Abstract

Efficient provisioning of resources is a challenging problem in cloud computing environments due to its dynamic nature and the need for supporting heterogeneous applications. Even though VM (Virtual Machine) technology allows several workloads to run concurrently and to use a shared infrastructure, still it does not guarantee application performance. Thus, currently cloud datacenter providers either do not offer any performance guarantee or prefer static VM allocation over dynamic, which leads to inefficient utilization of resources. Also, the workload may have different QoS (Quality Of Service) requirements due to the execution of various types of applications such as HPC and web, which makes resource provisioning much harder. Earlier work either concentrate on single type of SLAs (Service Level Agreements) or resource usage patterns of applications, such as web applications, leading to inefficient utilization of datacenter resources. In this paper, we tackle the resource allocation problem within a datacenter that runs different types of application workloads, particularly non-interactive and transactional applications. We propose an admission control and scheduling mechanism which not only maximizes the resource utilization and profit, but also ensures that the QoS requirements of users are met as specified in SLAs. In our study, we find that it is important to take care of various types of SLAs along with applicable penalties and the mix of workloads for better resource allocation and utilization of datacenters. The proposed mechanism provides
substantial improvement over static server consolidation and reduces SLA violations.

References

conference on autonomic computing, Washington, USA; 2010.
- Zhang W, Qian H, Wills C, Rabinovich M. Agile resource management in a virtualized
engineering. California, USA; 2010.
- Soundararajan V, Anderson J. The impact of MNGT. Operations on the virtualized
datacenter. In: Proceedings of the 37th annual international symposium on computer
architecture. France; 2010.
resource allocation to virtual containers. In: Proceedings of the 10th IFIP/IEEE international
- Minarolli D, Freisleben B. Distributed resource allocation to virtual machines via artiﬁcial
neural networks. In: Proceedings of 22nd Euromicro international conference on parallel,
distributed and network-based processing (PDP), Turin, Italy; 2014.
- Casalicchio E, Menascé DA, Aldhalaan A. Autonomic resource provisioning in cloud
systems with availability goals. In: Proceedings of the 2013 ACM cloud and autonomic
computing conference, Miami, FL, USA; 2013.
CASCON &apos;09: Proceedings of the 2009 conference of the Center for Advanced Studies
on Collaborative Research, Ontario, Canada; 2009.

Index Terms

Computer Science                           Distributed Systems

Keywords

SLA  VM  HPC (High Performance Computing)