

**CHENNAI – PONDICHERRY**

**Optimizing for Tail Sojourn Times of Cloud Clusters**

**Abstract:**

User Request (UR) service scheduling is a process that significantly impacts the performance of a cloud data center. This is especially true since essential Quality-of-Service (QoS) performance metrics such as the UR blocking probability as well as the data center’s response time are tightly coupled to such a process. This paper revolves around the proposal of a novel Deadline-Aware UR Scheduling Scheme (DASS) that has the objective of improving the data center’s QoS performance in term of the above-mentioned metrics. A minority of existing work in the literature targets the formulation of mathematical models for the purpose of characterizing a cloud data center’s performance. As a contribution to covering this gap, this paper presents an analytical model, which is developed for the purpose of capturing the system’s dynamics and evaluating its performance when operating under DASS. The model’s results and their accuracy are verified through simulations. In addition, the performance of the data center achieved under DASS is compared to its counterpart achieved under the more generic First-In-First- Out (FIFO) scheme. The reported results indicate that DASS outperforms FIFO by 11% to 58% in terms of the blocking probability and by 82% to 89% in terms of the system’s response time.

**Existing System:**

Centers constitute a cloud’s infrastructure and can be accessed through a web browser from anywhere around the globe; hence, the signal importance of modelling and performance analysis of cloud data centers. However, unfortunately, the research community has invested minor effort in this direction. As a matter of fact, the formulation of mathematical models that are able to capture the dynamics of cloud data centers and, hence, accurately reflect their performance is quite challenging. This is especially true when virtualization is exploited for provisioning users with sufficient amounts of adequate computing resources for the purpose of accomplishing their placed requests. The level of complexity grows considerably as a function of the virtualization degree (i.e. the number of Virtual Machines (VMs) hosted over a single Physical Machine (PM)). Today’s state of technology allows for hosting hundreds of VMs over a single PM, [9]. At this level, a close survey of the literature reveals only very few work revolving around the performance evaluation of cloud computing infrastructure operating under the above conditions.

**Proposed System:**

The proposal of DASS which is a novel UR service scheduling scheme that accounts for the deadline of incoming URs and aims at improving the QoS performance of a cloud data center in terms of reducing both the UR blocking probability and the system’s response time.

The development of an analytical framework encompassing a tractable queueing model, which has the ability of capturing the dynamics of the LBS’s global queue with an elevated degree of virtualization. The model’s complexity is simplified through the adoption of approximating assumptions. However, it is proven herein that approximation errors are negligible and hence, the established model can be confidently utilized for the purpose of evaluating the performance of the cloud data center in terms of important QoS metrics such as: i) the system’s response time, ii) the matching service completion probability, iii) the deadline mismatch probability, iv) the premature blocking probability, and v) the service probability.