Abstract

Load balancing and Consolidation of Virtual Machines is a way which is effective to improve the utilization of resources and energy efficiency in Cloud data centers. Determining when it is best to reallocate Virtual Machines from an overloaded host is an aspect of dynamic Virtual Machine consolidation that directly influences the utilization of resource and Quality of Service which the system is delivering [1]. The influence on the Quality of Service is explained by the fact that server overloads cause the shortage of resources and the degradation of applications performance. The current solutions to the problem of host overload detection are generally relying on statistical analysis guided by nature inspired in order to find the optimal solution. The limitations of these techniques are that they lead to sub-optimal results and do not allow explicit specification of a Quality of Service goal. We propose a new approach that for any stationary workload which is known and a given state configuration solves the problem of detection of host overload by maximizing the mean inter-migration time under the specified Quality of Service goal optimally [2]. Through simulations with real-world workload traces from more than a thousand Virtual Machines, we show that our approach outperforms the best benchmark.
algorithm and provides almost 88% of the performance of the optimal offline algorithm.

References


**Index Terms**

Computer Science  
Distributed Systems

**Keywords**

Cloud computing, Distributed systems, dynamic consolidation, virtualization, host overload detection and energy efficiency.