 ISBN: 978-93-86005-02-1

A Survey on Models for Analysis of Data Centre Performance and QOS in IaaS Cloud

M. Karthik* and N. Rakesh

Department of Computer Science, Amrita School of Engineering, Amrita Vishwa Vidhyapeetham, Amrita University, Bengaluru – 560035, Karnataka, India; mkarthik8923@gmail.com, n_rakesh@blr.amrita.edu

Abstract

Recent years there is a huge migration of business applications to cloud. The main challenges faced are data centre management and providing a survey of QOS modeling approaches and using analytical model in which stochastic reward net model, that is efficient to model systems composed of several quality attributes like utilization, availability, waiting time is taken into consideration.

Keywords: Cloud Computing, Iaas, Quality of Service, Reward Net Model

1. Introduction

IaaS provides the utilities like virtual machine deployed in data centre. Quality of service is very important for provisioning service level agreements. Data centre performance and resource provisioning is essential. For performance analysis there are few system models has been used like queuing systems, queuing networks and Layered Queuing Networks (LQN's)[1] has to take into consideration. Several models can be used to provide quality of service in cloud. In this paper we mainly concentrate on queuing systems, queuing networks and layered queuing networks, however the other classes exist like stochastic reward net models analysed through probabilistic methods. The concept is that a given model can perform better than other models.

2. System Models

Out of many models, the paper has a tendency to survey queuing systems, queuing networks, and layered queuing networks. Whereas queuing systems are mainly used to model a system with single resource, queuing networks will model the interaction between and/or application parts. Layered queuing networks are modeled for key interaction between application mechanisms.

*Author for correspondence

2.1 Queuing Systems

Queuing theory is usually employed in modeling to describe the resource of the system. Many mathematical formulas exist, an example to check the request mean latency time, and probability of waiting buffer occupancy in one queuing system. In cloud systems, mathematical queuing formulas are normally included in improvement aspects and they are commonly used in analysis in what-if conditions. The general mathematical formulas that use queues with service and arrival rates with 1 server (M|M|1), m servers (M|M|m), and queues are used with (M|M|1), the scheduling is done using FCFS or processor sharing. Mainly the processor can be modeled by M|G|1 PS queue, used in many cloud research because it is simple and suitable to apply model with different variety of workloads. For example service level architecture aware allocations in clod applications are mainly obtained using M|G|1 PS queues as the quality of service. And in n-tier cloud applications resource provisioning is given by modeling processor as an M|G|1 PS queue. The M|M|1open queue with first come first serve approach has some difficulties with average response times of the cloud systems. For the customer level satisfaction M|M|K|m priority queue[2] can be used, providing better service times and inter arrival rates, no buffer. This is used to analyze rejection rate of jobs and improving cloud data centers. The other aspects depends on cloud models is used to model resources used in discrete time control problems, arrival time may change with respect to time and circumstances. But the limitation in queuing systems is used to model a cloud application with single resource.

2.2 Queuing Networks

Queuing networks are explained as a group of queues which will have arrival requests and departure of processed jobs. Every queue signifies a resource like processor, bandwidth or a software buffer. The queuing networks are used to visualize the interaction between cloud application layers. In cloud applications, cloud service can be designed as multiple levels single level single service queues. Scheduling will be predefined as the resources to design processor sharing. Each division of queues will model the different aspects of the cloud systems. The queuing networks are used to render good performance and resource allocation strategies. In cloud applications, the queuing networks are also used to model multi tier architectures integrated with cloud applications and to support service level resource allocation aspects. Every has processing times and scheduling aspects. The limitation with queuing systems is it is used to model only single resource. In cloud applications, the queuing networks are also used to model multi tier architectures integrated with cloud applications and to support service level resource allocation aspects. The limitation with queuing systems is it is used to model only single resource.

2.3 Layered Queuing Networks

Layered queuing networks are advanced techniques of networks to narrate layered software architecture system. Software engineering models can be made use to prepare LQN model of application automatically by using UML or Palladio Component Models (PCM)[3]. The advantageous position of LQN's over normal queuing networks in so for as it explain the dependencies occur during tedious flow of requests and as a results will process hardware and software resources. The impact of the latency on response time for different system applications came to light when authors on the subject tried to explored from the cloud systems by applying LQN. LQN's has its significance in handling the complexity of remote applications such as transactional and streaming workloads. LQN model is made use of o find out the performance of the Rubis benchmark application and in turn it is used as basis of optimization algorithm that aims at determining the installation of replication levels and application components in a best way. Though this work is not specific to the cloud it reveals the application of LQN's

to multitier applications which are commonly used often and often in such environments, enterprise applications adopted on the cloud in strict conformity of SLS requirements depending on historical data. The limitations of LQN's identifying main limitations for practical application in cloud systems information is available from the authors on the subject. The above process among others in inclusive of difficulties in modeling caching, absence of methods to calculate percentage of response time, variation between proposed system accuracy and speed[4]. From that juncture evaluation techniques for LQN's which allow determination of response time percentage has been presented.

