BioOpera: Cluster Computing in Bioinformatics

Gustavo Alonso, Information and Communication Systems Research Group, ETH Zürich.

Many scientific disciplines are shifting from in vitro to in silico research as more physical processes and natural phenomena are simulated in a computer (in silico) instead of being directly observed (in vitro). In these virtual laboratories, the computations involved are often a major bottleneck and a significant source of errors and inefficiencies. To spare scientists such limitations, at ETH Zürich we have developed BioOpera, an extensible process support management system for virtual laboratories.

In existing virtual laboratories, storing, manipulating, and keeping track of complex computations is done manually through ad-hoc pieces of code. The data processing logic is typically written using conventional programming languages (e.g., C or Fortran) for the basic algorithms and collections of operating system scripts (mainly Perl scripts) as the glue between the different components. Such an approach leads to logic that is extremely difficult to modify and rather primitive, unsystematic methods for driving and monitoring computations. Given the increasing relevance of the work done in virtual laboratories, better software tools are becoming critical to the success of any virtual experiment. As an example of the functionality such tools should provide, consider dependability. An adequate software environment for a virtual laboratory should dependably run computations for months at a time with minimal user intervention. This requires to automatically handle issues such as efficient scheduling of jobs, load balancing, tracking of progress and results of the computation, recovery from system errors and machine crashes, access to intermediate results as they are computed, automatic accounting of statistics concerning computing time, and a systematic method for storing all necessary meta-data.

The first step towards providing this functionality involves finding an appropriate representation for the computation. We have chosen the notion of process, similar to that used in workflow management systems (although the final implementation is rather different since workflow tools are not entirely adequate for virtual laboratories). A process is an annotated directed graph where the nodes represent tasks and the arcs represent the control/data flow between these tasks. The notion of process allows one to capture sequences of invocations of computer programs in a distributed and heterogeneous environment and the corresponding data exchanges between these programs. From here, the process can be encoded in such a way so as to allow its efficient storage in a database. Once in a database, this information is persistent, allowing us to both automatically manage the computation and increase its dependability.

We have implemented these and many other novel ideas in BioOpera, a process support system for virtual laboratories in bioinformatics. BioOpera is based on Opera, a workflow-like middleware tool that has evolved into a programming and runtime environment for cluster computing with the capability to define, execute, monitor and manage a broad range of large-scale, distributed computations. BioOpera is being developed at the Information and Communication Systems Research Group of ETH Zürich in collaboration with groups in bioinformatics at ETH Zürich and McGill University in Canada.

The best way to describe BioOpera is as a high-level distributed operating system managing the resources of a computer cluster. Using BioOpera, we have already achieved interesting results both in computer science and in bioinformatics. From the point of view of research in grid computing, we have shown how BioOpera can effectively sustain complex distributed computations during long periods of time. The system is capable of surviving total failures, software and hardware upgrades, and node crashes without manual intervention being required for the computation to resume after the failures have been repaired. We have also proven how BioOpera can duplicate the processing capacity of a grid cluster by making more efficient use of the available resources. In bioinformatics, several search algorithms have been considerably optimized using BioOpera, in some cases reducing the cost to less than 30 % of the original cost.
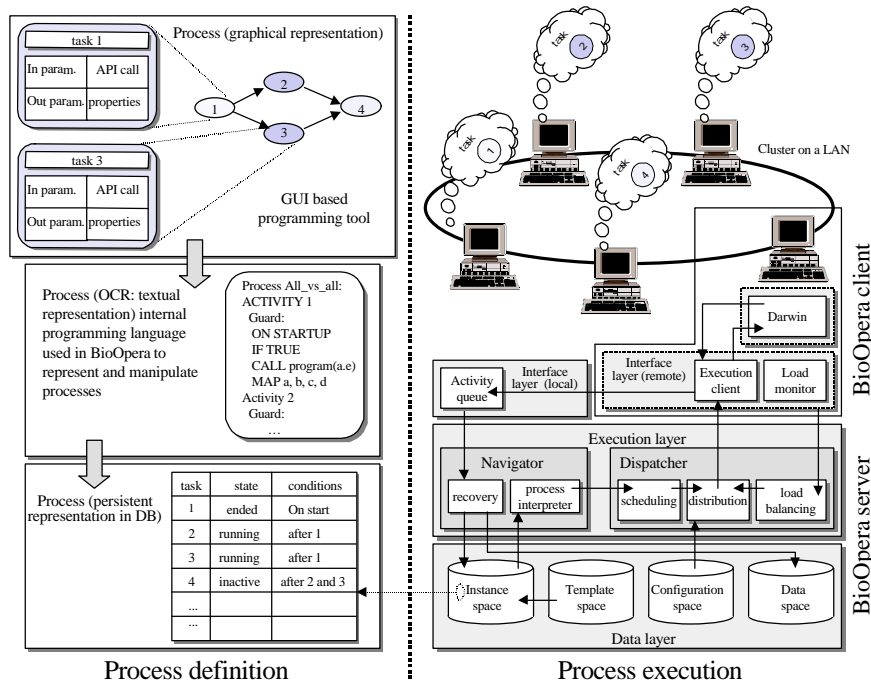
Contact information:
Prof. Dr. Gustavo Alonso
Institute of Information Systems (CLU E1)
Swiss Federal Technical University (ETH)
ETH Zentrum, 8092 Zürich
Switzerland
E-mail: alonso@inf.ethz.ch
Tel: +41 - 1 - 632 7306
WWW: www.inf.ethz.ch/personal/bausch/biopera/main.html

Process definition

Process execution