior of the BioDCV system on LCG/EGEE computational grids can be used in practical large scale experiments. The overall effort for gridification was limited to three months. We will investigate if substituting a model of one single job asking for N CPUs (MPI approach) with a model that submits N different single CPU jobs can overcome some limitations. Next step is porting our system under EGEE's Biomed VO.

BioDCV is an open source application and it is currently distributed under GPL (SubVersion repository at http://biodcv.itc.it).

Primary author: PAOLI, Silvano (ITC-irst)

Co-authors: BARLA, Annalisa (ITC-irst, Trento, Italy); FURLANELLO, Cesare (ITC-irst, Trento, Italy); ALBANESE, Davide (ITC-irst, Trento, Italy); JURMAN, Giuseppe (ITC-irst, Trento, Italy); REID, James F. (INT/IFOM-FIRC, Milano, Italy); FLOR, Roberto (ITC-irst, Trento, Italy); COZZINI, Stefano (INFM/Democritos, Trieste, Italy); MERLER, Stefano (ITC-irst, Trento, Italy)

Presenter: PAOLI, Silvano (ITC-irst)

Session Classification: 1a: Life Sciences

Track Classification: Life Science