A Status Report on Resource Allocation in Cloud Computing Using Queuing Theory

R.Murugesan[1], C.Elango[2], S.Kannan[3]

[1]Department of Computer Science, Cardamom Planters' Association College, Bodinayakanur, Tamilnadu
[2]Department of Mathematical Sciences, Cardamom Planters' Association College, Bodinayakanur, Tamilnadu
[3]Department of Computer Applications, Madurai Kamaraj University, Madurai, Tamilnadu

Abstract - **Cloud computing has emerged as an optimal way of sharing and providing resources over the internet. Resource Allocation is one of the pretentious concerns in the cloud. It allocates the resources to its consumer on demand. Because it offers dynamic flexible resource allocation for reliable and guaranteed services in pay as-you-use manner to public. In this paper, we surveyed and analyzed several resource allocation techniques and models which focus on Queuing Theory concepts related to the existing resource allocation in the cloud.**

Index Terms - **Cloud Computing, Resource Allocation, Queuing Theory, QoS**

I. CLOUD COMPUTING AND QUEUES

In this section we discuss some related work for resource allocation in cloud computing using Queuing theory.

In [1] **Jordi Vilaplana** et, al., author analyze, the design of a cloud architecture with QoS requirements. The combination of $M/M/_1$ and M/M/m queuing models in sequence was proposed to model the cloud platform. It shows that to provide good QoS in terms of response time, one has to determine where the system has a bottleneck and then improve the corresponding parameter. Finally, it concludes that the model can be very useful for tuning service performance, i.e., QoS (response time). Thus guaranteeing the SLA contract between the client and the service provider.

Hamzeh Khazaei, et, al., [2] presented a performance model suitable for analyzing the service quality of large sized IaaS clouds, using interacting stochastic models. The author examined the effects of various parameters including arrival rate, task service time, the virtualization degree, and service task size on task rejection probability and total response delay. The stable, transient and unstable regimes of operation for given configurations have been identified so that capacity planning is going to be a less challenging task for cloud providers.

Masoud Salehpour,et, al., [3] proposed different workload types with different characteristics that should be supported by cloud computing, but there is no single solution can allocate resources to all imaginable demands optimally. Consequently, it is necessary to design specific solutions to allocate resources for each workload type. This paper has focused on bag of tasks applications. It is proposed an idea to facilitate dynamic resource allocation this workload type. This approach has monitored server's traffic intensity to response users based on an appropriate resource selection and allocation in a reasonable time.

Mohamed Ben El, et, al.,[4] presented the technical benefits of cloud computing. They found processing overflow traffic incoming from customers resources to cloud centers, this overflowed traffic has specific characteristics. They designed an approximate model based on a Markov chain using an IPP/G/m/k queue. This model is more realistic since it considers the properties of arriving tasks and the characteristics of a cloud center. Authors described a new analytical approximation for performance evaluation of a cloud computing center and resolved it to get a very decent estimate.

In [5],**Lizheng Guo Tao Yan**, et, al., have studied the performance of the optimization and the parameter for evaluating the service in cloud computing. In order to analyze the performance of services in cloud computing they proposed a queuing model and developed a synthetically optimization method to optimize the performance. Further they simulated the system to validate our optimization method. Simulation results showed that the proposed method can allow less wait time, queue length and more customers gaining the service.

Sandeep K. Sood,[6] proposed a new approach, uses banker's algorithm for resource allocation. That mean, there is no possibility of deadlock. Also by restricting the number of login users, resources are not choked out even in case of heavy demand of resources.

Moreover, resource allocation matrix specifies the requirement of resources in advance to run that job. This model is an effective model that is efficient from other related existing dynamic resource provisioning model. It provides better response time to each request in real time interactive applications. Simulation of proposed model shows that results are good for dynamical allocation of resources.