3. Proposed System

The IaaS Cloud has N physical resources as shown in Figure 1 and the job requests are sent to system queue. The size of the system queue is fixed and denoted by 'Q'. If the maximum size is reached the upcoming job requests will be rejected the queue is maintained based on the system behavior. Once the resource is free, the job will be accepted and corresponding virtual machine will be instantiated we assume service time is exponentially varies with mean 1/μ. As per the virtual machine multiplexing, cloud system provides 'M' logical resources more than 'N' and multiple virtual machines are often used with same physical

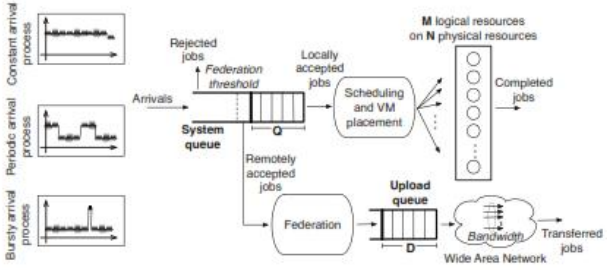


Figure 1. An infrastructure cloud.

machine. Many VM's using a same physical machine. Many VM's using a same physical machine leads to reduction of performance. The degradation issue will be considered as d = 0, the reduction is performance of multiplexed VM's depends on its techniques and VM placement aspects[5]. The system will optimally balance the load among for physical machines and resources required by virtual machine. Let us consider T = 1/μ.

Will be the expected time required to execute two virtual machines can be given by T2 = T.(1 + d).

The expected execution time of n virtual machines as:

$$T_i = T.(1 + d)i - 1 \quad (1)$$

The cloud managers can estimate the parameter 'd' shown in theoretical work and analytical observations. Cloud federation provides the resources given by alternative cloud through sharing and paying model. So whichever the jobs are pending, so these requests can be shifted to other cloud by transferring the particular virtual machine disk to cloud. We assume the following assumptions:

- let us consider the job arrives once queue is full.
- Availability is given by a_f.
- The quality level qf ($0 < qf \leq 1$) reached by request the virtual machine needs T = 1/μ time to work and service time is given by Tf = 1/(qf •μ) ≥ T).
- The job is sent to upload queue and waiting for virtual machine transfer completion.
- Size of upload queue is D.
- Bandwidth allows virtual machines to transmit 'm' virtual machines simultaneously.
- The transfer time is 1/η.

For arrival process three scenarios we will consider. In the constant arrival process, is a poisson process and its arrival rate is λ, but

practically it contains thousands of users. They depends on completion of the previous arrival processes. Arrival rate is constant, but latency will be more. In the periodic arrival process, the arrival process is a markov modulated poisson process. We will refer to an M M P P (λh, λl, λh2l, λl2h), λh and λl represents the arrival rate at high and low. 1/λh2l and 1/λl2h shows the duration of two load conditions. The bursty arrival process with fixed and short duration is used to analyze the system resiliency. We use SRN's to analyze the features of IaaS cloud. These are improved version of generalised SRN's. The number of tokens in P is given by P#.if load λh when Pmmpp# = 1.the low and high conditions are given by λh2l and λl2h respectively. The queue is modelled through Pqueue. Token is nothing but a job waiting in the queue. When tokens are greater than size Q, tdrop is enabled, thus the reject of request is modelled. The resources are modelled in the place Pres. Such tokens represent 'M' logical resources provided by system, if resource ia available and any pending requests is there, then these requests shifted from tlocal queue to place Prun if M is equal to N then service time is modelled by Tserv equal to μ formally, indicating with Rserv (Prun#) the rate of transition Tserv.

$$\text{Rserv (Prun\#)} = \text{Prun\#} \bullet \mu \text{ , one \< Prun\#} \leq N. \qquad (2)$$

The projected SRN cloud performance model is represented in Figure 2. Transition Tarr models the arrival method. If the load (with

