Identifying Overloaded Servers and Managing Dynamic Placement of Virtual machines in Cloud

Abstract

Cloud computing is becoming one of the most popular commercial infrastructure due to its little maintenance expense and on demand resource utilization. Cloud computing possesses many kinds of technical challenges such as fault tolerance, reliability, availability, integrity etc. due to its complex and distributed nature. But the main problem related to all those is overload incurred by Virtual Machines (VM). So, load balancing is one of the most significant issues that can help to gain rapid performance of cloud infrastructure. This research proposes algorithms for detecting failed servers due to overloaded VMs. The failure detection algorithm checks server status after a predefined time interval. This algorithm gives proactive technique to deal with overloaded VMs. When any failure in the server is found, the resource balancing algorithm migrates its VMs to an adequate healthy Physical Machine (PM). To distribute workload evenly, the resource utilization skew is measured. This VM to PM mapping is done in a way that every PM will do almost equal amount of work.

References


Index Terms

Computer Science  Distributed Systems

Keywords

Cloud Computing, Resource management, Skewness, Virtual machine migration, Overload Detection.