Mohamed Eisa, et, al., [7] propose the cloud computing model based on queuing system. He studied the routing of incoming requests to the queue with reduced workload, response time and the average length of the queue. These results indicate that the model increase utilization of global scheduler and decrease waiting time. The experimental results indicated that proposed model decrease waiting time at global scheduler in cloud architecture.

In [8] **Fatima Oumellal**, et, al., proposed an approximate model based on Markov chain to evaluate the performance cloud computing center using the queue MMPP/G/m/m+r. Due to the nature of the environment of cloud computing and the diverse needs and demands of users, a MMPP arrival process to considered which reflects the nature of arrivals in the cloud. A general service time, a number of servers and a finite buffer capacity are also assumed. The new analytical approximation for performance evaluation of a center of cloud computing has been evolved and resolved to get a very decent estimate. In this model performance indicators such as the average number of tasks in the system, blocking probability, probability of immediate service and the average of response time are calculated analytically.

The author**Hyacinth C**, et, al., [9] impose the nominal throughput of the workload, and waiting time in the queuing model, Server utilization and service time shows that the model will facilitate service efficiency and optimal performance. In the cloud queuing formulation, the feedback branching is probabilistic in nature. The model which was tested and implemented in the actual system shows that queuing theory will optimize network traffic. The model demonstrates that the direct proportion with the incoming job queues or arrivals since the buffer size is approximately infinite.

The author **Xiaoming Nan**, et, al., [10] mentioned that the data center consists of a master server, a computing server, and a transmission server. All these servers are virtual machine (VM) instances generated from physical computation resources. Thus, the schedule queue is modeled as a preemptive priority $M/M/_1$ queuing system with mean service rate of the master server. They employe the queuing model to study the service response time in each queue. The simulations results show that the proposed optimal resource allocation scheme can achieve minimal resource cost or minimal service response time in multimedia cloud.

Preeti kamble, et, al., [11] established that the performance and availability of cloud applications has a noticeable impact on user adoption and revenue of the cloud. The work on the performance analysis of the cloud using M/G/m/m+r queuing system till date gives the novel and approximate analytical solution. It gives the relationship between the input buffer size and number of servers available. It also gives the performance indicators like mean number of tasks in the system, task blocking probability and immediate service probability.

Satyanarayana, et, al., [12] approach a novel cloud computing model which is much useful for analyzing the cloud more effectively and efficiently to increase performance measures of cloud. The work presented in this paper focus on the improvement of allocation of resources dynamically following request dependent strategy under non homogeneous condition with time dependent arrival of jobs. It shows that dynamic allocation of resources can reduce mean delay and mean service time

Goswami.V et, al., [13] developed a model in which virtual machines are taken as service centers and the web applications are modeled as queues. The author employed the finite buffer multi server queuing system with queue dependent heterogeneous servers and the number of server's changes depending on the queue length. The service load in cloud computing is dynamically scaled up and down depending upon end users service requests. Steady state queue size distribution is obtained using a recursive method assuming Markovian arrival and service times to get a Markov chain. It has been shown that the queuing based model is effective in the web applications on cloud and that no VM live migration is involved. It will be also useful in the services performance prediction of cloud computing.

Bharathi M, et, al.,[14] presents the cloud data center modelled as M/G/m/m+r queuing system with a single task arrivals and a task request buffer of finite capacity. This gives s versatile model for performance evaluation of a cloud computing data center.

In [15]**Ani Brown Mary N** et, al., modeled the cloud center as an [(M/G/$_1$) : (∞/GD MODEL)] queuing system with single task arrivals and a task request buffer of infinite capacity. They evaluated the performance of queuing system using an analytical methods and solve it to obtain important performance factors like mean number of tasks. Mean as well as standard deviation of the number of tasks is computed. The blocking probability and probability of immediate service are also computed.

Mohamed Ben et, al., [16] proposed an analytical model for performance evaluation of a cloud computing data center using the queue GE/G/m/m+r. Owing to the nature of the cloud environment and the diversity of needs and demands of users, the proposed model uses a Generalized exponential (GE) arrival process that reflects the nature of arrivals in the cloud with geometrically distributed batch sizes, general service time, number of servers and a finite buffer capacity. In this model they calculated analytically the performance indicators such as the average number of tasks in the system, blocking probability, probability of immediate service and the average of response time.

