SUPPORTING DATA MINING APPLICATIONS ON CLUSTERGRID

Vida Gábor, vida@sztaki.hu
Dr. Podhorszki Norbert, pnorbert@sztaki.hu
Dr. Kacsuk Péter, kacsuk@sztaki.hu
MTA SZTAKI Laboratory of Parallel and Distributed Systems

Abstract:

The main objective of the GVOP project "Next Generation Data Mining on High-Performance Distributed Parallel Systems" is to develop a data mining software prototype running on high-performance distributed parallel systems and enabling the creation of data mining models that beat any previous models concerning their quality. The task of the system is to exploit the available resources in a most optimal way adapting itself to the network features of the system and at the same time hiding the details of parallel execution for the data mining specialists who can steer the execution mechanism according to the constantly received partial results. In the framework of the project the data and computation intensive data mining applications will be executed on various Grid systems.

In order to make the Grid systems transparent for the users and to provide high-level access to the important low-level Grid functionalities a new programming interface (called as Distributed Computing API, or shortly DC API) was defined. The implementation of the DC API should solve the problems of distributed system access, network and data management. It provides an easy-to-use interface for the upper software layers meanwhile hides the details of the implementations of the distributed parallel system. The current implementation of the DC API should provide the basic functionalities enabling the execution of a demo version of the distributed data mining algorithms.

The first implementation of the DC API was performed on the Hungarian ClusterGrid and the goal of the current paper is to describe this implementation. The paper explains the way of how the work units are created and distributed in the ClusterGrid as well as the methods of executing, suspending, interrupting and resuming the work units. During the implementation the biggest challenge was the handling of partial results and the realization of a dynamic, on-line steering of execution control according to partial results. Due to this challenging problems the paper describes in detail how these problems were solved on the ClusterGrid.