Abstract

In Infrastructure as a Service (IaaS) model of cloud computing paradigm, users acquire computing resources such as CPU, memory, storage and network bandwidth from an IaaS provider and these resources are used to deploy and run their applications. Cloud service providers share computing resources of a physical machine by running isolated Virtual Machines (VM) for web applications. As the load on web application increases, the associated VM must be able to scale up resources to support the increasing load. At the same time, VM should also be able to scale-down resources at light load. In this paper the novel architecture is proposed that provides the hybrid solution of vertical followed by horizontal scaling techniques of resource management in cloud data center. As per the dynamic load on web applications, the proposed algorithm takes the appropriate scaling decision to allocate resources from available pool of resources. Generally dynamic scaling is achieved by the conventional live VM migration technique to create additional VM instances, but VM migration spends CPU time and consumes large amount of IO and network traffic. The proposed technique postpones live VM migration as long as possible with the help of vertical scaling technique to improve the performance of
applications.

References

12. VMware, http://www.vmware.com/in


Index Terms

Computer Science Information Sciences

Keywords