Anupama, et, al., [17] used stochastic process to analyses the dynamic behavior of infinite severs over single server. They have studied the utilization factor, throughput, length of server, and waiting time of infinite server system. From the user point of view it gets service immediately there is no need to be in queue for service. With good selection of number of servers in infinite server system can reduce queue length and increase throughput and utilization.

In [18] **Xiaoming Nan** studied resource allocation problems for differentiated multimedia services. They propose a Queuing model to characterize the service process in cloud center. Based on the proposed Queuing model, they investigated the resource allocation in FCFS scenario and priority scenario, respectively. For each scenario they formulated and solved the optimal resource allocation problem to minimize resource cost under the response time constraints. Simulation results demonstrate that the proposed resource allocation schemes can optimally utilize cloud resources to provide satisfactory services for different classes of requests at the minimal resource cost.

Sai Sowjanya[19] proposed a M/M/$_S$ model with two servers which increases the performance over using one server by reducing the queue length and waiting time. Analysis and numerical results clearly shows that the M/M/$_2$ approaches reduce queue length and waiting time when compared to M/M/$_1$.

Xiaoming Nan,et, al., [20] proposed queuing model, to investigate resource optimization problems for multimedia cloud computing in three different scenarios: single-service scenario, multi-service scenario, and priority service scenario. The authors followed different techniques. They are M/M/$_1$,M/M/$_s$ and M/Hm/$_1$ Queuing system, where H_M represents the hyper-exponential-m distribution. In each scenario, they formulated and solved the response time minimization problem and the resource cost minimization problem.

Chandan Banerjee, et,al., [21] proposed a model, in which service requests have been executed using queues. Virtual machines are modeled as service centers using M/EK/$_1$ model and M/EK/$_2$ model. Multiple servers based scenario has improved the performance of the system by queue length reduction and waiting time optimization over single server based scenario.

Murugesan R, et al. [22] studied the Cloud Computing Network (CCN) with Poisson arrival process and exponentially service times with M/M/s queue. They found to mean number of requests waits and probability that the system is busy. This system performance measures are used to get the optimal resource allocation parameters.

The author **Murugesan R**, et al. [23] proposed, a stochastic model in which the resource allocation is modeled as queues with the virtual machines as service centers. They considered, M/G/s queue as a tool. Arrival Poisson and general service time for requests with single server and infinite waiting space. The author evaluated the performance measures of cloud server farms and they solved it to obtain accurate estimation of the complete probability distribution of the request response time and other important performance indicators. They also obtained, the expected waiting time in the system and expected number of tasks waiting for transmission in the system.

Murugesan R, et al., [24] proposed an approximate model to evaluate the performance of a cloud computing center using the G/M/s queue model method. They considered a G/M/s queuing system that reflects the general nature of BoT's arrivals in the cloud. This system has general inter-arrival time, more number of servers and a infinite buffer capacity. The author observed that when the arrival rate increases, the length of queue (queue size) also increases, and the waiting time of a customer increases linearly with the arrival rate.

