AN LP-BASED HEURISTIC FOR THE SERVERS LOAD BALANCING PROBLEM

N. A. Kochetova

Sobolev Institute of Mathematics,
Novosibirsk State University, Novosibirsk
e-mail: nkochet@math.nsc.ru

We consider the servers load balancing problem which is originated from optimal web hosting in cloud computing [1]. We are given a set of servers. Each server contains the set of disks (images of disks). Internet sites with heterogeneous information are distributed among the disks. A lot of users visit these sites and generate the load for the servers. The load is changed during time interval and can be characterized by some parameters: CPU, RAM and others. We assume that the activities of users are known for each site and as result for each disk and each server. If the load of the server does not exceed a given threshold, we say that the server is working in regular mode. Otherwise, it works in heavy mode. To avoid the heavy mode we can move the disks from one server to another. This moving requires some computational efforts. We will call them as overhead expenses. We assume that the expenses are known for each parameter if a disk is extracted from one server and moving to another one. Initial distribution of disks among the servers is given. The problem is to find a new distribution of the disks to minimize the total exceeding over the thresholds during the time interval subject to overhead expenses for each server.

This problem can be presented as a mixed integer linear program. Thus we can use commercial software to solve it. Unfortunately, the integrality gap is high for this formulation and the solvers can find optimal solution for small instances only. Therefore, we develop an LP-based approximate algorithm with a priori performance bound. We replace the Boolean variables by the continuous ones and solve the corresponding linear programming problem. This optimal solution allows us to fix some variables and reduces the dimension of the problem. We apply this approach until the reduced subproblem can be solved by CPLEX software. Computational results for 20 servers and 200 disks are discussed.

REFERENCES

1. Yu. A. Kochetov, N. A. Kochetova. The servers load balancing problem // Vestnik of Novosibirsk State University. Series Information Technology. 2013. Vol. 11, issue 4. P. 71-76. (in Russian)