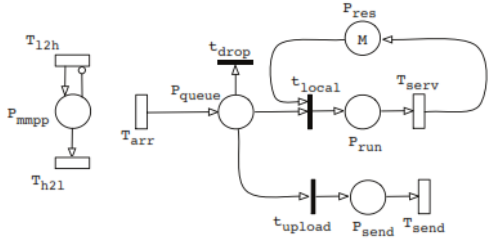
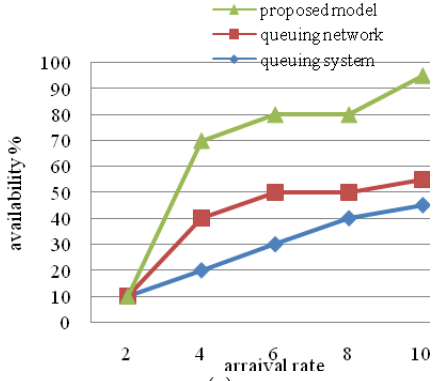


Figure 2. Projected SRN cloud performance model.

rate adequate λh) once Pmmpp# = one. The alter-nation between low and high load conditions is sculptured by exponentially distributed transitions Th2l and Tl2h with rates λh2l and λl2h, severally. The technique adopted to model the Bursty arrival method. The system queue is sculptured through place Pqueue. A token during this place represents employment waiting within the queue. Once the quantity of tokens in Pqueue is larger than the queue size letter of the alphabet, transition tdrop is enabled (see the corresponding guard operate in Figure 2) so modeling the rejection of an invitation. Cloud resources area unit sculptured by tokens in situ Pres. Such tokens represent the M logical resources offered by the system. Once a resource is offered and there area unit unfinished requests, a token is removed (through transition tlocal) from the system queue and is extra to put Prun. If VM multiplexing isn't allowed (M = N) service time is sculptured through transition Tserv with firing rate adequate μ whereas the VM parallel execution (one for every PM) is sculptured by setting transition Tserv with the infinite server linguistics [18] so as to extend its firing rate in proportion to the quantity of tokens within the sanctionative place Prun. additional formally, indicating with Rserv (Prun#) the rate of transition Tserv



(a)

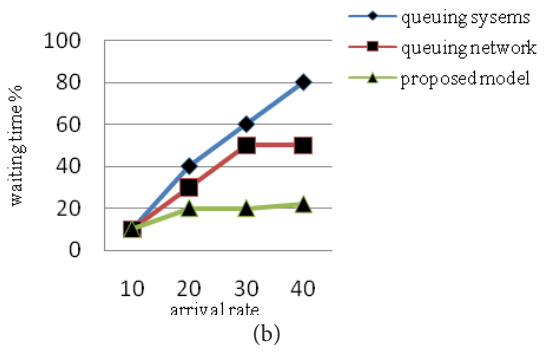


Figure 3. **(a)** Availability. **(b)** Waiting time.

$$\text{Rserv (Prun\#)} = \text{Prun\#} \bullet \mu, \text{ one \< Prun\#} \leq N. \quad (3)$$

Rejection of an invitation. Cloud resources area unit sculptured by tokens in situ Pres.

The graph (a) availability shows that compare to queuing systems and the queuing networks model, the proposed stochastic reward net model will provide more availability for the cloud systems. The graph (b) waiting time shows that compare to queuing systems and queuing networks, the proposed model depicts that waiting time will be less. Here the waiting time is (delay + service time).

4. Conclusion

This paper has presented a stochastic model using reward nets approach which is an analytical model, and easy to model the cloud computing and its strategies[6]. Several performance parameters like availability and utilization and service time are taken into consideration. A general approach is used for the aspect of system capacity like allocating bandwidth, memory, creation of virtual machines under different strategies in the business aspect such as cloud systems for an exact analysis of factors is required to provide Quality and manage service level agreements. Future work is involved in the exclusive techniques to reset the system configurations for the different system behaviors. The stochastic model for the IaaS cloud, so the same model can be used to represent Platform and Software as a service cloud systems and we can include VM migration concepts and various energy saving policies can be included.

5. References

1. Buyya R, et al. Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Gener Comput Syst. 2009 Jun; 25:599–616.
2. Meng X, et al. Efficient resource provisioning in compute clouds via vm multiplexing. Proceedings of the 7th ACM International Conference on Autonomic Computing, Ser, ICAC'10; New York, NY, USA. 2010. p. 11–20.
3. Liu H, et al. Live virtual machine migration via asynchronous replication and state synchronization. IEEE Transactions on Parallel and Distributed Systems. 2011 Dec; 22(12):1986–99.
4. Buyya R, Ranjan R, Calheiros R. Modeling and simulation of scalable cloud computing environments and the cloudsim toolkit: Challenges and opportunities. International Conference on High Performance Computing Simulation, HPCS '09; 2009 Jun. p. 1–11.
5. Iosup A, Yigitbasi N, Epema D. On the performance variability of production cloud services. 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid); 2011 May. p. 104–13.
6. Ostermann S, et al. A performance analysis of EC2 cloud computing services for scientific computing. Cloud Computing Ser Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. 2010; 34(9):115–31. Springer Berlin Heidelberg.