Abstract

Considering the growing interest in using cloud services, the accessibility and the effective management of the required resources, irrespective of the time and place, seems to be of great importance both to the service providers and users. One of the best ways for increasing utilization and improving the performance of the cloud systems is the auto-scaling of the applications; this is because of the fact that, due to the scalability of cloud computing, on the one hand, cloud providers believe that sufficient resources have to be prepared for the users, and on the other, the users also have a tendency towards the "pay as you go" system of payment for the resources. This paper seeks to offer an approach, based on the learning automata, for the scalability of the web applications, which combines virtual machine clusters and the learning automata in order to provide the best possible way for the scaling up and scaling down of the virtual machines. The results of this study indicate that the proposed approach has decreased the number of SLA violations (in percentage), while it has a smaller load of scalability compared to the other approaches in this regard.
Index Terms

Computer Science

Distributed Systems

Keywords

Cloud computing  Auto scaling  Learning automata  SLA violation