**International Journal of Grid and High Performance Computing**
Volume 12 • Issue 1 • January-March 2020

# A Hybrid Multiple Parallel Queuing Model to Enhance QoS in Cloud Computing

Shahbaz Afzal, B. S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India
https://orcid.org/0000-0002-1217-9357

G. Kavitha, B. S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India

## ABSTRACT

Among the different QoS metrics and parameters considered in cloud computing are the waiting time of cloud tasks, execution time of tasks in VM's, and the utilization rate of servers. The proposed model was developed to overcome some of the pitfalls in the existing systems among which are sub-optimal markdown in the queue length, waiting time, response time, and server utilization rate. The proposed model contemplates on the enhancement of these metrics using a Hybrid Multiple Parallel Queuing approach with a joint implementation of M/M/1: ∞ and M/M/s: N/FCFS to achieve the desired objectives. A neoteric set of mathematical equations have been formulated to validate the efficiency and performance of the hybrid queuing model. The results have been validated with reference to the workload traces of Bit Brains infrastructure provider. The results obtained indicate the significant reduction in the queue length by 60.93 percent, waiting time in the queue by 73.85 percent, and total response time by 97.51%.

## KEYWORDS

Cloud Computing, CSC, CSP, Hybrid Multiple Parallel Queuing Model, Internet-as-a- Service, QoS, QoS Metrics, Queuing Theory

## INTRODUCTION

According to the NIST, cloud computing is elucidated as an automatic computing prototype for enabling pervasive, agreeable and on-demand service access to an infinite pool of customizable computing resources that can be easily supplied and released with minimal overhead (Zhang et al., 2010). It is a computing manifesto to manage and deliver services to a considerable number of people over a network, particularly on the internet (Pradhan et al., 2016). With the advances in internet and web service technologies, cloud computing has emerged as a potential internet-based utility during current technological era having capability of providing services to a vast collection of scalable users. It has provisions of providing both hardware and software utilities together with developmental platforms and testing tools as infrastructure as a service (IaaS), Software as a service (SaaS) and Platform as a service (PaaS), while internet-as a-service (iaaS) being a backbone. From technological perspective there are three primary actors on the scene - Cloud Service Provider (CSP),

DOI: 10.4018/IJGHPC.2020010102

Cloud Service Consumer (CSC) and the internet (Mishra et al., 2018; Buyya et al., 2010; Voorsluys et al., 2011). Internet being the communication medium, CSP provides necessary services in the form of rented scalable virtual machines to CSC at some nominal price on guaranteed levels of QoS as agreed between the two stakeholders in a proper documented service level agreement (SLA). The techniques of virtualization makes it possible to transform a single user physical machine into a multiple of shared multiuser virtual machines with variant of computing configurations in terms of operating system, memory, CPU power, Number of cores, storage, SSD, operating frequency, network bandwidth and much more, hence enabling better resource utilization with small wastages (Jain and Choudhary, 2016; Alouane and El Bakkali, 2016; Rimal et al., 2009).

Technically, when a huge number of user tasks requesting service arrive at cloud service provider, there is a direct consequence on the deliverable quality of service in terms of performance, deadline, cost, efficiency, profit, etc., for both provider and consumer. To preserve QoS according to SLA keeping in mind provider's turnover and other provider associated metrics it is essential to go for an efficient system model. This can be achieved by either developing an efficient queuing model or an efficient Scheduling and allocation algorithm (with or without load balancing features) or combination of both. While queuing models deal with enhancement of queuing metrics before actual execution of user tasks, scheduling and allocation algorithms deals with augmentation of scheduling metrics once the user tasks are sent for execution on virtual machines (VM's) (Vilaplana et al., 2014; Khazaei et al., 2011a; Takagi, 1991). The present paper focuses on the Queuing model approach of improving QoS in cloud computing. The paper does not deal with scheduling and allocation part of reinforcing the QoS metrics in cloud computing. From queuing model perspective, the generic cloud computing architecture is based on the queuing theory model (Vilaplana et al., 2014).

Though queuing model in cloud environment can be applied to any of the services in general, concentration is to focus on infrastructure-as-a-service utility model in specific, because IaaS represents actual service facilities required by cloud users for their service. Ghomi et-al proposed a generic model of cloud computing with versatile components in their cloud architecture (Ghomi et al., 2017). With analogy to queuing theory, each component of cloud computing model can be thought as a distinguished element of queuing model. A substantial number of user requests (demanding service) arrive at cloud are queued in data center controller and need to be completed on VMs in minimum finite time. From queue tasks are scheduled to one or more VMs for execution performed through task scheduling.

The objective of proposed approach is to minimize lifespan of a task. The lifespan of a task is the time it takes since it is generated in the queue to the time it has been accomplished and equals to sum of queue waiting time along with execution time neglecting transmission time in network media (which is a function of network bandwidth). This can be achieved by either minimizing queue waiting time or execution time (through scheduling) or both. The queue waiting time can be minimized by adopting an efficient queuing model and queuing discipline while execution time can be minimized by employing a robust scheduling algorithm. The waiting time for a task in a queue in turn depends on the length of the queue. So conversely to reduce the waiting time, queue length has to be reduced. Therefore, HMPQ model is developed to optimize the QoS metrics in cloud computing. The contributions of this research work are summarized as follows:

The minimal queue length, waiting time, and response time have been achieved due to addition of the factor $s^n$ in denominator of the base equation, where 's' is number of serving channels and 'n' is number of tasks in the queuing system. The work is based on pure analytical modeling and is hence computationally more efficient than simulation.

The paper is organized as follows. Section 2 summarizes related works; section 3 reviews the queuing system theory. The HMPQ model is presented in Section 4. Section 5 presents methodology and mathematical analysis of HMPQ model. The data collection and verification are given in Section 6 meanwhile Section 7 examines results and performance analysis of HMPQ approach. Section 8 concludes our research work and points out future work.