ABSTRACT

Software failures, workload-related failures and job overload conditions bring about Service Level Agreement (SLA) violations in software as-a-service (SaaS) systems. Existing work does not address elimination of SLA violations completely as (i) while some do not address software and workload-related failures, other approaches do not address the problem of target PM selection for workload migration comprehensively and (ii) a clear mathematical mapping between workload, resource demand and SLA is lacking.

In this work, a software framework is introduced or presented for the cloud controller that helps minimizing service failures due to SLA violation of availability, utilization and response time in SaaS cloud data centers. Though migration is considered to be the primary mitigation technique, still it has been tried to mitigate SLA violations without migration for better performance and taking some vital factors into account. It is achieved this by performing a capacity check on the host physical machine (PM) before the migration to identify if enough capacity is available on the current PM to address the upcoming SLA violations. As in restart/rebooting which is the mitigation technique against availability violations due to software failures in a VM the primary technique is to create a current state replica of the existing VM because of which an SLA violation may be triggered. In certain cases such as workload related failures due to corrupt files, workload rerouting is preferred to a replica VM over migration. The selection of a target PM is also formulated as a multi-objective optimization problem. For workload migration to mitigate the SLA violation needs a proper understanding of the underlying workload.

For managing workload it is extremely important for a Software as a Service (SaaS) provider to understand the characteristics of the business application workload in order to size and place the virtual machine (VM) containing the application. Using the knowledge of the application architecture and statistical analysis of the workload, one can obtain an appropriate capacity and a good placement strategy for the corresponding VM. In this work a technique is proposed that determines VM capacity and VM collocation possibilities for a given set of application workloads. An empirically study on workloads are performed for geographically distributed data centers. This technique will determine the fixed reserved capacity and a shared capacity of a VM which it can share with another collocated VM. Based on the workload variation, the tool determines if the VM should be statically allocated or needs a dynamic placement. Again a peak utilization analysis must be performed over the workloads. This approach is a precursor step that is necessary before the placement decision is made. This approach will be applicable in situations where Pattern of the resource utilization by each workload is known apriori. An optimal resource management technique is then proposed which relies on examining the time varying resource demands and variability of the workloads to determine most optimal placement. The behavior and resource demands of the workload with time do not need to be known apriori for this technique. An empirical study is then performed over workloads for geographically distributed data center.

Keywords- *workload analysis, virtual machine, VM placement, VM resizing, COV, multi-objective optimization, compatibility matrix, reserved capacity, shared capacity.*