

**CHENNAI – PONDICHERRY**

**BURSTINESS-AWARE RESOURCE RESERVATION FOR SERVER CONSOLIDATION IN COMPUTING CLOUDS**

**ABSTRACT**

In computing clouds, burstiness of a virtual machine (VM) workload widely exists in real applications, where spikes usually occur aperiodically with low frequency and short duration. This could be effectively handled through dynamically scaling up/down in a virtualization-based computing cloud; however, to minimize energy consumption, VMs are often highly consolidated with the minimum number of physical machines (PMs) used. In this case, to meet the dynamic runtime resource demands of VMs in a PM, some VMs have to be migrated to some other PMs, which may cause potential performance degradation. In this paper, we investigate the burstiness-aware server consolidation problem from the perspective of resource reservation, i.e., reserving a certain amount of extra resources on each PM to avoid live migrations, and propose a novel server consolidation algorithm, QUEUE. We first model the resource requirement pattern of each VM as a two-state Markov chain to capture burstiness, then we design a resource reservation strategy for each PM based on the stationary distribution of a Markov chain. Finally, we present QUEUE, a complete server consolidation algorithm with a reasonable time complexity. We also show how to cope with heterogenous spikes and provide remarks on several extensions. Simulation and testbed results show that, QUEUE improves the consolidation ratio by up to 45 percent with large spike size and around 30 percent with normal spike size compared with the strategy that provisions for peak workload, and achieves a better balance between performance and energy consumption in comparison with other commonly-used consolidation algorithms.