II. COMPARISON OF RESOURCE ALLOCATION MODELS USING QUEUING THEORY

S.No	Authors	Techniques	Methods	Parameters	Findings
1	Jordi Vilaplana, Francesc Solsona Ivan Teixidó Jordi Mateo Francesc Abella Josep Rius, **2014**	M/M/$_1$, M/M/m	Poisson Arrival Process	Arrival rate, service rate, file size server bandwidth, client bandwidth	Response time
2	Hamzeh Khazaei, Jelena Mi ˇ si´ c, and Vojislav B. Mi ˇ si ´c, **2012**	M/M/$_1$ (FIFO)	Poisson Arrival Process	Arrival rate, task service time, the virtualization degree, task rejection	Reliable response time and blocking probability avoidance
3	Masoud Salehpour, and Asadollah Shahbahrami, **2012**	M/G/m, FCFS	Poisson Arrival Process	Arrival rate, service rate	Number of tasks, Mean Response Time, Mean Waiting and Service Time
4	Mohamed BenN El Mohamed Hanini, Fatima Oumellal, Abdellah Zaaloul, Abdelkrim Haqiq., **2014**	IPP/G/m/m r	Poisson Arrival Process	Arrival rate, service rate	Number of tasks, waiting and response times, immediate service, Blocking probability
5	Lizheng Guo Tao Yan, Shuguang Zhao, Changyuan Jiang, **2013**	M/M/m	Poisson Arrival Process	Arrival rate, service rate	Mean Queue Size, Delay, Waiting Time
6	Sandeep K. Sood, **2013**	M/M/$_1$	Poisson Arrival Process	Arrival rate, service rate	Observation time, busy time and completion time
7	Mohamed Eisa, E. I. Esedimy,M. Z. Rashad **2014**	M/M/$_1$, M/M/s	Poisson Arrival Process	Arrival rate, service rate,	Queue length, Residence time, Utilization Throughput
8	Fatima Oumellal, Mohamed Hanini, Abdelkrim Haqiq, **2014**	MMPP/G/m/m+ r	Markov modulated Poisson Process (MMPP)	Arrival rate, service rate	Average number of tasks, blocking probability, probability of immediate service and the average response time.
9	Hyacinth C, Inyiama Å and Nkolika O. Nwazor , **2014**	G/G/$_1$	Poisson Arrival Process	Arrival rate, service rate	Throughput , waiting time, Server Utilization and service time
10	Xiaoming Nan, Yifeng He and Ling Guan	M/M/s	Poisson Arrival process	Arrival rate, service rate	Cost minimization and the service response time minimization
11	Preeti kamble, Hemlata channe, **2013**	M/g/m/m+r	Poisson Arrival Process	Arrival rate, service rate	Performances like mean number of tasks, task blocking probability and immediate service Probability
12	Satyanarayana .A, P. Suresh Varma M.V.Rama Sundari P Sarada Varma, **2013**	M/M/$_1$	Non-homogeneou Poisson Arrival Process	Arrival rate, service rate	Reduce mean delay and mean service time
13	Goswami.V, Patra, S. S, Mund G. B, 2012	M/M/s	Poisson Arrival Process	Arrival rate, service rate	Services performance prediction
14	Bharathi M, Sandeep Kumar P, Poornima G,V **2012**	M/G/m/m+r	Poisson Arrival rate, Service rate	Arrival rate, service time, queue capacity	Performance evaluation of data center.
15	N.Ani Brown Mary and K.Saravanan **2013**	[(M/G/$_1$) : (∞/GDMODEL)]	Poisson Arrival Process	Arrival rate, service time,	Standard deviation, The blocking probability and probability of immediate service

16	Mohamed Ben el aattar, Abdelkrim Haqiq **2012**	GE/G/m/k	Batch Poisson with geometrically distributed batch sizes	Arrival rate, service time,	Average number of tasks in the system, blocking probability, probability of immediate service and the average of response time
17	A.Anupama, G.Satya Keerthi **2014**	M/M/$_1$, M/M/∞	Poisson Arrival Process	Arrival rate, service time	Utilization factor, throughput, length of server, waiting time of infinite server
18	Xiaoming Nan, Yifeng He, and Ling Guan **2014**	M/M/$_1$/∞/FCFS	Poisson processes in FCFS	Arrival rate, service time	Minimize resource cost under the response time constraints
19	Sai Sowjanya. T, Praveen. D, Satish,A.Rahiman. K **2011**	M/M/s/∞	Poisson Arrival Process	The arrival processes service time, distribution number of servers, number of places in the system, calling population, queue's discipline	Reducing the queue length and waiting time
20	Xiaoming Nan, Yifeng He, Ling Guan **2014**	M/M/$_1$,M/M//$_1$& M/Hm /$_1$ queueing system, H_m represents the hyperexponentia l-m distribution.	Poisson Arrival Process	Arrival rate, service time	Response time minimization problem and the resource cost minimization problem
21	Chandan Banerjee, Anirban Kundu, Ayush Agarwal, Puja Singh, Sneha Bhattacharya, and Rana Dattagupta ., **2014**	M/EK/$_1$ model and M/EK/$_2$	Poisson Arrival Process	Arrival rate, service time	Queue length reduction and waiting time optimization
22	Murugesan R Elango C, and Kannan S, **2014**	M/M/s	Poisson Arrival Process	Arrival rate, service time	Mean number of requests waits and Probability that the system is busy
23	Murugesan R Elango C, and Kannan S, **2014**	M/G/s	Poisson Arrival Process	Arrival rate, service time	Expected waiting time in the system and expected number of tasks waiting
24	Murugesan R Elango C, and Kannan S, 2014	G/M/s	Poisson Arrival Process	Arrival rate, service time	The arrival rate increases, the length of queue also increases, waiting time of a customer increases linearly with arrival rate

III. CONCLUSION

Cloud computing is a boon for this modern technological world. It is one of most essential technology needed to share and allocate resources needed for data processing and computation in online. Cloud has permanent all time storage environment and flexible computing ability. It is portable and flexible to save and share the resources throughout the world. In this paper, we reviewed cloud computing models studied in the present decade using Queuing systems with resource allocation mechanisms. Even though the test of articles is not exhaustive, it represents wide variety of models. Hence this paper will hopefully motivate and guide the researchers to the smart resource allocation in cloud computing using Queuing theory.

REFERENCES

[1] Jordi Vilaplana Francesc SolsonaIvan Teixidó Jordi MateoFrancesc Abella Josep Rius, "A queuing theory model for cloud computing", J Supercomputer,© Springer Science+Business Media New York 2014.

[2] Hamzeh Khazaei, Jelena Mi ˇ si´ c, and Vojislav B. Mi ˇ si ´c, "A Fine-Grained Performance Model of Cloud Computing Centers", IEEE transaction on parallel and distributed systems , vol. X, no. Y, 2012.

[3] Masoud Salehpour and Asadollah Shahbahrami, "Alleviating Dynamic Resource Allocation for Bag of Tasks Applications in Cloud Computing", International Journal of Grid and Distributed Computing Vol. 5, No. 3, September, 2012.

[4] Mohamed BEN EL AATTAR, Mohamed HANINI, Fatima OUMELLAL, Abdellah ZAALOUL, Abdelkrim HAQIQ, "Analysis of Queuing System Model for Overflow Tasks

Routed to a Cloud Computing Center", International Journal of Applied Mathematics and Modeling IJA2M Vol.2, No. 3, 28-37, May, 2014, ISSN: 2336-0054.

[5] Lizheng Guo, Tao Yan,Shuguang Zhao, Changyuan Jiang," Dynamic Performance Optimization for Cloud Computing Using M/M/m Queueing System".

[6] Sandeep K. Sood, " Dynamic Resource Provisioning in Cloud based on Queuing Model", International Journal of Cloud Computing and Services Science (IJ-CLOSER) Vol.2, No.4, August 2013, pp. 314~320 ISSN: 2089-3337.

[7] Mohamed Eisa, E. I. Esedimy,M. Z. Rashad, "Enhancing Cloud Computing Scheduling based on Queuing Models", International Journal of Computer Applications (0975 – 8887)Volume 85 – No 2, January 2014.

[8] Fatima Oumellal, Mohamed Hanini and Abdelkrim Haqiq, "MMPP/G/m/m+r Queuing System Model to Analytically Evaluate Cloud Computing Center Performances", British Journal of Mathematics & Computer Science,4(10): 1301-1317, 2014.

[9] Hyacinth C. Inyiama Ȧ and Nkolika O. Nwazor,"Model of a Job Traffic Queue of a Cloud- Based Research Collaboration Platform", International Journal of Current Engineering and Technology E-ISSN 2277 – 4106, P-ISSN 2347 – 5161,(Oct 2014).

[10] Xiaoming Nan, Yifeng He and Ling Guan," Optimal Resource Allocation for Multimedia Cloud in Priority Service Scheme".

[11] Preeti kamble, Hemlata channe, "Performance analysis of cloud computing centers by Breaking-down response time", International Journal of Advanced Computational Engineering and Networking, ISSN (p): 2320-2106, Volume- 1, Issue- 8, Oct-2013.

[12] Satyanarayana A, Dr. P. Suresh Varma Dr. M.V.Rama Sundari Dr. P Sarada Varma," Performance Analysis of Cloud Computing under Non Homogeneous Conditions",International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013 ISSN: 2277 128X.

[13] Goswami V, Patra , S. S, Mund G. B, "Performance Analysis of Cloud with Queue Dependent Virtual Machines", 1st Int'l Conf. on Recent Advances in Information Technology | RAIT-2012.

[14] Bharathi M, Prof. Sandeep Kumar P, Prof. Poornima G V, "Performance factors of cloud computing data centers using M/G/m/m+r queuing systems", IOSR Journal of Engineering (IOSRJEN) e-ISSN: 2250-3021, p-ISSN: 2278-8719, Volume 2, Issue 9 (September 2012), PP 06-10.

[15] Ani Brown Mary N and K.Saravanan,"Performance factors of cloud computing Data centers using [(m/g/1) : (∞/gdmodel)] Queuing systems", International Journal of Grid Computing & Applications (IJGCA) Vol.4, No.1, March 2013.

[16] Mohamed Ben el aattar, Abdelkrim Haqiq,"Performance Modeling for a Cloud Computing Center Using GE/G/m/k Queuing System", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064,2012.

[17] Anupama A, G.Satya Keerthi, "Using Queuing theory the performance measures of cloud with infinite servers," Using Queuing theory the performance measures of cloud with infinite servers, ISSN : 2229-3345 Vol. 5 No. 01 Jan 2014.

[18] Xiaoming Nan, Yifeng He, and Ling Guan, "Towards Optimal Resource Allocation for Differentiated Multimedia Services in Cloud Computing Environment",2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)978-1-4799-2893-4

[19] Sai Sowjanya. T, Praveen. D, Satish,A .Rahiman. K,"The Queueing Theory in Cloud Computing to PReduce the Waiting Time", IJCSET | April 2011 | Vol 1, Issue 3,110-112.

[20] Xiaoming Nan, Yifeng He, Ling Guan."Queueing Model based Resource Optimization for Multimedia Cloud", Preprint submitted to Elsevier January 7, 2014.

[21] Chandan Banerjee, Anirban Kundu, Ayush Agarwal, Puja Singh, Sneha Bhattacharya, and Rana Dattagupta, "Priority based K-Erlang Distribution Method in Cloud Computing", Int. J. on Recent Trends in Engineering and Technology, Vol. 10, No. 1, Jan 2014.

[22] Murugesan R, Elango C, and Kannan S," Cloud Computing Networks with Poisson Arrival Process-Dynamic Resource Allocation", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. IV (Sep – Oct. 2014), PP 124-129.

[23] Murugesan, R,Elango C, and Kannan S," Resource Allocation in Cloud Computing with M/G/s – Queueing Model", Volume 4, Issue 9, September 2014 , ISSN: 2277 128X, International Journal of Advanced Research in Computer Science and Software Engineering, PP 443-447.

[24] Murugesan R, Elango C, and Kannan S, "Resource Allocation in Cloud Computing with General Classification Time and Exponential Service (G/M/s)", International Journal Of Engineering And Computer Science ISSN:2319-7242,Volume 3, Issue 10 October, 2014 Page No. 8905-8910.

